Expressive Speech Synthesis for Czech Limited Domain Dialogue System – Basic Experiments

Martin Grůber, Daniel Tihelka
Dept. of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Czech Republic
Email: gruber@kky.zcu.cz, dtihelka@kky.zcu.cz

***Abstract*—This paper describes a development of limited domain expressive speech synthesis for the Czech language. Our current speech synthesis system is based on unit selection methods and produces high quality speech in a neutral speaking style. This work focuses on modifications made in the synthesis algorithm to integrate expressivity into generated speech. There is also introduced a listening test, which should prove or disprove that expressivity in limited domain speech synthesis could be modelled by so-called communicative functions. A comparison between neutral and expressive speech synthesis is presented. This paper also briefly discusses the process of an expressive speech corpus recording and its annotating using the communicative functions by means of another listening test.**

I. Introduction

Current speech synthesis techniques are surely able to produce high quality and intelligible speech. However, if we are talking about artificial speech that should not be recognized from human speech, some kind of speaker's attitude have to be considered and incorporated in a speech production process. It means that some expressivity or emotions in accordance with a content of speech will certainly improve perception of communicated information by listeners. Perhaps, this issue is not so hot in terms of some information systems or call centers which also use synthesized speech but in tasks dealing with personal dialogues between a computer and a human it should be taken into consideration.

Some techniques incorporating expressivity into synthesized speech have been introduced so far. Some of them consist in modifications of acoustic parameters of synthesized speech, others produces special speech or non-speech expressions to evoke some expressivity. Methods using unit selection techniques consist in creating a unit inventory containing expressive speech units.

However, the task of completely natural expressive speech synthesis within unlimited domain is so extensive and complex that it is beyond present technical capabilities. Therefore we need to limit this task somehow. The first level limitation is a dialogue between a computer and a human but it is not restrictive enough. Since the task is solved within Companions Project (www.companions-project.org), it is determined as a dialogue between a senior and a computer and the topic for these conversations is set to reminiscing about seniors' personal photographs.

To understand the domain sufficiently, we recorded natural dialogues between seniors and the computer using the Wizard of Oz method. The data collection process is described in [1] and [2]. This way we obtained 65 real dialogues (approximately 60 hours of speech) that gave us knowledge of how conversations between the seniors and the computer develop and what the seniors like to talk about when reminiscing about photographs.

We suppose that this way the domain is limited enough to improve our current speech synthesis system and to create an expressive speech synthesizer. Since our current TTS system ARTIC [3] is corpus oriented and based on a unit selection algorithm without any signal modifications, the creation of the expressive speech synthesis system consists in speech corpus enrichment and in modifications of the unit selection algorithm.

Thus an expressive speech corpus was recorded and annotated using various categories of expressivity by means of a listening test [4]. Reliability of such annotation was proved using measures of inter-listeners agreement. In the annotated expressive corpus, each speech unit is marked with a specific feature — communicative function — indicating what kind of expressivity it conveys. The algorithm calculating a target cost for each candidate unit was modified and several settings of new feature weighting was tested. Finally, one experimental setting was used and expressive utterances for a listening test were generated.

The paper is organized as follows: In Section II, the expressive speech recording, the annotation of expressive recordings and the corpus creation are briefly described. Modifications made in the target cost calculation algorithm are outlined in Section III. Section IV deals with the background of the preliminary listening test, its results and evaluation with respect to credibility and reliability of the listeners. Conclusion and future work is presented in Section V.

II. Description of the Expressive Speech Corpus

The expressive speech corpus was created on the basis of an audiovisual database which contains natural dialogues between a computer and a human (this database was beforehand recorded using Wizard of Oz method and the data collection process is described in details in [1] and [2]). The expressive recordings were recorded by a professional female speaker in an anechoic room using high quality recording equipment and software specially designed for this purpose. Glottal signal was captured along with the speech to allow us to utilize algorithms for pitch-mark detection [5], which is further used in corpus creation process. In this way, we recorded more than 7200 expressive utterances, mostly short ones, total length of which is almost 4.5 hours.

The expressive recordings were later manually transcribed and annotated using so-called communicative functions by means of a listening test [4]. These functions are supposed to describe various categories of expressivity which can occur

TABLE I
The set of the communicative functions and probabilities of their occurrence in the expressive speech corpus.

