CLUSTERING METHODS USED TO OBTAIN TYPICAL SENTENCE MELODY CONTOURS FOR SLOVAK LANGUAGE TTS SYSTEM

ČERŇANSKÁ MÁRIA AND ŠKVAREK ONDREJ

Abstract

The Text-to-speech system developed at our department is based on the concatenative synthesis method. Sound recordings are saved in the sound database in normalized form (constant glottal frequency F0). According to the input text, sound elements (diphones) are chosen from the database and they are concatenated into the monotonous sentence waveform. Then prosody properties (overall melody contour, intensity contour…) are added. In our TTS system [2, 12] the overall melody is composed of sentence melody contour and shorter speech segments melodies.

This paper deals with obtaining typical sentence melody contours for different types of sentences in Slovak language (declarative, interrogative, exclamatory, clauses with final "," etc.).

Sentence melody contours were obtained in our previous work (voice recordings of 8000 sentences, F0-contours obtained by program Praat [10] which uses normalized autocorrelation function, then smoothed by weighted-MA method).

In this work we use data-mining (cluster analysis) to find groups of sentences (clauses) with similar melody contours. Among hierarchical clustering methods, available in R-software program, the Ward's method was chosen. Separate clustering was computed for different types of sentences and 1-, 2-, 3- … and 14-member clauses. Clusters were analysed and proper number of clusters was estimated. For each cluster typical melody contour was computed and mapping of text features to melody types for the TTS system was prepared.

Mathematics Subject Classification 2000: 62H30, 68T10, 91C20
General Terms: text-to-speech, prosody, data-mining
Additional Key Words and Phrases: melody contour, sentence types, clustering, R-software

1. INTRODUCTION

Modern areas of speech processing (human-computer communication, VoIP [1, 5], new sampling methods [4] etc.) include speech synthesis systems. We meet speech synthesis on web pages, computer translators, navigation systems, mobile devices, speaking timetables, e-learning systems [9] etc. The text-to-speech (TTS) system developed at the Department of InfoCom Networks [2, 12, 8, 6, and 11] is one of such systems. It uses concatenative method for speech synthesis. Sound recordings are stored in the TTS database in normalized form (constant glottal frequency). According to the input text, speech elements (diphones) are chosen from the database and they are concatenated into one waveform. Finally, prosody properties (overall melody contour and intensity contour) are added to the monotonous sound.

Our present effort is focused on finding proper melody contours of Slovak language sentences that can be used for speech synthesis.

There are two basic approaches to the sentence melody analysis:

1. Classical approach
2. Data mining approach

The classical approach is often used by language scientists [7]. This approach analyses structure of the sentence to find rules determining sentence melody contours (see paragraph 2).

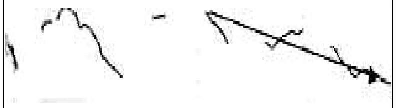
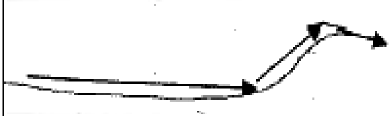
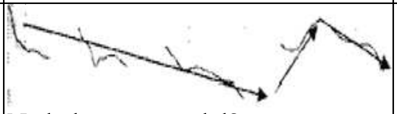
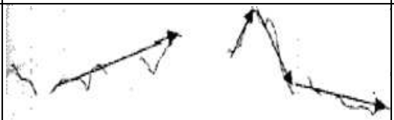
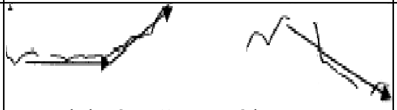
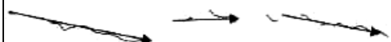
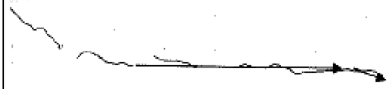
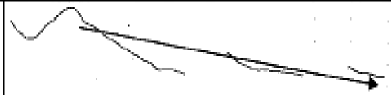
The data mining approach is an opposite approach. Melody contours are extracted from real speech recordings and they are treated as vectors of numbers. Similar vectors are grouped into clusters and common properties of corresponding sentences are searched. Rules determining sentence melody contours are formulated. In this work we use data mining approach (see paragraph 3).

