Czech Expressive Speech Synthesis in Limited Domain
Comparison of Unit Selection and HMM-Based Approaches*

Martin Grůber and Zdeněk Hanzlíček

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

Abstract. This paper deals with expressive speech synthesis in a limited domain restricted to conversations between humans and a computer on a given topic. Two different methods (unit selection and HMM-based speech synthesis) were employed to produce expressive synthetic speech, both with the same description of expressivity by so-called communicative functions. Such a discrete division is related to our limited domain and it is not intended to be a general solution for expressivity description. Resulting synthetic speech was presented to listeners within a web-based listening test to evaluate whether the expressivity is perceived as expected. The comparison of both methods is also shown.

Keywords: expressive speech synthesis, unit selection, HMM-based speech synthesis, communicative functions.

1 Introduction

Nowadays, production of synthetic expressive speech is a hot topic in the field of speech synthesis research. Modern text-to-speech (TTS) systems are able to produce high quality, intelligible and naturally sounding speech. The TTS systems can be used in our everyday life e.g. for reading emails and web pages, to make text documents available to vision impaired people [1] or to help hearing impaired people to better understand television programs [2]. Another application of TTS systems is in smart phones [3], navigation systems, dialog systems [4], or as a component of audiovisual speech synthesis systems [5]. However, for the human–computer interaction in a dialogue system, the naturalness and intelligibility are not a sufficient criterion to rate the quality of synthetic speech. In the direct contact with a human user, who expects to be served by another human, as presented e.g. in [6], the TTS system should express also some human's feeling and attitudes.

Thus, some kind of expressivity is necessary to be incorporated in the synthetic speech. That way the listeners could completely understand the information and its nature that is communicated with them. Various techniques incorporating expressivity into

* This research was supported by the Technology Agency of the Czech Republic, project No. TA01011264 and by the grant of the University of West Bohemia, project No. SGS-2010-054. The access to the MetaCentrum computing facilities provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" LM2010005 funded by the Ministry of Education, Youth, and Sports of the Czech Republic is highly appreciated.

P. Sojka et al. (Eds.): TSD 2012, LNCS 7499, pp. 656–664, 2012.

synthetic speech have been introduced so far. Some of them perform the modification of acoustic parameters of synthesized speech [7] or voice conversion [8], others produce special speech or non-speech expressions or use emphasis [9] to evoke some expressivity. Methods using unit selection techniques are based on the creating of a unit inventory containing expressive speech units (or also other non-speech events) [10,11,12].

Since the general expressive speech synthesis is a very complex task, it is usually somehow limited (as well as limited domain speech synthesis systems are). In our work, the domain was restricted to conversations between seniors and a computer about their personal photographs. The work started as a part of a major project whose objective was to develop a virtual senior companion. The detailed motivation is described in [13,14].

To incorporate expressivity into our current Czech unit selection TTS system [15] and newly developed HMM-based system [16], two basic steps had to be performed.

First, an expressivity description was proposed. Many approaches have been suggested in the past, e.g. continuous description using a 2-dimensional space with two axis, one for positive/negative and one for active/passive determination of expressivity position in this space [17]. Another option is a discrete division, e.g. corresponding to the fundamental emotions like happiness, sadness, anger, joy, etc. Within our limited domain we decided to employ a slightly different approach, similar to the one described in [18], where so-called dialogue acts are proposed. A set of communicative functions was designed to fit our limited domain. This set is not a general solution for the problem of speech expressivity description. The communicative functions are assumed to describe more the part of a dialogue than the attitude of the speaker.

Next, the communicative functions need to be somehow incorporated into a speech corpus that is used in our speech synthesis systems. Thus, an expressive speech corpus was recorded. The process of the data preparation and the corpus recording is in more details described in [13]. In the resulting corpus, all utterances are labeled by a feature indicating the appropriate communicative function. This corpus is completed by merging with a part of neutral corpus (which is normally used for neutral speech synthesis) to cover all possible speech units that can occur in any input text. In the case of unit selection method, this corpus was used to build a speech unit inventory, in the case of HMM-based synthesis, corresponding statistical models were trained. These processes are described in Section 2. Thus, the systems were ready to be used for synthetic expressive speech generation.

