elektrorevue
ISSN 1213 - 1539

Statistical Approach for Prosody Contour Modeling based on Sentence Classification

Anna Kondelová, Ján Tóth, Gregor Rozinaj

Fakulta elektrotechniky a informačných technológií STU v Bratislave
Email: {anna.kondelova,jan.toth}@stuba.sk

Abstract *– In this article need of complete prosody contour model in Slovak language are described. Prosody is important part of speech synthesis, especially nowadays, when computer systems are taking control of many activities. In this article the process to model missing prosody contour is mentioned and why is it necessary. At the end final prosody contours and results of subjective evaluation are shown. For analysis different statistic methods are used.*

1 Introduction

Prosody contour modeling is a part of huge process called speech synthesis. The speech synthesis converts written text to spoken form. This conversion has more levels of quality based on actual needs. It depends on using intension. Part of synthesizer dealing with prosody modeling is not necessary for base information transmission, though to sound natural or to present emotions and mood. There are some languages (e.g. Chinese language) with need of varying prosody for recognition of word's meaning [1].

Prosody on continuous speech melody is simplified. In digital speech processing prosody represents three variable values F_0 (fundamental frequency or vocal fundamental frequency), duration and energy [2]. We are focused on modeling prosody contour in this paper so fundamental frequency is considered as the most suitable value for modeling.

Additional information is provided by the prosodic features. Two aspects of prosodic markers are, as described in single sentence comprehension: distribution of intonational phrase (IPh) boundaries and the positions and type of accents [3].

Prosody on speech part ended with punctuation mark is observed, simply called sentence. Sentence prosody has a decreasing character, caused also by decreasing amount of air in lungs [3]. Every language has own sentence typology. It is necessary to have prosody contour for each type, which represents all sentences in corresponding group. Shorter prosody phenomenon is important to recognize specific contour, e.g. intonation peak for imperative sentence. Increasing melody in some types of questions at the end of sentence is observed.

This article deals with sentence typology for Slovak language, prosody contour modeling for missing sentence types (imperative sentences) and changing prosody contour according to modeled prosody contours. Theory is mentioned first and results are concluded at the end of the article. In section Results final prosody contours and also success of changed prosody contours are shown. Evaluation by subjective testing was performed.

2 Prosody Contour Modeling

In TTS (Text-to-Speech) Synthesizer developed on Institute of Telecommunication prosody contours for three types of questions are designed. We decided to expand field of implemented sentence types. Here the process of acquiring final prosody contours of imperative sentences is described.

2.1 Need Analysis

Current speech synthesizers are able to synthesize text. A difference in final quality is observed. Synthesizers by blind people to read websites for example are used. The area for our research stands in making speech more natural. Possible is to change voices, according to male or female voice, child or adult sounding voices up to user's wish. Human answer according to mood or importance of information is variable. Repeating of the same answer because of not understanding is connected with melody changing. Similar announcements, like information in which city stands train (speech synthesizer used in train to inform travelers about train position), is natural to say with accent on city or other similar situations. Here is necessary to know context of synthesized text. Good reason for changing sentence melody is in case of repeating commands, when GPS device repeatedly corrects the journey.

2.2 Sentence Typology

Slovak language knows three grammatical sentence types: declarative (Tomorrow I will go to the store.), imperative (Get me some water!) and interrogative sentence (What did the teacher say to you yesterday?). Recognition character is a punctuation mark at the end of each sentence: dot (.) for declarative sentence, exclamation point (!) for imperative sentence and question mark (?) for interrogative sentence. Grammatical types are further divided. The wish and exclamatory sentences also belong to the sentences with exclamation mark. We know more types of questions in Slovak language, e.g. declaratory question, yes/no question (question where the answer is only yes or no), etc.

In speech synthesis or text to speech process is appropriate to know form of text. Well known are communication systems, where computer gives questions and on the other side user answers. The module for generating prosody will work

elektrorevue
ISSN 1213 - 1539

with questions (interrogative sentences) and answers (declarative sentences) in such case. Another possibility is GPS Navigator with specific commands for reaching the destination. Here just imperative sentences with possible duplicating when needed are used.