2. MELODY CONTOURS DESCRIBED BY LANGUAGE SCIENTISTS

Melody of human speech is formed by variations of glottal frequency (F0) in time. Language scientists described several basic melody patterns typical for different types of Slovak language sentences and clauses (parts of sentences) [7]. We summarize these melody types in Table I. As mentioned in [7], the most significant part of melody contour is called „melodeme“. Melodeme exhibits outstanding changes in pitch frequency that distinguish different types of sentences. For example, substitution for melodeme can change interrogative sentence into declarative one. Also location of the melodeme can change meaning of the sentence (can inquire different subjects in the sentence).

Melody of sentence can be influenced also by other factors: stress, accent and emphasis. Also these phenomena can be located in different positions in the sentence. (Accent is usually placed in the „core of expression“, where the new information is contained).

Table I. Summary of basic sentence melody types of Slovak language sentences described by language scientists [7].

Melody type	**Sentence type**	**Melody contour example**	**Identified in the text by**
1. Satisfying ending (conclusive cadence)	A. **Declarative Exclamatory Wh-question** (“doplňovacia otázka” in Slovak)	 Zavolal všetkých priateľov.	„.“, „!“, „?“ and the sentence begins with an interrogative word (“čo, kto, ako …”)
2. Non-satisfying ending (anticadence)	A. **Yes–no question** (zisťovacia otázka)	 Ty?	„ ? “ and the sentence does not begin with an interrogative word
	B. **Question to myself** (rozvažovacia otázka)	 Na koho som myslel?	„ ? “ and the sentence does not begin with an interrogative word
	C **Alternative question** (offers a choice of answer) (rozlučovacia otázka)	 Priznávaš chybu alebo ju popieraš?	„ ? “ and the word „ alebo”
3. Non-satisfying non-ending (semicadence)	A. **Rising** (stúpavá)	 Povedal nám, že nepríde.	„ , “
	B. **Flat** **- Raised** **- Not-raised** (rovná - zdvihnutá - nezdvihnutá)	 Oprávnená je aj otázka, čo budeme robiť.	„ , “
		 Urobím to, ak budem môcť	„ , “
	C. **Falling** (klesavá)	 Neviem, čo je to.	„ ,“

3. DATA-MINING - OBTAINING TYPICAL SENTENCE MELODY CONTOURS

Our data mining approach consists of two steps:

1. Melody contour extraction
2. Cluster analysis

In the first step input data are prepared. We extract melody contour from speech recordings by program Praat [10]. Praat uses normalised autocorrelation function and best path selection algorithms to detect glottal frequency contour (Figure 1). Then we remove short-term variations by “weighted moving average” smoothing method [3]. Obtained equidistant melody contour values are prepared for clustering.

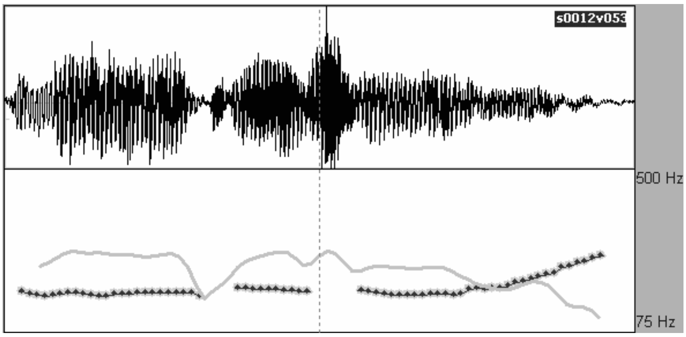


Fig. 1. Melody contour extracted by program Praat (Upper part – waveform of Slovak language sentence “Bolo to zlé?”. Lower part: dotted line – pitch contour, solid line – intensity contour.)

In the second step - “cluster analysis”, we look for groups of similar melody contours and for each group we compute the representative contour of the group. Then waveforms belonging to obtained groups are analysed and melody contours for speech synthesis (needed in the TTS system) are chosen.

