Improvements in Czech Expressive Speech Synthesis in Limited Domain *

Martin Grůber and Jindřich Matoušek

Department of Cybernetics, Faculty of Applied Sciences
University of West Bohemia, Czech Republic
http://www.kky.zcu.cz
{gruber,jmatouse}@kky.zcu.cz

Abstract. In our recent work, a method on how to enumerate differences between various expressive categories (communicative functions) has been proposed. To improve the overall impact of this approach to both the quality of synthetic expressive speech and expressivity perception by listeners, a few modifications are suggested in this paper. The main ones consist in a different way of expressive data processing and penalty matrix calculation. A complex evaluation using listening tests and some auxiliary measures was performed.

Keywords: expressive speech synthesis, unit selection, target cost, communicative functions

1 Introduction

At present, research in the field of expressive speech is very interesting topic for many scientists. The reason is that naturally sounding speech can be used in various complex systems, especially when considering dialogue systems focused on human-computer interaction. For such systems that attempt to "replace" a human in personal dialogues, there is even more need for incorporating expressivity in speech. Current TTS systems are for sure able to produce high quality speech. However, without any sign of expressivity the listeners (human partners in dialogues) always know that they are communicating with just a machine "pretending" to be a human.

To synthesize expressive speech, an expressivity description has to be designed. Many approaches have been suggested in the past. Continuous descriptions using multidimensional space with several axes to determinate "expressivity position" were described e.g. in [1]. Another option is a discrete division into various groups, for emotions e.g. happiness, sadness, anger, joy, etc. [2]. The discrete description is the most commonly used method and various sets of expressive categories are used, e.g. dialogue

* This work was supported by the European Regional Development Fund (ERDF), project "New Technologies for Information Society" (NTIS), European Centre of Excellence, ED1.1.00/02.0090. The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005) is highly appreciated.

acts [3], emotion categories [4] or categories like good news and bad news [5]. The systems dealing with expressive speech synthesis are often focused on a specific limited domain, e.g. [6].

In our work, we restricted the domain to conversations between seniors and a computer. As the topic for these discussions, personal photographs were chosen since the work started as a part of a major project whose aim was to develop a virtual senior companion with an audiovisual interface [7] (the more detailed background is described in [8–10]). Such a companion should help the elderly people when they are alone and want to talk to someone.

To describe expressivity within our limited domain, we decided to employ a set of so-called communicative functions (see Section 2). Even though the set is not a general solution for the expressivity description issue, a similar approach (with different expressive categories) might be used when designing a dialogue system for a different domain.

Since our current neutral TTS system ARTIC [11] is a data-driven system based on a unit selection method with a huge neutral speech corpus, there was a need to collect expressive data, i.e. an expressive speech corpus. It can be merged with the neutral one to obtain a robust system being able not only to produce expressive speech but also to keep the ability of synthesizing general texts. The description of the expressive speech corpus recording process and its annotation in terms of expressivity (communicative functions) can be found e.g. in [10]. The modifications of the unit selection algorithm that were performed to enable expressive speech synthesis are described in [9, 12]. The modifications mainly consisted in an adjustment of a target cost function. In the unit selection approach, the target cost is used to measure a suitability of a speech unit (a candidate) from a unit inventory (a database of candidates) for a target utterance (an utterance that is requested to be synthesized; it consists of so-called target units) in terms of prosodic features. One of the features is named communicative function (hereinafter referred to as CF), and a penalty is given to a candidate if its CF label does not meet the target unit requirement. In our recent work [12], suggestions about the penalty settings have been presented and in the current work we try to improve these settings to obtain synthetic speech of a better quality and with more expressed expressivity.

To evaluate achieved results, listening tests were performed to asses both the synthetic speech quality and the expressivity perception by listeners. In addition, overall impression of the expressive speech synthesis system was evaluated in dialogues from our limited domain.

The paper is organized as follows. The expressivity description and the set of CFs is briefly described in Section 2, short background of the target cost calculation is shown in Section 3. Modifications in expressive data processing and acoustic penalty matrix calculation (when compared to [12]) are presented in Section 4 and an evaluation is shown in Section 5. Conclusions are outlined in Section 6.

