Dialogue Act-Based Expressive Speech Synthesis in Limited Domain for the Czech Language

Martin Grůber, Jindřich Matoušek, Zdeněk Hanzlíček and Daniel Tihelka
University of West Bohemia
Faculty of Applied Sciences
NTIS – New Technologies for the Information Society, Department of Cybernetics
Univerzitní 8, Pilsen, Czech Republic
E-mail: gruber@ntis.zcu.cz

Keywords: speech synthesis, unit selection, HMM, expressivity, dialogue act, limited domain

Received: October 30, 2018

This paper deals with expressive speech synthesis in a dialogue. Dialogue acts – discrete expressive categories – are used for expressivity description. The aim of the work is to create a procedure for development of expressive speech synthesis for a dialogue system in a limited domain. The domain is here limited to dialogues between a human and a computer on a given topic of reminiscing about personal photographs. To incorporate expressivity into synthetic speech, modifications of current algorithms used for neutral speech synthesis are made. An expressive speech corpus is recorded, annotated using a predefined set of dialogue acts, and its acoustic analysis is performed. Unit selection and HMM-based methods are used to synthesize expressive speech, and an evaluation using listening tests is presented. The listeners asses two basic aspects of synthetic expressive speech for isolated utterances: speech quality and expressivity perception. The evaluation is also performed for utterances in a dialogue to asses appropriateness of synthetic expressive speech. It can be concluded that synthetic expressive speech is rated positively even though it is of worse quality when comparing with the neutral speech synthesis. However, synthetic expressive speech is able to transmit expressivity to listeners and to improve the naturalness of the synthetic speech.

Povzetek: Razvita je metoda za izrazno govorno sintezo v češčini.

1 Introduction

Nowadays, speech synthesis techniques produce high quality and intelligible speech. However, to use synthetic speech in dialogue systems (ticket booking [1], information on restaurants or hotels [2], flights [3], trains [4] or weather [5]) 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. Synthetic speech cannot sound completely natural until it expresses a speaker's attitude. Thus, expressive (or emotional) speech synthesis is a frequently discussed topic and has become a concern of many scientists. Even though some results have already been presented, this task has not been satisfactorily solved yet. Some papers which deal with this problem include, but are not limited to [6, 7, 8, 9, 10, 11, 12].

To reduce the complexity of the general expressive speech synthesis, the task is usually somehow limited (as well as limited domain speech synthesis systems are) and focused on a specific domain, e.g. expressive football announcements [13], sport commentaries [14] or dialogue system in a tourism domain [15]. In this work, we limited the domain to conversations between seniors and a computer. Personal photographs were chosen as the topic for these discussions since the work started as a part of a major project aiming at developing a virtual senior companion with an audiovisual interface [16].

Once the specific limited domain is defined, the task of expressive speech synthesis becomes more easily solvable. However, this work tries to propose a general methodology for designing an expressive speech synthesizer in a limited domain. Thus, it should be possible to create a synthesizer for various limited domains following the procedure described herein.

In the first phase of our research, becoming acquainted with the defined domain was the main goal. Thus, an extensive audiovisual database containing 65 natural dialogues between humans (seniors) and a computer (represented by a 3D virtual avatar) was created using the Wizard-of-Oz method which was proposed in [17] and used e.g. in [18, 19]. Afterwards, the dialogues were manually transcribed so that the text could be used later. The process of the database recording is described in Section 2.

Next, on the basis of these dialogues (the texts and the audio recordings), an expressive speech corpus was designed and recorded. The recording of the expressive corpus was performed in the form of a dialogue between a professional female voice talent and a computer. The dialogues were designed on the basis of the natural dialogues recorded in the previous phase. Thus, the voice talent (acting as the virtual avatar now) was recording predefined sentences as responses to the seniors' speech that the voice talent was listening to. The expressive speech corpus recording process is in more details described in Section 3.

To synthesize expressive speech, an expressivity description has to be defined. Many approaches have been suggested in the past. Continuous descriptions using multidimensional space with several axes to determinate "expressivity position" were described e.g. in [20, 21]. Another option is a discrete division into various groups, for emotions e.g. happiness, sadness, anger, joy, etc. [22]. The discrete description is the most commonly used method and various sets of expressive categories are used, e.g. dialogue acts [23, 15], emotion categories [24, 7, 25] or categories like good news and bad news [8, 26]. Thus, a set of expressive categories was defined and used to annotate the expressive speech corpus. The expressive categories used in our work are presented in Section 4 and annotation of the expressive speech corpus is described in Section 5.

There are various methods to produce synthetic speech, the mostly used are unit selection [27], HMM-based methods [28], DNN-based methods [29] or other methods based on neural networks [30, 31]. These methods can be certainly used also for the expressive speech synthesis. In addition, a method for voice conversion [32] can be taken into consideration. Although this method is primarily used for a conversion of source voice to a target voice in the process of the speech synthesis, it can be also used to convert one speaking style to another [33]. DNN-based approaches then allows e.g. adaptation of an expressive model to a new speaker [34].

To incorporate expressivity into speech using unit selection method, the baseline algorithm used e.g. in [35, 36] was slightly modified. The main modification consists in a different target cost calculation. A prosodic feature representing an expressive category is considered in addition to the current set of features used for the cost calculation. To get specific penalties for speech units labelled with an expressive category different from the requested one, enumerated differences between various expressive categories are used. To compute the penalties, a penalty matrix based on perception and acoustic differences is used. The complex acoustic analysis of the expressive speech corpus along with the unit selection method modifications is described in Section 6.

Even though this work is mainly focused on using the unit selection method for expressive speech synthesis, a brief description of preliminary experiments with HMM-based method is also presented. The HMM-based TTS system settings is described in Section 7.

The results and evaluation are presented in Section 8. The expressivity perception ratio is investigated for natural speech and for synthetic speech generated by both the unit selection based TTS system and the HMM-based TTS system. The synthetic speech quality is also discussed in that section. As the results of this work are to be used in a dialogue system, the suitability of produced expressive synthetic speech is evaluated also directly in dialogues.

2 Natural dialogues

To become acquainted with the limited domain, an extensive audiovisual database[1] of natural dialogues was created using the Wizard-of-Oz method. This means that each dialogue was recorded as a dialogue between a human (senior) and a computer (avatar) which was allegedly controlled only by the human voice. However, the computer was covertly controlled by human operators from another room. Thus, the operators were controlling the course of the dialogue whereas the recorded human subjects thought they are interacting with an independent system based on artificial intelligence. The avatar was using neutral TTS system ARTIC [35] to speak to the human subjects. The recording procedure is described in [37] in more details.

2.1 Recording setup

A soundproof recording room has been established for the recording purposes (the scheme is shown in Figure 1). In the recording room, the human subject faces an LCD screen and two speakers. The speech is recorded by two wireless high-quality head microphones (one for the human subject and one for the computer avatar), and the video is captured by three miniDV cameras. A surveillance web-camera was placed in the room to monitor the situation, especially the senior's state. The only contact between a user and the computer was through speech, there was no keyboard nor mouse on the table.

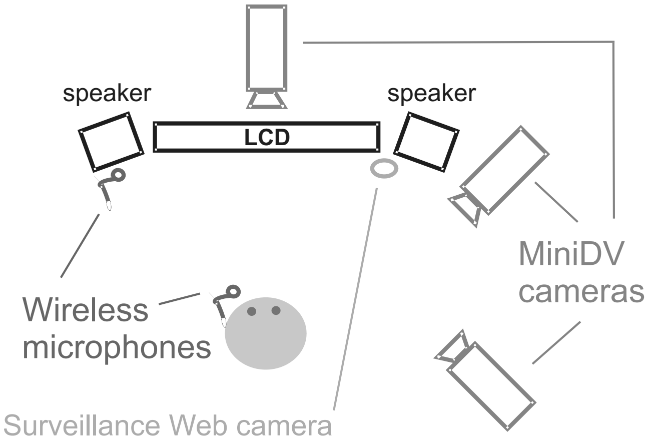


Figure 1: Recording room setup.