At the beginning each contour is normalized to zero mean value. Then “R-software” program is used to compute clustering. R-software with packages offers many clustering procedures. We chose “hclust” procedure, which accomplishes hierarchical clustering. “Ward’s” method was selected to compute inter-cluster distance.

The hierarchy of computed clusters is represented by the tree structure called “dendrogram”. Dendrogram for 5-member (word) clauses ended with question mark is depicted on the Figure 2. Melody contours corresponding to individual clusters obtained during computations are shown on the Figure 3. Representative contours are drawn as a thick line.

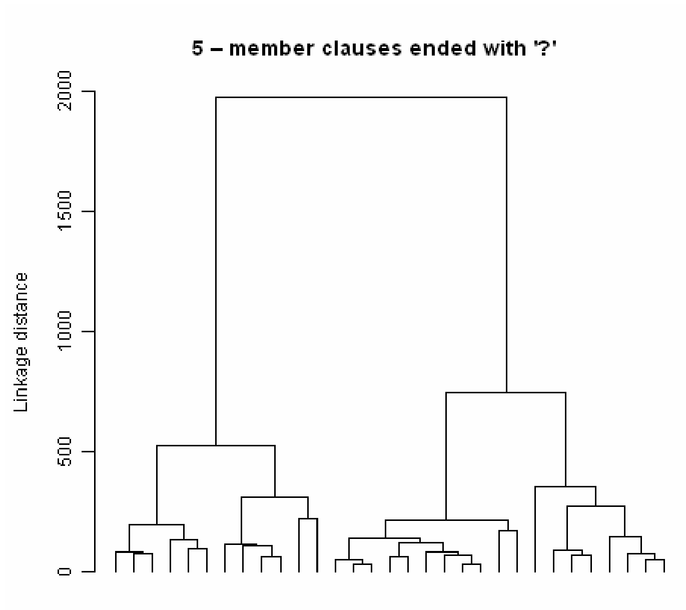


Fig. 2. Dendrogram tree computed for melodies of 5-word clauses ended with question mark.

Clustering was computed separately for each type of sentence (declarative, interrogative, exclamatory, clause ended with “,”, clause ended with “a” etc.) and for each set of 1-, 2-, 3- … and 14-member clauses (sentences).

Results obtained by clustering were analysed. We investigated clusters, starting from the top of dendrogram, finishing at the moment, when next splitting of cluster did not increase the number of significantly different representative contours.

Ones the investigation is finished, the final number of clusters is known. Final clusters are further analysed to find:

1. Clusters corresponding to basic contour types (see Table I)
2. Clusters with non-standard melody
3. Clusters for further study (accent put on different locations in the sentence, clusters caused by other phenomena)

Correspondence between contour types described by language scientists (see paragraph 2) and contours obtained by clustering is searched.

For example, bottom left contour (see Figure 3) corresponds to the melody type “1A Wh-question” (Table I). Bottom right contour (Figure 3) corresponds to the melody type “2A Yes–no question” (Table I).

Some clusters are composed of sentences pronounced in a non-standard way:

1. Sentences pronounced with very gentle melody variations
2. Sentences pronounced in a whisper
3. Sentences with accent on several words in the sentence

Bottom middle cluster (Figure 3) contains this kind of sentences. We cannot detect from the text when this type of melody should be used. So these clusters were excluded from further considerations.

We also found other type of clusters - "clusters for further study". In this case, sentence melody sounds naturally, however cluster representative contour differs from basic melody contours summarized in the Table I.

