# Listening-test-based annotation of communicative functions for expressive speech synthesis*

Martin Grůber and Jindřich Matoušek

Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Czech Republic
{gruber,matousek}@kky.zcu.cz

**Abstract.** This paper is focused on the evaluation of listening test that was realized with a view to objectively annotate expressive speech recordings and further develop a limited domain expressive speech synthesis system. There are two main issues to face in this task. The first matter in issue to be taken into consideration is the fact that expressivity in speech has to be defined in some way. The second problem is that perception of expressive speech is a subjective question. However, for the purposes of expressive speech synthesis using unit selection algorithms, the expressive speech corpus has to be objectively and unambiguously annotated. At first, a classification of expressivity was determined making use of communicative functions. These are supposed to describe the type of expressivity and/or speaker's attitude. Further, to achieve objectivity at a significant level, a listening test with relatively high number of listeners was realized. The listeners were asked to mark sentences in the corpus using communicative functions. The aim of the test was to acquire a sufficient number of subjective annotations of the expressive recordings so that we would be able to create "objective" annotation. There are several methods to obtain objective evaluation from lots of subjective ones, two of them are presented.

**Key words:** expressive speech synthesis, listening test, communicative functions, inter-rater agreement measure

## 1 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

* This work was partially funded by the Companions project IST-FP6-034434. This research 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.

be considered and incorporated in the speech production process. It means that some expressivity or emotions in accordance with the content of speech will for sure improve the perception of the 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.

The task of general 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. We are speaking on dialogues between a computer and a human but it is not restrictive enough. Our task was determined as a dialogue between a senior and a computer and theme for conversations was set to reminiscing about personal photographs. It will be shown that this way the domain is limited enough to improve our current speech synthesis and to create an expressive speech synthesizer.

Since our current TTS system ARTIC [1] is corpus oriented and based on unit selection algorithms [2], the improvement of speech synthesis consists in speech corpus enhancement. Thus an expressive speech corpus was recorded and annotated using various categories of expressivity by means of a listening test. Reliability of such annotation was proved using measures of inter-listeners agreement.

The paper is organized as follows: In Section 2 the expressive speech recording process is briefly described. The background, preparation works and settings of the performed listening test is shown in Section 3. In Section 4 we focused on the evaluation of the listening test with respect to credibility and reliability of the listeners. Finally, in Section 5, conclusions arising from the listening test results are drawn and future work is outlined.

## 2 On the Expressive Speech Corpus

To incorporate expressivity into our current TTS system, an expressive speech corpus was recorded and merged together with existing neutral one. Issues in this task include but are not limited to corpus design, corpus recording, description of expressivity that is supposed to be contained in the corpus and annotation of the expressive recordings using the defined expressivity categories. These issues are more discussed in the following sections.

### 2.1 Corpus Design

Since we are dealing with limited domain expressive speech synthesis, definition of the domain is necessary. The domain in this task was restricted to dialogues between seniors and a computer. Theme for these conversations was set to reminiscing about seniors' photographs. This limitation is already sufficient enough. To become familiar with these conversations, an extensive audiovisual database containing natural dialogues between seniors and a computer (applying 3D avatar - "talking head" [3] with a neutral TTS system) was recorded using

Wizard of Oz method and manually transcribed. Process of database recording is presented in [4]. On the basis of the recorded database we got knowledge about how the natural dialogues develop, what the seniors like to talk about and what kind of expressivity is expected to be conveyed within the synthesized speech.

### 2.2 Corpus recording

We have decided to proceed with the expressive corpus creation as follows. First, we hired a professional female speaker (stage-player) and instructed her not to express a specific emotions but just to put herself in the place of a partner for seniors in a dialogue – pretend to be an avatar. In order to facilitate such an empathy, a special software application was developed (see Figure 1) - it played back the parts of the natural dialogues when the senior was speaking (to provide the speaker with the relevant context) and at the time when the avatar have originally spoken, the dialogue was paused and the speaker was prompted to record the avatar's sentence herself. The text of the actual sentence was displayed on the screen even when the real (context) dialogue was being played so that the speaker had enough time to get acquainted with it before the recording. Also time remaining to the recording of the next utterance was displayed. Controlling of the application was designed to be very easy for the speaker so that she could have been fully concentrated on the recording.

