Maria Cernanska – Ondrej Skvarek *

CLUSTERING OF SLOVAK SENTENCE MELODY – METHODS AND RESULTS

Modern speech synthesis systems implement prosody features to achieve more naturally sounding voices, trying to produce sounds similar to the human speech. This article deals with obtaining one of prosody characteristics (sentence melody contour) from human speech recordings. Similarities of melody contours are studied using cluster analysis. Several clustering methods are evaluated for this purpose, estimation of proper number of clusters is described and typical melodies for different types of Slovak language sentences (declarative, interrogative, exclamatory, clauses with final "," etc.) are found and recommended for implementation in our text-to-speech system.

Keywords: speech synthesis, prosody, melody contour, cluster analysis, R-software

1. Introduction

Progress in the area of speech processing enables new ways of communication between people and computers. Reading a text from the computer screen can be extended or replaced by a text to speech synthesis, writing a text on the keyboard or choosing commands using the mouse can be enriched by voice recognition systems. A text-to-speech synthesizer (TTS) converts a text from documents into audio voice files. One of such synthesizers (TTS-KIS) [10, 3] is developed in the Department of Information Networks at the Faculty of Management and Informatics of the University of Zilina.

Our TTS system implements a concatenative method of speech synthesis. The method is based on concatenation of speech elements "diphones" [3]. Diphones are selected from the TTS sound database at the time of speech synthesis.

Diphones in the database are stored in a normalised form with monotonous melody. Consequently a monotonous waveform corresponding to the input text is produced, and hence a proper melody modification is needed to achieve a naturally sounding sentence.

Our TTS system applies composed melody contour to the synthesized sentence waveform. The contour is composition of "sentence melody contour" (long-term melody trend spanning along the whole sentence or clause) and of "short-term melody contours" (related to shorter speech segments, words, bars, syllables). Our present work is focused on obtaining the long-term "sentence melody contour".

We use terms "word" and "bar" interchangeably to denote the word with neighbouring proclitics and enclitics. Terms "sentence" and "clause" are interchangeably used to denote smaller sentence parts of simple, complex or compound sentences separated by punctuation marks and conjunctions (see Table 2).

The TTS-KIS system analyses text of each input sentence and chooses melody contour corresponding to the characteristics found in the text: the sentence type (recognized by punctuation marks, interrogative words ...) and the number of words (bars) in the sentence. Our aim is to find typical "sentence (clause) melody contour" for each group of sentences characterised by the sentence type and by the number of words. Sentence types recognised by the TTS system are listed in Table 3.

Proposed method obtains original sentence melody contours using software Praat [11].

Then weighted-MA method [4] defined in formula (1) is used to remove short-term melody variations and to obtained sentence melody contour values. The smoothed values are computed in equidistant time events along the whole sentence (clause) waveform. The number of melody contour values is set proportionally to the number of words (bars) in the sentence. The number is equal for all sentences in the same sentence group, which makes comparison of different length waveforms possible.

Sentence melody contours are further centred to 0 Hz mean frequency value. Contour dynamics remains unchanged.

Similar melodies are grouped into separate clusters (the best clustering method for this purpose is found). Cluster with melodies best suited for analysed sentence type is selected and typical melody of this cluster is recommended for our TTS system.

* **Maria Cernanska, Ondrej Skvarek**
Department of Information Networks, Faculty of Management and Informatics, University of Zilina, Slovakia,
E-mail: maria.cernanska@fri.uniza.sk