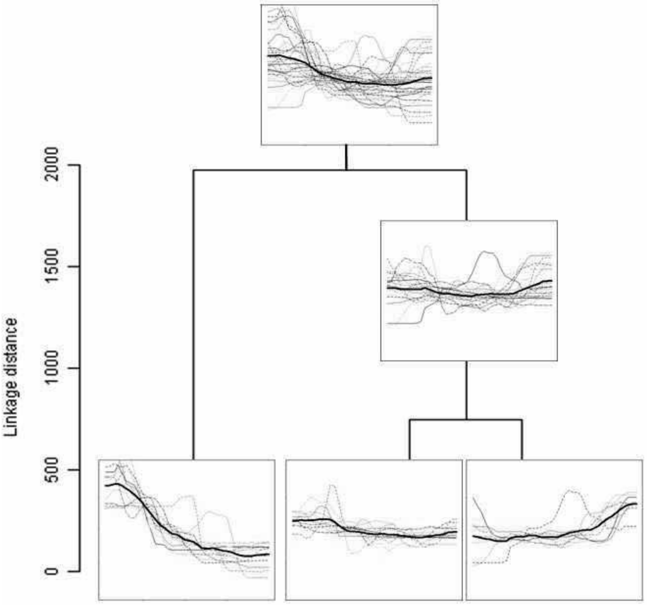


Fig. 3. Hierarchy of contour clusters computed by "hclust" procedure. (Melodies of 5-word clauses ended with question mark.)

One reason for this kind of clusters is placing the accent on various positions in the sentence. Sentences belonging to the same cluster have accent on the same relative position (see Figure 4). Also, these melodies we do not use in current version of our TTS system. Putting accent on different words in the sentence is left for further study.

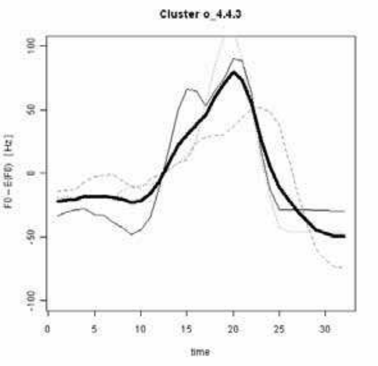


Fig. 4. Melody cluster comprised of sentences with strong accent placed on the same relative position (For example, the sentence: ”A kde budeš bývať?” has strong accent on the first syllable of the word “bývať”)

There are also other phenomena causing different clusters that could be studied. For example, the interrogative sentence starting with short words exhibit accent moving to the second syllable (see Figure 5).

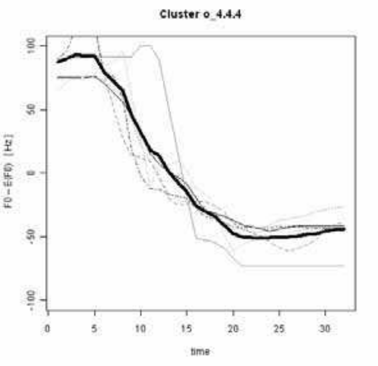


Fig. 5. Cluster comprised of interrogative sentences starting with short words
(For example, in the sentence “A čo je viac?” the accent is placed on the syllable “čo”.)

4. MAPPING TEXT CHARACTERISTICS INTO MELODY CONTOURS FOR THE TTS SYSTEM

Our TTS system performs input text analysis. According to obtained text characteristics (punctuation marks, nouns, words) proper melody contour will be chosen and added to the synthesized speech. In the Table I we appended the last column, which describes text characteristics of different sentence types. For example, “1A Wh–question” (see Table 1) starts with an interrogative word (like: čo, kto, kedy, komu, prečo, načo, začo, ako, však, koľko ..., also with: keby ...)” and ends with question mark.

For some sentence types, the same text characteristics are obtained by the text analysis, so we designed mapping of sentence type to melody contour to be used in our TTS system (see Table II).

Table II. Mapping of sentence types to melodies used in the TTS system (see Table I for notation of sentence types)

Sentence type	**Melody contour to be used in our TTS system**	**Identified in the text by**
1A	1A	„ . “, „ ! “ „ ? “ when beginning with an interrogative word
2A, 2B	2A	„ ? “ when does not begin with an interrogative word
2C	2A (first part) 1A (second part)	„ ? “ and the word „alebo”
3A, 3B, 3C	3B (Flat - rising)	„ , “

5. CONCLUSION

We present our data-mining approach to the sentence melody analysis. This approach consists of two steps: melody contour extraction and cluster analysis. Contour extraction was described in detail in [3]. This paper deals with the second step – cluster analysis. Slovak language sentences are analysed – typical melody contours are looked for. R-software's "hclust" procedure with Ward's method is used to find clusters of sentences with similar melodies. Clusters were analysed and several types of clusters were found: clusters corresponding to melody types described by language scientists, clusters with non-standard melody (these clusters were excluded), clusters of sentences with accent put on various locations in the sentence, and clusters reflecting other phenomena. Last two types of clusters are left for further study. Finally, mapping of text characteristics to melody contours that can be used in our TTS system was prepared.