communicative function (symbol)	*occurr. prob.*	*example*
directive (DIRECTIVE)	0.0236	Tell me that. Talk.
request (REQUEST)	0.0436	Let's get back to that later.
wait (WAIT)	0.0073	Wait a minute. Just a moment.
apology (APOLOGY)	0.0059	I'm sorry. Excuse me.
greeting (GREETING)	0.0137	Hello. Good morning.
goodbye (GOODBYE)	0.0164	Goodbye. See you later.
thanks (THANKS)	0.0073	Thank you. Thanks.
surprise (SURPRISE)	0.0419	Do you really have 10 siblings?
sad empathy (SAD-EMPATHY)	0.0344	I'm sorry to hear that. It's really terrible.
happy empathy (HAPPY-EMPTATHY)	0.0862	It's nice. Great. It had to be wonderful.
showing interest (SHOW-INTEREST)	0.3488	Can you tell me more about it?
confirmation (CONFIRM)	0.1319	Yes. Yeah. I see. Well. Hmm.
disconfirmation (DISCONFIRM)	0.0023	No. I don't understand.
encouragement (ENCOURAGE)	0.2936	Well. For example? And what about you?
not specified (NOT-SPECIFIED)	0.0736	Do you hear me well? My name is Paul.

in the utterances. The information about the communicative function is assigned to all speech units coming from a particular sentence which is marked with this function. The set of communicative functions proposed for this task is presented in Table I.

As it is obvious, the most often appearing communicative functions in the annotated expressive corpus were *SHOW-INTEREST* (in 35% of sentences), *ENCOURAGE* (29%) and *CONFIRM* (13%). Frequencies of the others were less than 10%. The annotators were allowed to mark one sentence with more than one communicative function during the annotation. However, only one communicative function with the highest score (calculated during objective annotation assessment) was taken into account for this preliminary experiment.

Since the expressive corpus itself did not contain all possible speech units occurring in the Czech language, a part of our current neutral corpus was merged with the expressive one. Only the sentences containing missing units were chosen to be integrated into the expressive corpus (both corpora were recorded under the same conditions by the same female speaker with relatively short time interval in between). This way we made up a complete expressive speech corpus which can be used for TTS system.

III. Target Cost Calculation

Using a unit selection algorithm [6], speech units forming resulting synthesized speech are selected from a list of corresponding candidate units. These candidates are stored in a unit inventory which is built up on the basis of a speech corpus. The speech unit selection process respects two various groups of candidates' features.

Features in one group are used for a concatenation cost computation. This cost reflects continuity distortion, i.e. how smoothly each candidate for unit u_{i-1} will join with each candidate for unit u_i in the sequence. The lower the cost is, the less the unit boundaries are noticeable. In this group of features there are usually included mostly ordinal values (acoustic and spectral parameters), e.g. some acoustic coefficients, energy values, F0 values, their differences, etc. The concatenation cost for candidate u_i is then calculated as follows:

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

where C_i is the concatenation cost of a candidate for unit u_i, 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 corresponding features of two potentially adjacent candidates for units u_{i-1} and u_i — for unit u_i the features from the beginning of the unit are compared with features from the end of unit u_{i-1}.

Features in the other group are used for a target cost computation. This cost reflects the level of approximation of a target unit by any of candidates, in other words, how a candidate from the unit inventory fits a corresponding target unit — a theorethical unit whose features are specified on the basis of the sentence to be synthesized. In this group there are usually included mostly nominal features, e.g. phonetic context, prosodic context, position in word, position in sentence, position in syllable, etc. The target cost for candidate u_i is then calculated as follows:

$$T_i = \frac{\sum_{j=1}^{n} w_j d_j}{\sum_{i=j}^{n} w_j} \tag{2}$$

where T_i is the target cost of a candidate for unit u_i, 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* features of a candidate for unit u_i and target unit t_i. The differences of particular features (d_j) will be further referenced as penalties.

When using the expressive speech corpus, the set of the features used for the target cost computation is extended with one more feature. With regard to what was mentioned above, it is naturally called *communicative function*. The penalty d_{cf} between candidate u_i and target unit t_i is calculated as follows:

$$d_{cf} = \begin{cases} 1 & \text{if } cf_t = cf_c \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where d_{cf} is a difference (penalty), cf_t is a communicative function of target unit t_i and cf_c is a communicative function of a candidate for unit u_i.

Finally, we need to set a weight for this penalty since the target cost is calculated as a weighted sum of particular penalties. For preliminary experiments, the weight of the communicative function penalty was determined ad-hoc and its value is almost the highest among the other weights (e.g. it is 4x higher than a value of a phonetic context weight). It reflects its function and also our assumption that this feature is intended to influence the overall cost considerably.

For further improvement, the weight setting is going to be more explored and some weight adjusting is going to be made. Some suggestions are proposed in Section V.

IV. Listening Test Background, Evaluation and Results

To rate the quality of expressive speech synthesis, we decided to perform a listening test. For this purpose, 68 utterances were synthesized using two different methods and two different corpora — using the current neutral synthesis system with the neutral corpus on one hand and the new expressive speech synthesis system with the expressive corpus making use of the communicative functions on the other hand. Thus we got two different versions of each sentence.

