ARCHIVES OF ACOUSTICS
32, 1, 25–40 (2007)

PROSODY ANNOTATION FOR UNIT SELECTION TTS SYNTHESIS

Grażyna DEMENKO, Agnieszka WAGNER

Adam Mickiewicz University
Institute of Linguistics
Międzychodzka 5, 60-371 Poznań, Poland
e-mail: {lin, wagner}@amu.edu.pl

(*received October 16, 2006; accepted December 27, 2006*)

This paper concerns prosody annotation and intonation modeling, especially for the application in a corpus based speech synthesis. In order to establish the rules of the automatic intonation modeling, a four hour fully annotated speech database has been acoustically and perceptually analyzed. The speech material included different text types, dialogs and prosodically rich phrases.

As the result of these analyses, a basic prosodic annotation including 6 pitch accent types and 5 types of prosodic phrases have been distinguished. Moreover, the analyses made it possible to define rules for a semi-automatic stylization and parametrization of intonation contours for the application in text-to-speech and speech recognition systems. The assumptions behind the stylization method and results of the quantitative and qualitative evaluation of the stylization accuracy based on the speech consisting of ca. 1000 phrases coming from a literary text read by female and male speakers are discussed. Finally, a classification of pitch accents and boundary tones based on the parameterization is presented.

Keywords: speech synthesis and recognition, segmental and suprasegmental (prosodic) annotation, intonation modeling, intonation stylization, pitch accents, boundary tones.

1. Introduction

Among many factors affecting the quality of speech recognition and synthesis, the prediction and modeling of prosody plays an essential role. Irrespective of the speech synthesis type (i.e. diphone or unit selection), this is important for several reasons: a) intonation has discourse functions (e.g. signaling given/new information, focus), b) errors in the segmental structure are accepted by the listeners to a greater degree than those in the suprasegmental structure of the utterance, c) an erroneous placement of accent or an incorrect accent type may change the meaning of the utterance significantly or create the impression of unnaturalness.

In speech recognition systems, suprasegmental features are essential as sources of information on the syntactic and semantic structures of utterances [16], but their ex-

traction is difficult and prone to errors ([14, 19]). The verification of the nuclear accent position in a phrase and finding the most essential (from the *informative point of view*) fragments of an utterance makes it possible to reduce the time spent on examining the lexicon. Paralinguistic and nonlinguistic aspects of suprasegmental features play a secondary role if a rapid adaptation of the system is essential and the necessity of the initial decoding of the signal (e.g. hoarseness) or voice identification is avoided.

Speech Technology systems are based on machine learning techniques and rely heavily on training a speech material representative of the specific task. As far as intonation modeling is concerned, many different approaches have been proposed for the representation and prediction of intonation events, usually pitch accents and edge tones (which can be represented either in terms of continuous acoustic parameters or discrete phonological categories as well as by generation of intonation from a given representation ([2, 9, 12, 17, 18, 21, 26]). The quality of the synthesized speech depends heavily on the text type and the synthesis domain: intonation is very natural for a restricted domain, e.g. news or weather forecast, and prosodically stable speech (read or dictated texts) which is distinguished by quite flat intonation, a stable voice quality and easily predictable duration of the speech units. Although there exists such a large number of different approaches to prosody modeling, no universal methodology has been worked out so far. As far as annotation and modeling of Polish intonation is concerned, mostly general representations of intonation for the needs of Speech Technology were developed [6]. There exists, as yet, no method of the intonation stylization which could provide a parametrization of the intonation contours in terms of acoustic variables (e.g. amplitude and duration of commands of functions realizing prosodic constituents ([9, 17]) from which a higher level of the representation of intonation could be derived. The development of a stylization method will be one of the subjects of this paper.

For the purpose of prosody modeling in the Polish module of BOSS (Bonn Open Source Synthesis), which is a corpus based speech synthesis system [3], only fundamental types of prosodic information were distinguished such as lexical stress, the pitch accent type and prosodic phrase type.

For the duration modeling we used the CART algorithm with 52 input factors among which there were the following ones: the identity of the current, the preceding and following phoneme, phoneme start position, phoneme position in the phrase and foot length [4]. This produced an RMSE of 25.86 ms and a mean correlation of 0.62 and gave a relatively correct time structure of the synthesized speech. But with respect to intonation, the synthesized speech was rather of poor quality (examples are available at: http://main.amu.edu.pl/~fonetyka/synthesis_examples.html).

Therefore the aim of the current research was: a) to establish an inventory of distinctive (with respect to realization and perception) intonation events to be annotated in a speech database, especially for the application in speech synthesis, b) to develop rules for the approximation of intonation contours on the basis of these findings, c) to classify intonation events on the basis of the parameterization and to compare the results with existing representations of Polish intonation.

2. Speech database annotation

The entire linguistic material (4 hours) was read by a professional radio speaker during several recording sessions and supervised by an expert phonetician. The annotation of the speech database for the speech synthesis application consists in providing information on the segmental and suprasegmental structure of utterances such as phone, syllable and word boundaries, lexical stress, pitch accent type and prosodic phrase type (according to the European Centre of Excellence for Speech Synthesis guidelines, see http://www.ecess.eu/). Most of these factors were annotated automatically using speech transcription and segmentation software (Polphone [8], CreatSeg, AnnotationEditor [24, 25]).