2 Communicative Functions

In the first phases of our research in this field, a set of CFs was designed to describe expressive categories appearing in the given dialogues of the limited domain. The set of CFs was inspired by dialogue acts proposed in [13] and more detailed description is in [9, 14, 12]. The process of expressive corpus recording and annotation is presented in [8] or [10]. Thus, in this work we present only a list of used CFs labels along with their relative occurrence in the expressive corpus:

- DIRECTIVE (2.4%)
- REQUEST (4.4%)
- WAIT (0.7%)
- APOLOGY (0.6%)
- GREETING (1.4%)
- GOODBYE (1.6%)
- THANKS (0.7%)
- SURPRISE (4.2%)
- SAD-EMPATHY (3.4%)
- HAPPY-EMPATHY (8.6%)
- SHOW-INTEREST (34.9%)
- CONFIRM (13.2%)
- DISCONFIRM (0.2%)
- ENCOURAGE (29.4%)
- NOT-SPECIFIED (7.4%)

Most of the CFs were detected only sparsely in the expressive corpus. Since we need to create a robust TTS system, all the CFs must be used. There might be some mistakes when representing distinctions between the sparsely appearing CFs but we believe that this effect does not influence the overall synthetic speech quality so much. Nevertheless, only the most appearing CFs are used for an evaluation to avoid result distortions caused by usage of not very well represented expressive categories.

It should be noted that the sum of all relative occurrence rates is greater than 100% in our case. This is caused by the fact that during the expressive corpus annotations by CFs, the annotators were allowed to label any sentence from the corpus with more than one CF if necessary. Thus, this is also reflected in the final annotations [12]. However, such sentences have been omitted from the experiments.

Speech units coming from the original neutral speech corpus were marked as *NEUTRAL* for the further processing. It should represent neutral speaking style (i.e not expressing any kind of expressivity).

3 Target cost for expressive speech

In the unit selection method, target cost C^t is a function that is used to measure a suitability of a speech unit u for a target unit t in terms of prosodic features. The target cost can be calculated as follows:

$$C^t = \frac{\sum_{j=1}^{n} w_j \cdot d_j}{\sum_{j=1}^{n} w_j}, \tag{1}$$

where C^t is the target cost of candidate u for target unit t, n is a number of features under consideration, w_j is a weight of *j-th* feature and d_j is an enumerated difference between *j-th* feature of candidate u and target unit t. The differences of particular features (d_j) will be further referred to as penalties.

For synthesis of expressive speech, the set of prosodic features is extended so it includes the feature of CF. This means that a measure enumerating a difference (penalty) between various CFs must be developed. In our recent work [12], a penalty matrix determining such penalties has been proposed. It is based on:

1. perceptual similarities revealed during annotation of expressive speech corpus [8] – perceptual penalty matrix **P**;
2. acoustic analysis that was performed on this corpus [15] – acoustic penalty matrix **A**.

Coefficients m_{ij} of the final penalty matrix **M** are calculated as

$$m_{ij} = \frac{w_p \cdot p_{ij} + w_a \cdot a_{ij}}{w_p + w_a}, \tag{2}$$

where p_{ij} and a_{ij} represent coefficients from matrices **P** and **A**, w_p and w_a are corresponding weighs.

Several combinations of weighs w_p and w_a were examined. Finally, $w_p = 3$ and $w_a = 1$ setting was used. Using this setting, the best results were achieved when subjectively comparing resulting synthetic speech. We also believe that the perceptual part should be emphasized.

In this work, more acoustic data preprocessing techniques were employed and more enhanced description of acoustic parameters was used to improve the acoustic penalty matrix **A**.

4 Acoustic penalty matrix enhancement

To enhance the acoustic penalty matrix (when comparing with [12]), data coming from the acoustic analysis of expressive speech [15] were preprocessed using outliers detection techniques. To describe various acoustic parameters such as $F0$, phoneme duration and RMS values, several statistical characteristics were employed.

4.1 Outliers detection

Results of the acoustic analysis of all voiced segments from the expressive speech corpus (in terms of various CFs) were used to create an acoustic penalty matrix. For these segments, 3 acoustic parameters were measured: $F0$, phoneme duration and RMS. It means that each voiced segment is represented by a 3 dimensional vector. For outliers detection, technique [16] based on Wilks method [17] was used. Using this approach, the outliers can be identified in a multidimensional space. The detected outliers were removed. Thus, for each CF a representative set of data was available.

