Single Speaker Acoustic Analysis of Czech Speech for Purposes of Emotional Speech Synthesis

Martin Grůber [1] and **Milan Legát** [2]

Abstract. This paper deals with an acoustic analysis of the sets of Czech sentences uttered by single speaker. The data used in this analysis consists of both emotional and neutral sentences. We have been especially interested in some features which are supposed to influence the perception of speech, such as F0, phoneme duration, formant frequencies or energy. The analyzed sets of sentences were composed of utterances expressing various speaker's attitude. We have tried to reveal some acoustically measurable differences among various speaker's attitudes so that we could incorporate this knowledge into our speech synthesis system [8] to obtain emotional synthetic speech.

1 Introduction

Without question, contemporary speech synthesis techniques produce high quality and intelligible speech. However, the synthetic speech cannot sound completely natural until it expresses a speaker's attitude. Thus, emotional speech synthesis is a frequently discussed topic and has become a concern of many scientists. In spite of the fact that some results have already been presented, this issue has not been satisfactorily solved yet. Some papers which deal with this problem include, but are not limited to [1] [2] [3] [4] [5] [9] [11] [13].

To incorporate some expressivity into the synthetic speech, we firstly need to find out which factors are important for listeners to perceive spoken speech as expressive speech. In the first phase of this research, we have focused on acoustic characteristics of the expressive speech. The results of our analysis cannot be generalised as we have analysed sentences uttered by single speaker and due to this reason they are not statistically representative. Nevertheless, we can utilize the revealed acoustic characteristics for incorporation of emotions into our speech synthesis system.

This paper is organised as follows. Section 2 deals with the description of the data used in the analysis. In section 3 the acoustic analysis as such is described. In this section we list the features that were measured on the data and the techniques which were used for their acquisition. Section 4 is dedicated to an overview of the attained results. Some conclusions and future work are also presented in this section.

2 Speech material used in analysis

Just for experimental purposes, we have recorded a database of utterances containing various speaker's attitudes. The speech data were uttered in an anechoic room by a semi-professional female speaker with some radio-broadcasting experience. Before the recording of emotional sentences, the speaker was instructed to try her best to portray the emotions.

The database is composed of four sets of sentences uttered in neutral speaking style - 100 wh questions (referred to as *whQuest*), 97 yes-no questions (*ynQuest*), 91 imperative sentences (*imperSen*) and 100 indicative sentences (*indicSen*). We consider these four sets of sentences to be emotionally neutral and we have used them as referential ones in our analysis.

In addition, the database contains six sets of sentences in which two emotions are expressed by the speaker - happiness and sadness. These two contrasting emotions have been chosen because they are supposed to be well distinguishable, according to [4], [10] and [12]. Another reason for the selection of these two emotions is that the emotional speech synthesis is a very complex task and our short-term plan is to enable our synthesis system to use sad, neural and happy speaking style. For each emotion, three sets of utterances were analysed.

The first pair of sets (*happyHC* / *sadSC*) contains sentences with emotionally dependent content corresponding with the particular emotion, in each of these sets there are 100 sentences. The second pair of sets (*happySel* / *sadSel*) is a selection from the first one. These sets contain the amount of 30 and 20 items, respectively. The selection was made by a few listeners, whose task was to mark sentences which seemed to them to correspond perfectly with the given emotion. The last pair of sets (*happyNC* / *sadNC*) is similar to the first one but the content is emotionally neutral and identical for both emotions. Again, both of these sets contain 100 sentences. The speaker was instructed to utter the same sentences using both happy and sad speaking style.

This division of emotionally uttered sentences has been intended to show whether the content of a sentence affects the speaker when portraying the emotion. Further, a comparison between two given emotions was required. We have decided to compare the sets of sentences with neutral content for exclusion of an influence of the content. The reason for making a selection of some emotional sentences by listeners was to find out whether these perceptively slightly different sentences differ also from other emotional utterances in terms of their acoustic characteristics.

3 Acoustic analysis

In the following subsections, there are presented the results of the acoustic analysis of expressive speech. Each subsection corresponds with one feature which is supposed to influence the perception of speech significantly. These are the fundamental frequency F0 (in sec-

[1] University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Cybernetics, Czech Republic, email: gruber@kky.zcu.cz

[2] University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Cybernetics, Czech Republic, email: legatm@kky.zcu.cz

tion 3.1), duration of phonemes (in section 3.2), values of the formant frequencies F1, F2 and F3 (in section 3.3) and values of the RMS energy (in section 3.4).