We aimed to expand the set of prosody contours to expand the work that has been done on our department. Three types of interrogative sentences and respective prosody contours were identified in past years. We found two another different sentence types after analyzing the whole sentence typology. First type covers declarative sentences. The melody has decreasing character. This phenomenon in Figure 1 on declarative sentence is shown and is called cadency. Another known phenomenon is called anticadency and in Figure 2 on interrogative sentence is shown.

Figure 1: Sentence "*A man has a freewill.*" with shown phenomenon called cadency

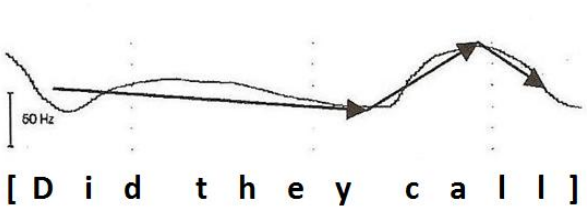

Figure 2: Sentence "*Did they call?*" with shown phenomenon called anticadency

Second type covers all sentences ended with exclamation mark (imperative, wish and exclamatory sentences). We focused on second type in this article.

2.3 **Creating Database**

In order to obtain the prosody contour it is important to analyze the speech database. Databases using corpora are described in [4] for English, in [5] for Japanese, English and Chinese. Databases using elicited speech are described in [6] for Swedish, in [7] for Hebrew. Databases using acted speech are described in [8] for Burmese and Mandarin, in [9] for Japanese, in [10] for Russian, in [11] for Finnish, in [12] for Spanish.

The first step was to create database of imperative or other sentences. We picked up imperative sentences containing one word, two words, three words, four words and five words. In all groups remained the most proper candidates. The whole database consists of at least 200 different sentences. One of groups (one word sentence) in Table 1 is shown.

Table 1: One word group members

imperative sentences		
1 word		
Stop!	Attention!	Repeat!
Stay!	Don't sass!	Run away!
Don't repeat!	Run!	Don't stop!
Open!	Eat!	Work!
Cut!	Drink!	Chop!
Go!	Count!	etc.

The whole database was recorded and post processed using the software Audacity 2.0.3. For recording two speakers were chosen and as recording device external microphone was used.

Segmented audio files in WAV format to get prosody contour of fundamental frequencies were than processed. We used freely available program PRAAT. The fundamental frequencies acquired from PRAAT are written in text files and ready for further processing. One word sentence prepared for processing in Table 2 is shown.

Table 2: One word sentence (in Slovak language) with extracted F_0 in corresponding times

Sentence: „Go!“			
time	**F_0**	**time**	**F_0**
0,124	165,28	0,204	138,39
0,134	166,31	0,214	145,51
0,144	164,54	0,224	148,28
0,154	160,83	0,234	146,83
0,164	156,91	0,244	147,63
0,174	151,47	0,254	144,4
0,184	148,12	0,264	137,29
0,194	143,47	0,274	135,5

2.3.1 PSOLA

The program PRAAT is a robust tool for audio signal processing. The functionalities related to get and to change prosody contour were used by our research. Algorithm PSOLA (Pitch-Synchronous Overlap-Add) to change prosody contour is used.

The PSOLA technique in speech processing to change the pitch of a speech audio signal without affecting its duration is often used. A very simple method to modify the fundamental frequency is to change the duration of the speech signal. We decrease the fundamental frequency by adding more periods and otherwise it is possible to increase fundamental frequency by removing relevant number of periods [13]. The feature that speech signal is quasiperiodic in PSOLA algorithm is used. The period by using Hamming window is isolated (another possibility is to use Hannig, Blackman, Barlett or Kaiser window). Such isolated periods from parts where is necessary to change fundamental frequency are added or removed. The

elektrorevue
ISSN 1213 - 1539

increasing fundamental frequency process in Figure 3 is shown [14].