A. Background

The sentences to be synthesized for purposes of this listening test were basically chosen from the audiovisual database of natural dialogues along with their textual context. However, the content of these sentences was modified to avoid just playing back utterances from the expressive corpus. The selection was made in order that the communicative functions of the synthesized sentences represent all available communicative functions equally (except communicative function *NOT-SPECIFIED* which was not synthesized — this one should represent a neutral speaking style and it was not our objective).

The listening test was organized on the client-server basis using specially developed web application. This way listeners were able to work on the test from their homes without any contact with the test organizers. The listeners were required to have only an internet connection, any browser installed on their computers and some device for audio playback (using headphones was recommended). Various measures were undertaken to detect possible cheating, carelessness or misunderstandings.

In the listening test, the listeners were asked to decide which spoken form of the same text is more suitable in a particular part of a dialogue. Corresponding textual context and conveyed (synthesized) communicative function was displayed for each query of the listening test.

The listeners were instructed to read textual context of a natural dialogue, to listen to recordings very carefully and subsequently to mark the most appropriate option — answer to a question: *"Which utterance suits the best the defined textual context and the specified communicative function?"*. The listeners were also instructed to take their decisions mainly in terms of expressivity being perceived. However, the overall quality of synthesized speech should not have been ignored completely.

Basically, there were two possible options for two versions of synthesized sentences. However, the listeners were allowed to mark both the versions at once in case they were not able to take a decision.

Finally, 10 listeners have successfully finished the listening test. However, this way we obtained 10 subjective assessments that can vary across the listeners. To obtain an *objective assessment* of the listening test and thus to be able to compare neutral and expressive speech synthesis correctly, an evaluation based on a maximum likelihood method was made. The process of evaluation is presented in the next section.

B. Evaluation

There are several ways to be used for deduction of the objective assessment. We decided to utilize an approach that is based on a maximum likelihood method. The maximum likelihood estimation is a statistical method used for fitting a statistical model to data and providing estimates for the model's parameters. Under certain conditions, the maximum likelihood estimator is consistent. The consistency means that having a sufficiently large number of observations (listeners' assessments in our case), it is possible to find the values of statistical model's parameters with an arbitrary precision. The parameters calculation is implemented using the EM algorithm [7]. This evaluation is an asymptotically consistent, asymptotically normal and asymptotically efficient estimate. We have also successfully used this approach in recent works regarding speech synthesis research, see [8], [9] or [4].

The statistical model can be viewed as a model listener who assesses the listening test with the meaning of the maximum likelihood method. Having this model listener we are able to deduce true observation which we call the objective assessment. The precision of the estimate is one of the outputs of the statistical model. Using the precision, any untrustworthy assessment can be eliminated. The untrustworthiness can be viewed as an inability of the model listener to decide which choice in a query of the listening test is more suitable. In other words, the options varies very much among the real listeners and the decision about the objective assessment cannot be taken.

Further, we need to confirm that the listeners achieved some measure of agreement. Otherwise the subjective assessments could be considered as accidental and thus the acquired objective assessment would be false. For this purpose, we make use of two statistical measures for assessing the reliability of agreement among listeners.

One of the measures used for such evaluation is Fleiss' kappa κ_F. It is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items. We calculated a mean value of this measure among all listeners.

Another measure used here is Cohen's kappa κ_C. It is a statistical measure of inter-rater agreement for categorical items and takes into account the agreement occurring by chance as well as Fleiss' kappa. However, Cohen's kappa measures the agreement only between two listeners. We decided to measure the agreement between each listener and the objective assessment obtained by the maximum likelihood method. Calculation of Cohen's kappa is done separately for each listener. Thus we can find out whether particular listener was in agreement with the objective assessment. Finally, mean of Cohen's kappas of all listeners was calculated.

Values of Fleiss' and Cohen's kappa vary between 0 and 1, the higher value the better agreement. More detailed interpretation of measure of agreement is in [10]. The resultant value $\kappa_F = 0.44$ means that the moderate inter-rater agreement was achieved. The resultant value $\kappa_C = 0.41$ means that the moderate agreement was achieved also between the listeners and the objective assessment. Thus we proved that the results

TABLE II
Results of the listening test.

preferred synthesis	*No. utterances*	*ratio*
neutral	17	25%
expressive	45	66%
no preference	6	9%

presented in the next section are not accidental and can be viewed as trustful.

C. Results

As it was mentioned above, 68 utterances in two various versions were presented to the listeners to assess them. The listeners were supposed to choose that version of each utterance which is more suitable in a particular situation. The certain situation was slightly described using a textual information taken from a natural dialogue. The results based on the model listener's assessment are presented in Table II.