The part of the database containing emotional utterances was recorded at another time and with slightly different settings of recording equipment (assembling/disassembling of the recording devices because of sharing the anechoic room with other projects) than the part containing questions, indicative and imperative sentences. This is due to the fact that emotional utterances were recorded only for experimental reasons whereas the other sentences were selected from the huge speech corpus which was recorded in neutral speaking style and which is currently used by our speech synthesizer. Unfortunately, the different settings of recording equipment resulted in a slight difference in the intensity level of these two sets of utterances. Because of this fact, we have not performed the analysis of RMS energies of these two groups of sentences.

3.1 F0 analysis

To determine F0 contours, we took advantage of having corresponding glottal signals recorded along with speech signals. We have used Robust Multi-Phase Pitch-Mark Detection Algorithm [7] for marking of pitch pulses in speech and derived the F0 contour from this sequence. First, we obtained local F0 estimates calculated as median of inverse values of distances between four consecutive pitch marks. Then, the sequence of these local F0 estimates was smoothed by median filter of order 3 (see Fig.1).

Table 1. Mean values and standard deviations of the F0.

set of sentences	mean value [Hz]	standard deviation [Hz]
sadSC	184.82	28.55
sadSel	181.27	28.31
sadNC	**181.32**	**29.01**
happyHC	202.40	44.73
happySel	209.57	49.34
happyNC	**203.62**	**46.28**
indicSen	**193.76**	**36.63**
whQuest	188.96	43.67
ynQuest	197.72	32.94
imperSen	198.78	39.15

The results summarized in Tab. 1 show that all the sentences representing happiness have higher F0 than the sentences representing sadness. The F0 mean value of the neutral utterances is in the middle of the values for two emotional sets.

A major difference between *happySel* and the other sets for happiness could be also noticed. It could suggest, that the listeners' selection may express the given emotion more than the sets containing all sentences. However, the same conclusion cannot be drawn for the sentences representing sadness.

Some differences were also found among the sets expressing various speaker's attitude, i.e. *indicSen*, *whQuest* and *ynQuest*. Note that these sets of sentences were all uttered in a neutral speaking style.

In Fig. 1, there is shown the F0 contour for a neutral sentence and for sentences representing sad and happy emotion. All three sentences have the same content - *"A připíjí vínem"* `[a pQ\ipi:ji: vi:nem]` - according to the Czech version of SAMPA phonetic alphabet. The difference in mean values and variances of F0 is visible as is the different duration of the whole sentence, as described in 3.2.

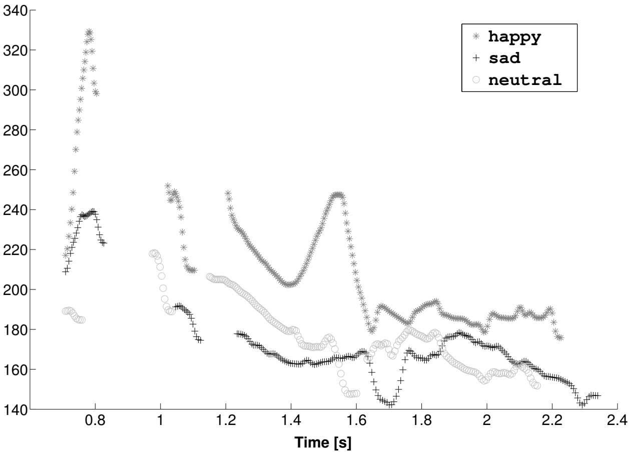

Figure 1. F0 contour of neutral sentence and sentences expressing emotions (selected from *sadNC* and *happyNC* sets).

3.2 Duration analysis

For the determination of durations of phonemes, an automatic segmentation technique using HTK Tools improved by a statistic approach [6] was utilized. To calculate the duration of a phoneme, the time of its end and its beginning were simply subtracted. The results of this analysis are shown in Tab. 2.