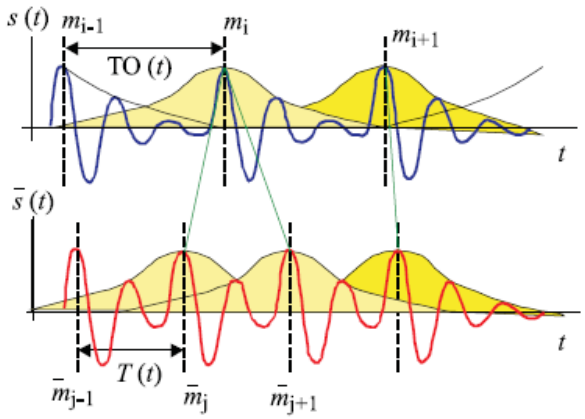

Figure 3: The representaion of increasing fundamental frequency

There are other algorithms to change prosody, e.g. OLA (Overlap-Add), MBROLA (Multi band Resynthesis Overlap-Add) and sinusoidal models [15].

2.4 Median Filtering

The text files with extracted fundamental frequencies to personal correction of two correctors were submitted. The aim of correction was to control the individual values. If values were extremely high (i.e. 553 Hz instead of maximal value 203 Hz) they were removed and replaced with new optimal values. We decided to set optimal values with median filter. The window on size three was set. All prosodic contours by median filter with size three were smoothed without any critical effects on prosodic phenomenon. The size was set up on three to avoid changing prosodic phenomenon and to smooth uneven curves.

Median filter uses windows with various sizes. If the window size is three, then filter takes three values in row. Values in order upward are sorted. Finally the middle value is picked. Median filtrated 1D signal is shown below. Input signal is *x[n]* and output signal is *y[n]*, where *n* is discrete time [16]. Individual steps of median operation are listed below.

x[n] = [3 45 8 16]

y[1]=Median[3 3 45] = 3
y[2] = Median[3 45 8] = Median[3 8 45] = 8
y[3] = Median[45 8 16] = Median[8 16 45] = 16
y[4] = Median[8 16 16] = Median[8 16 16] = 16

y[n] = [3 8 16 16]

2.5 Normalization

It is necessary to normalize the whole database after applying median filter. Every contour in every group (e.g. three words or four words sentences) needs to be normalized as seen in Figure 4. The optimal value of normalization samples on 1000 after analyzing all parameters was set.

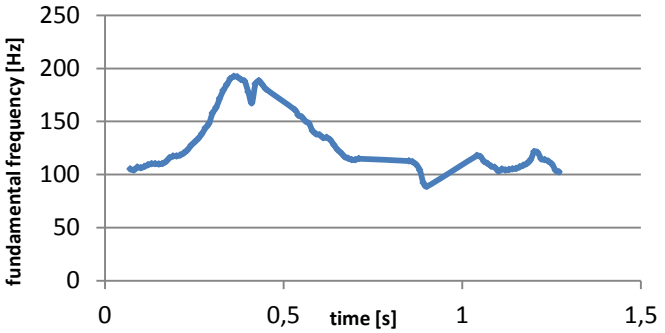

Figure 4: Sentence built-up of four words before normalization

Acquired prosody contour is linear interpolation of discrete values seen in formula (1).

$$\frac{y - y_1}{y_2 - y_1} = \frac{x - x_1}{x_2 - x_1} \quad (1)$$

$[x_1, y_1]$ are coordinates for start point and $[x_2, y_2]$ are coordinates for end point of interpolation. Point $[x, y]$ is actual position in linear dependency.

In Figure 5 normalized prosody contour of the same sentence is shown. On *x*- axis are normalized samples and on *y*-axis is fundamental frequency.