2.1. Segmental annotation

The computer coding conventions were drawn up in SAMPA for Polish [27] with revisions and extensions and from the IPA alphabet (IPA Homepage). Two sets of characters were precisely defined for the exact GTP mapping of the Polish language – an input set of characters and an output phonetic/phonemic alphabet. The input set of symbols for Polish was defined here as a set of the following symbols: X = {a, ą, b, c, ć, d, e, ę, f, g, h, i, j, k, l, ł, m, n, ń, o, ó, p, q, r, s, ś, t, u, v, w, x, y, z, ź, ż, #, ##}. One hash substitutes interword spaces in the string of orthographic symbols, and two hashes denote sentence final punctuation marks. An inventory of 39 phonemes was employed for the broad transcription (cf. Appendix) and a set of 87 allophones was established for the narrow transcription of Polish. The authors verified the existing phonetic notations and transcription rules for Polish on the basis ofthe literature describing the Polish phonological system and Polish rules of pronunciation and the results of the acoustic segmentation for a few hundred utterances produced by 50 speakers from two Polish cities. The following modifications were made to the original Polish version of SAMPA:

1. The palatal phonemes /c/ (as in kiedy) and /J/ (as in giełda) are necessary for describing accordingly the acoustical differences between velar /k/ (as in: kat) and /c/; velar /g/ (as in gad) and /J/. /c/ is similar to English “k” in “ski” (never aspirated). Possible before /j/, /i/, /e/ and /o/ only. /k/ like English “k” in “sky” (never aspirated). /J/ similar to English “g” in “geese”. The voiced counterpart to /c/. Possible before /j/, /i/, /e/ and /o/ only. /g/ like English “g” in “go”. The voiced counterpart to /k/.

2. The rules for the transcription of Polish graphemes: ą and ę have not been defined precisely for a long time and the existing system has made assumptions according to the synchronic or a synchronic pronunciation based mainly on theoretical considerations.

3. For ę: the possible transcriptions are: /ew̃/ /e/ /em/ /en/ /eń/ /eŋ/. For ą: the possible transcriptions: /ow̃/ /o/ /om/ /on/ /oń/ /oŋ/. j~ nasal counterpart to /j/. It may replace /w̃/ before palatalised spirants, especially after /e/. w~ nasal counterpart to /w/. Word-finally only after /o/ (spelt “ą”) and /e/ (spelt “ę” – when emphatic). Besides, before spirants only.

The phonetic labeling was done automatically with the program CreatSeg (described in the following sections) which uses HMM models (see e.g. [20]). The example strategy yielded 75.4% reduction of gross errors while requiring 13.6% of the database to be manually segmented.

We have used the software (CreatSeg [24, 25]) developed for the automatic segmentation of speech. Its features include:

a. calculating segment (usually phoneme) boundaries based on phonetic transcription,
b. context-dependent phoneme duration models,
c. considering "forced" transition points for semi-automatic segmentation,
d. accepting triphone statistical models trained with HTK tools,
e. tools for the duration models calculation,
f. orthographic-to-phonetic conversion,
g. evaluation of decision trees to synthesis of unknown triphones,
h. accepting wave or MFCC files (plus several label formats) as input,
i. posterior triphone-to-monophone conversion.

The segmentation quality was approx. 80–90% depending on the type of the transcription. Because an accurate phoneme segmentation has fundamental significance for speech processing, some manual corrections were needed.

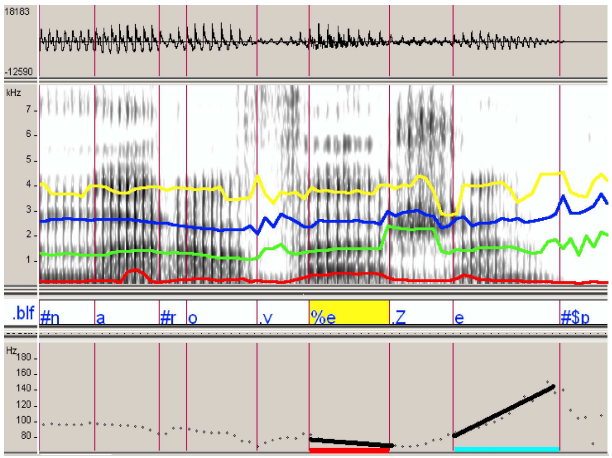
2.2. Suprasegmental annotation

The annotation system for unit selection needs information about the segmental and suprasegmental structure, such as phone, syllable and word boundaries, syllable stress, phrase boundaries of different types and strength. The rules of syllabifying in our research were based on the assumption that there is a relationship between the *sonority* and *syllable structure*. According to this, a continuous phoneme string can be converted into syllables by locating the *syllable boundaries* in the string at positions immediately preceding a point of minimum sonority. Slavic languages are known for their large variety of word-initial consonant clusters. This problem is important for the syllabic theory because most theories of the syllable are based on the Maximal Onset rule (e.g. [5]) which tells to syllabify as many consonants as possible into Onsets. The definition of a "possible Onset" is governed by 1) the Sonority-Sequencing Principle ("within an Onset, the sonority must not decrease") and 2) a language-specific feature that defines allowed Onset-clusters according to the existing word-initial sequences ("CC is a possible Onset-cluster if it occurs word-initially"). For establishing syllable boundaries for Polish, the rules based on the lexicon of 10 million items have been set by an expert and fully automatically implemented in the software program (Annotation Editor).