A snapshot captured by the miniDV cameras during a recording session is presented in Figure 2. The cameras

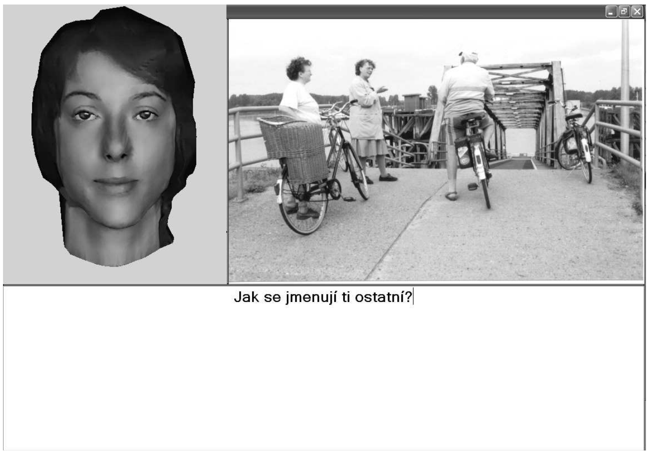
[1] The video recordings are not used for the purposes of the expressive TTS system design. They were just archived and are intended for future use in audiovisual speech recognition, emotion detection, gesture recognition, etc.

were positioned to be able to capture the subject from three different views to provide data usable in various ways.

Figure 2: Screenshot captured by the miniDV cameras during a recording session.

2.2 Recording application description

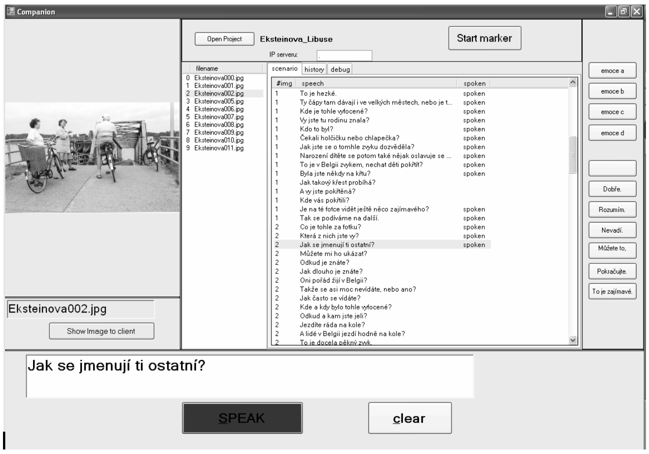
A snapshot of the screen presented to human subjects is shown in Figure 3 ("Presenter" interface). On the left upper part of the LCD screen, there is visualized 3D model of a talking head. This model is used as the avatar, the impersonate companion that should play a role of the partner in the dialogue. Additionally, on the right upper part, there is shown a photograph which is currently being discussed. On the lower half of the screen, there is a place used for displaying subtitles (just in case the synthesized speech is not intelligible sufficiently). The subtitles were displayed only during the first few sessions and then they were switched off as the generated speech turned out to be understandable enough.



Jak se jmenují ti ostatní?

Figure 3: Snapshot of the WoZ system interface - the user's side.

In Figure 4, a screen of the operator's part of the recording application is shown ("Wizard" interface). The interface provides the human operators with possibilities of dialogue flow controlling. The middle part of the screen serves to display the pre-prepared scenario for a dialogue. Note that the wizards could select the sentences from the scenario, the assumption on how the dialogue could develop, by clicking on them. Each sentence of the scenario was given a number related to the picture displayed on the left. This enabled the orientation in large pre-prepared scenarios. Under the picture there is a button for displaying the picture on the "Presenter" screen. Once a sentence is selected by clicking on the list, it appears in the bottom edit box just above the buttons "SPEAK" and "clear". The displayed sentence can be modified before pressing "SPEAK" button and also an arbitrary text can be typed into the edit box. The right part of the screen is intended for displaying buttons bearing non-speech acts (smile, laughter, assentation, hesitation) and quick phrases (Yes. No. It's nice. Alright. Doesn't matter. Go on; etc.).



Eksteinova_Libuse

IP serveru: .

	filename
0	Eksteinova000.jpg
1	Eksteinova001.jpg
2	Eksteinova002.jpg
3	Eksteinova005.jpg
4	Eksteinova006.jpg
5	Eksteinova007.jpg
6	Eksteinova008.jpg
7	Eksteinova009.jpg
8	Eksteinova010.jpg
9	Eksteinova011.jpg

#img	speech	spoken
1	To je hezké.	spoken
1	Ty čápy tam dávají i ve velkých městech, nebo je t...	spoken
1	Kde je tohle vyfocené?	spoken
1	Vy jste tu rodinu znala?	spoken
1	Kdo to byl?	spoken
1	Čekali holčičku nebo chlapečka?	spoken
1	Jak jste se o tomhle zvyku dozvěděla?	spoken
1	Narození dítěte se potom také nějak oslavuje se ...	spoken
1	To je v Belgii zvykem, nechat děti pokřtít?	spoken
1	Byla jste někdy na křtu?	spoken
1	Jak takový křest probíhá?	
1	A vy jste pokřtěná?	
1	Kde vás pokřtili?	
1	Je na té fotce vidět ještě něco zajímavého?	spoken
1	Tak se podíváme na další.	spoken
2	Co je tohle za fotku?	spoken
2	Která z nich jste vy?	spoken
2	Jak se jmenují ti ostatní?	spoken
2	Můžete mi ho ukázat?	
2	Odkud je znáte?	
2	Jak dlouho je znáte?	
2	Oni pořád žijí v Belgii?	
2	Takže se asi moc nevídáte, nebo ano?	
2	Jak často se vídáte?	
2	Kde a kdy bylo tohle vyfocené?	
2	Odkud a kam jste jeli?	
2	Jezdíte ráda na kole?	
2	A lidé v Belgii jezdí hodně na kole?	
2	To je docela pěkný zvyk.	

Eksteinova002.jpg

Jak se jmenují ti ostatní?

Figure 4: Snapshot of the WoZ system interface - the operator's side.

2.3 Audiovisual database statistics

Almost all audio recordings are stored using 22kHz sample rate and 16-bit resolution. The first six dialogues were recorded using 48kHz sample rate, later it was reduced to the current level according to requirements of the cooperating team dealing with ASR (automatic speech recognition). The total number of recorded dialogues is 65. Based on gender, the set of speakers can be divided into 37 females and 28 males. Mean age of the speakers is 69.3 years; this number is almost the same for both male and female speakers. The oldest person was a female, 86 years old. The youngest one was also a female, 54 years old. All the recorded subjects were native Czech speakers; two of them (1 male and 1 female) spoke a regional Moravian dialect. This dialect differs from regular Czech language in pronunciation and also a little in vocabulary. Approximately one half of the subjects stated in the after recording form that they have a computer at home. Nevertheless, most of them do not use it very often. Almost all the dialogues were rated as friendly and smooth. And even more, the users were really enjoying reminiscing on their photos, no matter that the partner in the dialogue was an avatar. Duration of each dialogue was limited to 1 hour, as this was the capacity of tapes used in miniDV cameras, resulting in average duration 56 minutes per dialogue. During the conversation, 8 photographs were discussed in average (maximum was 12, minimum 3).