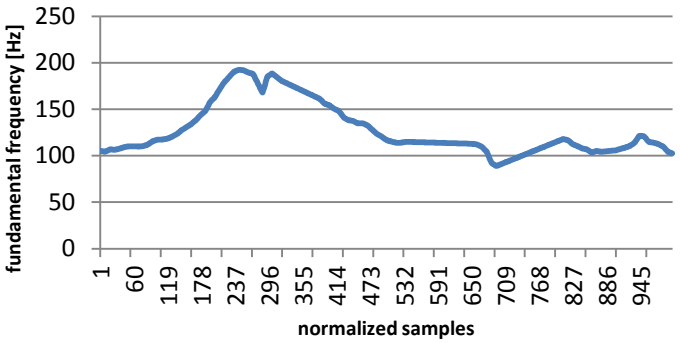

Figure 5: The same four words sentence as in Figure 3 after normalization

2.6 Low Pass Filtering

For smoothing the prosody contour low pass filter is used [17]. We used low pass filter with window size five. Low pass filter has dimensions [1 2 3 2 1]. The dimensions as well as window sizes were tested to get optimal influence on contour. E.g. window size three is too low to smooth the curve. The prosody contour of the same sentence as mentioned above after applying low pass filter with window size five in Figure 6 with red color is shown. The blue one is before applying low pass filter.

elektrorevue
ISSN 1213 - 1539

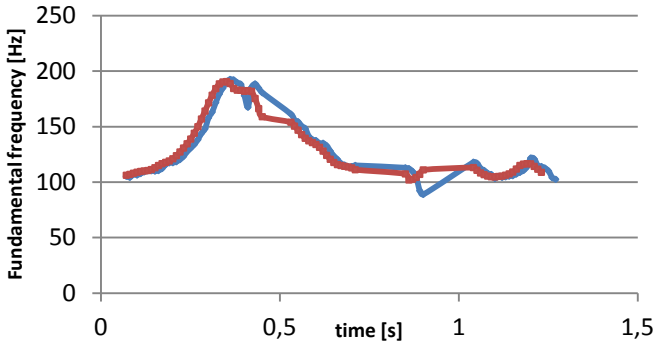

Figure 6: Prosody contour of the same sentence as in Figure 3 and Figure 4 with applied low pass filter with window size 5 is shown with red line. Prosody contour before applying low pass filter is shown with blue line

2.7 Average Normalized Prosody Contour

For acquiring general prosody contour for each group statistical method was used. Every representative in each group has 1000 normalized values which are summarized. Then average normalized values $f_{aver\_norm}(i)$ with formula (2) are calculated and average normalized prosody contour is created. In formula (2) is N number of normalized samples (as mentioned above N=1000) and $f_0(n,i)$ are fundamental frequencies according to specific speaker. The variable values $f_{aver\_norm}(i)$ assures the speaker independence [13].

$$f_{aver\_norm}(i) = \frac{\sum_{n=1}^{N} f_0(n,i)}{N} \qquad (2)$$

2.8 Changing Prosody Contour

We need to proof rightness of changed prosody in imperative sentences with modeled prosody contours. Prosody changing by our earlier module used in Modular synthesizer designed on Institute of Telecommunication on STU was performed. The whole process in [13] is described.

3 Subjective Testing

The evaluation was taken by 35 users. We created website to collect all data from 35 users. Questionnaire by three questions was made. In each question: "Which imperative sentence sounds more natural?" were two possibilities. The mentioned sentences are: "Attention!"(in Slovak language has the sentence one word "Pozor!"), "Clean up the room!"(in Slovak language has the sentence two words "Uprac izbu!"), "Listen to good music!"(in Slovak language has the sentence three words "Počúvame dobrú hudbu!") and the last "Stop doing any mess at home!" (in Slovak language has the sentence four words "Prestaň doma robiť neplechu!"). In each question are two WAV files one with monotonous melody and the second adapted to the predicted modeled contour.

4 Results

In this section the final prosody contours for imperative sentences are summarized. The method for subjective testing to proof the naturalness of synthesized speech was designed.