4.2 Statistical characteristics

After outliers removal, the probability distribution of values of each acoustic parameter (in terms of various CFs) was described using 4 statistical measures: mean, standard

deviation, skewness and kurtosis (only mean was used in [12] but results of expressive speech acoustic analysis [15] suggest that other statistical measures might be influenced by expressivity too, as confirmed also by other studies [18, 19]). For each CF we obtained a 12-dimensional feature vector $\mathbf{x}_i$ (3 acoustic parameters $\times$ 4 statistical characteristics), where *i* represents *i*-th CF.

4.3 Enumerating differences

To enumerate differences between various CFs, suppression of absolute differences of various statistical characteristics of various acoustic parameters was needed. Thus, normalization was applied to the feature vectors as follows:

$$\forall i : \mathbf{x}_i^N = \frac{\mathbf{x}_i - \min_{\mathbf{x}}}{\max_{\mathbf{x}} - \min_{\mathbf{x}}}, \tag{3}$$

where $\mathbf{x}_i$ is a feature vector representing *i*-th CF, $\min_{\mathbf{x}}$ is a vector consisting of minimal values of all $\mathbf{x}_i$ and $\max_{\mathbf{x}}$ is a vector consisting of maximal values of all $\mathbf{x}_i$. Resulting values of vectors $\mathbf{x}_i^N$ are in the range of $\langle 0, 1 \rangle$.

To find coefficients a_{ij} of the acoustic penalty matrix **A**, Euclidean distance was used. The calculation of coefficients was performed in two steps:

1. Obtaining coefficients a'_{ij} as the Euclidean distance of normalized feature vectors:

$$\forall i, j : a'_{ij} = d(\mathbf{x}_i^N, \mathbf{x}_j^N), \tag{4}$$

   where *i* and *j* represent *i*-th and *j*-th CF, $\mathbf{x}_i^N$ is the normalized feature vector obtained from (3) and d represents the Euclidean distance;
2. Normalization of coefficients a'_{ij} to get the values into the range $\langle 0, 1 \rangle$ again:

$$\forall i, j : a_{ij} = \frac{a'_{ij}}{\max_{a'}}, \tag{5}$$

   where *i* and *j* represents *i*-th and *j*-th CF and $\max_{a'}$ is maximum value of all a'_{ij}. This is the same normalization as (3) but the $\min_{a'}$ can be omitted since it is always 0.

The perception penalty matrix **P** remains the same as proposed in [12]. The final penalty matrix **M** is then created as described in Section 3 using matrices **P** and **A** and keeping the same weighs. An excerpt from the matrix **M** is depicted in Table 1.

5 Evaluation

To evaluate an impact of our modifications on synthetic expressive speech, several views were used. At first, isolated utterances were presented to listeners for evaluation in terms of speech quality and expressivity perception. Next, part of dialogues between a computer and a human in two versions (expressive vs. neutral speech synthesis for the computer responses) were created and presented to listeners to obtain their preferences.

Table 1. Excerpt from the final penalty matrix **M**.

	CONFIRM	ENCOURAGE	HAPPY EMPATHY	NOT SPECIFIED	SAD EMPATHY	SHOW INTEREST	NEUTRAL
CONFIRM	0.00	0.71	0.40	0.48	0.41	0.50	0.72
ENCOURAGE	0.50	0.00	0.35	0.31	0.39	0.14	0.55
HAPPY-EMPATHY	0.25	0.24	0.00	0.21	0.29	0.27	0.58
NOT-SPECIFIED	0.26	0.12	0.15	0.00	0.22	0.13	0.46
SAD-EMPATHY	0.28	0.26	0.33	0.28	0.00	0.25	0.67
SHOW-INTEREST	0.53	0.15	0.43	0.27	0.41	0.00	0.45
NEUTRAL	0.72	0.55	0.58	0.46	0.67	0.45	0.00

5.1 Isolated utterances

Seven CFs (including *NEUTRAL*) were selected from the whole set to evaluate the performance of expressive speech synthesis. The selection was necessary for two reasons. First, some of the CFs occurred only sparsely in the expressive corpus and thus the coefficients of the penalty matrix might be a little distorted. Next, there are two main requirements for the listening tests that must be met: sufficient number of examples for each CF and sufficient number of listeners. Thus, there is a need to reduce the number of test queries to an acceptable level. However, full final penalty matrix was used during the synthesis (not only the excerpt shown in Table 1), i.e. speech units labelled with any CF (all speech units from both corpora) could be used to produce synthetic speech.