3 Expressive corpus recording

3.1 Texts preparation

For developing a high-quality expressive speech synthesis system, an expressive speech corpus has to be created. Such a corpus can be then merged or just enhanced by a neutral one to create a robust corpus containing neutral speech as well as expressivity while keeping a maximum speech units coverage (phonetic balance). The process of designing texts for the expressive corpus recording is very important. The real natural dialogues and their transcriptions were taken as a basis for such a design. Thus, almost all the texts (more than 7000 sentences) uttered by the computer avatar during the natural dialogues were used. Texts containing unfinished phrases due to e.g. speakers overlapping were omitted. These texts form a set of sentences to be recorded.

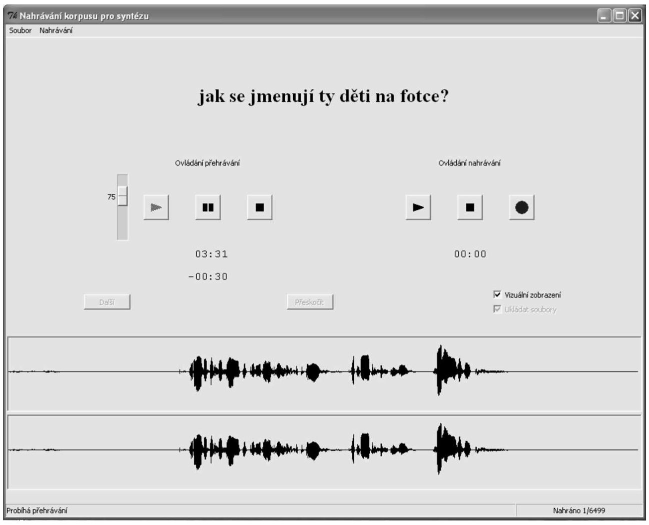
3.2 Recording process

For the expressive corpus recording, a method using so-called scenarios was applied. A scenario in our case can be viewed as a natural dialogue whose course is prepared in advance, just with missing audio of one of the participants (the avatar). This means that the parts of the dialogues to be uttered by a voice talent represent the computer avatar responses and order of these parts is fixed. The parts also follow the natural dialogues and are accompanied with the other participant's original speech to provide the voice talent with information about the context. Actually, the recording was a simulation of the natural dialogues where the voice talent was standing for the computer avatar and was pronouncing its sentences. This should stimulate the voice talent to became naturally expressive while recording.

As the voice talent, a female professional stage-player experienced in speech corpora recording was chosen. The voice talent had already recorded the neutral speech corpus for our neutral TTS system. This corresponds with the intention suggested in Section 3.1 that the expressive corpus should be enhanced by the neutral one to keep the speech units coverage. To improve the performance of tools processing the recorded speech corpora, glottal signal was captured along with the speech signal during the recording.

3.3 Recording application description

To record the expressive corpus using the above described method, a special recording application was developed. The application interface is depicted in Figure 5.



jak se jmenují ty děti na fotce?

Figure 5: Interface of the application for expressive corpus recording.

On the upper part of the application window, the text to be recorded is displayed. However, the voice talent was allowed to change the exact sentence wording if unclear[2] while keeping the same meaning. On the middle part, there are, among other things, control buttons for recording and listening. On the bottom, the waveform of the just recorded sentence is shown. The application can be also controlled via keyboard short-cuts to make it more comfortable for the voice talent.

4 Expressivity description

To incorporate expressivity in synthetic speech, some kind of its description is necessary. A general description of expressivity is a very complex task that has not been satisfactorily solved yet even though there are some studies (e.g. [38]) dealing with this topic. For various research fields and their tasks, there are various possibilities of expressivity description. In our work, a description making use of so-called dialogue acts was used. It is a categorical description based on a classification of expressivity into pre-defined classes (used also in [39, 23, 15]).

Although there are several schemas describing expressivity using dialogue acts (including DAMSL [40, 41], SWBD-DAMSL [42], VERBMOBIL [43, 44] or AT&T schema [39]), a new schema was employed to describe expressivity in our limited domain in question. The set of proposed dialogue acts is shown in Table 1 along with a few examples.

The definition of the dialogue acts was based on the audiovisual database of the natural dialogues (described in Section 2) and on the expressive speech corpus (described

[2] Since the texts for the recording were prepared automatically and were not manually checked due to their high number, they could contain some typos, unintelligibilities or unclarities.

dialogue act	*example*
directive	Tell me that. Talk.
request	Let's get back to that later.
wait	Wait a minute. Just a moment.
apology	I'm sorry. Excuse me.
greeting	Hello. Good morning.
goodbye	Goodbye. See you later.
thanks	Thank you. Thanks.
surprise	Do you really have 10 siblings?
sad empathy	I'm sorry to hear that. It's really terrible.
happy empathy	It's nice. Great. It had to be wonderful.
showing interest	Can you tell me more about it?
confirmation	Yes. Yeah. I see. Well. Hmm.
disconfirmation	No. I don't understand.
encouragement	Well. For example? And what about you?
not specified	Do you hear me well? My name is Paul.

Table 1: Set of dialogue acts.

in Section 3). These dialogue acts are than used for expressive corpus annotation (Section 5) and also in the process of the expressive speech synthesis (Section 6).

The need for a new dialogue act schema was driven by a definition of our specific limited domain. Most of the dialogue acts are intended to encourage the (human) partner in a dialogue to talk more about a topic while keeping the computer dialogue system to behave more like a patient listener.

Even though the dialogue acts schemas are generally supposed to describe various phases of dialogues, we assume that in various dialogues' phases a speaker can present his state of mind, mood or personal attitude in a specific way. We believe that the proposed set of dialogue acts can be used not only for description of various dialogue phases but that it also represents the speaker's attitude and affective state expressed by expressive speech. Using these dialogue acts in this limited domain, the synthetic speech is supposed to become more natural for the listeners (seniors in this case).

5 Expressive corpus annotation

The expressive speech corpus was annotated by dialogue acts using 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 a particular listener. Preparation works, listening test framework, evaluation of listening test result and a measure of inter-rater agreement analysis is presented in the following paragraphs.

5.1 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 were instructed to listen to the recordings very carefully and subsequently mark dialogue acts that are expressed within the sentence. The number of possibly marked dialogue acts for one utterance was just upon the listeners, they were not limited anyhow. Few sample sentences labelled with dialogue acts were provided and available to the listeners on view at every turn. If any listener marked one utterance with more than one dialogue act, he was also required to specify whether the functions occur in that sentence consecutively or concurrently. If the dialogue acts are marked as consecutive in a particular utterance, this utterance is omitted from further research for now. These sentences should be manually reviewed later and either divided into more shorter sentences or omitted completely.

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

5.2 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 dialogue act that was marked by the majority of the listeners. In case of less then 50% of all listeners marked any dialogue act, 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 [45]. Knowing the model parameters, it is possible to deduce true observation which is called 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 dialogue act 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 dialogue act 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. This approach was also successfully used in other works regarding speech synthesis research, see [46].

Further, we need to confirm that the listeners marked the sentences with dialogue acts consistently and achieved some measure of agreement. Otherwise the subjective annotations could be considered as accidental or the dialogue acts 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 [47, 48]. 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 dialogue act. Computation of overall Fleiss' kappa is impossible because the listeners were allowed to mark more than one dialogue act for each sentence. However, the overall value can be evaluated as the mean of Fleiss' kappas of all dialogue acts.

Another measure used here is Cohen's kappa [49, 48]. 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 dialogue act separately. Thus, we can find out whether particular listener was in agreement with the objective annotation for certain dialogue act. Finally, the mean of Cohen's kappas of all dialogue acts can be calculated.

Results of agreement measures are presented in Table 2. Values of Fleiss' and Cohen's kappas vary between 0 and 1, the higher value the better agreement. More detailed interpretation of measure of agreement is e.g in [50].