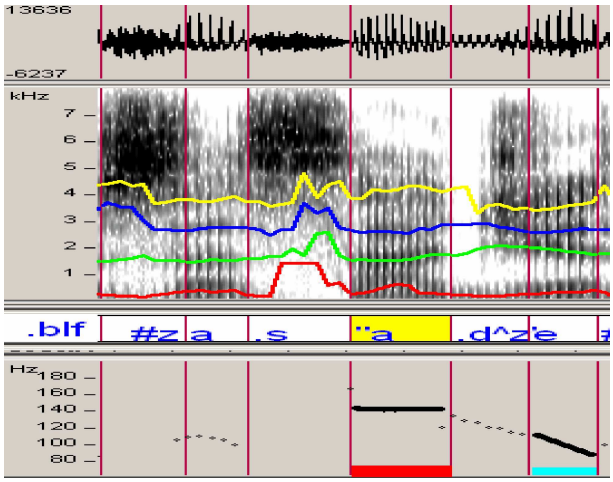
For prosody modeling, only fundamental types of prosodic structures were distinguished, such as the word and phrase accents placement and the accent type or phrase boundary type according to the BOSS label format BLF [3].

Our automatically and phonetically labeled speech database was annotated using suprasegmental features by 4 experts on the basis of perceptual and acoustic analyses of

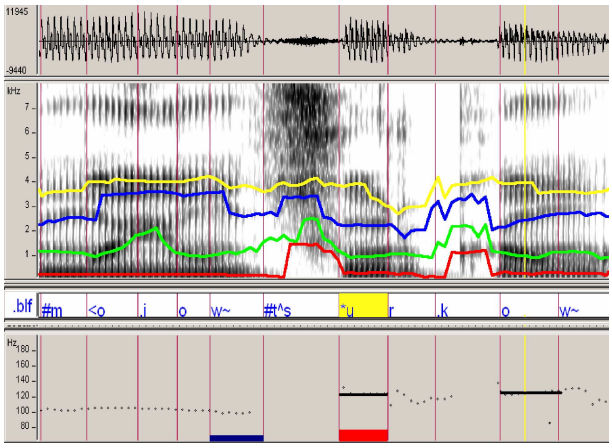
a)


18183
-12590
kHz
7
6
5
4
3
2
1
.blf
#n
a
#r
o
v
%e
Z
e
#$p
Hz
180
160
140
120
100
80

b)


13636
-6237
kHz
7
6
5
4
3
2
1
.blf
#z
a
.s
"a
.d^z'e
Hz
180
160
140
120
100
80

c)


11945
-9440
kHz
7
6
5
4
3
2
1
.blf
#m
<o
.i
o
w~
#^s
"u
r
.k
o
w~
Hz
180
160
140
120
100
80

[Fig. 1. a, b, c]

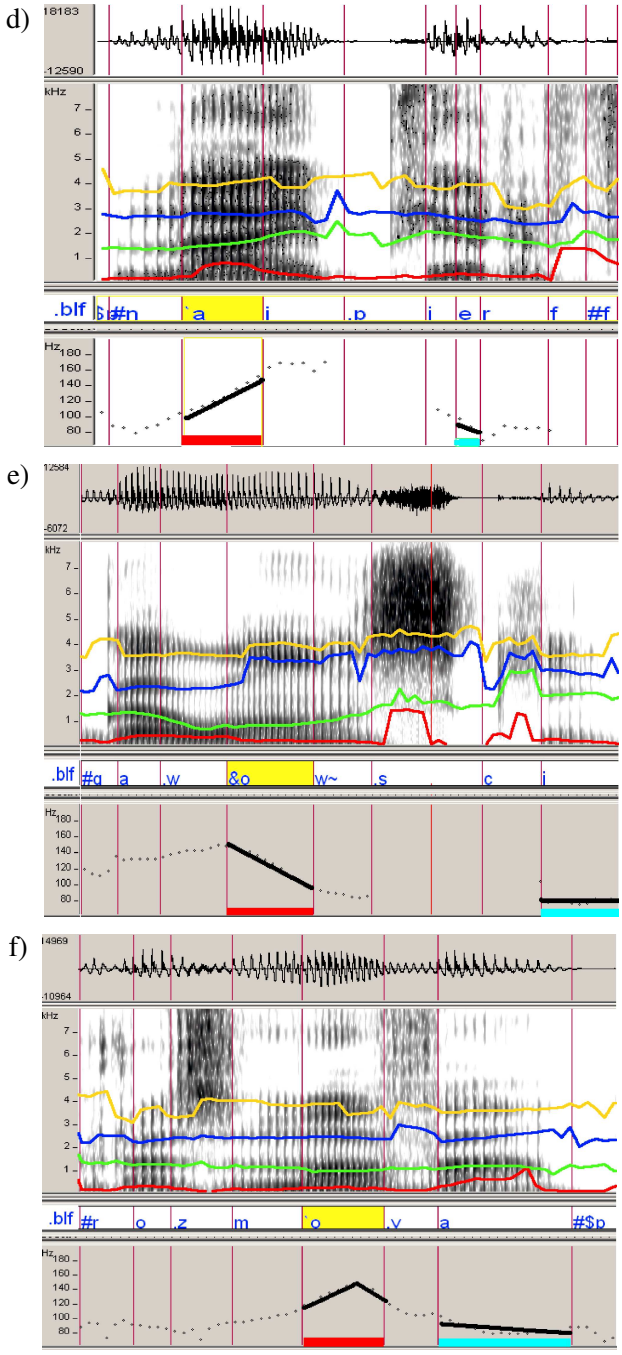


Fig. 1. The inventory of pitch accents: a) pitch movement with rising intonation R (on postaccented syllable – LH), b) falling intonation F (on postaccented syllables HL), c) level intonation, d) rising intonation on accented syllable, e) falling intonation on accented syllable, f) rising – falling intonation on accented syllable. Accented syllables are marked in colour.