The following CF labels were used when synthesizing expressive speech for the evaluation: *SHOW-INTEREST*, *ENCOURAGE*, *CONFIRMATION*, *HAPPY-EMPATHY*, *SAD-EMPATHY* (that was chosen mainly to complete the set with supposedly contradictory pair of happy vs. sad empathy). We also used *NOT-SPECIFIED* and *NEUTRAL* which usage is assumed to produce neutral speech.

The evaluation is divided into two parts: synthetic speech quality and expressivity perception. In both parts, 13 listeners assessed 30 utterances (4 for each CF and 2 natural neutral utterances – to compare the synthetic speech quality with the natural speech).

The test stimuli were the same for both parts and were prepared as follows: random sentences with required CFs were selected from the corpora and content (text) of these sentences was modified — similar meaning was retained. This approach ensures that the sentences will be really synthesized and not only replayed. Before the synthesis, each sentence was tagged with the required CF label.

In addition, two auxiliary measures were employed to evaluate the speech quality and the expressivity perception. The first one is relative ratio of so-called "smooth joints". Smooth joint is a concatenation of two speech units that were originally adjacent in a speech corpus. The next one is relative ratio of speech units used for the synthesis being labelled with such CF that is required to be synthesized (hereinafter referred to as RRSU measure).

Speech quality During the listening test, the listeners were asked to assess the synthetic speech quality using 5-point MOS scale. In the following evaluation, we would like to present the comparison with the previous system presented in [12] which is called as *baseline system*.

The results of two independent evaluations are presented in Table 2: evaluation performed for the new system proposed in this work and former evaluation of baseline system from [12]. In addition to the absolute values of MOS score (mean values of all CFs in evaluation), a relative comparison with natural speech is presented in both cases. It allows us to compare results of various MOS tests.

Table 2. Results of MOS test.

Settings	new system	natural speech	baseline system	natural speech
MOS Score	3.5	4.6	3.4	4.7
Relative	69%	100%	65%	100%

We might conclude that the synthetic speech quality has improved from 65% (for the baseline system) to 69% (for the new proposed system). The difference is statistically significant – confirmed by ANOVA test (with $\alpha = 0.05$).

The auxiliary measure of relative ratio of smooth joints in synthetic speech is shown in Table 3. We can observe an improvement in smoothness when compared with the baseline system for almost all CFs. This is consistent with results of MOS evaluation.

Expressivity perception Beside the speech quality evaluation, the listeners were asked to mark if they are able to perceive any kind of expressivity in the presented utterances. They were not instructed to mark any specific CF since the main objective of the test was just to generally evaluate a difference in speech perception when comparing an expressive TTS and a neutral (mainstream) TTS approach. This way we also tried to avoid any forced-choice evaluation. The listeners were provided with a few samples of expressive and neutral sentences to outline a definition of expressivity in speech. The results of the evaluation are shown in Table 4.

The mean value of expressivity perception ratio is 54%, the mean value of auxiliary RRSU measure is 50%. It is remarkable that the expressivity perception ratio of 42% was achieved in natural neutral utterances. This could mean that utterances in neutral

Table 3. Relative occurrence of "smooth joints" in the resulting synthetic speech.

CF label	new system	baseline system
CONFIRM	80%	80%
ENCOURAGE	76%	70%
HAPPY-EMPATHY	77%	67%
SAD-EMPATHY	80%	69%
SHOW-INTEREST	82%	75%
mean	**79%**	**72%**
NOT-SPECIFIED	82%	82%
NEUTRAL	82%	not available

Table 4. Expressivity perception ratios and values of RRSU measure in terms of various CFs.

