Enumerating Differences Between Various Communicative Functions for Purposes of Czech Expressive Speech Synthesis in Limited Domain

Martin Grůber

NTIS, Faculty of Applied Sciences,
University of West Bohemia, Pilsen, Czech Republic
gruber@kky.zcu.cz

Abstract

This paper deals with determination of a penalty matrix that should represent differences between various communicative functions. These are supposed to describe expressivity that can occur in expressive speech and were designed to fit a limited domain of conversations between seniors and a computer on a given topic. The penalty matrix is assumed to increase a rate of the expressivity perception in synthetic speech produced by unit selection method. It should reflect both acoustic differences and differences based on human perception of expressivity.

Index Terms: expressive speech synthesis, unit selection, target costs, communicative functions

1. Introduction

Currently, there is a boom in the field of expressive speech research, both the expressive speech synthesis/analysis and expressivity/emotion recognition. Text-to-speech (TTS) systems used for producing synthetic speech are at a high level, i.e. they are able to produce high quality and naturally sounding speech. However, to use synthetic speech in dialogue systems (restaurant reservation, information on flights, trains or weather) or in any other human–computer interactive systems (virtual computer companions, computer games), the voice interface should be more friendly to make the user to feel more involved in the interaction or communication.

Thus, some kind of expressivity or speaker's attitude is necessary to be incorporated in the synthetic speech. That way, the listeners could completely understand the information and its nature that is communicated. 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, 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 [1]. The more detailed background is described in [2] or [3].

To synthesize expressive speech, an expressivity description has to be designed. Many approaches have been suggested in the past, e.g. a continuous description using a 2-dimensional space with two axes, one for positive/negative and one for active/passive determination of expressivity position in this space. Another option is a discrete division, e.g. into various groups like happiness, sadness, anger, joy, etc. Within our limited domain, we decided to employ a little bit different approach, similar to the one that was described in [4], where so-called dialogue acts are proposed. A set of communicative functions (see Section 2) was designed to describe expressivity in our limited domain. However, the set is not a general solution for the expressivity description issue.

To incorporate expressivity into our current TTS system ARTIC [5] based on a unit selection method, expressive speech data was collected [2] and modifications of the unit selection algorithm were made [3]. The modifications consisted in an adjustment of a target cost function. In the unit selection approach, it is used to measure a suitability of a speech unit (a candidate) from a unit inventory (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 (referred to as CF), and a penalty is given to a candidate if its CF label does not meet the target unit requirement. So far, we have used a simple penalty function in the form of

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

where d_{cf} is a difference (penalty), cf_t is a CF of a target unit and cf_c is a CF of a candidate for this target unit.

This penalty assumes that a difference between various CFs is equal. It means that if we are synthesizing a sentence, e.g. in a *HAPPY-EMPATHY* manner and there is no suitable unit with the *HAPPY-EMPATHY* label in the unit inventory, all CFs are considered equally regardless of how similar they are to the required one. In this work, we would like to change this approach and to try to determine some similarity measure between various CFs. We believe that this way an improvement in the synthetic expressive speech can be achieved in terms of the expressivity perception. However, the synthetic speech quality might be also affected, both negatively or positively. Thus, the evaluation of changes in quality was performed too.

The paper is organized as follows. Expressivity description and set of CFs is described in Section 2. Enumeration of differences between various CFs is described in Section 3 and an evaluation is presented in Section 4. Conclusions are outlined in Section 5.

2. Communicative Functions

The proposed CFs are designed to describe expressivity in dialogues of our limited domain. The set of CFs (shown in Table 1 along with their occurrence rate in the expressive corpus) was inspired by so-called dialogue acts presented in [4] and was obtained using expressive speech corpus annotation described in [6].

As it is obvious from the statistics, most of the CFs were detected only sparsely in the corpus. However, if we want to successfully produce expressive synthetic speech in terms of

Table 1: The set of the CFs (their symbols) and relative occurrence rate in the expressive speech corpus.