Table 2. Mean values and standard deviations of the phonemes duration.

set of sentences	mean value [ms]	standard deviation [ms]
sadSC	96.5	60.3
sadSel	98.7	63.6
sadNC	**97.6**	**62.0**
happyHC	91.6	48.6
happySel	92.7	50.8
happyNC	**85.5**	**43.2**
indicSen	**84.2**	**47.1**
whQuest	77.0	47.7
ynQuest	79.4	44.9
imperSen	81.3	44.6

The average duration of the phonemes appearing in neutral sentences is 84.2 ms. For sadness, the duration is longer, for *sadSC* set the mean value is 96.5 ms (by 15% above the neutral set). The differences among three sets with sentences expressing sadness are not statistically significant, according to the performed t-test.

For happiness, quite surprising results were obtained. The average duration for *happyHC* and *happySel* was about 92 ms, that is longer than the average duration in neutral sentences (by 9%), but the mean value in the *happyNC* set is almost at the same level as the mean value in the neutral one, no statistically significant difference was detected between these two sets. It suggests that the speaker was not able to portray the happy emotion in the sentences with neutral content, in terms of phone duration.

As mentioned above, the listeners' selections have almost the same average phonemes duration as the sets containing all emotional sentences with emotionally dependent content. The average phone durations for sentences expressing various speaker's attitudes were also different. The mean value for *whQuest* set was the lowest across all the sets.

3.3 Formant analysis

To obtain formant frequency estimates, we used Speech Filing System[3]. We employed the `formanal` program which is referred to as currently the best one in SFS to perform fixed-frame formant analysis. This program was originally implemented in the Entropic Signal Processing System and it is used under licence from Microsoft. The formant estimates which are presented in Tab. 3 come from the middle parts of vowels which were found by cutting the initial and the final quarter of the vowel length.

Table 3. Mean values and standard deviations of the formant frequencies for Czech short vowels.

a	**F1 [Hz]**		**F2 [Hz]**		**F3 [Hz]**	
sets	mean	std	mean	std	mean	std
indicSen	640	124	1377	223	2623	380
happyNC	695	126	1405	270	2657	441
sadNC	649	140	1320	235	2598	375
e	**F1**		**F2**		**F3**	
sets	mean	std	mean	std	mean	std
indicSen	487	96	1859	352	2697	224
happyNC	513	116	1949	313	2754	209
sadNC	492	137	1803	350	2634	250
i	**F1**		**F2**		**F3**	
sets	mean	std	mean	std	mean	std
indicSen	387	61	2054	445	2739	167
happyNC	371	59	2180	355	2782	226
sadNC	375	77	1975	381	2657	214
o	**F1**		**F2**		**F3**	
sets	mean	std	mean	std	mean	std
indicSen	444	57	1050	197	2719	222
happyNC	443	74	1026	211	2741	387
sadNC	421	74	1063	183	2683	275
u	**F1**		**F2**		**F3**	
sets	mean	std	mean	std	mean	std
indicSen	377	109	985	267	2698	184
happyNC	344	76	933	257	2634	391
sadNC	339	71	969	280	2544	396

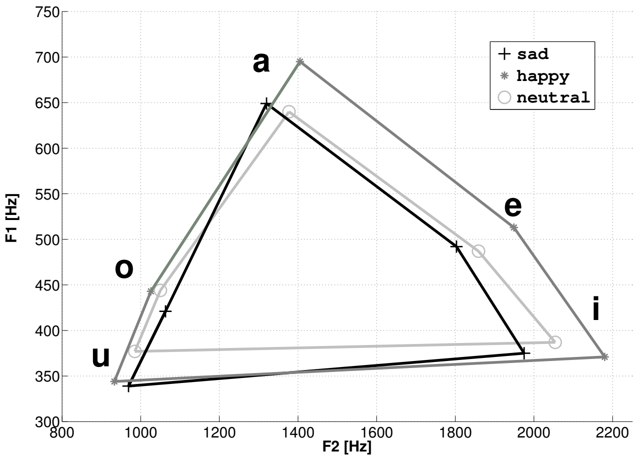

Figure 2. F1-F2 diagram for 5 Czech short vowels. A comparison of vocalic triangles for sentences uttered in neutral speaking style and happy and sad emotion (values measured for *happyNC* and *sadNC* sets).