the speech signals. On the phrase level information on the sentence and intonation type was provided. On the syllable level, pitch accent types have been marked. On the acoustic level, pitch accents are determined by pitch variations occurring on the successive vowels/syllables and pitch relations between syllables. The pitch accent type annotation can be complex because it may include combinations of many acoustic features (e.g. the pitch movement direction, range of the pitch change, pitch peak position).

With a view to simplifying the annotation of the pitch accents, only two features have been taken into account: the direction of the pitch movement and its position with respect to the accented syllable boundaries (Fig. 1). The resulting inventory of the pitch accent labels includes: two labels reflecting the pitch movement direction, i.e. the falling intonation (HL) and rising intonation (LH). In both cases the movement is realized on the post-accented syllable and the maximum/minimum occurs on the accented syllable. Another three labels also reflect the pitch movement direction (falling, rising and level), but the pitch movement is fully realized on the accented syllable. The level accent is realized by duration. A special label describes the rising-falling intonation on the accented syllable (RF).

3. Stylization of intonation – towards automatic description of intonation

3.1. Assumptions

The main assumption of the method is that the intonation phrase is a sequence of intonation events, i.e. of pitch accents and boundary tones. Stretches of unaccented syllables between subsequent accented syllables are called connections. This approach is very similar to those used in the Tilt [26] and PaIntE [18] models of intonation.

We use three different types of approximation functions: R – rising, F – falling, and C – connection. The R and F functions can be either straight lines or parabolas (with a positive and negative slope, respectively), C is always a straight line. In contrasta to the Tilt and PaIntE models, we parameterize F0 curves not only on accented syllables and phrase boundaries but also on postaccented syllables.

The functions are given by the following equations (γ is the shape coefficient of the curve):

$$0 < x < 1 \qquad y = x^{\gamma},$$
$$1 < x < 2 \qquad y = 2 - (2 - x)^{\gamma}.$$

The result of the stylization the F0 curves on accented and post-accented syllables and phrase boundaries are described by a set of acoustic parameters: the slope (describing the steepness of the F0 curve), Fp (F0 value at the point of the alignment of the approximation function), the range of the F0 change and the shape coefficient of the curve (further referred to as the bend). The stylization accuracy can be evaluated by measuring the NMSE (i.e. the normalized mean square error) between the original and stylized F0 contours.

3.2. Speech material

The speech material used in the experiments consisted of 1000 phrases read by female and male speakers. They have been taken from a chapter of M. Bulhakow's "The Master and Margaret" (translated by I. Lewandowska and W. Dąbrowski) which included various sentence modes and a large amount of dialogues. The speakers were instructed to read in a moderate tempo and to convey the emotional load of the text. The recordings were done in a professional radio studio. The resulting material presents a collection of expressive speech. For each phrase the phonetic, syllabic and word segmentations have been carried out using a special software programme called Annotation Editor and have been manually corrected. Accented syllables received a $A label and the post-accented syllables a $PA label. The suprasegmental annotation was carried out according to the Boss Label Format specifications [3]. Intermediate tones have been labeled as: 2,? or 2,. For minor intonation phrase boundaries the following labels have been used: 5,? (for questions), 5,! (for exclamations) and 5,. (for statements). The pitch was extracted every 10 milliseconds using the ESPS method available in WaveSurfer [23]. Most pitch tracking errors (e.g. faulty F0 values occurring at the transitions from unvoiced to voiced parts and vice versa, and in the context of voiced plosives and fricatives, pitch doubling or halving) were eliminated by setting an appropriate pitch range and an analysis window length. The pitch contours obtained have been smoothed, but some errors still remained (e.g. faulty F0 values affected by laryngelization at the end of phrases, F0 values tracked in unvoiced regions) and had to be manually corrected on the basis of the findings presented in [11].

3.3. The stylization method

A special software program called Pitch Line was developed to perform the F0 stylization. It was built-in in the Borland C++ Builder environment on the Win32 platform. At the input, the program required a wave file, `a.pitch` file including the F0 values, `a.lab` file with the phoneme boundaries, `a.syl` file containing syllable labels and `a.break` file with the boundary tone labels. The stylization using Pitch Line is done semi-automatically and a decision must be made what kind of the functions should be used in the approximation. In the Pitch Line editing window, the approximation function can be chosen by clicking in the appropriate area on the pitch contour pane and controlling the NMSE. The editing window makes it also possible to shift the phoneme boundaries and to store the acoustic parameters, the stylized F0 and the snapshot of the current view. Figure 2 presents a fragment of the editing window of Pitch Line. The upper window contains the waveform; the bottom window presents the original F0 contour (dotted line), the stylized contour (solid line), approximation functions (R, F, C) used for stylization of the intonation events and NMSE. The vertical lines show approximate phoneme boundaries, the accented syllables are marked in grey, the bottom window includes the SAMPA transcription of the utterance.