symbol of comm. function	*example*	*occurr. rate*
DIRECTIVE	Tell me that. Talk.	2.36%
REQUEST	Let's get back to that later.	4.36%
WAIT	Wait a minute. Just a moment.	0.73%
APOLOGY	I'm sorry. Excuse me.	0.59%
GREETING	Hello. Good morning.	1.37%
GOODBYE	Goodbye. See you later.	1.64%
THANKS	Thank you. Thanks.	0.73%
SURPRISE	Do you really have 10 siblings?	4.19%
SAD-EMPATHY	I'm sorry to hear that. It's really terrible.	3.44%
HAPPY-EMPATHY	It's nice. Great. It had to be wonderful.	8.62%
SHOW-INTEREST	Can you tell me more about it?	34.88%
CONFIRM	Yes. Yeah. I see. Well. Hmm.	13.19%
DISCONFIRM	No. I don't understand.	0.23%
ENCOURAGE	Well. For example? And what about you?	29.36%
NOT-SPECIFIED	Do you hear me well? My name is Paul.	7.36%

our limited domain, we have to take into consideration all the CFs. There might be some mistakes when representing distinctions between the sparsely appearing CFs but we believe that this effect will not influence the overall synthetic speech quality so much. Nevertheless, only the most appearing CFs were later used for an evaluation to avoid result distortions caused by usage of not very well represented expressive categories. It means that all the CFs were used for creating the penalty matrix but only the following ones were used when synthesizing expressive speech: *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* which usage is assumed to produce neutral speech.

It should be noted that the sum of all relative occurrence rates in Table 1 is greater than 100% in our case. This is caused by the fact that during the expressive corpus annotation by CFs, the listeners were allowed to label a sentence from the corpus with more than one CF if necessary. However, such sentences have been omitted from this preliminary experiments.

3. Enumeration of Differences

When using unit selection methods for expressive speech synthesis, one of the most important factors influencing the synthetic speech quality is a proper selection of speech units from a unit inventory and forming a sequence of these units with the least number of deterioration factors as possible [7]. The decision on which units are selected from the unit inventory is made on the basis of a cost function. This cost function usually consists of two subfunctions – a concatenation cost and a target cost. The former represents how smoothly consecutive units are joint together; the latter represents how the unit from the inventory fits the required target unit from the input text that is about to be synthesized.

The target cost is mostly computed on the basis of prosodic features like phonetic context, position in word or syllable, etc. and various features can have various weighs. To incorporate expressivity into the synthetic speech, another feature was taken into account – the communicative function. The task is to select the most suitable unit with the required CF if possible. However, if there is no such unit (e.g. the concatenation of two consecutive units would not be smooth) other units with the most similar CF should be considered. To measure these similarities a penalty matrix should be designed to enumerate differences between various CFs. The definition of the similarity is assumed to be coded in the penalty matrix of the following form:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots \\ a_{21} & a_{22} & a_{23} & \dots \\ a_{31} & a_{32} & a_{33} & \dots \\ \dots & \dots & \dots & \end{pmatrix}$$

where a_{ij} represents dissimilarity (a penalty) between a CF *i* and a CF *j*.

To create such a penalty matrix, we need to know how the units (or speech in general) labeled with various CFs differ. Both acoustic measures and human perception are taken into consideration because both these views are supposed to influence the overall difference. In the following sections, the process of constructing two penalty matrices is shown. The first one is based on a listening test that was performed to annotate the CFs in the expressive speech corpus [2], the second one is based on results of an acoustic analysis of expressive speech [8]. Finally, we tried to combine these two penalty matrices to create a final penalty matrix that should represent the overall differences between various CFs.

We also have to point out that so far a simple penalty function has been used for the expressive speech synthesis in our system [3] as it was mentioned in Section 1. The expressive TTS system using this simple setting is taken as a baseline for the final evaluation and a comparison with the new experimental one.

3.1. Listening Test Based Differences

Assuming the expressive recordings annotations presented in [6], a penalty matrix in the form described in Section 3 was created. The coefficients a_{ij} were calculated as follows:

$$a_{ij} = abs(\log \frac{num_{ij}}{max_i}) \qquad (2)$$

where num_{ij} represents how many times recordings with CF *i* (according to the objective annotation as presented in [6]) were labeled with CF *j* (calculated over all listeners and all recordings) and max_i represents the maximum value of num_{ij} for fixed *i*. For situation where the *log* is not defined, the a_{ij} was set to a default value which was higher than any other value in the matrix. The *log* was used to emphasize differences between calculated ratios and we also assumed that the human perception is logarithmic-based (as suggested e.g. by The Weber-Fechner Law).

3.2. Acoustic Analysis Based Differences

On the basis of the acoustic analysis of all speech units coming from the expressive corpus [8], a penalty matrix was created in the form as it was mentioned in Section 3. In this case, the coefficients a_{ij} were calculated as an Euclidean distance between numeric vectors representing the CFs *i* and *j* in a 3-dimensional space represented by the following axes: F0 value, RMS energy and unit duration. Each component of the vector was calculated as a mean value of one of the aforementioned acoustic features for all speech units of a particular CF.