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, dialogue acts *OTHER* and *NOT-SPECIFIED* should be considered as poorly recognizable. It is understandable when taking into consideration their definitions. After eliminating values of these dialogue acts 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 dialogue acts *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 a substantial agreement.

As it is shown in Table 2, agreement among listeners regarding classification of consecutive dialogue act was measured too. The listeners agreed on this label moderately among each other and substantially with the objective annotation. There are also shown ratios of the particular dialogue acts occurrence when maximum likelihood method was used for the objective annotation obtaining. It is obvious that dialogue acts *SHOW-INTEREST* and *ENCOURAGE* are the most frequent.

6 Unit selection

6.1 General unit selection approach

In general, a unit selection algorithm (for our system described e.g. in [51]) is used to form resulting synthetic speech from speech units that 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 unit selection process usually respects two various groups of candidates' features.

6.1.1 Concatenation cost

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 of the speech signal), 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 end of the originally preceding (adjacent in the original corpus) unit are compared with the same features from the end of unit u_{i-1}.

dialogue act	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 DA	0.5138	Moderate	0.6570	0.2443	Substantial	0.0374

Table 2: Fleiss' and Cohen's kappa and occurrence ratio for various dialogue acts and for the "consecutive DAs" 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.

6.1.2 Target cost

Features in the other group are used for a target cost computation. This cost reflects the level of an approximation of a target unit by any of the candidates; in other words, how a candidate from the unit inventory fits a corresponding target unit — a theoretical 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* feature of a candidate for unit u_i and target unit t_i. The differences of particular features (d_j) can be also referred to as penalties.

For our ARTIC TTS system, the features that are considered when calculating the target cost are shown in Table 3.

feature	weight
position in a prosodic word	7.0
left phoneme context	3.0
right phoneme context	3.0
prosodeme type	14.0
voicing – at the beginning	8.5
voicing – at the end	8.5

Table 3: Prosodic features along with their weights used for target cost calculation in the ARTIC TTS system.

6.2 Basic target cost for expressive speech synthesis

When using the expressive speech corpus, the set of the features used for the target cost computation is extended with one more feature. Regarding the aforementioned expressivity description, it is called *dialogue act*. The penalty d_{da} between a candidate u_i of a target unit t_i can be in the easiest way calculated as follows:

$$d_{da} = \begin{cases} 0 & \text{if } da_t = da_c \\ 1 & \text{otherwise} \end{cases}, \tag{3}$$

where d_{da} is a difference (penalty), da_t is a dialogue act of the target unit t_i and da_c is a dialogue act of the candidate u_i.

Finally, a weight for this penalty needs to be set since the target cost is calculated as a weighted sum of particular penalties.

6.3 Advanced target cost for expressive speech synthesis

The target cost calculation presented in equation 3 is very simple and it assumes that penalties for different expressive categories (represented by the dialogue acts) are the same. However, this is not true in most cases. For instance, the difference between *SAD-EMPATHY* and *HAPPY-EMPATHY* should be probably greater than a difference between *SAD-EMPATHY* and *NEUTRAL* — this means that when synthesizing a sentence in the *SAD-EMPATHY* manner and there is no available or suitable candidate labelled with this dialogue act, it is probably better to consider a candidate labelled with *NEUTRAL* dialogue act than considering a candidate labelled as *HAPPY-EMPATHY*. Therefore, it is necessary to enumerate differences between various dialogue acts and use them for the target cost calculation. The basics of the procedure are described in [52], a bit enhanced version is presented here.

6.3.1 General penalty matrix

The differences are assumed to be coded in a penalty matrix **M**, where coefficients m_{ij} represents a difference (a penalty) between a dialogue act *i* and a dialogue act *j*.

To determine coefficients of the matrix, i.e. the differences in dialogue acts, two aspects should be considered: human perception of the speech and acoustic measures calculated from the signal. Thus, two separate matrices are created and then combined. Coefficients of the first matrix **P** are calculated on the basis of a listening test that was performed to annotate the dialogue acts in the expressive speech corpus [37] (see Section 6.3.2). The second matrix **A** is then based on results of an acoustic analysis of expressive speech [53] (see Section 6.3.3). The combined final penalty matrix **M** represents the overall differences (penalties) between various dialogue acts.

6.3.2 Listening test based differences

Given the annotations of the expressive recordings presented in Section 5, a penalty matrix **P** was created. Its coefficients p_{ij} were calculated according to the following equation:

$$p_{ij} = \frac{\text{abs}(\log(\frac{num_{ij}}{max_i}))}{K}, \tag{4}$$

where num_{ij} represents how many times recordings with dialogue act *i* (according to the objective annotation as presented in Section 5.2) were labelled with dialogue act *j* (calculated over all listeners and all recordings), max_i represents the maximum value of num_{ij} for fixed *i* and *K* is a constant defined as $K \geq K_{min}$ where:

$$K_{min} = \max_{\forall i,j}(\text{abs}(\log(\frac{num_{ij}}{max_i})), \tag{5}$$

where $\max_{\forall i,j}$ is the maximum value for all i, j for which the *log* is defined. For situations where the *log* is not defined, the p_{ij} was set as $p_{ij} = K$. In our experiments, the $K = 5 \approx 2 \times K_{min}$. 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).

6.3.3 Acoustic analysis based differences

An extensive acoustic analysis of the expressive corpus was performed in [53]. On the basis of this analysis, a penalty matrix **A** was created. Its coefficients a_{ij} were calculated as the Euclidean distance between numeric vectors representing the dialogue acts *i* and *j* in a 12-dimensional space. The components of the vector consist of normalized values of 4 statistical characteristics (mean value, standard deviation, skewness, kurtosis) for 3 acoustic parameters (F0 value, RMS energy and unit duration). The acoustic analysis proved that these features can be used as acoustic distance measures for this purpose. It is likely that there other features not considered in this work which may affect the measure in any way and whose influence should be explored in the future.

6.3.4 Final penalty matrix

The final penalty matrix containing numeric differences between various dialogue acts is an appropriate combination of two separate penalty matrices (matrix **P** based on the annotations and matrix **A** based on the acoustic analysis). The coefficients m_{ij} of matrix **M** can be calculated as follows:

$$m_{ij} = \frac{w_p \cdot p_{ij} + w_a \cdot a_{ij}}{w_p + w_a}, \tag{6}$$

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

After several experiments, values $w_p = 3$ and $w_a = 1$ were used as the weights. Using this setting, the best results were achieved when subjectively comparing resulting synthetic speech. We also believe that the perceptual part should be emphasized. The final penalty matrix is depicted in Table 4.

6.3.5 Weight tuning for dialogue act feature

Proper setting of a weight for any of the features is not an easy task. Some techniques for automatic settings have also been developed [54, 55]. However, in our system the settings shown in Table 3 is used as it was proved to be appropriate in applications of our TTS.

To set the weight for the dialogue act feature, sets of synthetic utterances were generated for various settings. Using a subjective evaluation (a brief listening test) and considering weights for other features, the final weight was defined as $w_{DA} = 12.0$. When compared with Table 3, this weight is one of the highest among others.

7 HMM algorithm modification/training

Along with the concatenative unit selection method, statistical parametric speech synthesis based on using hidden Markov models (abbreviated as HMM-based speech synthesis) is one of the most researched synthesis methods [28]. Several experiments on using this synthesis method for generating expressive speech are described in [14]. In the HMM approach, statistical models (an extended type of HMMs) are trained from natural speech database. Spectral parameters, fundamental frequency and eventually some excitation parameters are modelled simultaneously by the corresponding multi-stream HMMs.

The variability of speech is modelled by using models with large context description, i.e. individual models are defined for various phonetic, prosodic and linguistic contexts, that are described by so-called contextual factors. The contextual factors employed in our experiments are listed in Table 5. For more details, see e.g. [56].