4.1 The acquired Prosody Contours

The general contours for each group are shown below. In Figure 7 general prosody contour for one word imperative sentence is shown.

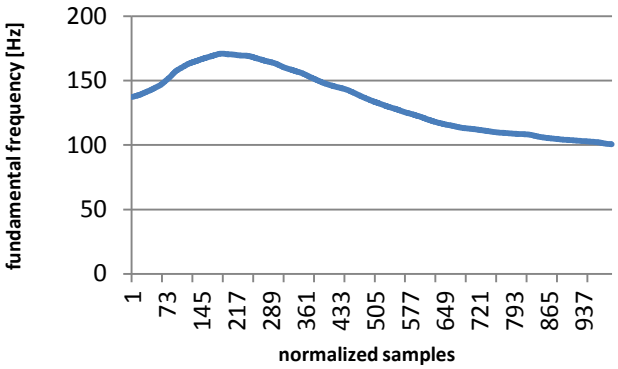

Figure 7: General prosody contour for one word sentence

In Figure 8 general prosody contour for two words imperative sentence is shown.

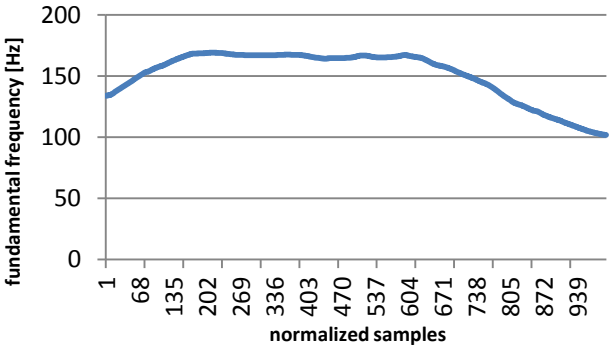

Figure 8: General prosody contour for two words sentence

In Figure 9 general prosody contour for three words imperative sentence is shown.

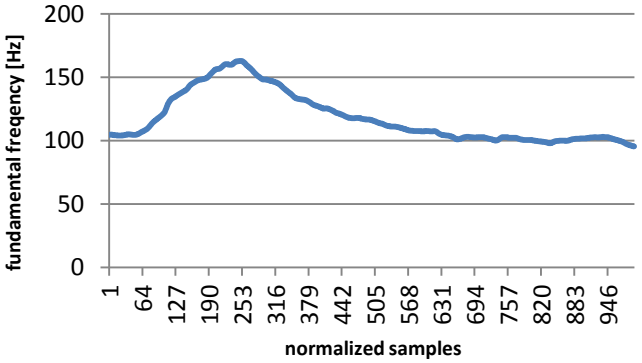

Figure 9: General prosody contour for three words sentence

elektrorevue
ISSN 1213 - 1539

In Figure 10 general prosody contour for four words imperative sentence is shown.

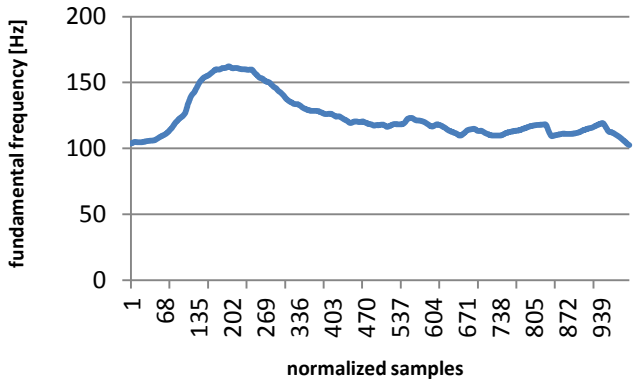

Figure 10: General prosody contour for four words sentence

In Figure 11 general prosody contour for five words imperative sentence is shown.