The relevance of these features as an acoustic distance measure is proven by the results of the acoustic analysis. It should be noted that there might also be other features that are not considered in our work and that may in any way affect the measure.

3.3. Penalty Matrix

Having two penalty matrices, one based on the annotations and one based on the acoustic analysis, the final matrix representing the numeric differences between various CFs could be created. The final coefficients a_{ij} of this matrix were calculated as

$$a_{ij} = \frac{3 * a^l_{ij} + a^a_{ij}}{K} \tag{3}$$

where a^l_{ij} and a^a_{ij} represent coefficients obtained from the listening test based and acoustic analysis based penalty matrix and K is a constant, for our preliminary experiments ad-hoc set as $K = 6$.

Obviously, we put more weigh on the annotation-based coefficients since we believe that the human perception is more important in our task. The final penalty matrix is depicted in Table 2.

4. Evaluation

To evaluate an impact of our modifications on synthetic expressive speech in terms of both the expressivity perception and the speech quality, several views were used. The same texts were synthesized using the simple penalty function described in Section 1 and the newly created penalty matrix based on the differences between CFs presented in Section 3.3. Sentences with both neutral and expressive content were synthesized to show that the text content is a very important factor when using a limited domain speech synthesizer. For CF *NOT-SPECIFIED*, the texts were neutral in all cases.

Firstly, we employed a measure that is believed to reflect a level of expressivity being expressed by the synthetic speech. It is based on calculating relative occurrence of units with the required CF (compared to the total number of used units). The results are shown in Table 3 separately according to the text content.

Table 3: Relative occurrence of units with appropriate communicative function in the resulting synthetic speech.

CF	neutral content		expressive content	
	baseline	new	baseline	new
CONFIRM	4%	12%	84%	90%
ENCOURAGE	53%	84%	85%	90%
HAPPY-EMPATHY	11%	39%	65%	84%
SAD-EMPATHY	6%	36%	63%	71%
SHOW-INTEREST	38%	61%	79%	77%
mean	**22%**	**46%**	**75%**	**82%**
NOT-SPECIFIED	100%	100%	100%	100%

Obviously, when synthesizing expressive texts, more units with the appropriate CF were selected – it is probably related to the fact that the expressive texts were similar to those that appeared in the expressive corpus (but not equal). The results for CF *NOT-SPECIFIED* were not considered when calculating the mean value to avoid any result distortion. The results show that we achieved 109% improvement when considering neutral text content and 9% improvement for expressive texts. However, the number of units with the appropriate CF might be increased also by different settings of feature weighing mentioned in Section 3 Although the weighs remained equal in our case, the maximum value of the penalty as such has increased from 1.0 in the baseline system to 4.0 in the experimental settings. This could also contribute to the increase of the number of the units with the required CF.

Next, a measure indicating a smoothness level was applied. This measure is based on computing a relative number of natural concatenation points, i.e. relative number of speech units selected from the inventory that were originally adjacent in the speech corpus. The results are presented in Table 4 separately according to the text content. The results for CF *NOT-SPECIFIED* were not considered when calculating the mean value.

Table 4: Relative occurrence of natural concatenation points in the resulting synthetic speech.

CF	neutral content		expressive content	
	baseline	new	baseline	new
CONFIRM	70%	65%	79%	80%
ENCOURAGE	58%	45%	72%	70%
HAPPY-EMPATHY	69%	51%	72%	67%
SAD-EMPATHY	69%	49%	73%	69%
SHOW-INTEREST	60%	48%	76%	75%
mean	**65%**	**52%**	**74%**	**72%**
NOT-SPECIFIED	71%	71%	82%	82%

We can observe noticeable deterioration in smoothness when using the new experimental penalty matrix for neutral texts. However, the deterioration is almost imperceptible for expressive texts.

Last, a listening test was used to assess an impact of our modifications on the synthetic expressive speech quality. During this listening test, 9 listeners were presented with 52 isolated utterances and rated them according to the standard MOS (mean opinion score) 5-point scale (1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent). The test also contained several natural utterances for a comparison. We have to point out that this assessment was performed only for utterances with expressive text content in order to reduce the amount of test queries. Since we are more interested in limited domain expressive texts, this limitation is not very crucial. The results (mean values) are presented in Table 5.

We might conclude that the synthetic speech quality deteriorated by 3%. However, according to the performed t-test, the difference between the mean values is not statistically significant (*p-value* > 0.05; performed only on the ratings for the baseline and the experimental system). This suggests that the quality remains almost at the same level.

Table 5: Results of MOS test.