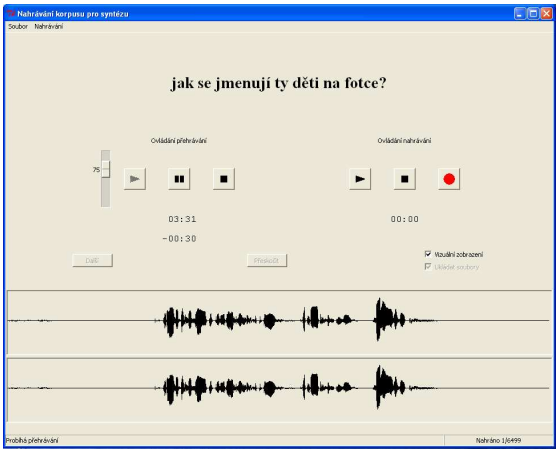


**Fig. 1.** Software interface for expressive corpus recording with the use of real dialogues.

The recording equipment was carefully selected and set-up in order to ensure the highest possible technical quality of the corpus - the speaker was placed in the anechoic room and the recording was done using a professional mixing desk. The glottal signal was captured along with the speech. That way we have recorded more than 7,000 of (mostly short) sentences. Those were carefully transcribed.

### 2.3 Expressivity description

The issue of expressivity description is very complex. In the past, several techniques were proposed and are divided into two main groups. One basic approach is to use continuous representation in two-dimensional space introduced in [5]; any kind of expressivity is referenced as a point with specific coordinates in that space. The other alternative is a categorical view; any kind of expressivity is classified into one (or more) of predefined classes.

For purposes of expressive speech synthesis and machine processing, the categorical classification of expressivity seems to be more suitable. Therefore we decided to utilize this approach. Moreover, since unit selection algorithms are applied, the expressivity class can be used as a feature for each particular unit which is stored in a unit inventory.

Specification of an appropriate set of classes for the limited domain defined above was based on dialogue acts proposed in [6]. This set was modified for our purposes and is shown in Table 1. Since each of the categories expresses function of a sentence in communication, it is called *communicative function*.

**Table 1.** Set of communicative functions.

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

### 2.4 Annotation

The expressive speech corpus was annotated using communicative functions by means of a listening test. The test was aimed to determine objective annotation

on the basis of several subjective annotations as the perception of expressivity is always subjective and may vary depending on particular listener. Preparation works and listening test framework are described in the following section. Evaluation of listening test result and a measure of inter-rater agreement analysis is presented in Section 4.

## 3 Listening Test Background

The listening test was organized on the client-server basis using a 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. Various measures were undertaken to detect possible cheating, carelessness or misunderstandings.

Potential test participants were addressed mostly among university students from all faculties and the finished listening test was financially rewarded (to increase motivation for the listeners). The participants have been instructed to listen to the recordings very carefully and subsequently mark communicative functions that are expressed within the sentence. The number of possibly marked communicative functions for one utterance was just upon the listeners, they were not limited anyhow. Few sample sentences labelled with communicative functions were provided and available to the listeners on view at every turn. If any listener marked one utterance with more than one communicative function, he was also required to specify whether the functions occur in that sentence consecutively or concurrently. If the communicative functions are marked as consecutive in a particular utterance, this utterance is omitted from further research for the present. These sentences should be later manually reviewed and either divided into more shorter sentences or omitted completely.

Finally, 12 listeners have successfully finished the listening test. However, this way we obtained subjective annotations that vary across the listeners. To objectively annotate the expressive recordings, proper combination of the subjective annotations was needed. Therefore an evaluation of the listening test was made.

## 4 Listening Test Evaluation

### 4.1 Objective annotation

We utilized two ways to deduce the objective annotation.