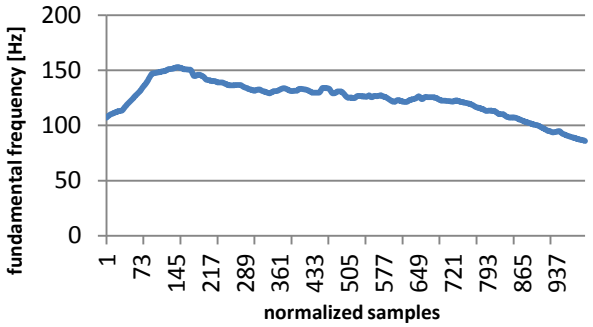

Figure 11: General prosody contour for five words sentence

4.2 The Evaluation of Subjective Testing

The subjective testing according to recommendations ITU-R BT.500-13 and ITU-R BS.1116-1 was designed. In evaluation following results were obtained. From 35 respondents were 30 satisfied with prosody change in first one word sentence (85,71%), in the second two words sentence were satisfied 32 respondents (91,43%), in the third three words sentence were satisfied 30 respondents (85,71%) and in last four words sentence were satisfied again 30 respondents (85,71%).

5 Conclusion

The intonational peak at the beginning of imperative sentence in every type of sentence group was observed. The prosody contour begins with increase of fundamental frequency up to maximum and then decreases to the end of sentence. The width of intonational peak decreases with the rising count of words in imperative sentence. This assumption in groups of one word, three words, four words and five words except group of two words sentences was confirmed. This fact leads us to test this phenomenon with subjective testing. The results, as shown in section above, prove naturalness of two words prosody contour.

Our explanation is that Slovak language has sentence accent on first word and word accent on first syllable.

We consider his phenomenon as very interesting and here we see the possibility to continue with research.

The general prosody contour has two ways of being modeled. The difference is in order of all mentioned processes. The first prosody contour (red line) is acquired in order of three processes: first normalization of all representatives, than average normalization and at the end low pass filtration of eventual average normalized contour. The second prosody contour (blue line) is acquired in order of applying at first low pass filter on every representative in a whole group, than normalization of each low pass filtered representatives and at the end calculating average normalized prosody contour. The comparison of two curves in Figure 12 is shown. On *x*-axis are normalization samples and on *y*-axis is fundamental frequency.

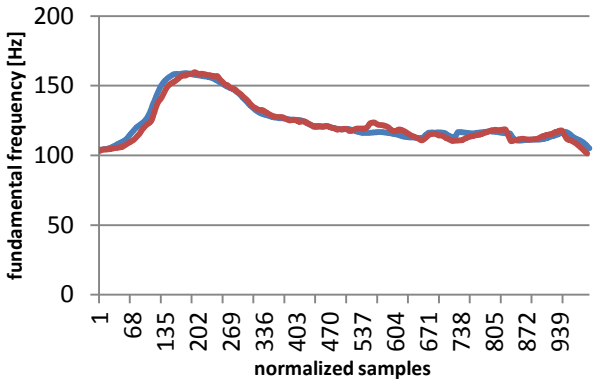

Figure 12: The comparison of two prosody contours calculated in two different ways

Difference is not very significant. It result that using low filtration once, not on every representative, is not so distorting. So the first sequence was picked up: normalization, calculation of average normalized prosody contour and final low pass filtering.

From acquired responses in questionnaire results that all the modeled prosody contours for imperative sentences are appropriate.

Using appropriate and natural prosody contour especially for imperative sentences is very useful for GPC devices or other navigations. Devices generate many statements and commands for user. Lexical meaning is clear (nothing changes in command) but acting like real human behind navigation device makes it comfortable and convenient for the user.

Acknowledgement

This work has been supported by HBB-Next - Next-Generation Hybrid Broadcast Broadband FP7-ICT-2011-7 - 287848-06, VEGA 1/0961/11, VEGA 1/0708/13 and under the program to support young researchers on STU.

elektrorevue
ISSN 1213 - 1539

References

[1] KAČUR, J., ČEPKO, J., PÁLENÍK, A. A*utomatic labeling Schemes for Concatenative Speech Synthesis*. In Proceedings Elmar 50th International Symposium ELMAR. Zadar: Croatian Society Electronics in Marine, September 2008. ISBN 978-953-7044-09-1, p. 639-642.