The produced synthetic speech has to be evaluated to determine whether the expressivity was incorporated successfully or not, and whether the quality and intelligibility of the synthetic speech is still at an acceptable level. Both criterion were evaluated by using listening tests. The backgroud of those tests and the achieved results are described in Section 3. Finally, some conclusions and future work is suggested in Section 4.

2 Expressive Speech Synthesis

Expressive speech can be produced by using various synthesis methods. In our work, we focused on two of them that are currently widely used: the unit selection method and the HMM-based speech synthesis. For both approaches, expressive speech data need to be obtained and labelled with appropriate expressivity descriptors. The process of

Table 1. The set of the communicative functions and their occurrence rate in the expressive speech corpus

Communicative function (symbol)	*Occurrence rate*	*Examples*
directive (DIRECTIVE)	2.36%	Tell me that. Talk.
request (REQUEST)	4.36%	Let's get back to that later.
wait (WAIT)	0.73%	Wait a minute. Just a moment.
apology (APOLOGY)	0.59%	I'm sorry. Excuse me.
greeting (GREETING)	1.37%	Hello. Good morning.
goodbye (GOODBYE)	1.64%	Goodbye. See you later.
thanks (THANKS)	0.73%	Thank you. Thanks.
surprise (SURPRISE)	4.19%	Do you really have 10 siblings?
sad empathy (SAD-EMPATHY)	3.44%	I'm sorry to hear that. It's really terrible.
happy empathy (HAPPY-EMPTATHY)	8.62%	It's nice. Great. It had to be wonderful.
showing interest (SHOW-INTEREST)	34.88%	Can you tell me more about it?
confirmation (CONFIRM)	13.19%	Yes. Yeah. I see. Well. Hmm.
disconfirmation (DISCONFIRM)	0.23%	No. I don't understand.
encouragement (ENCOURAGE)	29.36%	Well. For example? And what about you?
not specified (NOT-SPECIFIED)	7.36%	Do you hear me well? My name is Paul.

recording natural expressive speech corpus and its annotation are presented in [13,19][1] where so-called communicative functions were used to describe the expressivity. The set of communicative functions is listed in Table 1 along with their occurrence rate in the expressive corpus.

Obviously, a lot of communicative functions occurred only sparsely in the corpus. For that reason, we decided to use only the most frequent ones in our experiments to obtain representative results – *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 communicative function *NOT-SPECIFIED* that is related to the neutral speech.

The process of production of expressive speech using the aforementioned methods is described in the following subsections.

[1] Since the expressive corpus did not sufficiently cover the speech units that can occure in the Czech language, a part of the neutral corpus was merged with the expressive one. Only the sentences containing missing units were chosen and integrated into the expressive corpus (both corpora were recorded under the same conditions by the same female speaker with a relatively short time lag). This way a complete expressive speech corpus for employing in our TTS systems was created.

2.1 Unit Selection Method

Modifications of the standard unit selection algorithm are in details described in [14]. Thus, the process will be described only briefly.

The main issue of the unit selection method is to select the most appropriate sequence of speech units from the speech corpus (speech unit inventory). This sequence should form speech as smooth and natural as possible (in view of several prosodic and acoustic features). This is ensured by considering two criteria – target cost and concatenation cost.

The target cost reflects the level of approximation of a target unit by any of candidates from the speech unit inventory, in other words, how a candidate from the unit inventory fits the required target unit — a theoretical unit whose features are specified on the basis of the text to be synthesized.

The features affecting the target cost value are usually nominal, e.g. phonetic context, prosodic context, position in word, position in sentence, position in syllable, etc. Currently, these features are chosen manually, based on previous experience; in the future, an algorithm choosing them automatically on the basis of given data (or a speaker) nature might be suggested if it is possible [20]. The value of the target cost T_i for the unit candidate u_i is calculated as follows:

$$T_i = \frac{\sum_{j=1}^{n} w_j d_j}{\sum_{i=j}^{n} w_j} \tag{1}$$

where n is a number of features under consideration, w_j is a weigh of *j-th* feature and d_j is an enumerated difference between *j-th* features of a candidate for unit u_i and target unit t_i. The differences of particular features (d_j) will be further referred to as penalties.

In the case of expressive speech synthesis, the set of the features used within the target cost is extended with an additional feature, a communicative function. The penalty d_{cf} between candidate u_i and target unit t_i is calculated as follows:

$$d_{cf} = \begin{cases} 1 & \text{if } cf_t = cf_c \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where cf_t is the communicative function of target unit t_i and cf_c is the communicative function of the candidate for unit u_i.