The first way is a simple majority method. Using this easy and intuitive approach, each sentence is assigned a communicative function, that was marked by the majority of the listeners. In case of less then 50% of all listeners marked such communicative function, the classification of this sentence is considered as untrustworthy.

The second approach is based on maximum likelihood method. 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 (annotations in our case), it is possible to find the value of statistical model parameters with arbitrary precision. The parameter calculation is implemented using the EM algorithm [7]. Knowing the model parameters we are able to deduce true observation which we call objective annotation. Precision of the estimate is one of the outputs of this model. Using the precision, any untrustworthy assignment of a sentence with a communicative function can be eliminated.

Comparing these two approaches, 35 out of 7287 classifications were marked as untrustworthy using maximum likelihood method and 571 using simple majority method. The average ratio of listeners who marked the same communicative function for particular sentence using simple majority approach was 81%, when untrustworthy classifications were excluded. Similar measure for maximum likelihood approach cannot be easily computed as the model parameters and the estimate precision depend on number of iteration in the EM algorithm.

We decided to use the objective annotation obtained by maximum likelihood method. It 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].

Further, we need to confirm that the listeners marked the sentences with communicative functions consistently and achieved some measure of agreement. Otherwise the subjective annotations could be considered as accidental or the communicative functions inappropriately defined and thus the acquired objective annotation 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. 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 this measure among all listeners separately for each communicative function. Computation of overall Fleiss' kappa is impossible because the listeners were allowed to mark more than one communicative function for each sentence. However, the overall value can be evaluated as the mean of Fleiss' kappas of all communicative functions.

Another measure used here is Cohen's kappa. 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 annotation obtained by maximum likelihood method. Again, calculation of Cohen's kappa was made for each communicative function separately. Thus we can find out whether particular listener was in agreement with the objective annotation for certain communicative function. Finally, the mean of Cohen's kappas of all communicative functions was calculated.

**Table 2.** Fleiss' and Cohen's kappa and occurrence probability for various communicative functions and for the "consecutive CFs" label. For Cohen's kappa, mean value and standard deviation is presented, since Cohen kappa is measured between annotation of each listener and the reference annotation.

| communication function | Fleiss's kappa | Measure of agreement | Cohen's kappa | Cohen's kappa SD | Measure of agreement | Occurr. probab. |
|---|---|---|---|---|---|---|
| DIRECTIVE | 0.7282 | Substantial | 0.8457 | 0.1308 | Almost perfect | 0.0236 |
| REQUEST | 0.5719 | Moderate | 0.7280 | 0.1638 | Substantial | 0.0436 |
| WAIT | 0.5304 | Moderate | 0.7015 | 0.4190 | Substantial | 0.0073 |
| APOLOGY | 0.6047 | Substantial | 0.7128 | 0.2321 | Substantial | 0.0059 |
| GREETING | 0.7835 | Substantial | 0.8675 | 0.1287 | Almost perfect | 0.0137 |
| GOODBYE | 0.7408 | Substantial | 0.7254 | 0.1365 | Substantial | 0.0164 |
| THANKS | 0.8285 | Almost perfect | 0.8941 | 0.1352 | Almost perfect | 0.0073 |
| SURPRISE | 0.2477 | Fair | 0.4064 | 0.1518 | Moderate | 0.0419 |
| SAD-EMPATHY | 0.6746 | Substantial | 0.7663 | 0.0590 | Substantial | 0.0344 |
| HAPPY-EMPATHY | 0.6525 | Substantial | 0.7416 | 0.1637 | Substantial | 0.0862 |
| SHOW-INTEREST | 0.4485 | Moderate | 0.6315 | 0.3656 | Substantial | 0.3488 |
| CONFIRM | 0.8444 | Almost perfect | 0.9148 | 0.0969 | Almost perfect | 0.1319 |
| DISCONFIRM | 0.4928 | Moderate | 0.7153 | 0.1660 | Substantial | 0.0023 |
| ENCOURAGE | 0.3739 | Fair | 0.5914 | 0.3670 | Moderate | 0.2936 |
| NOT-SPECIFIED | 0.1495 | Slight | 0.3295 | 0.2292 | Fair | 0.0736 |
| OTHER | 0.0220 | Slight | 0.0391 | 0.0595 | Slight | 0.0001 |
| *mean* | *0.5434* | *Moderate* | *0.6632* | | *Substantial* | |
| consecutive CF | 0.5138 | Moderate | 0.6570 | 0.2443 | Substantial | 0.0374 |

Results of agreement measures are presented in Table 2. Value 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 [9].