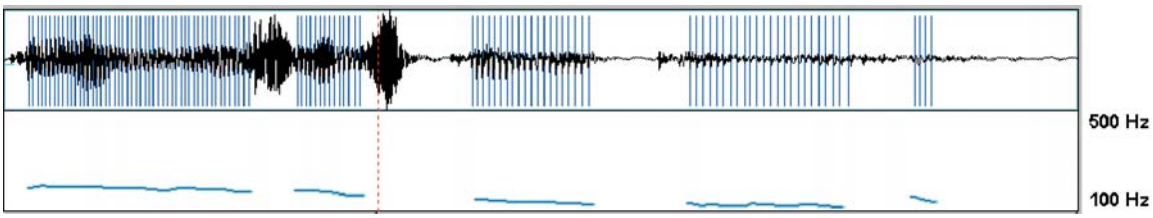


Fig. 1 Waveform (upper part) and melody contour (lower part) of sentence „Lenze zostalo to v nej.“

2. Speech material

We examined sound recordings of the novel [16]. The story was narrated by a female speaker in a faster speaking rate with an emotive accent (noticeable changes in the intonation, loudness and tempo).

Recordings were cut into smaller parts corresponding to the clauses. See Table 2 for punctuation marks and conjunctions used to determine clause boundaries. (Simple sentence and parts of complex or compound sentences were taken as separate clauses.)

Sound files were assigned names corresponding to the page number, sentence number, number of bars in the sentence and the type of the starting and ending punctuation marks. About 8000 sentence sounds (see Table 1) of the PCM format (22050 Hz, 16 bit/sample, mono) were stored in the 500 MB disk space.

3. Obtaining melody contours from voice recordings

We use the program Praat [11, 2] to obtain the glottal frequency contour F0. Praat implements a normalised autocorrelation function and best path selection algorithms. Proper settings of parameters are needed to obtain real F0 values [4]. To process the large number of voice records, the script of Praat commands was programmed.

4. Preparing melody contours for cluster analysis

Melody contours obtained by the program Praat are composed of equidistant values located inside detected voiced intervals (see Fig. 1).

To compare melodies of sentences with different time duration and with different distribution of voiced intervals, a smoothing method “weighted moving average” is used [4]. Weighted-MA values $y'(t)$ are computed according to the formulas (1) and (2).

$$y'(t) = \frac{\sum_{i=1}^{n} y_i * w_i}{\sum_{i=1}^{n} w_i} \tag{1}$$

where

- n denotes the number of original melody values taking into computation ($n/2$ values to the left and $n/2$ values to the right from the time event t),
- y_i denotes the original melody values,
- t_i denotes corresponding time events,
- w_i denotes weight assigned to y_i values, see the formula (2).

$$w_i = a * |t - t_i| + b \tag{2}$$

Weights decrease with raising distance from the time event t. Parameters were set to values $n = 12$, a $= -2$, b $= 1$.

Number of analyzed n-member clauses — Table 1

Ending punctuation mark or conjunction	n														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	∑
a, i, aj, ani, či	87	154	201	191	115	67	55	21	2	5	1	1	0	0	900
„ . “	141	388	519	489	425	300	155	88	27	19	15	5	2	1	2574
„ , “	502	607	582	540	373	201	112	52	21	12	5	1	2	0	3010
„ : “	4	10	16	9	2	2	0	0	0	0	0	0	1	0	44
„ ... “	145	118	124	144	108	85	40	20	9	7	2	0	0	1	803
„ ? “	54	78	76	65	31	24	10	3	2	0	0	0	0	0	343
„ ?! “	7	8	3	10	8	3	0	1	0	0	0	0	0	0	40
„ - “	16	4	5	3	2	0	3	0	0	2	2	0	0	0	37
„ ! “	37	30	33	28	18	11	6	1	0	0	0	0	0	0	164

The method removes short-term variations from the overall sentence melody and computes "smoothed sentence melody contour".

We compute melody values in equidistant time intervals. The number of contour points is set to 8-th multiply of number of words in the sentence. Each contour is further centred to 0 Hz mean frequency value. Dynamics of the contour remains unchanged.

5. Description of clustering methods

Once melody contours are represented by equal length vectors (values determined in equidistant time intervals) the cluster analysis of melody contour similarities is possible. Clustering methods found in the R-software's "hclust" method [12] were taken and fitness for melody contour clustering was examined. All the methods (ward, single, complete, average, mcquitty, median, centroid) perform hierarchical agglomerative clustering. Initially, each object (contour) is assigned to its own cluster, then (iteratively) distances between clusters are computed and the two closest clusters are joined.