Finally, the weigh of this penalty has to be specified since the target cost is calculated as a weighed sum of particular penalties. For our preliminary experiments, the weigh of the penalty for communicative function was determined ad-hoc; its value is one of the highest from all the weighs (e.g. it is 4x higher than the weigh for the phonetic context). It reflects our assumption that this feature should considerably influence the overall criterion.

2.2 HMM-Based Method

Nowadays, beside concatenative unit selection method, HMM-based speech synthesis is one of most researched synthesis methods [21]. In this method, statistical models (an extended type of HMMs) are trained from natural speech database. Spectral parameters, fundamental frequency and eventually some excitation parameters are modeled simultaneously by the corresponding multi-stream HMMs.

Table 2. A list of contextual factors and their values

Factors	Possible values
Previous, current and next phoneme	Czech phoneme set
Phone position in prosodic word (forward and backward)	1, 2, 3, 4, 5 ...
Prosodic word position in clause (forward and backward)	
Prosodeme type	terminating satisfactorily/unsatisfactorily, non-terminating, formal null
Communicative function	see begin of Section 2

To model the variation of spectral and other speech parameters, so-called contextual factors are defined. They describe the general context of speech units within the utterance regarding the phonetic, prosodic and linguistics characteristics of speech for the given language, speaker, etc. Contextual factors utilized in our experimental TTS system are listed in Table 2. For a more thorough explanation see e.g. [16].

For a more robust estimation of model parameters, HMMs are clustered by using decision tree-based context-clustering algorithm, that analyzes the similarity of speech units in different contexts. As a result, similar units share one common model. The clustering trees created during that process are also employed for synthesis of speech units unseen within the training stage. In synthesis stage, trajectories of speech parameters are generated from these trained models in the maximum likelihood sense.

Within the HMM-based speech synthesis, several different methods for modeling of the expressivity or speaking styles have been introduced. The most simple approach is an independent training of HMMs for each expression (so-called style dependent model [22]). An evident disadvantage of that approach is a quite large amount of speech data needed for sufficient training of models for particular expressions.

A better solution is to train one set of HMMs for data for all expressions together, the particular expressions are distinguished by addition of a new contextual factor (so-called style mixed model [22]). In this approach, models corresponding to speech units, which are (almost) identical for more expressions, will be clustered and the common model will be trained from all corresponding data. On the other hand, in the case of substantial differences between speech units belonging to particular expressions, corresponding models remain independent to preserve the variability of those expressions.

Recently, methods based on model adaptation [23,24] are preferred because they allow to control the speech style or expression more precisely and require less training data. However, for our first experiments we decided to use the style mixed model with an additional contextual factor for communicative function. Advanced methods for modeling expression will be an objective of our future work.

3 Experiments and Results

To evaluate our modified TTS systems, two listening tests were performed – the first was focused on the perception of expressivity and the second assessed the quality of resulting synthetic speech. Both test were organized on the client-server basis using

a specially developed web application. Thus, listeners were able to work on the test from their homes without any contact with the test organizers.

3.1 Comparison in Terms of Expressivity

The first listening test focused on the perception of expressivity in synthetic speech was carried out by 8 listeners. They listened to a big set of 170 synthetic utterances — 85 of them were synthesized by using the unit selection method (referred to as *USEL*) and 85 by using the HMM-based approach (referred to as *HMM*). The text contents of sentences were equal for both *USEL* and *HMM*. For each synthesis method, 35 sentences were of a neutral content and the remaining 50 sentences were of an expressive content. Moreover, disregarding the text content, 2 versions of both TTS-systems were evenly employed for synthesis of particular utterances: the default version (producing neutral speech) and the modified version employing communicative functions (producing expressive speech).

In this test, the listeners were asked to indicate whether they percieve any kind of expressivity or speaker's affective state in the presented utterance. In Table 3 and Table 4 the expressivity perception improvement is shown for both speech synthesis methods. Evidently, there is almost no improvement for the neutral text content whereas a slight improvement can be observed for the expressive text content.