CF label	expressivity perception	unable to decide	RRSU measure
CONFIRM	69%	4%	75%
ENCOURAGE	42%	8%	68%
HAPPY-EMPATHY	50%	10%	35%
SAD-EMPATHY	63%	4%	33%
SHOW-INTEREST	46%	4%	40%
mean	**54%**	**6%**	**50%**
NOT-SPECIFIED	10%	0%	4%
NEUTRAL	15%	0%	100%
natural speech	42%	4%	–

corpus are not expressively neutral as it was supposed or that the listeners are very sensitive in expressivity perception. For synthetic neutral speech, the results mean that the listeners perceived almost no expressivity. It can be also observed that the RRSU measure does not correspond with the expressivity perception very much.

The value 4% of RRSU for *NOT-SPECIFIED* was further inspected. We found out that for synthesis of sentences tagged with this CF, speech units coming from neutral corpus (labelled with *NEUTRAL*) were mostly selected. This might suggest that these two CFs are very similar (neither should express any kind of expressivity).

To prove that the results are different from those that would be achieved by chance, several measures were used: precision, recall, F1 measure and accuracy. These are often used in classification tasks to evaluate classifiers. However, the listeners can be also viewed as classifiers classifying into two classes: perceive or do not perceive expressivity ("unable to decide" responses were not considered here). Thus, these measures were calculated for our results and for results of a random simulation. We simulated a situation when the listeners evaluate the listening test randomly. The measured values are shown in Table 5.

Table 5. Measures for classification of expressivity in synthetic speech; comparison of the real results and the results achieved by the random simulation.

measure	listeners	simulation
precision	0.92	0.72
recall	0.58	0.50
F1 measure	0.71	0.59
accuracy	0.66	0.50

It can be concluded that the results achieved using the listening test are above the chance level. It means that the expressivity is quite recognizable in the synthetic speech.

Unlike the speech quality evaluation, comparison with [12] is not possible for expressivity perception because of missing results in that previous work. On the other hand, a comparison of RRSU measures would be distorted since in [12] different penalty matrix coefficients were used (not in range of $\langle 0, 1 \rangle$ and thus influencing the target cost calculation significantly).

5.2 Dialogues

For evaluation of an overall impact of synthetic expressive speech in dialogues, test stimuli were prepared as follows:

- 6 parts of natural dialogues[1] between a human and a computer avatar (approximately 1 minute in length) were randomly selected (referred to as *mini-dialogues*);
- texts of avatar responses were extracted from the mini-dialogues and were modified in order to avoid just replaying from the corpus during the following synthesis;
- the modified texts were synthesized using neutral TTS system ARTIC [11] and the new system proposed in this work;
- the original avatar responses in mini-dialogues were replaced by the newly synthesized utterances producing two versions for each mini-dialogue: one with neutrally synthesized responses and one with expressively synthesized responses.

Each mini-dialogue contained 4 avatar responses in average expressing various CFs, mostly *SHOW-INTEREST* or *ENCOURAGE*. However, all CFs in evaluation were used at least once. The mini-dialogues (both version at once) were then presented to listeners to mark which version is more pleasant, more natural and preferred. The results are shown in Table 6.

The expressive speech were much more preferred to the neutral one (83%). This result is one of the most important findings since the expressive speech synthesis proposed in this work is supposed to be used in similar dialogues.

[1] The process of natural dialogues collection is described e.g. in [10].

Table 6. Evaluation of expressive speech synthesis in dialogues.

synthesis method	preference
neutral	8 %
expressive	83 %
unable to decide	9 %

6 Conclusions & future work

In this work, improvements in Czech expressive speech synthesis in limited domain were shown in comparison with neutral synthesis and with our previous work in this field. Benefits of the penalty matrix coefficients calculation enhancement were presented in the form of listening test results and an auxiliary measure. The results show that the proposed system improved the synthetic speech quality when compared to the previous work and that expressive speech is preferred to the neutral one by listeners in dialogues.

For the future work, other modifications of penalty matrix approach should be considered. The main challenge for the near future is to create a phoneme-dependent acoustic penalty matrix since differences of acoustic parameters might vary in terms of various phonemes or phoneme groups (like vowels/consonants).