Initial distances between clusters (dissimilarity matrix) are computed by the R-software's "dist" method as the Euclidean (geometric) distance.

The second and further iterations of a dissimilarity matrix are recomputed by the Lance–Williams dissimilarity update formula [1, 6]. The formula parameters are set to values [9] according to the method chosen (ward, single ...).

Each of the mentioned methods evaluate the distance between clusters in a different way, hence clusters of different properties are produced.

Single Linkage method In this method, the distance between two clusters is computed as the *smallest distance* of two objects, where the first object belongs to the first cluster and the second object belongs to the second cluster. Resulting clusters tend to represent long, straggly "chains" [13].

Complete Linkage method The distance between two clusters is computed as the *largest distance* of two objects, where the first object belongs to the first cluster and the second object belongs to the second cluster. This method tends to find extremely compact clusters [5]. The method is usually suitable when the objects form naturally distinct clusters. If the clusters tend to be somehow elongated or "chain" type nature then this method is inappropriate [13].

Average Linkage method The distance between two clusters is computed as the *average distance* of two objects, where the first object belongs to the first cluster and the second object belongs to the second cluster. The method tends to find globular clusters [5]. It is very efficient when the objects form natural distinct clusters and performs equally well with elongated "chain" type clusters [13]. This method is also called "the group average linkage algorithm" or "the unweighted pair group method average" (UPGMA) [14].

McQuitty's method In this method, when two clusters are merged into a new one, the distance from the new cluster to the old one is computed as an average of distances between two merged clusters and the old cluster [14]. Such rule corresponds to the weighted average computation, where *objects in small clusters have a larger weight* than those in large clusters. This method is also known as "the weighted average linkage algorithm" or "the weighted pair group method average" (WPGMA). This method (rather than the previous method) should be used when the cluster sizes are suspected to be greatly uneven [13].

Centroid Linkage method In this method, centroid (average of objects) of each cluster is computed, then the distance between two clusters is determined as the distance between centroids representing the two clusters.

Median Linkage method This method extends Centroid method by weights to consider different numbers of objects in clusters. This method is preferable to the Centroid method, when considerable differences in cluster size are expected [13].

Ward's minimum variance method In this method, the distance between two clusters is evaluated as the growth of total dispersion of objects around their respective cluster centroids. This method minimises clusters heterogeneity [8] and it tends to find globular clusters [5]. In general, it is regarded as very efficient [13].

6. Selection of the best method for melody contour clustering

We used two criteria to select the best method for melody contour clustering:

1. Correct separation of melody contours according to their similarities (similar melodies grouped in the same cluster, different melodies separated into different clusters).
2. Minimal number of clusters needed to accomplish the criterion 1.

Seven clustering methods (ward, single, complete, average, mcquitty, median, centroid) were evaluated. 4-member (4-word) declarative and interrogative sentences (clauses) were taken and clustering was computed by each of the methods (see corresponding dendrograms in Figs. 2 and 3). For each case criteria were evaluated and the method that best met the criteria was recommended to cluster all other sentence groups.

We evaluated Criterion 1 using:

- dendrograms obtained by clustering
- drawings of melody contour clusters
- hearing sounds grouped into clusters

Dendrograms depict inter-cluster distances and the size of clusters formed during computations. Figs. 2 and 3 shows that the