[2] SCHERER, K.R. *Vocal communication of emotion: review of research paradigmas.* In Speech Communication, vol. 40(1-2), 2003, p. 227–256.

[3] HRUSKA, C., ALTER, K. *Prosody in Dialogues and Single Sentences: How Prosody can influence Speech Perception.* In Language Context and Cognition. Berlin: Walter de Gruyter, 2004. ISBN 3-11-017934-2, p. 211-227.

[4] DOUGLAS-COWIE, E., COWIE, R., SCHRÖDER, M. *A new emotion database: Considerations, sources and scope*. In Proc. ISCA Workshop on Speech Emotion, Belfast, Northern Ireland, 2000, p. 39–44.

[5] CAMPBELL, N. *Building a corpus of natural speech—And tools for the processing of expressive speech—the JST CREST ESP project*. In Proc. 7th Eur. Con. Speech Commun. Technol., Aalborg, Denmark, 2001, p.1525–1528.

[6] KARLSSON, I., BANZIGER, T., DANKOVICOVA, J. et al. *Speaker verification with elicited speaking-styles in the verivox project.*In Speech Commun., vol. 31, no. 2, 3, 2000 p. 121–129.

[7] AMIR, N., RON, S., LAOR, N. *Analysis of an emotional speech corpus in Hebrew based on objective criteria*. In Proc. ITRW Speech Emotion, Newcastle, Northern Ireland, 2000, p. 29–33.

[8] LAY NEW, T., FOO, S. W., DE SILVA, L. *Speech emotion recognition using hidden Markov models*. In Speech Commun., vol. 41, no. 4, 2003, p. 603–623.

[9] IIDA, A., CAMPBELL, N. *A database design for a concatenative speech synthesis system for the disabled*. In Proc. 4th ISCA Workshop Speech Synth., Edinburgh, United Kingdom, 2001, p. 189–194.

[10] MAKAROVA, V., PETRUSHIN, V. *RUSLANA: Adatabase of Russian emotional utterances*. In Proc. ICSLP, Denver, Colorado, United States of America, 2002, p. 2041–2044.

[11] SEPPÄNEN, T., TOIVANEN, J., VÄYRYNEN, E. *MediaTeam speech corpus: A first large Finnish emotional speech database*. In Proc. 15th Int. Congr. Phonetic Sci., Barcelona, Spain, 2003, p. 2469–2472.

[12] MONTERO, J. M. et al. *Emotional speech synthesis: From speech database to TTS.* In Proc. ICSLP, vol. 3, Sydney, Australia, 1998, p. 923–926.

[13] KONDELOVA, A., TOTH, J., ROZINAJ, G. *Analysis of Prosody Features of Slovak.* In Proceedings ELMAR 52nd International Symposium ELMAR. Zadar: Croatian Society Electronics in Marine, September 2010. ISBN 978-953-7044-11-4, p. 371-374.

[14] MOUSA, A. *Voice Conversion using Pitch Shifting*. In Journal of ELECTRICAL ENGINEERING, vol. 61, no. 1, 2010, p. 57-61.

[15] QUATIERI, T. F., MCAULAY, R. J. *Shape invariant time-scale and pitch modification of speech.* In IEEE Trans. Signal Process., vol. 40, no. 3, 1992, p.497–510.

[16] GABBOUJ, M. et al. *Weighted median filters:a tutorial.* In IEEE Transactions on Circuits and Systems II Analog and Digital Signal Processing. 1996. ISSN 1057-7130, p. 157-192.

[17] FU-JUN, H. et al. *The aplication of low-pass filtering to pretreatment in thermal wave NDT.* In International Conference on Measurement, Information and Control (MIC), vol. 2, 2012, p. 590 – 594.