As it is obvious from the results, the utterances generated using expressive speech synthesis system were mostly preferred (66%) to those produced by neutral synthesis (25%). In some cases (9%) the listeners were not able to decide which version is more suitable. It is remarkable that the ratio varies across the communicative functions. For example: for communicative functions *APOLOGY*, *CONFIRM*, *HAPPY-EMPATHY* or *SAD-EMPATHY*, most of the listeners preferred expressive speech synthesis ($70\% - 100\%$); for *DISCONFIRM* or *THANKS*, 80% of listeners preferred neutral synthesis. It might indicate that the expressivity is not suitable in these situations or that in a particular situation the expressivity is not expressed so much as it is in other situations, i.e. it may be not necessary for the speaker to convey an emotional state of his mind in these parts of the dialogue.

V. Conclusions and Future Work

We performed the listening test which was intended to compare neutral speech synthesis with expressive synthesis. Expressive speech synthesis uses the expressive speech corpus – speech units from this corpus are assigned with communicative functions. These functions are assumed to describe or classify the expressivity contained in speech.

The results of the listening test show that expressive speech synthesis is among the listeners generally preferred to neutral synthesis. However, in some rare situations neutral synthesis is evaluated equally or even more suitable than expressive synthesis with corresponding communicative functions.

In future work we plan to further improve the present expressive speech synthesis system. For example, the evaluation of the target cost is crucial and can be done in various ways for various unit features. For the communicative functions there should be designed a penalty (distance) matrix. It is assumed to represent distances between various communicative functions. It is obvious that a penalty between two contradictory functions should be greater than a penalty between two partly compatible functions. The differences among the functions have to be obtained somehow, e.g. by means of another listening test.

Using the expressive speech corpus we can produce mainly expressive speech. However, the dialogue system within the Companions project has to be designed to produce both expressive and neutral high quality speech. To achieve this objective, we have to consider the neutral speaking style and use also the neutral speech corpus in the dialogue system. Therefore the neutral corpus is going to be merged with the expressive one. The speech units coming from the neutral corpus are going to be assigned with the communicative function *NOT-SPECIFIED* as the recordings in this corpus should not convey any communicative function used in our limited domain dialogue system. This function should represent the neutral style of speaking. Thus we can perform another listening test to compare neutral and expressive speech synthesis which is going to utilize the merged speech corpus.

As it was mentioned above, the presented expressive speech synthesis methods are being developed within the limited domain dialogue system. This system is supposed to play a role of a partner in personal dialogues between computers and seniors. Therefore we also have to deal with a social aspect of such a human-computer interaction.

Acknowledgements

This work was partially funded by the Companions project sponsored by the European Commission (grant number IST-FP6-034434). It was also supported by the Grant Agency of the Czech Republic, project No. GAČR 102/09/0989, and partly also by the University of West Bohemia, project No. SGC-2010-054.. The access to the METACentrum computing facilities provided under the research intent MSM6383917201 is highly appreciated.

References

[1] M. Legát, M. Grůber, and P. Ircing, "Wizard of Oz data collection for the Czech senior companion dialogue system," in *Fourth International Workshop on Human-Computer Conversation*. Bellagio, Italy: University of Sheffield, 2008, pp. 1–4.

[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," in *Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznan, Poland: Wydawnictvo Poznanskie, 2009, pp. 266–269.

[3] J. Matoušek, D. Tihelka, and J. Romportl, "Current state of Czech text-to-speech system ARTIC," in *Text, Speech and Dialogue, proceedings of the 9th International Conference TSD 2006*, ser. Lecture Notes in Artificial Intelligence, vol. 4188. Berlin, Heidelberg: Springer, 2006, pp. 439–446.

[4] M. Grůber and J. Matoušek, "Listening-test-based annotation of communicative functions for expressive speech synthesis," submitted to TSD2010.

[5] M. Legát, J. Matoušek, and D. Tihelka, "A robust multi-phase pitchmark detection algorithm," in *Proceedings of Interspeech*, Antwerp, Belgium, 2007, pp. 1641–1644.

[6] D. Tihelka and J. Matoušek., "Unit selection and its relation to symbolic prosody: a new approach," vol. 1, pp. 2042–2045, 2006.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977, with discussion.

[8] J. Romportl, "Prosodic phrases and semantic accents in speech corpus for Czech TTS synthesis," in *Text, Speech and Dialogue, proceedings of the 11th International Conference TSD 2008*, ser. Lecture Notes in Artificial Intelligence, vol. 5246. Berlin–Heidelberg, Germany: Springer, 2008, pp. 493–500.

[9] ——, "Statistical evaluation of prosodic phrases in the Czech language," in *Proceedings of the Speech Prosody 2008 Conference*. Campinas, Brazil: Editora RG/CNPq, 2008, pp. 755–758.

[10] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data." *Biometrics*, vol. 33, no. 1, pp. 159–174, March 1977.