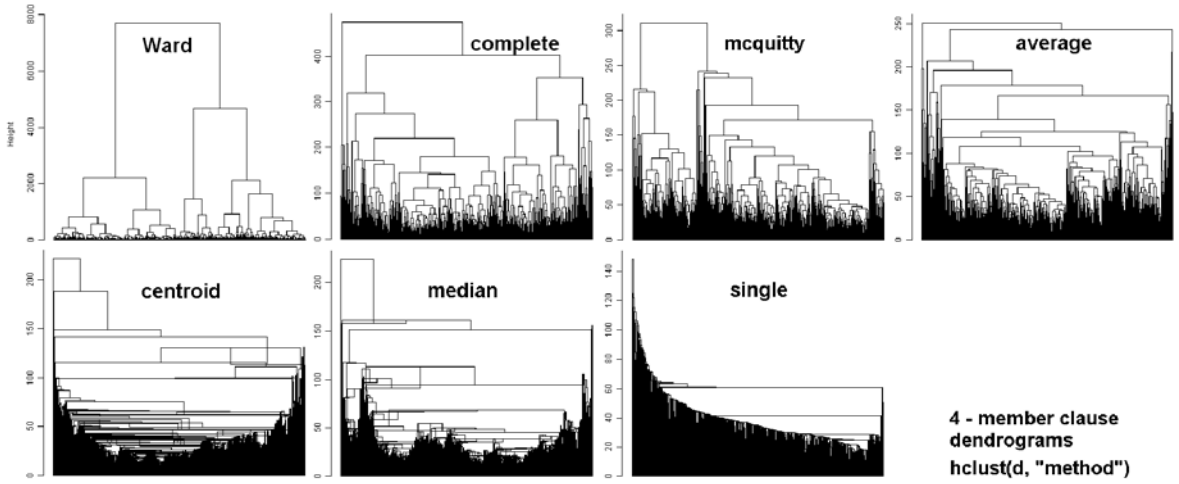


Fig. 2 Declarative 4-member clause dendrograms (various clustering methods)

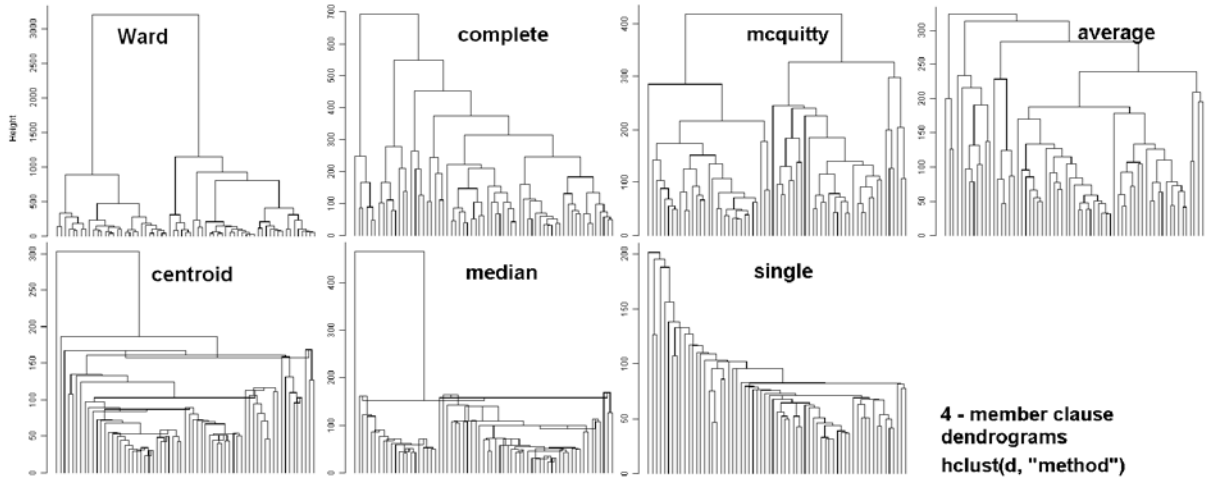


Fig. 3 Interrogative 4-member clause dendrograms (various clustering methods)

Ward's method tend to create clusters of similar size. Comparing these clusters to Fig. 4 we can see good separation of different melody shapes corresponding to these clusters. On the other hand the Single method separates very small clusters while retaining different melody shapes in one large cluster – not capable to meet the Criterion 1. The other methods exhibit properties between the Ward's and the Single method.