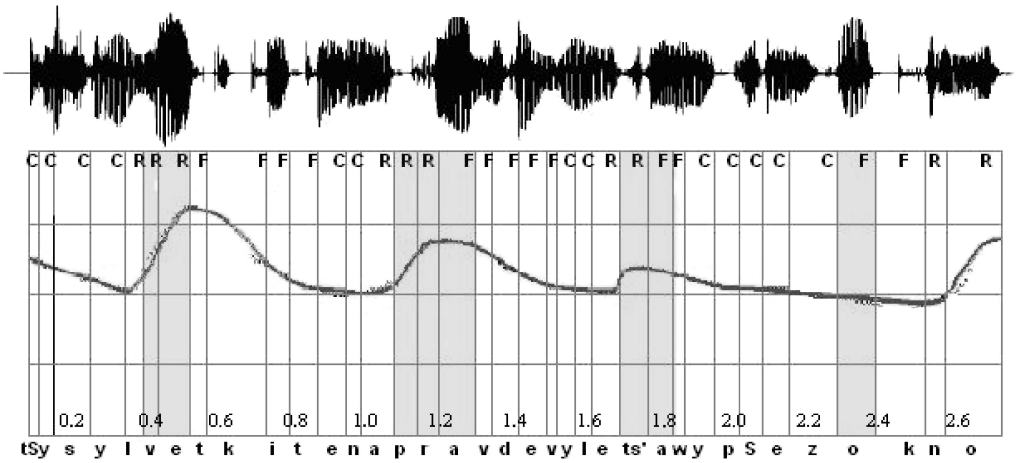


Fig. 2. The sentence: "Czy sylwetki te naprawdę wyleciały przez okno?" (Whether the figures really flew out of the window?), female speaker.

3.4. Stylization rules

Special attention has to be paid to very accurate pitch modeling on vowels because they are regarded as basic units of perception of the speech melody [15]. Pitch accents are associated with accented syllables, but may span over the syllable boundaries ([9, 12, 17, 18, 26]). Therefore the stylization rules should allow to locate the starting and ending points of the approximation functions before/after the syllable start and before/after the syllable end. The pitch on the post-accented syllables can be modeled by two different functions (e.g. as a sequence of R, F or C according to the inventory of 6 pitch accents). We introduced the L (level) intonation event. It was associated only with accented syllables and realized by a flat pitch with a near zero slope. The treatment of these level accents, on which no F0 change on the syllable can be observed but which acoustic feature give the perception of accentuation, needs further investigation.

This enables to model pitch peaks which occur on the post-accented syllables. It is debatable whether pitch accents of different peak positions belong to different phonological classes or not. Nevertheless, listeners are sensitive to whether the peak is early or late [10]. Therefore the stylization method should be able to capture this phenomenon, but at the same time it should be sensitive to microprosody. In the previous stylization scheme, the boundary tones did not constitute a separate class of the intonation events as it is in other intonation models ([2, 21]) and were stylized according to the same rules as the post-accented syllables. However such differentiation should be introduced because boundary tones are manifested by pitch movements independent of the pitch accents and have discourse functions (e.g. signaling the sentence mode). Boundary tones can be modeled not only by the F and R functions and their combinations, but also by straight lines. The stylization rules given above are based on the findings presented in the literature on intonation stylization and modeling, and on previous results of the evaluation of the stylization quality (mentioned at the beginning of this section).

3.5. Results

The stylization accuracy was evaluated in an objective manner by measuring the NMSE between the original and stylized F0 contours and in a subjective manner by a perception study. The parameterization was carried out for 750 accented syllables, 526 post-accented syllables and 298 boundary tones in the female voice speech database, and for 791 accented syllables, 545 post-accented syllables and 370 boundary tones in the male voice speech database. For both the speakers a very low NMSE (mean = 0.003) for each of the acoustic parameters was obtained. This suggests that the stylization was very accurate. However, the correspondence of the low NMSE between the original and stylized F0 contours and their perceptual similarity is not straightforward ([10, 15]). Therefore an informal perception test was carried out, in which five subjects (with no linguistic background) participated. They listened to the original and stylized phrases and assessed not only their perceptual similarity but also the naturalness of the stylized phrases. According to the listeners, all the stylized phrases sounded very natural and were very similar to the original phrases, insignificant differences (e.g. in the pitch height at the end of the phrase) were perceived only in 5% of the phrases and had no influence on the meaning. In the previous perception test [7], only 64% of the stylized phrases read by the male speaker were assessed as identical to the original phrases, which proves that the revised stylization rules are very efficient in extracting the macroprosodic component of the F0 contours.

4. Clustering of intonation events

On the basis of the acoustic parameters obtained in the stylization of the female voice speech database a classification of pitch accents and boundary tones was carried out. The k-means clustering algorithm available in *Statistica* was used for this task. The method requires the user to define the number of clusters. On the basis of analyses of Polish intonation in ([6, 13, 22]) and a visual analysis of the pitch accents present in the current speech material, six clusters were determined. The resulting classification had a lowest variance among objects within the same group and a maximal variance among the groups. All the acoustic parameters significantly differentiate the groups.

Figure 3 shows the prototypical pitch accents found in each cluster; Table 1 presents the mean values of the parameters in each cluster.

Table 1. Mean values of the acoustic parameters for each of the pitch accent clusters (slope Hz/s, range Hz, bend-shape curve).