Regarding the analysis of formant frequencies, various differences were detected across all the sets in terms of various vowels. Results for *happyNC*, *sadNC* and *indicSen* (as a referential set) are shown in Tab. 3. Since tendencies of formant frequencies shifts are not clear from this table, the vocalic triangle is depicted in Fig. 2. It represents the distribution of Czech short vowels in the F1-F2 space for various speaker's attitudes. Unfortunately, the influence of different emotions on F3 is not visible from this figure.

It seems that the vocalic triangle is expanded for values measured in sentences portraying happiness, it means that low formant frequencies are lowered and high ones are increased, in comparison with neutral speech. This phenomenon applies to both F1 and F2. The formant frequencies obtained for sentences conveying sadness cause a counter-clockwise rotation of the vocalic triangle. Again, these results apply only to our speaker and further analysis of more speakers would be necessary to generalise this phenomenon.

There were detected also differences among the mean values for various sets representing the same emotion. The differences were not tested by any test in order to find out whether they are statistically significant or not, but it could be said that the emotions are well represented by sets *happyNC* and *sadNC*. The differences between particular emotions were greater than the differences among the sets representing the same emotion.

Regarding the neutral sets of our database, no considerable differences were found.

3.4 RMS analysis

RMS[4] energy is a value that characterizes the intensity of a speech signal. Using this feature, the differences of intensity level in different sets of sentences can be measured. For this analysis, we had to divide our speech material into two parts and analyze them separately. One group contains the emotional sentences and the other one contains the neutral sentences. This separation was necessary due to slightly different settings of technical equipment for recording, as explained in the second paragraph of the section 3.

For the calculation of the RMS energy (1) of a sentence, initial and final pauses were cut off.

$$RMS = \sqrt{\frac{\sum_{i=1}^{n} s(i)^2}{n}}, \qquad (1)$$

where $s(i)$ is $i - th$ sample of the signal and n is the length of the signal.

The results obtained for the emotional part of the corpus are shown in Tab. 4. It is obvious that there is a difference between given emotions. The RMS energy of the sentences portraying happiness is higher than for the sentences portraying sadness. It means that the happy sentences are spoken louder. The difference is statistically significant which was proved by t-test. On the other hand, the differences between sets representing the same emotion are not statistically significant, except the sets *happyNC* and *happyHC*. In this case, the p-value reached the value 0.0407, which means that these two sets can be regarded as equal in terms of the mean value of the RMS energy considering lower significance level, e.g. $\alpha = 0.01$.

The results for the neutral part of our database are presented in Tab. 5. Comparing *indicSen* vs. *whQuest* and *ynQuest* vs. *imperSen*,

[3] Speech Filing System – http://www.phon.ucl.ac.uk/resource/sfs

[4] RMS = Root Mean Square, also known as the quadratic mean; a statistical measure of the magnitude of a varying quantity. It is especially useful when variates are positive and negative, e.g. waves.

Table 4. Mean values and standard deviations of the RMS energy (signal range $\langle -1,\ 1 \rangle$).

set of sentences	mean value	standard deviation
sadSC	0.0232	0.0039
sadSel	0.0232	0.0031
sadNC	**0.0224**	**0.0038**
happyHC	0.0307	0.0050
happySel	0.0299	0.0055
happyNC	**0.0294**	**0.0039**

Table 5. Mean values and standard deviations of the RMS energy (signal range $\langle -1,\ 1 \rangle$).

set of sentences	mean value	standard deviation
indicSen	0.1333	0.0179
whQuest	0.1282	0.0217
ynQuest	0.1480	0.0393
imperSen	0.1518	0.0373

no statistically significant differences can be observed, the other cases seem to be different.

4 Conclusions & future work

In this study, we have compared and contrasted some emotional and neutral utterances in terms of F0, phoneme duration, formant frequencies and RMS energies. Some results are briefly summed up in Tab. 6, where the analysed emotional speech is compared with the referential one in terms of F0, duration and RMS. Initially, this paper was intended to cover single speaker analysis of both emotional sentences and sentences expressing various speaker's attitude in spite of being uttered in neutral speaking style. However, we are currently more concerned with the synthesis of emotions which is why we decided to prefer analysis of emotional utterances to obtain some results useful for speech synthesis.