We investigated melody contour drawings of two, three and more clusters (starting from the top of the dendrogram). We stopped at the number of clusters when further cluster division would not give new clusters with significantly different melody shapes. Obtained number of clusters is taken as the Criterion 2. Proper separation of melody contours we verified by hearing of sounds from corresponding clusters.

Ward's method created clusters of similar size (see Figs. 2, 3 and 4) and postpone separation of single or small groups of contours to the later iterations. The proper melody separation was achieved at the number of clusters equal to five (Criterion 2 value). The *Complete method* put small groups of melodies into separate clusters in earlier iterations. In the case of interrogative clauses, at the moment of five clusters, different kinds of melodies (falling and rising) were still included in the same cluster. *McQuitty's method* kept two large clusters while separating small size clusters. The *Average method* separated even smaller clusters, still keeping one or two large clusters. The *Median method* separated objects into clusters of a very small size. The *Centroid method* separated melodies nearly one-by-one. The *Single method* also separated single objects; keeping dissimilar melodies in one large cluster (compare Figs. 2, 3 and 4).

The *Ward's method* was chosen as the most suitable for clustering melody contours. It performs *proper separation of melody contours* using the smaller number of clusters. All the methods

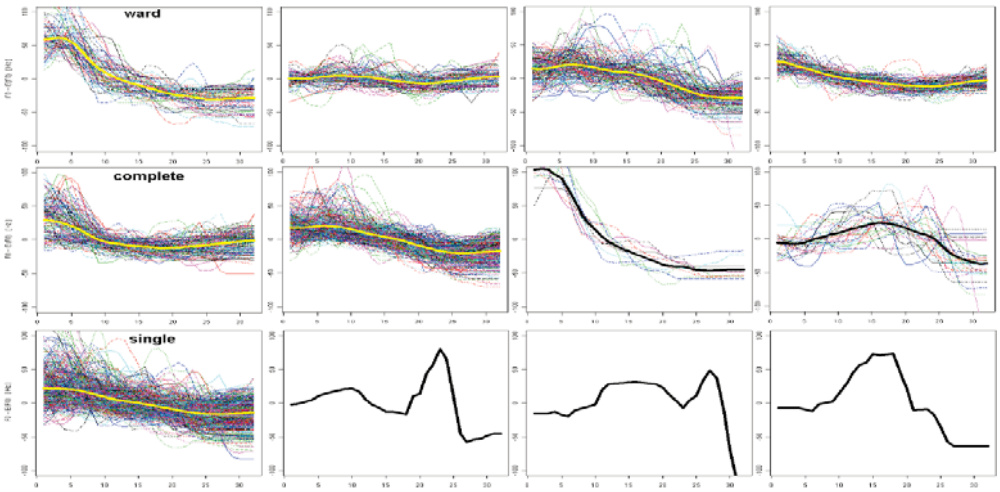


Fig. 4 Clusters of 4-member declarative clauses computed by three clustering methods: Ward's, Complete and Single method.

were ordered according to the suitability: Ward's, Complete, McQuitty, Average, Median, Centroid and Single.

7. Identifying melody contours for the TTS system

We used Ward's method (the best clustering method found in the previous paragraph) to compute clustering for all sentence types and different number of words. For each clustering we determined clusters with proper melody separation (also described in the previous paragraph). Then we determined the proper melody for particular groups of sentences using following criteria:

- melody features described by Slovak language scientists in [7]
- sentences uttered in a neutral way (without strong word accent)
- number of contours in the cluster