	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**p**
SLOPE	327.45	128.9	608.63	1145.64	−113.34	−407.005	0.000
RANGE	60.08	23.56	88.57	130.69	−25.74	−83.265	0.000
BEND	1.88	2.35	1.79	1.54	2	1.484	0.003

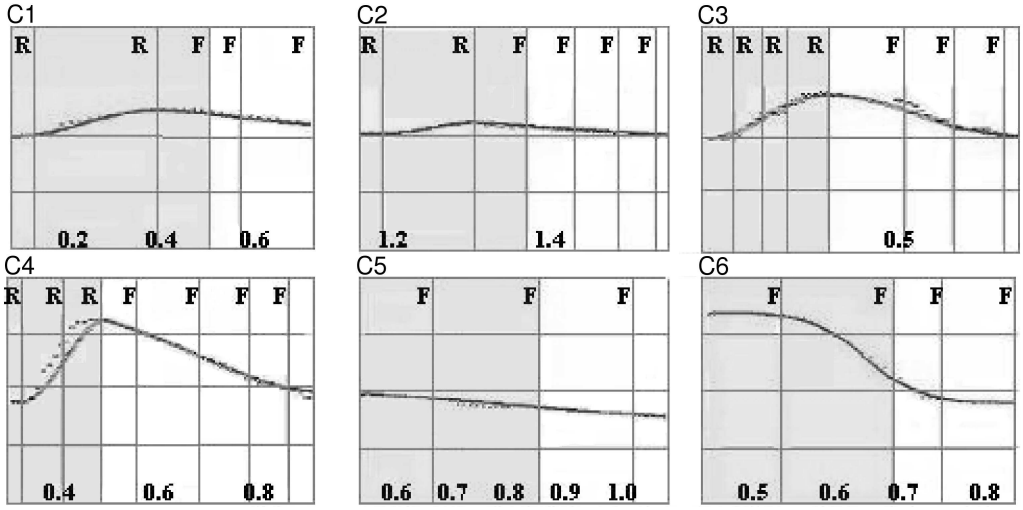


Fig. 3. Pitch accent classes. Accented syllables are marked in grey.

Pitch accents grouped in the cluster 1 (C1, 176 instances) are characterized by a gentle rise followed by a gentle fall starting on the accented syllable and by a significant range of the F0 change. They usually occur at the beginning or in the middle of continuation phrases (56%) and statements (ca. 30%). Pitch accents grouped in the second cluster (C2, 202 instances) can be found in the middle or towards the end of continuation phrases (50%), statements (35%) and exclamations (10%). The slopes of the rising and falling F0 curves are very gentle as the range of the F0 change is small. Cluster 3 (C3) also groups rising (or rising-falling) pitch accents (110 instances). They are realized by a very steep rising F0 curve followed by a steep fall starting on the accented syllable or on the post-accented syllable. These pitch accents almost always occur at the beginning of continuation phrases (55%), statements (20%) and exclamations (25%).

27 rising pitch accents have been classified in the Cluster 4 (C4). They have a very steep F0 slope and are followed by a very steep fall on the post-accented syllable; there is also a very big range of the F0 change. 37% of these pitch accents occur at the beginning of exclamations. Pitch accents grouped in the cluster 5 (C5, 183 instances) occur at the end of statements (48%) and are realized by a gently falling F0 curve on the accented syllables and followed by a very gently falling boundary tone. A smaller percentage of these pitch accents occurs at the end of continuation phrases (35%) and questions (5.4%, see Fig. 4, C4). The cluster 6 (C6) has 51 members; these are sharply falling pitch accents of a final position in exclamations and continuation phrases (see Fig. 4, C1). Depending on the phrase type, they may be followed by a falling or rising pitch. The k-means clustering presented in this section resulted in a pitch accent classification similar to that one presented in [6] and the pitch accents shown in Fig. 2 could receive proper labels as defined in [6], e.g. pitch accents in C1 could be labeled LM, the pitch accents grouped in C4 could receive the LH label. Boundary tones were grouped into

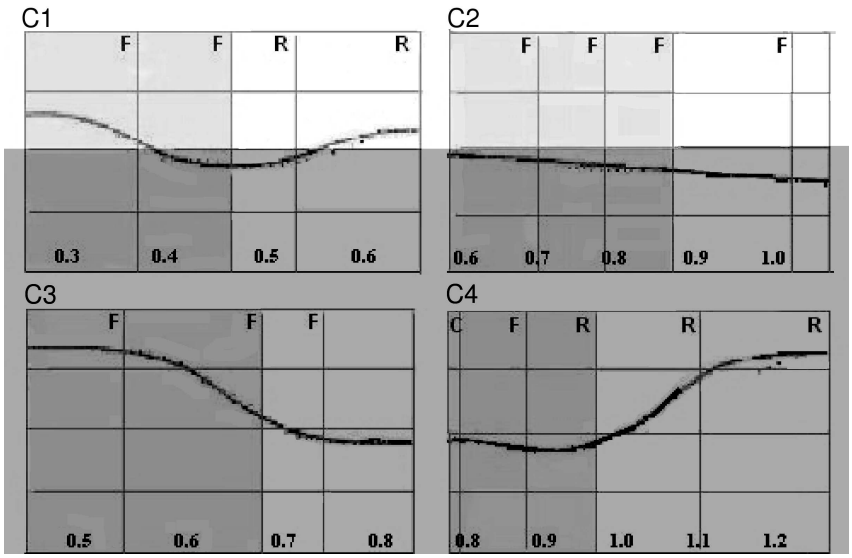


Fig. 4. Boundary tone classes. Accented syllables are marked in grey.