ACKNOWLEDGMENTS

We would like to thank Prof. Ing. Martin Klimo, PhD, from University of Zilina, for discussion and comments on our work.

REFERENCES

[1] BACHRATÁ, K., BOROŇ, J., ROSIVAL, B. 2007. New methods proposal for VoIP. *Scientific Papers of the University of Pardubice: 12 (2006): Series B*, 201-210.

[2] ČAKY, P., KLIMO, M., MIHÁLIK, I., MLADŠÍK, R. 2004. Text-to-speech for Slovak language. In proceedings of the *Text, Speech and Dialogue: 7th International Conference, TSD 2004, Brno, Czech Republic, September 8-11, 2004,* Berlin: Springer, 291-298.

[3] ČERŇANSKÁ, M., ŠKVAREK, O. 2009. Sentence melody analysis for speech production in the TTS system. In proceedings of the *TRANSCOM 2009: 8-th European conference of young research and scientific workers: Žilina June 22-24, 2009*, *Slovak Republic. Section 3: Information and communication technologies.* Published by University of Žilina, 25-30.

[4] KLIMO, M., BACHRATÁ, K., 2004. Sampling of function generating shift invariant spaces. *Scientific papers of the University of Pardubice: The Jan Perner Transport Faculty. 10 (2004): Series B.* Pardubice: Univerzita Pardubice, 165-175.

[5] KLIMO, M., BACHRATÁ, K., SMIEŠKO, J., URAMOVÁ, J. 2009. *Teória IP telefónie.* Žilinská univerzita. (In Slovak)

[6] KLIMO, M., MIHÁLIK, I. 2007. Extending annotational SSML into structural content for speech synthesizer. *Scientific Papers of the University of Pardubice: 12 (2006): Series B,* 193-199.

[7] KRÁĽ, A. 1996. *Pravidlá slovenskej výslovnosti.* Slovenské pedagogické nakladateľstvo Bratislava. (In Slovak)

[8] MIHÁLIK, I. 2008. *Kompresia reči v TTS* [Ph.D. thesis]. Žilinská univerzita v Žiline, Fakulta riadenia a informatiky, Katedra informačných sietí. Žilina. (In Slovak)

[9] MIKUŠ, Ľ., IVANIGA, P. 2007. E-learning. *Svet komunikácie - Sila e-informácií.* Internet journal. http://www.svet-komunikacie.sk/index.php?ID=4117. (Taken: July 23 2009.)

[10] Program Praat main web page. http://www.fon.hum.uva.nl/praat/. (Taken: March 18 2009.)

[11] SELEP, P. 2007. Diphone concatenation method for speech synthesis. In proceedings of the *TRANSCOM 2007: 7-th European conference of young research and science workers: Section 3: Information and communication technologies.* Žilina: University of Žilina, 2007, 221-224.

[12] TTS system version 1 demo (developed at the Department of InfoCom Networks, Faculty of Informatics and Management Science, University of Zilina, Slovakia). http://tts.kis.fri.utc.sk/. (Taken: March 18 2009.)

ČERŇANSKÁ MÁRIA
Department of InfoCom Networks
Faculty of Informatics and Management Science
University of Zilina, Univerzitna 8215/1, 01026 Zilina, Slovak Republic
e-mail: maria.cernanska@fri.uniza.sk

ŠKVAREK ONDREJ
Department of InfoCom Networks
Faculty of Informatics and Management Science
University of Zilina, Univerzitna 8215/1, 01026 Zilina, Slovak Republic
e-mail: ondrej.skvarek@fri.uniza.sk

Received July 2009