To increase the robustness of the estimated model parameters, models of acoustically similar units are clustered by a decision-tree-based context-clustering algorithm. As a result, similar units share one common model.

Within the HMM-based speech synthesis, various methods for modelling the expressivity or speaking styles have been introduced. The simplest one uses so-called style dependent models [59], i.e. an independent set of HMMs is trained for each expression. An obvious drawback of this approach is a large amount of training data required for particular expressions.

A better solution are so-called style mixed models [59], where one set of HMMs is trained for all expressions together and particular expressions are distinguished by introducing an additional contextual factor. Then, models of units that are acoustically similar for more expressions are clustered. Independent models are trained only when there is a significant difference between particular expressions.

Another option of modelling expressions are methods based on model adaptation [60, 61]; they are usually preferred because they allow to control the speech style or expression more precisely and require less training data. However, the style mixed model utilizing an additional contextual factor for dialogue act was used in this work.

8 Evaluation & results

This section deals with an evaluation of the procedure described in this paper to verify that it fulfils the goals which were specified at the beginning. Especially, it should be verified that listeners perceive the synthetic speech produced by the developed system as expressive (Section 8.2.1) and also how the quality of synthetic speech changed in comparison with the baseline system (Section 8.2.2). Since the proposed TTS system is focused on a usage in a specified dialogue system, the suitability of the

	APOLOGY	CONFIRM	DIRECTIVE	DISCONFIRM	ENCOURAGE	GOODBYE	GREETING	HAPPY-EMPATHY	NOT-SPECIFIED	OTHER	REQUEST	SAD-EMPATHY	SHOW-INTEREST	SURPRISE	THANKS	WAIT	NEUTRAL
APOLOGY	0.00	0.58	0.39	0.40	0.83	0.25	0.90	0.48	0.36	0.50	0.84	0.17	0.45	0.48	0.83	0.47	0.78
CONFIRM	0.96	0.00	0.71	0.58	0.71	0.95	0.92	0.40	0.48	0.72	0.93	0.41	0.50	0.53	0.74	0.70	0.72
DIRECTIVE	0.90	0.56	0.00	0.58	0.26	0.86	0.46	0.50	0.36	0.65	0.33	0.49	0.38	0.81	0.92	0.59	0.41
DISCONFIRM	0.34	0.33	0.84	0.00	0.43	0.87	0.86	0.28	0.23	0.40	0.44	0.28	0.42	0.42	0.90	0.45	0.54
ENCOURAGE	0.83	0.50	0.48	0.64	0.00	0.83	0.60	0.35	0.31	0.43	0.27	0.39	0.14	0.27	0.75	0.52	0.55
GOODBYE	0.30	0.53	0.34	0.87	0.47	0.00	0.59	0.19	0.17	0.50	0.82	0.25	0.42	0.82	0.25	0.55	0.61
GREETING	0.90	0.63	0.86	0.86	0.58	0.90	0.00	0.52	0.31	0.68	0.89	0.90	0.53	0.87	0.93	0.54	0.42
HAPPY-EMPATHY	0.44	0.25	0.58	0.41	0.24	0.39	0.89	0.00	0.21	0.45	0.54	0.29	0.27	0.28	0.46	0.54	0.58
NOT-SPECIFIED	0.39	0.26	0.25	0.35	0.12	0.19	0.26	0.15	0.00	0.34	0.20	0.22	0.13	0.18	0.32	0.38	0.46
OTHER	0.95	1.00	0.29	0.97	0.94	0.36	0.97	0.96	0.30	0.00	0.94	0.99	0.94	0.95	0.92	0.89	0.84
REQUEST	0.84	0.93	0.30	0.88	0.17	0.82	0.35	0.45	0.31	0.40	0.00	0.51	0.28	0.80	0.87	0.58	0.59
SAD-EMPATHY	0.28	0.28	0.37	0.41	0.26	0.32	0.90	0.33	0.28	0.50	0.49	0.00	0.25	0.33	0.89	0.59	0.67
SHOW-INTEREST	0.88	0.53	0.39	0.62	0.15	0.85	0.87	0.43	0.27	0.58	0.32	0.41	0.00	0.34	0.90	0.57	0.45
SURPRISE	0.86	0.29	0.47	0.40	0.05	0.82	0.87	0.15	0.14	0.37	0.35	0.24	0.06	0.00	0.59	0.51	0.51
THANKS	0.83	0.54	0.92	0.90	0.53	0.46	0.93	0.88	0.91	0.92	0.87	0.89	0.90	0.89	0.00	0.88	0.86
WAIT	0.47	0.58	0.24	0.89	0.32	0.93	0.88	0.55	0.56	0.89	0.29	0.57	0.39	0.90	0.88	0.00	0.63
NEUTRAL	0.78	0.72	0.41	0.54	0.55	0.61	0.42	0.58	0.46	0.84	0.59	0.67	0.45	0.51	0.86	0.63	0.00

Table 4: Final penalty matrix **M**.

Contextual factor	Possible values
Left and right phonetic context	Czech phonetic alphabet [57]
Phone position in prosodic word (forward and backward)	1, 2, 3, 4, 5 ...
Prosodic word position in clause (forward and backward)	
Prosodeme	terminating satisfactorily, terminating unsatisfactorily, non-terminating, null
Dialogue act	see Section 4

Table 5: A list of contextual factors and their values. Prosodic words, clauses and prosodemes are thoroughly described in [58].

expressive speech synthesis in such a dialogue system is also evaluated (Section 8.4).

During the design of the expressive TTS system, it turned out that some of the dialogue acts (further referred to as DAs) appear much more frequently than others, some of them are very rare. Thus, only the most frequent DAs were used to evaluate the system and they were divided into two separate groups:

Expressive dialogue acts:

- *SHOW-INTEREST* – relative frequency 34.9 %;
- *ENCOURAGE* – relative frequency 29.4 %;
- *CONFIRM* – relative frequency 13.2 %;
- *HAPPY-EMPATHY* – relative frequency 8.6 %;
- *SAD-EMPATHY* – it was added because it is considered to be an opposite to *HAPPY-EMPATHY* dialogue act; relative frequency 3.4%;

Neutral dialogue acts:

- *NOT-SPECIFIED* – besides it is one of the most frequently occurring DAs, it should also represents the neutral synthetic speech; relative frequency 7.4 %;
- *NEUTRAL* – this is not a DA per se, it is defined here to represent the neutral speech produced by the current baseline TTS system for the purposes of the evaluation.

All the listening tests described further were performed using the same system as it was used for the expressive corpus annotation (described in Section 5.1). Of course, the questions and options were different within this evaluation but the core of the system is the same. The majority of listening tests participants were experts in speech or language processing, some of them were university students. Texts of synthesized utterances were not a part of the corpora, new texts were created for this purpose. The content of the texts corresponds to the dialogue act to be synthesized (for expressive synthesis), or it is neutral (for neutral synthesis).

8.1 Expressivity perception in natural speech

Before assessing the synthetic expressive speech, a listening test focused on expressivity perception in natural speech was performed. This gives us a brief overview of how the listeners are able to perceive the expressivity and later a comparison between expressivity perception in natural and synthetic speech can be presented.

All the listeners were assessing randomly selected utterances form the natural corpora (neutral and expressive) and their task was to mark if they perceive any kind of expressivity or not or if they are not able to make a decision. 14 listeners participated in this test, each listener was presented with 34 utterances – 4 for each expressive dialogue act being evaluated and 7 for each dialogue act considered as neutral (i.e. *NOT-SPECIFIED* and *NEUTRAL*). The results are depicted in Figure 6 and also shown in Table 6.