Settings	baseline	experimental	natural speech
Score	3.5	3.4	4.7
Std. dev.	1.0	1.0	0.8

Table 2: The penalty matrix for all communicative functions.

required - available	APOLOGY	CONFIRM	DIRECTIVE	DISCONFIRM	ENCOURAGE	GOODBYE	GREETING	HAPPY-EMPATHY	NOT-SPECIFIED	OTHER	REQUEST	SAD-EMPATHY	SHOW-INTEREST	SURPRISE	THANKS	WAIT
APOLOGY	0.00	1.90	1.01	2.87	3.54	0.54	3.09	2.61	0.91	2.31	3.44	1.57	2.46	1.36	2.80	2.38
CONFIRM	2.67	0.00	2.43	2.96	4.03	2.75	2.41	1.54	1.77	3.56	4.11	1.68	3.27	2.06	1.93	2.35
DIRECTIVE	2.20	1.94	0.00	3.45	2.31	2.28	2.05	2.56	1.26	2.99	2.43	2.56	2.68	2.33	2.60	2.57
DISCONFIRM	2.66	2.11	3.82	0.00	4.63	4.11	2.94	1.28	2.50	3.95	4.53	1.25	4.47	3.22	3.22	1.78
ENCOURAGE	3.54	3.32	3.04	5.34	0.00	3.45	4.25	3.85	2.23	1.02	0.83	3.99	0.40	2.15	4.46	4.11
GOODBYE	0.70	1.87	1.04	4.11	2.77	0.00	2.63	1.91	0.40	2.10	3.35	2.08	2.37	2.05	1.22	2.61
GREETING	3.09	1.94	2.88	2.94	4.16	3.17	0.00	1.58	1.85	3.93	4.53	2.35	3.85	3.22	2.28	1.59
HAPPY-EMPATHY	2.47	1.03	2.84	1.71	3.50	2.59	2.29	0.00	1.99	3.49	4.40	0.87	3.40	2.29	1.70	1.37
NOT-SPECIFIED	1.00	1.05	0.90	2.92	1.62	0.46	1.67	1.76	0.00	1.61	1.81	2.00	1.48	0.46	1.43	2.12
OTHER	3.31	3.99	1.79	5.34	2.22	1.65	4.40	4.70	1.48	0.00	2.12	4.76	2.06	3.18	4.12	4.62
REQUEST	3.44	4.11	2.34	5.47	0.51	3.35	3.22	4.12	2.15	0.83	0.00	4.30	0.79	3.30	4.24	4.16
SAD-EMPATHY	1.94	1.26	2.18	1.69	3.56	2.31	2.35	1.00	2.17	3.63	4.24	0.00	3.35	2.43	2.64	1.55
SHOW-INTEREST	3.38	3.37	2.73	5.16	0.45	3.30	4.47	3.94	1.93	1.38	0.90	3.89	0.00	2.23	4.18	4.13
SURPRISE	2.13	1.28	1.68	3.14	1.43	2.05	3.22	1.86	0.33	1.75	2.33	2.15	1.30	0.00	2.43	2.64
THANKS	2.80	1.28	2.60	3.22	3.74	1.93	2.28	2.58	2.90	4.12	4.24	2.64	4.18	2.93	0.00	2.50
WAIT	2.39	1.95	1.40	2.72	3.42	3.39	2.21	1.40	2.73	4.62	3.23	1.46	3.54	3.44	2.50	0.00

5. Conclusions

In this work, we utilized annotations and acoustic analysis of expressive recordings in terms of various communicative functions to obtain a penalty matrix that should represent differences between such defined categories of expressivity. The final penalty matrix that reflects both the expressivity perception and acoustic measures was created and then used for expressive speech synthesis. A comparison of the new experimental setting and the baseline system was presented. The results show that the experimental system achieved higher rate in selection of units with the correct communicative function label while keeping the smoothness of concatenation points of units at an acceptable level. This was also confirmed by the listening MOS test. For the future work, we plan to perform another listening test to assess the level of expressivity as perceived by humans. Such a test will be more complex and its description would be outside the scope of this paper.

6. Acknowledgements

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 and also by the grant of the University of West Bohemia, project No. SGS-2010-054. 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.

7. References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] D. Tihelka, J. Matoušek, and J. Kala, "Quality deterioration factors in unit selection speech synthesis," in *Text, Speech and Dialogue, proceedings of the 10th International Conference TSD*, ser. Lecture Notes in Artificial Intelligence, vol. 4629. Berlin-Heidelberg, Germany: Springer, 2007, pp. 508–515.

[8] 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.