The Fleiss' kappa mean value of 0.5434 means that the measure of inter-listeners agreement is moderate. As it is obvious from Table 2, communicative functions *OTHER* and *NOT-SPECIFIED* should be considered as poorly recognizable. It is understandable when taking into consideration their definitions. After eliminating values of these communicative functions the mean value of 0.6191 is achieved, which means substantial agreement among the listeners.

The Cohen's kappa mean value of 0.6632 means that the measure of agreement between listeners and objective annotation is substantial. Moreover, we can again eliminate communicative functions *OTHER* and *NOT-SPECIFIED* as they were poorly recognizable also according to Cohen's kappa. Thus, mean value of 0.7316 is achieved. However, it is still classified as substantial agreement.

As it is shown in Table 2, agreement among listeners regarding classification of consecutive communicative function was measured too. The listeners agreed on this label moderately among each other and substantially with the objective annotation. There are also shown probabilities of the particular com-

municative functions occurrence when maximum likelihood method was used for the objective annotation obtaining. It is obvious that communicative functions *SHOW-INTEREST* and *ENCOURAGE* are the most frequent.

## 5 Conclusion and future work

In this work we have created an objectively annotated expressive speech corpus. The subjective annotations of expressivity was made by means of listening test, where listeners marked each sentence from the corpus with communicative functions. The objective annotation was deduced from the subjective ones using maximum likelihood method. The inter-listeners measures of agreement confirmed that the objective annotation is trustworthy.

Appropriate combination of the expressive speech corpus and current neutral corpus will allow us to create an expressive speech synthesizer. Its development is our objective for future work. The synthesizer is planned to be used in a limited domain dialogue system, which is going to serve elderly people to discuss their personal photographs with computer. We should also deal with social issues regarding such a human-computer interaction.

## References

1. Matoušek, J., Tihelka, D., Romportl, J.: Current State of Czech Text-to-speech System ARTIC. In: Proceedings of TSD2006. LNCS 4188, pp. 439–446. Springer, Heidelberg, Germany (2006)
2. Tihelka, D. and Romportl, J. : Exploring Automatic Similarity Measures for Unit Selection Tuning. In: Proceedings of Interspeech, pp. 736–739. ISCA, Brighton, Great Britain (2009).
3. Železný, M., Krňoul, Z., Císař, P., Matoušek, J.: Design, Implementation and Evaluation of the Czech Realistic Audio-visual Speech Synthesis. Signal Processing, vol. 12, pp. 3657–3673 (2006)
4. Grůber, M., Legát, M., Ircing P., Romportl, J., Psutka, J.: 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, pp. 266–269. Wydawnictvo Poznanskie, Poznan, Poland (2009)
5. Russel, J. A.: A Circumplex Model of Affect. Journal of Personality and Social Psychology, vol. 39, pp. 1161–1178 (1980)
6. Syrdal, A. K., Kim, Y.-J.: Dialog Speech Acts and Prosody: Considerations for TTS. In: Proceedings of Speech Prosody, pp. 661–665. Campinas, Brazil (2008)
7. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B, vol. 39, no. 1, pp. 1–38 (1977)
8. Romportl, J.: Prosodic Phrases and Semantic Accents in Speech Corpus for Czech TTS Synthesis. In: Proceedings of TSD 2008. LNCS 5246, pp. 493–500. Springer, Berlin–Heidelberg, Germany (2008)
9. Landis, J. R., Koch, G. G.: The Measurement of Observer Agreement for Categorical Data. In: Biometrics, vol. 33, no. 1, pp. 159–174 (1977)