5 classes. The means of the parameters in each group are given in the Table 2. Figure 4 shows the prototypical boundary tones found in each cluster.

Table 2. Mean values of the acoustic parameters for each of the boundary tone clusters (slope Hz/s, range Hz, bend-shape curve).

	C1	**C2**	**C3**	**C4**	**p**
SLOPE	234.79	−64.06	−408.39	607.59	0.00
RANGE	64.55	−21.9	−124.15	121.93	0.00
BEND	2.32	1.67	3.04	1.88	0.00

The 95 members of the first cluster (C1) are characteristic of continuation phrases. The second cluster (C2) counts 106 members which usually occur at the end of statements (45%) and sometimes exclamations (21%). The boundary tones grouped in the third cluster (C3, 37 instances) are characteristic of exclamations (54%) and statements (30%). All questions were realized by boundary tones from the cluster 4 (C4); they may also signal continuation. The results of the k-means clustering of the intonation events presented in this section will be verified in detailed acoustic and statistic analyses carried out on a bigger speech database with the purpose of establishing a phonological description of Polish intonation.

5. Conclusions and future research

This paper presents methods of segmental and suprasegmental annotation of speech databases and stylization and parameterization of the intonation contours. The annotation method will be verified while testing the quality of the speech synthesis in the Polish module of the BOSS text-to-speech system. The proposed stylization method brought good results: quantitative and qualitative differences between the original and stylized phrases were insignificant. The classification of the intonation events gave results similar to those presented in the literature on Polish intonation ([6, 13, 22]). This proves that the parameters resulting from the approximation of the intonation contours are adequate for the description and differentiation of classes of pitch accents and boundary tones. This will be the subject of future research aimed at the establishment of a phonological description of intonation on the basis of Pitch Line phonetic-acoustic representation. In the near future the issues of pitch normalization across speakers and voice quality phenomena will be dealt with. Another task will be to develop the prosody prediction module based on the Pitch Line prosody annotation. Prosody prediction will be partly rule-based and partly data-driven (i.e. using decision trees or neural networks). The results of such comprehensive analysis of the intonation tunes in Polish may be used directly, above all, in the ASR and TTS systems.

Acknowledgments

This research has been carried out under grant 3T11C 038 27 received from the Polish Ministry of Scientific Research and Information Technology.

Appendix

SAMPA	ORTOGRAPHY	TRANSCRIPTION	SAMPA EXTENSION	ORTOGRAPHY	TRANSCRIPTION
VOWELS					
i	pit	pit	i	PIT	
I	*typ*	*tIp*	*y*	*typ*	*typ*
e	test	test	e	test	test
a	pat	pat	a	pat	pat
o	pot	pot	o	pot	pot
u	puk	puk	u	puk	puk
e~	gęś	ge~s’			
o~	wąs	vo~s			
CONSONANTS					
p	pik	pik	p	pik	pik
b	bit	bit	b	bit	bit

t	test	test	t	test	test
d	dym	dIm	d	dym	dym
k	kit	kit	k	kat	kat
g	gen	gen	g	gen	gen
			c	kiedy	cedy
			J	giełda	Jewda
f	fan	fan	f	fan	fan
v	wilk	vilk	v	wilk	vilk
s	syk	sIk	s	syk	syk
z	zbir	zbir	z	zbir	zbir
S	szyk	SIk	S	szyk	Syk
Z	żyto	ZIto	Z	żyto	Zyto
s’	świt	s’vit	s’	świt	s’fit
z’	źle	z’le	z’	źle	z’le
x	hymn	xImn	x	hymn	hymn
ts	cyk	tsIk	*tˆs*	cyk	*tˆsyk*
dz	dzwon	dzvon	*dˆz*	dzwon	*dˆzwon*
tS	czyn	tSIn	*tˆS*	czyn	*tˆSyn*
dZ	dżem	dZem	*dˆZ*	dżem	*dˆZem*
ts’	ćma	ts’ma	*tˆs’*	ćma	*tˆs’ma*
dz’	dźwig	dz’vik	*dˆz’*	dźwig	*dˆz’wig*
m	mysz	mIS	m	mysz	myS
n	nasz	naS	n	nasz	naS
n’	koń	kon’	n’	koń	kon’
N	pęk	peNk	N	pęk	peNK
l	luk	luk	l	luk	luk
r	ryk	rIk	r	ryk	ryk
w	łyk	wIk	w	łyk	wyk
j	jak	jak	j	jak	jak
			w~	ciążą	*tˆs’*o w~Za
			j~	więź	vjej~s’

Computer coding convention

The computer coding conventions were drawn up in SAMPA for Polish created by J.C. Wells 1996 (SAMPA, homepage) with the revisions and extensions and from the IPA alphabet (IPA homepage).

The purpose is:

1. For the vowel y instead /I/ using script /y/ – practical reason – mark y is more readable.
2. For the palatal phonemes k (as in kiedy) and g (as in giełda) – using as in IPA scripts – accordingly /c/ and /J/.
3. For the affricates c, ć, dz, dź, dż, cz, using grouping mark ˆ as in: /tˆs/, /dˆz/, /dˆZ/, / tˆs’/, /dˆz’/, /tˆS/ – uniform description – practical reason. The clusters “dz”, “dź”, “dż” means affricates, but in some words they denote two sounds.
4. For the nasal phonemes ą (as in ciążą) and ę (as in więź) – using as in IPA scripts – accordingly /w~/ and /j~/.