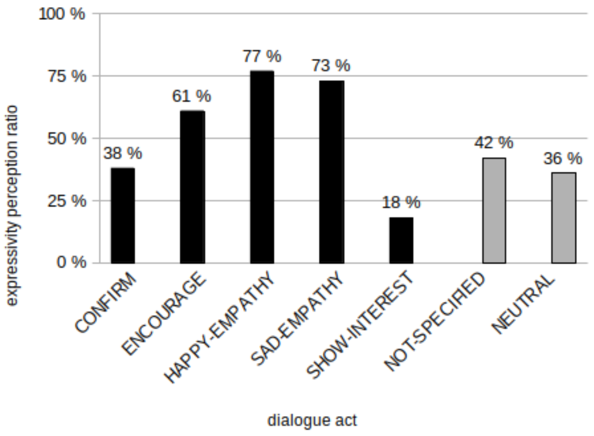


Figure 6: Expressivity perception in natural speech.

dialogue act	expressivity perception ratio	cannot decide
CONFIRM	38 %	3 %
ENCOURAGE	61 %	7 %
HAPPY-EMPATHY	77 %	4 %
SAD-EMPATHY	73 %	6 %
SHOW-INTEREST	18 %	11 %
mean	**53** %	**6** %
NOT-SPECIFIED	42 %	13 %
NEUTRAL	36 %	3 %
mean	**39** %	**7** %

Table 6: Expressivity perception in natural speech.

The results are quite surprising, especially for neutral speech. In 39 % of neutral natural utterances (in average, including *NOT-SPECIFIED*), the listeners perceived an expressivity. It seems that some kind of expressivity is included even in the neutral corpus and the listeners are very sensitive to that, and they are able to perceive it. This fact can be related to the content of speech since as it was described in [62], the content as such might also influence the listeners' expressivity perception.

The results for the expressive DAs depends on a particular DA. For instance, utterances marked as *HAPPY-EMPATHY* and *SAD-EMPATHY* are mostly recognized as expressive whereas utterances marked as *SHOW-INTEREST* are not.

These results give us a baseline for the evaluation of expressive synthetic speech. Since for some DAs the listeners don't perceive expressivity even in natural speech, it's unlikely that they will perceive it in synthetic speech.

8.2 Evaluation of the unit selection based expressive speech synthesis

During the evaluation of expressive synthetic speech, two main factors were investigated – expressivity perception and speech quality. It's supposed that the quality of synthetic speech will be affected by the expressivity integration as the expressive speech is much more dynamic and thus more artificial artifacts may occur. This section deals with the evaluation of expressive synthetic speech produced by the unit selection TTS system. The evaluation of HMM-based TTS system is presented in section 8.3.

In the listening tests regarding expressive synthetic speech evaluation, 13 listeners assessed 30 utterances – 4 for each DA in question and 2 for natural neutral speech (so that a comparison of speech quality can be performed).

8.2.1 Expressivity perception in synthetic speech

The results for expressivity perception in synthetic expressive speech are depicted in Figure 7 and presented in Table 7.

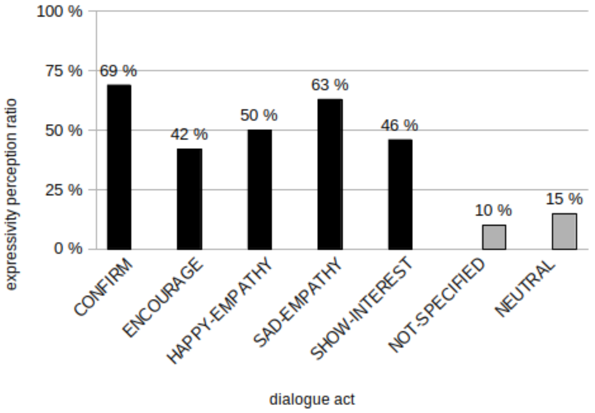


Figure 7: Expressivity perception in synthetic speech (unit selection).

Again, a surprising result can be observed for natural neutral speech as an expressivity was perceived at a quite high ratio (42 %). However, it is consistent with the previous results presented in Table 6 (39 %).

dialogue act	expressivity perception ratio	cannot decide
CONFIRM	69 %	4 %
ENCOURAGE	42 %	8 %
HAPPY-EMPATHY	50 %	10 %
SAD-EMPATHY	63 %	4 %
SHOW-INTEREST	46 %	4 %
mean	**54** %	**6** %
NOT-SPECIFIED	10 %	0 %
NEUTRAL	15 %	0 %
mean	**13** %	**0** %
natural speech (neutral)	42 %	4 %

Table 7: Expressivity perception in synthetic speech (unit selection).

For synthetic speech generated as *NOT-SPECIFIED* and for baseline neutral synthetic speech (marked as *NEUTRAL*), almost no expressivity was perceived. On the other hand, for expressive DAs, the expressivity perception ratio was quite high (mean value 54 %) and it was even slightly higher than for expressive natural speech (mean value 53 %, see Table 6).

To verify that the achieved results are not random, a statistical measure for listeners agreement (the Fleiss' kappa was used here) was calculated. Its value varies in the range $< -1, 1 >$ and a positive value indicates an agreement above the chance level. In our experiment, the Fleiss' kappa was calculated as $\kappa_F = 0.37$ which means a moderate agreement.

In addition, other measures might be used to verify the results; for instance *precision*, *recall*, *F1* and *accuracy* measures which are mostly used for evaluation of classifiers in classification tasks. However, the presented listening test can be also viewed as a classification task where the listeners as classifiers classify into two distinct classes: *perceive* and *do not perceive* expressivity (the *cannot decide* answers were not considered in this verification). The measure are determined as follows:

$$P = \frac{t_p}{p_p}, \quad R = \frac{t_p}{a_p}$$

$$F1 = \frac{2 * P * R}{P + R}, \quad A = \frac{t_p + t_n}{a_p + a_n}$$

where P is *precision*, the ability of a listener not to perceive a neutral sentence as expressive; R is *recall* (also *sensitivity*), the ability of a listener not to perceive expressive sentences as neutral; A is *accuracy*, the ability of a listener to perceive expressivity in expressive sentences and not to perceive it in neutral sentences; $F1$ is the harmonic mean of precision and recall; t_p means "true positives" (i.e. the number of expressive sentences correctly perceived as expressive); t_n means "true negatives" (i.e. the number of neutral sentences correctly perceived as neutral); p_p stands for "predicted positives" (i.e. the number of all sentences perceived as expressive); a_p stands for "actual positives" (i.e. the number of all actual expressive sentences); a_n means "actual negatives" (i.e. the number of all actual neutral sentences).

The calculated values of these measures are presented in Table 8 altogether with values that would be achieved in case the expressivity perception is assessed completely at random.

measure	real listeners	random assessment
precision	0.92	0.72
recall	0.58	0.50
F1 measure	0.71	0.59
accuracy	0.66	0.50

Table 8: Statistical measures for expressivity perception listening test and comparison with completely random assessment.

As the verification indicates, the expressivity perception ratio in synthetic speech is not a result of a random process. It's necessary to note that there are two main facts which affect the expressivity perception. The first one is the TTS system and the synthetic speech whose evaluation is the main goal. The second fact is the listeners – each of them might perceive (assess) various intensity of various expressivity categories differently. However, the main task here is not to evaluate the listeners and if they are or they are not able to perceive an expressivity (which is basically impossible). The listeners are just believed to and the only thing that can be done is to perform some kind of agreement measure calculation.

In synthetic expressive speech generated with a particular DA in mind, the relative ratio between units originally coming from utterances labelled with this DA and units coming from other utterances can be measured. The ratio might vary depending on setting of the weight for the *dialogue act* feature. The calculated ratios for the current weight settings (as designed in Section 6.3.5) are shown in Figure 8.