For example, analysing Slovak 4-clause determination sentences we stopped clustering at five clusters (see Fig. 5, clusters b1-b5). *Cluster b1* exhibits the rise at the beginning of the clause followed by steep melody fall. The hearing of sentences confirmed sentences with very strong accent on the first word in the sentence expressing emotional speech. *Cluster b2* contains flat melody contours, corresponding to repose even phlegmatic atmosphere, without conspicuous emotions. *Cluster b3* exhibits a rise at the middle part. Stress is heard on the third word or on the fourth word. *Cluster b4* has a tendency similar to the cluster b1, with slower decrease. The beginning of the sentence is less stressed, which adds moderate dynamics to the speech. Representative melody contour of this cluster was chosen as the "sentence melody contour" for 4-clause determination sentences in our TTS system (see Table 3). *Cluster b5* contours exhibit a slow rise in the melody at the end of the clause. Hearing it we found that the cluster contains: sentences with

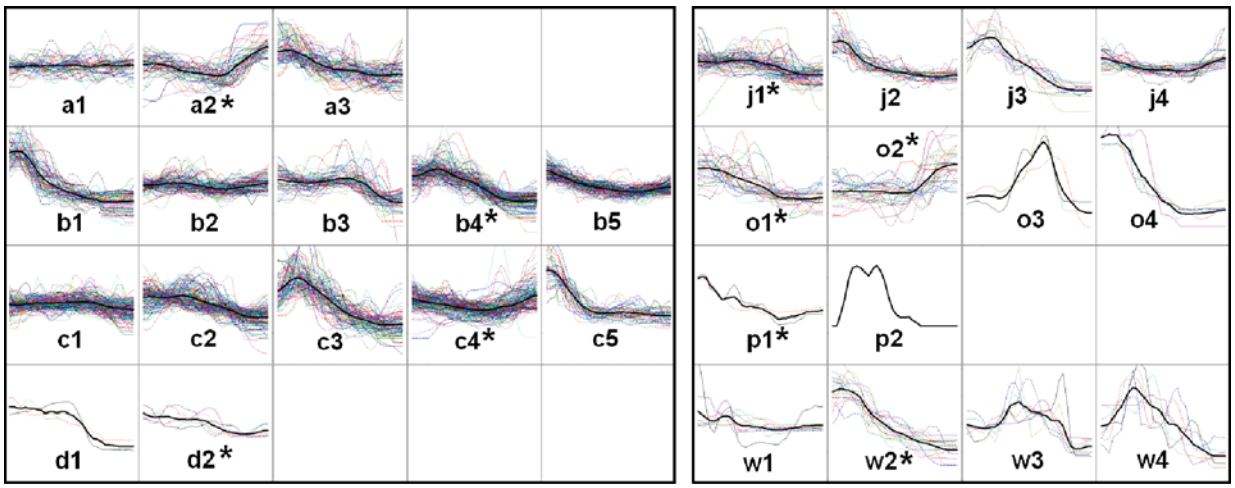


Fig. 5 Results of cluster analyses (Ward's method) of 4-member clauses. Letters "a, b, c, d, j, o, p, w" denote ending of the sentence with different punctuation marks (see Table 2).

Sentence ending labels and corresponding punctuation marks or conjunctions placed at the end of clauses

Table 2

Label	*a*	*b*	*c*	*d*	*j*	*O*	*p*	*w*
Ending punctuation mark or conjuction	"a, i, aj, ani, či, alebo" not preceded by comma mark ","	" . "	" , "	" : "	" ... "	" ? "	" – "	" ! "

melodies of non-ended sentences (similar to melody of comma "," ended sentences), sentences expressing theatricality of the story and sentences with a quiet ending (causing non-precise F0 calculation at the end of the sentence).

The clusters with melodies suitable for TTS systems are marked by asterisk "*" (see Fig. 5). Representative contours (arithmetic mean of contours) of these clusters are recommended for implementation in our TTS-KIS system. Alternation of chosen melodies with more expressive melodies (e.g. alternation of b4 and b1, Fig. 5) can be implemented to achieve more dynamic utterance production (see description of non-satisfying non-ending sentence melodies in [7]).

Another example – clustering of 4-member declarative clauses of different lengths is shown in Fig. 6.