References

[1] ADELL J., BONAFONTE A., *Towards phone segmentation for concatenation speech synthesis*, 5th Speech Synthesis Workshop, Pittsburgh 2004.

[2] BECKMAN M. E., AYERS ELAM G., *Guidelines for ToBI labelling*, available at: http://www.ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf

[3] BREUER S., STOBER K., WAGNER P., ABRESCH J., *Dokumentation zum Bonn Open Synthesis System BOSS II*, Unveroffentliches Dokument, IKP, Bonn 2000.

[4] BREUER S., FRANCUZIK K., DEMENKO G., *Analysis of Polish Segmental duration with CART*, Proceedings of Speech Prosody 2006, pp. 137–140, Dresden 2006.

[5] CLEMENTS G. N., KEYSER S. J., *A three-tiered theory of the syllable*, Technical Report, Massachusetts Institute of Technology, 1981.

[6] DEMENKO G., *Analysis of Polish suprasegmentals for needs of speech technology*, edited by UAM, Poznań 1999.

[7] DEMENKO G., WAGNER A., *The stylization of intonation contours*, Proceedings of Speech Prosody 2006, Dresden, pp. 141–144, 2006.

[8] DEMENKO G., WYPYCH M., BARANOWSKA E., *Implementation of grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis*, Speech and Language Technology, ed. PTFON, **7,** 79–97, Poznań 2003.

[9] FUJISAKI H., *Dynamic characteristics of voice fundamental frequency in speech and singing*, [in:] The Production of Speech, P.F. MacNeilage [Ed.], pp. 39-47, Springer-Verlag, 1983.

[10] T'HART J., COLLIER R., COHEN A., *A perceptual study of intonation*, Cambridge University Press, Cambridge 1990.

[11] HESS W., *Pitch determination of speech signals*, Springer Verlag, New York 1983.

[12] HIRST D., VÉRONIS J., IDE N., *Analysis of fundamental frequency patterns for multi-lingual synthesis using INTSINT*, Proceedings of 2nd ESCA/IEEE Workshop on Speech Synthesis, pp. 77–80, New Paltz, September, 1994.

[13] JASSEM W., *Fundamentals of the Polish phonetics* [in Polish], PWN, Warszawa 1973.

[14] MATOUSEK J., TIHELKA D., PSUTKA J., *Automatic segmentation for Czech concatenative speech synthesis using statistical approach with boundary-specific correction*, Proceedings of Eurospeech, 2003.

[15] MERTENS P., *The prosogram: Semi-automatic transcription of prosody based on a tonal perception model*, B. Bel and I. Marlien [Eds.], Proceedings of Speech Prosody 2004, pp. 549–552, Nara, Japan 2004.

[16] MILONE D. H., RUBIO A. J., *Prosodic and accentual information for automatic speech recognition*, Proceedings of IEEE, **11,** 4, 321–333 (2003).

[17] MIXDORFF H., *A novel approach to the fully automatic extraction of Fujisaki model parameters*, Proceedings of ICASSP 2000, **3,** 1281–1284, Istanbul 2000.

[18] MÖHLER G., *Describing intonation with a parametric model*, Proceedings of ICSLP98, pp. 2581–2584, Sydney 1998.

[19] NARAYANAN S., ALWAN A., *Text to speech synthesis, new paradigms and advances*, IMSC Press Multimedia Series, New Jersey 2004.

[20] OSTENDORF M., DOGALAKIS V. V., KIMBALL O. A., *From HMM's to segment models: A unified view of stochastic modeling for speech recognition*, IEEE Trans. on Speech and Audio Proc., **4**, 5, 360–378 (1996).

[21] PIERREHUMBERT J., *The phonology and phonetics of English intonation*, PhD dissertation, MIT, 1980.

[22] STEFFEN–BATOGOWA M., *Accentual structure of Polish* [in Polish], Wydawnictwo Naukowe PWN, Warszawa 2000.

[23] SJOLANDER K., BESKOW J., *WaveSurfer – An open source speech tool*, Proceedings ICSLP'00, Beijing, **4,** 464–467, 2000.

[24] SZYMANSKI M., GROCHOLEWSKI S., *Semi-automatic segmentation of speech: manual segmentation strategy; problem space analysis*, [in:] Advances in Soft Computing, Computer Recognition Systems: Proceedings of 4th Int. Conference on Computer Recognition, M. Kurzyński [Ed.], pp. 747–755, Springer Verlag, 2005.

[25] SZYMANSKI M., GROCHOLEWSKI S., *Transcription-based automatic segmentation of speech*, Archives of Control Sciences, **15**, 465–472 (2005).

[26] TAYLOR P., *Analysis and synthesis of intonation using the tilt model*, J. Acoust. Soc. Am., **107**, 3, 1697–1714 (2000).

[27] WELLS J. C., *SAMPA computer readable phonetic alphabet*, [in:] *Handbook of Standards and Resources for Spoken Language Systems* (Part IV, Section B), D. Gibbon, R. Moore and R. Winski [Eds.], Mouton de Gruyter, Berlin and New York 1998;
Available at: www.phon.ucl.ac.uk/home/sampa/polish.htm