It's worth noting that the measure is very low for *NOT-SPECIFIED* DA. However, after further investigation, it turned out that when synthesizing utterances for this DA, units coming from the neutral corpus (*NEUTRAL*) were mostly selected. It supports the assumption that the *NOT-SPECIFIED* DA represents neutral speech (although in the final penalty matrix **M** the distance between *NOT-SPECIFIED* and *NEUTRAL* was calculated as 0.46 which is quite high). It also seems that there is no strong relation between this measure and the expressivity perception results presented in Table 7.

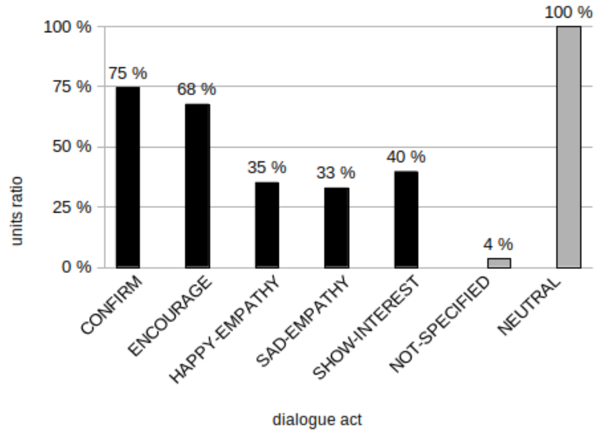


Figure 8: Relative ratio of units coming from utterances labelled with the DA which was intended to be synthesized.

8.2.2 Quality evaluation

To investigate whether the synthetic speech quality deteriorated by adding the expressivity, a MOS test evaluation was performed. In the MOS test, the listeners assess the speech quality using a 5-point scale where, in theory, the natural speech should be evaluated as 5 ($100\ \%$) and a very unnatural speech as 1 ($0\ \%$). The test was running along with the expressivity perception test, i.e. the test conditions, test utterances and the listeners were the same as for the evaluation that is presented in Section 8.2.1. The results of this MOS test are shown in Figure 9 and also in Table 9 altogether with a relative comparison with the natural speech (whose result is evaluated as $100\ \%$).

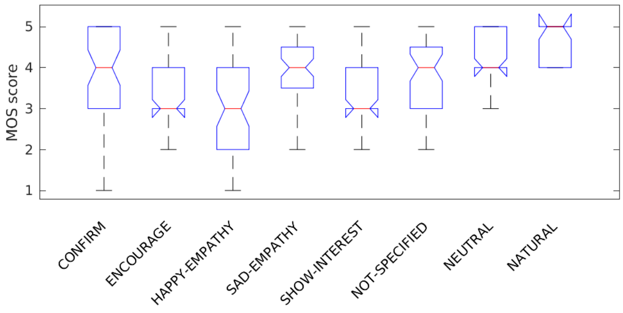


Figure 9: Evaluation of speech quality using a MOS test (unit selection).

The results suggest that the quality of expressive synthetic speech is worse than the quality of neutral synthetic speech by 0.49 of the MOS score ($13\ \%$) in average. It is almost the same difference as between natural speech and neutral synthetic speech (0.65 of the MOS score). This deterioration is probably caused by greater variability of the acoustic signal of expressive speech. Thus, the artifacts might occur more often than in neutral synthetic speech.

An auxiliary measure called *smooth joints* can be also calculated. A smooth joint is a concatenation point of two

dialogue act	MOS score		comparison with
	mean	std	natural speech
CONFIRM	3.87	1.11	79 %
ENCOURAGE	3.48	0.97	68 %
HAPPY-EMPATHY	3.10	1.00	58 %
SAD-EMPATHY	3.87	0.94	79 %
SHOW-INTEREST	3.25	0.92	62 %
mean	**3.51**	**0.99**	**69** %
NOT-SPECIFIED	3.92	0.78	81 %
NEUTRAL	4.08	0.78	83 %
mean	**4.00**	**0.78**	**82** %
natural speech	4.65	0.48	100 %

Table 9: Evaluation of speech quality using a MOS test (unit selection).

speech units that were originally adjacent in the speech corpus and thus their concatenation is natural. The smooth joints measure indicates the relative ratio of such joints with respect to the number of all concatenation points. The calculated values are presented in Figure 10. It is assumed that the less smooth joints in synthetic speech, the more artifacts can occur, causing the synthetic speech quality to be worse.

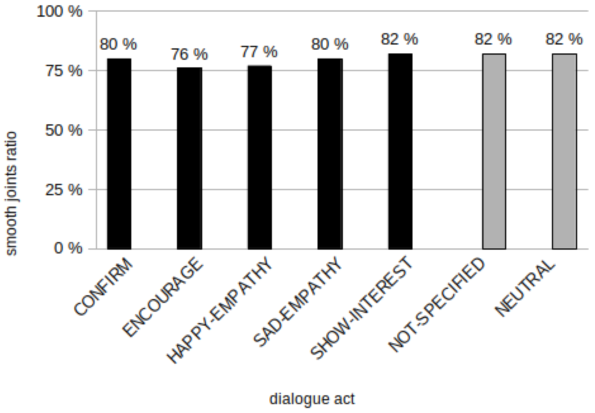


Figure 10: Relative ratio of smooth joints.

It is obvious that the relative ratio of smooth joints is almost the same regardless of the DA (mean 79 %) and also in comparison with neutral synthetic speech (mean 82 %). Also, this measure seems to be unrelated to the expressivity perception measure or the MOS score.

8.3 Evaluation of the HMM-based expressive speech synthesis

Even though this work deals mostly with the unit selection speech synthesis, the results of an experiment with the HMM-based expressive speech synthesis are to be briefly discussed in this section. The used method is based on the HTS system [63] and adapted to the Czech language [56]. The experiment is described in more details in [62] and the HMM approach is also briefly presented in Section 7. The aim is to evaluate the capability of the HMM-based TTS system to produce expressive speech (shown in Table 10) and to evaluate its quality (Table 11). The presented results are summarized and various DAs are not differentiated. There were 12 listeners participating in these listening tests.

dialogue act	expressivity perception ratio	cannot decide
expressive	15 %	5 %
NOT-SPECIFIED	8 %	3 %

Table 10: Expressivity perception in synthetic speech (HMM).

dialogue act	MOS score	comparison with
	mean	natural speech
expressive DAs + NOT-SPECIFIED	2.71	50 %
natural speech	4.44	100 %

Table 11: Evaluation of speech quality using a MOS test (HMM).

The expressivity perception ratio in synthetic speech produced by the HMM-based expressive TTS system is at a very low level (15 %) in comparison with the unit selection TTS system (54 %). Also the quality of synthetic speech is much worse, 2.7 of the MOS score (50 % of natural speech) for the HMM-based system and 3.5 (69 %) for the unit selection system. Generally, the HMM-based speech synthesis for the Czech language is not yet at such a high level as the unit selection approach is. Moreover, by adding expressive speech into this process, the trained HMM models may in fact mix natural and expressive acoustic signal depending on how the decision trees were created. Thus, in such synthetic speech of a lower quality, it is probably hard to identify any kind of expressivity.

8.4 Evaluation of the expressivity in dialogues

Since the unit selection expressive speech synthesis is going to be used in a specific dialogue system (conversations between seniors and a computer; see Section 1 and 2), it is necessary to evaluate it also with respect to this purpose. A preference listening test was used to perform this kind of evaluation. The test stimuli were prepared as follows:

- 6 appropriate parts of the natural dialogues (see Section 2), each approximately 1 minute in length, were randomly selected. The appropriateness were determined on the basis of sufficiency of the avatar's interactions within the dialogues. These parts will be further referred to as *minidialogues*.