8. Mapping of text characteristics into melodies recommended for TTS-KIS system

The cluster analysis showed the same results (shapes of melody contours) as language scientists described in [7] and we summarized in [15]. We recommend these melody contours for our TTS system and we have prepared mapping of the input text characteristics into recommended contours (e.g. see the 4-member clauses case in the Table 3). At the time of speech syntheses the contour is stretched to the length of the sentence waveform. The stretched contour and short-term melody contours are added to the synthesized sound.

9. Conclusion

In this article we described method of obtaining sentence melody contours for our TTS system. Cluster analysis is used to analyse 8000 melody sentence contours of Slovak language speech recordings. Melody contours were obtained from recordings by software Praat, prepared by the smoothing method and centred to the zero mean frequency. Then cluster analyses were performed. Seven clustering methods (implemented in "hclust" method of R-software) were compared. The Ward's method was chosen as the best one for the purpose of melody contour clustering. This method was used for clustering of different types of sentences with different numbers of words (bars). For each group of sentences clusters with correct separation of melody contours were found and melodies of individual clusters were analysed. The results of the analysis were formulated as recommended melody contours for individual sentence types. Text characteristics recognised by our TTS system were mapped to recommended sentence melodies.

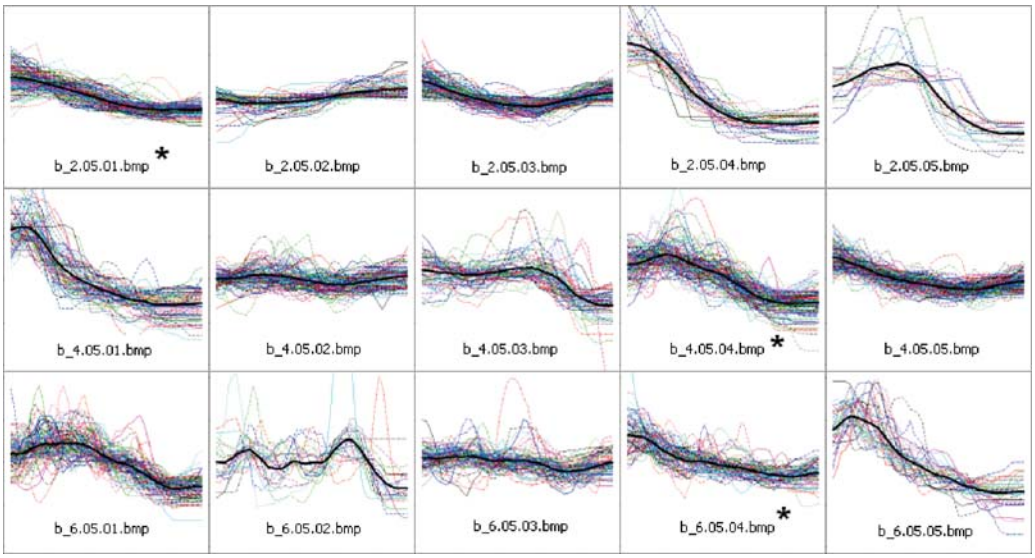
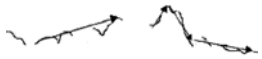
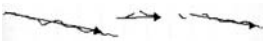


Fig. 6 Ward's method clustering of 2-, 4- and 6-member declarative clauses. Number of analysed sentences: 2-member (607), 4-member (540) and 6-member clauses (201)

Mapping of text characteristics into recommended sentence melodies and melodies expected by language scientists (4-member clauses case) Table 3