Table 3. Comparison of used approaches in terms of expressivity perception considering neutral text content

approach	*expressivity ratio using CFs*	*expressivity ratio w/out using CFs*
HMM-based	4%	3%
Unit selection	10%	10%

Table 4. Comparison of used approaches in terms of expressivity perception considering expressive text content

approach	*expressivity ratio using CFs*	*expressivity ratio w/out using CFs*
HMM-based	15%	8%
Unit selection	45%	39%

Obviously, the difference between the perception of expressivity for the *USEL* and *HMM* is remarkable. Comparison of both synthesis methods disregarding the sentence content is presented in Table 5. We can conclude that expressivity is better perceived in speech synthesized by *USEL*. This can be also related to the result presented in Section 3.2, i.e. that the synthetic speech produced by unit selection method is of a better quality.

Table 5. Overall comparison of used approaches in terms of expressivity perception

approach	*expressivity ratio using CFs*
HMM-based	10%
Unit selection	28%

The overall improvement in the expressivity perception regardless of the used synthesis method achieved by employing communicative functions in the synthesis process is presented in Table 6. Considering neutral text content, no improvement was unfortunately achieved. However, for texts with an expressive content, the perception of the expressivity was more evident.

Table 6. Overall evaluation of improvement in perception of expressivity by listeners regardless of the approach

text content	*expressivity ratio using CFs*	*expressivity ratio w/out using CFs*
neutral	7%	7%
expressive	30%	23%

This result might suggest that not only the usage of communicative functions might influence the expressivity perception in the synthetic speech. The text content seems to be also important for the listeners when decide whether they feel any kind of expressivity or not.

3.2 Comparison in Terms of Quality

For the absolute evaluation of the speech quality, MOS (mean opinion score) listening test was organized. 12 participants of that test listened to 40 isolated utterances and rated them according to the standard MOS 5-point scale (1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent). Besides sentences synthesized by the *USEL* (18) and *HMM* (18), the test also contained several natural utterances (4). The results are presented in Table 7.

Table 7. Results of MOS test

Method	HMM-based synthesis	Unit selection	Natural speech
Score	2.71	3.51	4.44

4 Conclusion and Future Work

In this paper, our first experiments on expressive speech synthesis using HMM-based method were presented. We also compared the level of the perception of expressivity in synthetic speech produced by using that method and unit selection. The quality of

synthetic expressive speech was also assessed for both methods. For the evaluation, web-based listening tests were performed.

From the achieved results we can conclude that the level of the expressivity perception increased when using expressive speech synthesis (employing the communicative functions), regardless of the used method. However, the improvement is apparent only for sentences with an expressive content. For those sentences, the expressivity perception rate increased from 23% to 30% while for sentences with a neutral content the rate remained almost unchanged at 7%. It is also remarkable that 23% of listeners perceived some kind of expressivity in utterances that were synthesized neutrally (i.e. without using communicative functions) but their content was expressive.

When comparing both synthesis methods, the expressivity is more perceivable for synthetic speech produced by unit selection algorithms (28% *USEL* vs. 10% *HMM*).

When comparing the quality of expressive speech, the *USEL* is preferred to *HMM*. This might be caused by the fact that our HMM-based speech synthesis system is only at an early stage of development and is supposed to be improved in the future. An overall improvement in the quality of synthetic expressive speech is also our future task. To improve the level of the expressivity perception using *USEL*, some modifications of the selection algorithm are planned to be done. For example, the similarity of speech features related to various communicative functions should be taken into account when a unit with required communicative function is not available.