References

1. J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.
2. R. R. Cornelius, *The science of emotion: Research and tradition in the psychology of emotions*. NJ, USA: Prentice-Hall, Englewood Cliffs, 1996.
3. A. K. Syrdal, A. Conkie, Y.-J. Kim, and M. Beutnagel, "Speech acts and dialog TTS," in *Proceedings of the 7th ISCA Speech Synthesis Workshop – SSW7*, Kyoto, Japan, 2010, pp. 179–183.
4. E. Zovato, A. Pacchiotti, S. Quazza, and S. Sandri, "Towards emotional speech synthesis: A rule based approach," in *Proceedings of the 5th ISCA Speech Synthesis Workshop – SSW5*, Pittsburgh, PA, USA, 2004, pp. 219–220.
5. W. Hamza, R. Bakis, E. M. Eide, M. A. Picheny, and J. F. Pitrelli, "The IBM expressive speech synthesis system," in *Proceedings of the 8th International Conference on Spoken Language Processing – ISCLP*, Jeju, Korea, 2004, pp. 2577–2580.
6. S. Krstulovic, A. Hunecke, and M. Schroder, "An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements," in *Proceedings of Interspeech*, Antwerp, Belgium, 2007, pp. 1897–1900.
7. P. Ircing, J. Romportl, and Z. Loose, "Audiovisual interface for Czech spoken dialogue system," in *IEEE 10th International Conference on Signal Processing Proceedings*. Beijing, China: Institute of Electrical and Electronics Engineers, Inc., 2010, pp. 526–529.
8. M. Grůber and J. Matoušek, "Listening-test-based annotation of communicative functions for expressive speech synthesis," in *Text, Speech and Dialogue, proceedings of the 13th International Conference TSD*, ser. Lecture Notes in Computer Science, vol. 6231. Berlin-Heidelberg, Germany: Springer, 2010, pp. 283–290.

9. M. Grůber and D. Tihelka, "Expressive speech synthesis for Czech limited domain dialogue system – basic experiments," in *IEEE 10th International Conference on Signal Processing Proceedings*, vol. 1. Beijing, China: Institute of Electrical and Electronics Engineers, Inc., 2010, pp. 561–564.
10. M. Grůber, M. Legát, P. Ircing, J. Romportl, and J. Psutka, "Czech Senior COMPANION: Wizard of Oz data collection and expressive speech corpus recording and annotation," in *Human Language Technology. Challenges for Computer Science and Linguistics*, ser. Lecture Notes in Computer Science, Z. Vetulani, Ed., vol. 6562. Berlin-Heidelberg, Germany: Springer, 2011, pp. 280–290.
11. D. Tihelka, J. Kala, and J. Matoušek, "Enhancements of Viterbi search for fast unit selection synthesis," in *Proceedings of Interspeech*, Makuhari, Japan, 2010, pp. 174–177.
12. M. Grůber, "Enumerating differences between various communicative functions for purposes of Czech expressive speech synthesis in limited domain," in *Proceedings of Interspeech*, Portland, Oregon, USA, 2012, pp. 650–653.
13. A. K. Syrdal and Y.-J. Kim, "Dialog speech acts and prosody: Considerations for TTS," in *Proceedings of Speech Prosody*, Campinas, Brazil, May 2008, pp. 661–665.
14. M. Grůber and Z. Hanzlíček, "Czech expressive speech synthesis in limited domain: Comparison of unit selection and HMM-based approaches," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin-Heidelberg, Germany: Springer, 2012, vol. 7499, pp. 656–664.
15. M. Grůber, "Acoustic analysis of Czech expressive recordings from a single speaker in terms of various communicative functions," in *Proceedings of the 11th IEEE International Symposium on Signal Processing and Information Technology*. 345 E 47TH ST, NEW YORK, NY 10017, USA: IEEE, 2011, pp. 267–272.
16. A. Trujillo-Ortiz, R. Hernandez-Walls, A. Castro-Perez, and K. Barba-Rojo. (2006) MOUTLIER1: Detection of outlier in multivariate samples test. A MATLAB file. [online; cited 2012-10-29]. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=12252
17. S. S. Wilks, "Multivariate statistical outlier," *The Indian Journal of Statistics*, vol. 25, no. 4, pp. 407–426, 1963.
18. J. Přibil and A. Přibilová, "Statistical analysis of spectral properties and prosodic parameters of emotional speech," *Measurement Science Review*, vol. 9, pp. 95–104, 2009.
19. ——, "Statistical analysis of complementary spectral features of emotional speech in czech and slovak," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, I. Habernal and V. Matoušek, Eds. Berlin-Heidelberg, Germany: Springer, 2011, vol. 6836, pp. 299–306.