Melody type	Sentence type	Characteristics recognised by the TTS system in the input text	4-member clause melodies (Fig. 5) recommended for the TTS system	Examples of melody contours expected by language scientists (see [7, 15])
1. Satisfying ending (conclusive cadence)	Declarative Exclamatory Wh-question ("doplnovacia otazka" in Slovak)	Final " . ", " ! "; " ? " when beginning with an interrogative word	b4, o1, w2	 "Zavolal vsetkych priatelov."
2. Non-satisfying ending (ant-icadence)	Yes–no question (zistovacia otazka) Question to myself (rozvazovacia otazka)	" ? " when it does not begin with an interrogative word	o2	 Ty?
	Alternative question (offers a choice of answer) (rozlucovacia otazka)	" ? " and the word "alebo"	a2 (& o1)	 "Priznavas chybu alebo ju popieras?"
3. Non-satisfying non-ending (semi-cadence)	Rising (stupava) – Flat (rovna) – Raised (zdvihnuta) – Not-raised (nezdvihnuta) – Falling (klesava)	" . "	c4 (& b4)	 "Opravnena je aj otazka, co budeme robit."

We found slight differences between expectations (sentence melodies described by language scientists) and melodies obtained from real speech recordings of the book [16]. The speaker often uses "non-satisfying non-ending" melody for sentences ended with the period, where "satisfying ending" melody was assumed. This attracts attention of the listener by signalling the next action of the story.

The investigation of clusters also showed strong influence of word melody accent even after the smoothing of the overall melody contour. So the melody of shorter speech segments (words, bars, syllables) should be studied.

References

[1] ANDERBERG, M. R.: *Cluster Analysis for Applications,* Academic Press, New York, 1973.

[2] BOERSMA, P.: *Accurate Short-term Analysis of the Fundamental Frequency and the Harmonics-to-noise Ratio of a Sampled Sound*, Inst. of Phonetic Sciences 17, 97–110. Univ. of Amsterdam, 1993.

[3] CAKY P., KLIMO, M., MIHALIK, I., MLADSIK, R.: *Text-to-speech for Slovak Language,* In proc. of the 7th international conference Text, speech and dialogue, TSD 2004, Brno - Berlin: Springer, 2004.

[4] CERNANSKA, M., SKVAREK, O.: *Sentence Melody Analysis for Speech Production in the TTS system.* In proc. of the TRANSCOM 2009: 8th European conference of young research and scientific workers, University of Zilina, 2009, Section 3: Information and communication technologies, p. 25–30.

[5] DOWNS, G. M., BARNARD, J. M.: *Hierarchical and non-Hierarchical Clustering.* Poster presented at Daylight EUROMUG meeting. GlaxoWellcome, Stevenage, 1995. BCI Barnard Chemical Information Ltd. Sheffield S6 6BX.

[6] KOWALSKI, G.: *Information Retrieval Systems,* Theory and Implementation, Kluver Academic Publishers, 1999.

[7] KRAL, A.: *The Pronunciation Rules of Slovak (in Slovak)*, Systematika a ortoepicky slovnik, Neografia, Martin 2005.

[8] MELOUN, M., MILITKY, J.: *Statistical Analysis of Experimental Data (in Czech)*, ACADEMIA, 2004

[9] MURTAGH, F.: *Correspondence Analysis and Data Coding with R and Java,* Chapman and Hall/CRC Press, 2005.

[10] TTS system version 1 demo, http://tts.kis.fri.utc.sk (taken: August 23 2009)

[11] Program Praat main web page, http://www.fon.hum.uva.nl/praat (taken: August-23 2009.)

[12] The R Project for Statistical Computing, http://www.r-project.org/

[13] Electronic Textbook StatSoft, http://statsoft.com/textbook (taken: August -18 2009)

[14] XU, R., WUNSCH, D. *Clustering.* Wiley-IEEE Press, 2009.

[15] CERNANSKA, M., SKVAREK, O.: *Clustering Methods used to Obtain Typical Sentence Melody Contours for Slovak Language TTS system.* Accepted at the conference: The Second International Scientific Conference on Applied Natural Sciences, 2009, Trnava,

[16] KELEOVA, VASILKOVA, T. *Cukor a sol.* Ikar, Bratislava, 2004. (novel written in Slovak language)