References

1. Matoušek, J., Hanzlíček, Z., Campr, M., Krňoul, Z., Campr, P., Grůber, M.: Web-Based System for Automatic Reading of Technical Documents for Vision Impaired Students. In: Habernal, I., Matoušek, V. (eds.) TSD 2011. LNCS (LNAI), vol. 6836, pp. 364–371. Springer, Heidelberg (2011)
2. Matoušek, J., Vít, J.: Improving automatic dubbing with subtitle timing optimisation using video cut detection. In: Proceedings of ICASSP, Kyoto, Japan, pp. 2385–2388 (2012)
3. Tihelka, D., Stanislav, P.: ARTIC for assistive technologies: Transformation to resource-limited hardware. In: Proceedings of World Congress on Engineering and Computer Science 2011, San Francisco, USA, Newswood Limited, International Association of Engineers, pp. 581–584 (2011)
4. Švec, J., Šmídl, L.: Prototype of Czech Spoken Dialog System with Mixed Initiative for Railway Information Service. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 568–575. Springer, Heidelberg (2010)
5. Krňoul, Z., Železný, M.: A development of Czech talking head. In: Proceedings of ICSPL 2008, pp. 2326–2329 (2008)
6. Ptáček, J., Ircing, P., Spousta, M., Romportl, J., Loose, Z., Cinková, S., Gil, J.R., Santos, R.: Integration of speech and text processing modules into a real-time dialogue system. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS (LNAI), vol. 6231, pp. 552–559. Springer, Heidelberg (2010)
7. Přibilová, A., Přibil, J.: Harmonic model for female voice emotional synthesis. In: Fierrez, J., Ortega-Garcia, J., Esposito, A., Drygajlo, A., Faundez-Zanuy, M. (eds.) BioID MultiComm 2009. LNCS, vol. 5707, pp. 41–48. Springer, Heidelberg (2009)
8. Přibil, J., Přibilová, A.: Application of expressive speech in TTS system with cepstral description. In: Esposito, A., Bourbakis, N.G., Avouris, N., Hatzilygeroudis, I. (eds.) HH and HM Interaction. LNCS (LNAI), vol. 5042, pp. 200–212. Springer, Heidelberg (2008)

9. Hamza, W., Bakis, R., Eide, E.M., Picheny, M.A., Pitrelli, J.F.: The IBM expressive speech synthesis system. In: Proceedings of the 8th International Conference on Spoken Language Processing, ISCLP, Jeju, Korea, pp. 2577–2580 (2004)
10. Iida, A., Campbell, N., Higuchi, F., Yasumura, M.: A corpus-based speech synthesis system with emotion. Speech Communication 40, 161–187 (2003)
11. Hofer, G., Richmond, K., Clark, R.: Informed blending of databases for emotional speech. In: Proceedings of Interspeech, Lisbon, Portugal, International Speech Communication Association, pp. 501–504 (2005)
12. Bulut, M., Narayanan, S.S., Syrdal, A.K.: Expressive speech synthesis using a concatenative synthesiser. In: Proceedings of the 7th International Conference on Spoken Language Processing, ICSLP, Denver, CO, USA, pp. 1265–1268 (2002)
13. Grůber, M., Legát, M., Ircing, P., Romportl, J., Psutka, J.: Czech Senior COMPANION: Wizard of Oz Data Collection and Expressive Speech Corpus Recording and Annotation. In: Vetulani, Z. (ed.) LTC 2009. LNCS, vol. 6562, pp. 280–290. Springer, Heidelberg (2011)
14. Grůber, M., Tihelka, D.: Expressive speech synthesis for Czech limited domain dialogue system – basic experiments. In: 2010 IEEE 10th International Conference on Signal Processing Proceedings, vol. 1, pp. 561–564. Institute of Electrical and Electronics Engineers, Inc., Beijing (2010)
15. Tihelka, D., Kala, J., Matoušek, J.: Enhancements of Viterbi search for fast unit selection synthesis. In: Proceedings of Interspeech, Makuhari, Japan, pp. 174–177 (2010)
16. Hanzlíček, Z.: Czech HMM-Based Speech Synthesis. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 291–298. Springer, Heidelberg (2010)
17. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178 (1980)
18. Syrdal, A.K., Kim, Y.J.: Dialog speech acts and prosody: Considerations for TTS. In: Proceedings of Speech Prosody, Campinas, Brazil, pp. 661–665 (2008)
19. Grůber, M., Matoušek, J.: Listening-Test-Based Annotation of Communicative Functions for Expressive Speech Synthesis. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 283–290. Springer, Heidelberg (2010)
20. Tihelka, D., Romportl, J.: Exploring automatic similarity measures for unit selection tuning. In: Proceedings of Interspeech, Brighton, Great Britain, ISCA, pp. 736–739 (2009)
21. Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Communication 51, 1039–1064 (2009)
22. Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T.: Modeling of various speaking styles and emotions for HMM-based speech synthesis. In: Proceedings of Eurospeech 2003, pp. 1829–1832 (2003)
23. Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T.: A style control technique for HMM-based speech synthesis. In: Proceedings of Interspeech 2004, pp. 1437–1440 (2004)
24. Nose, T., Kobayashi, Y.K.,, T.: A speaker adaptation technique for MRHSMM-based style control of synthetic speech. In: Proceedings of ICASSP 2007, pp. 833–836 (2007)