Table 6. Brief overview of the acoustic analysis.

set of sentences	**F0**		**duration**		**RMS**
indicSen	194	•	84.2	•	—
happyNC	204	⇑ 5%	85.5	⇑ 2%	0.0294
sadNC	181	⇓ 7%	97.6	⇑ 16%	0.0224

The discussion of the results of the formant analysis seems to be too complex and it is out of scope of this paper. Moreover, at the present time, incorporation of the results into our speech synthesis system would require more modifications of the current approach in comparison with incorporation of F0, duration and RMS energy results. However, in Fig. 2 there is depicted an influence of emotion being present in the spoken speech on the formant frequencies.

The results reached confirmed that all the features measured on the speech signal are important acoustic correlates of various speaker's attitudes. Nevertheless, it cannot be concluded that these features are sufficient for the distinction of all emotions from speech signal. In the future a similar analysis should be performed on more extensive database containing more emotions, e.g. anger, boredom and contentment.

The results found in this analysis could be confirmed by classification task using F0, duration, RMS energy and formant frequencies as predictors for determining emotion from speech signal. The reference data would be obtained by means of more complex listening tests. In the case that any classification model were able to give good results using these predictors, we could conclude that it would be sufficient to modify these characteristics of neutral speech to obtain emotional output.

Our future work will be focused on the incorporation of the obtained results into our speech synthesis system. It includes the modelling of prosodic features based on emotionally recorded data for single instance concatenation approach and the extension of a feature set for the unit selection approach.

ACKNOWLEDGEMENTS

This work was funded by the Ministry of Education of the Czech Republic, project No. 2C06020, and in part by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

The access to the METACentrum computing facilities provided under the research intent MSM6383917201 is highly appreciated.

REFERENCES

[1] A.W. Black, 'Unit selection and emotional speech', *Proc. of Eurospeech 2003*, 1649–1652, (2003).

[2] M. Bulut, S.S. Narayanan, and A.K. Syrdal, 'Expressive speech synthesis using a concatenative synthesiser', *Proc. of the 7th International Conference on Spoken Language Processing*, 1265–1268, (2002).

[3] W. Hamza, R. Bakis, E.M. Eide, M.A. Picheny, and J.F. Pitrelli, 'The ibm expressive speech synthesis system', *Proc. of the 8th International Conference on Spoken Language Processing*, 2577–2580, (2004).

[4] G. Hofer, K. Richmond, and R. Clark, 'Informed blending of databases for emotional speech synthesis', *Proc. of Interspeech 2005*, 501–504, (2005).

[5] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, 'A corpus-based speech synthesis system with emotion', *Speech Communication*, **40 n. 1-2**, 161–187, (2003).

[6] Matoušek J., Tihelka D., and Psutka J., 'Automatic segmentation for czech concatenative speech synthesis using statistical approach with boundary-specific correction', *Proc. of Eurospeech 2003*, 301–304, (2003).

[7] M. Legát, J. Matoušek, and D. Tihelka, 'A robust multi-phase pitch-mark detection algorithm', *Proc. of Interspeech 2007*, 1641–1644, (2007).

[8] J. Matoušek, J. Romportl, D. Tihelka, and Z. Tychtl, 'Recent improvements on artic: Czech text-to-speech system', *Proc. of Interspeech 2004 - ICSLP, 8th International Conference on Spoken Language Processing*, **III**, 1933–1936, (2004).

[9] J.M. Montero, J. Gutiérrez-Ariola, S. Palazuelos, E. Enríquez, S. Aguilera, and J.M. Pardo, 'Emotional speech synthesis: From speech database to tts', *Proc. of the 5th International Conference of Spoken Language Processing*, 923–926, (1998).

[10] J.A. Russell, 'A circumplex model of affect', *Journal of Personality and Social Psychology*, **39**, 1161–1178, (1980).

[11] M. Schroder, 'Emotional speech synthesis: A review', *Proc. of Eurospeech 2001*, 561–564, (2001).

[12] C.M. Whissell, *The Dictionary of Affect in Language*, 113–131, Robert Plutchik and Henry Kellerman (Ed.), Emotion: Theory, Research, and Experience, Academic Press, New York, 1989.

[13] E. Zovato, A. Pacchiotti, S. Quazza, and S. Sandri, 'Towards emotional speech synthesis: A rule based approach', *Proc. 5th ISCA Speech Synthesis Workshop*, 219–220, (2004).