- The acoustic signal of each minidialogue was splitted into parts where the person is speaking and parts where the avatar responses are expressed by the neutral speech synthesis.
- The text contents of the avatar's responses were slightly modified so that the newly generated responses are really to be synthesized and not only played back. The sense of the utterances was of course kept the same so that the dialogue flow is not disrupted.
- The new texts (avatar's responses) were synthesized using both the baseline neutral TTS system and the newly developed expressive TTS system – before the expressive speech synthesis, the texts were labelled by presumably appropriate DAs.
- In some parts of the minidialogues where the person is originally speaking, little modifications were done so that the length of the person's speech was shortened – for instance, the parts where the person was speaking for a long time or where a long silence was detected were removed. Again, the natural dialogue flow was not disrupted.
- the parts of the minidialogues were joint together so that two versions of each minidialogue were created – the first one with the avatar's responses with neutral synthetic speech and the second one with the avatar's responses with expressive synthetic speech.

Each of the 6 minidialoges contains 4 avatar's responses in average expressing various DAs, mostly *SHOW-INTEREST* or *ENCOURAGE*. However, each evaluated DA was included at least once in the responses. The minidialogues were then presented to the listeners within a listening test, both minidialogue's variants in a single test query. The task for the listeners was to decide which variant is more natural, more pleasant and which one would they prefer when being in place of the human minidialogue participant. The results of this evaluation are presented in Table 12; there were 11 listeners participating in this listening test.

synthesis variant	preference
neutral	8 %
expressive	83 %
cannot decide	9 %

Table 12: Evaluation of neutral vs. expressive speech synthesis in dialogues.

It's obvious that the listeners preferred the expressive speech synthesis to the neutral one (83 % preference ratio). This is one of the most important results indicating that the developed system increases the user experience with the TTS system for this limited domain task.

To verify that the avatar's responses were indeed synthesized and not only played back, the measure of smooth joints can be used. The mean value of this measure for the expressive avatar's responses is 86 % which is slightly higher than it was measured in Figure 10 of Section 8.2.1 (mean 82 % for neutral speech and 79 % for expressive speech). However, it still means that the responses were really synthesized.

9 Conclusion

It is necessary to incorporate some kind of expressivity into synthetic speech as it improves the user experience with systems using speech synthesis technology. Expressive speech sounds more naturally in dialogues between humans and computers. There are several ways to make the synthetic speech sound expressively. In this work, expressivity described by dialogue acts was employed and the algorithms of the TTS system were modified to use that information when producing synthetic speech.

The results presented in Section 8 suggest that in speech produced by the expressive TTS system the listeners perceived some kind of expressivity. More importantly, it was also confirmed that in the dialogues within the discussed limited domain, expressive speech is more suitable and preferred than the pure neutral speech produced by the baseline TTS system even though its quality is little bit worse.

Although the development of the expressive TTS system was done within a limited domain of conversations about personal photos between humans and a computer, the whole procedure – data collecting, data annotation, expressive corpus preparation and recording, expressivity description and TTS system modification – can be used within any other limited domain if appropriate expressivity definition is used. Thus, an expressivity can be incorporated to any other dialogue system with a similar structure.

Acknowledgement

This research was supported by the Czech Science Foundation (GA CR), project No. GA19-19324S and by the Ministry of Education, Youth and Sports of the Czech Republic project No. LO1506.

References

[1] J. D. Williams, S. Young, Partially observable Markov decision processes for spoken dialog systems, Computer Speech and Language 21 (2) (2007) 393–422.
`https://doi.org/10.1016/j.csl.2006.06.008`

[2] O. Lemon, K. Georgila, J. Henderson, M. Stuttle, An ISU dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling

in the TALK in-car system, in: Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations, EACL '06, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006, pp. 119–122.
https://doi.org/10.3115/1608974.1608986

[3] X. Wu, M. Xu, W. Wu, Preparing for evaluation of a flight spoken dialogue system, in: Proceedings of ISCSLP, 2002, paper 50.

[4] J. Švec, L. Šmídl, Prototype of Czech spoken dialog system with mixed initiative for railway information service, in: P. Sojka, A. Horák, I. Kopecek, K. Pala (Eds.), Text, Speech and Dialogue, Vol. 6231 of Lecture Notes in Computer Science, Springer, Berlin-Heidelberg, Germany, 2010, pp. 568–575.
https://doi.org/10.1007/978-3-642-15760-8_72

[5] A. Meštrović, L. Bernić, M. Pobar, S. Martinčič-Ipšić, I. Ipšić, Overview of a croatian weather domain spoken dialog system prototype, in: 32nd International Conference on Information Technology Interfaces (ITI), Cavtat, Dubrovnik, 2010, pp. 103–108.

[6] A. W. Black, Unit selection and emotional speech, in: Proceedings of Eurospeech, Geneva, Switzerland, 2003, pp. 1649–1652.

[7] M. Bulut, S. S. Narayanan, A. K. Syrdal, Expressive speech synthesis using a concatenative synthesiser, in: Proceedings of the 7th International Conference on Spoken Language Processing – ICSLP, Denver, CO, USA, 2002, pp. 1265–1268.

[8] W. Hamza, R. Bakis, E. M. Eide, M. A. Picheny, J. F. Pitrelli, The IBM expressive speech synthesis system, in: Proceedings of the 8th International Conference on Spoken Language Processing – ISCLP, Jeju, Korea, 2004, pp. 2577–2580.
https://doi.org/10.1109/tasl.2006.876123

[9] I. Steiner, M. Schröder, M. Charfuelan, A. Klepp, Symbolic vs. acoustics-based style control for expressive unit selection, in: Seventh ISCA Tutorial and Research Workshop on Speech Synthesis, Kyoto, Japan, 2010, pp. 114–119.

[10] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, Y. Ochiai, Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis, Speech Communication 99 (2018) 135–143.
https://doi.org/10.1016/j.specom.2018.03.002

[11] S. An, Z. Ling, L. Dai, Emotional statistical parametric speech synthesis using LSTM-RNNs, in: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2017, pp. 1613–1616.
https://doi.org/10.1109/apsipa.2017.8282282

[12] H. Li, Y. Kang, Z. Wang, EMPHASIS: An emotional phoneme-based acoustic model for speech synthesis system, in: Proceedings of Interspeech, 2018.
https://doi.org/10.21437/interspeech.2018-1511

[13] S. Krstulovic, A. Hunecke, M. Schroder, An HMM-based speech synthesis system applied to German and its adaptation to a limited set of expressive football announcements, in: Proceedings of Interspeech, Antwerp, Belgium, 2007, pp. 1897–1900.

[14] B. Picart, R. Brognaux, , T. Drugman, HMM-based speech synthesis of live sports commentaries: Integration of a two-layer prosody annotation, in: 8th ISCA Speech Synthesis Workshop, Barcelona, Spain, 2013.

[15] H. Yang, H. Meng, L. Cai, Modeling the acoustic correlates of dialog act for expressive Chinese TTS synthesis, IET Conference Publications 2008 (CP544) (2008) 49–53.
https://doi.org/10.1049/cp:20080758

[16] P. Ircing, J. Romportl, Z. Loose, Audiovisual interface for Czech spoken dialogue system, in: IEEE 10th International Conference on Signal Processing Proceedings, Institute of Electrical and Electronics Engineers, Inc., Beijing, China, 2010, pp. 526–529.
https://doi.org/10.1109/icosp.2010.5656088

[17] J. F. Kelley, An iterative design methodology for user-friendly natural language office information applications, ACM Transactions on Information Systems 2 (1) (1984) 26–41.
https://doi.org/10.1145/357417.357420

[18] S. Whittaker, M. Walker, J. Moore, Fish or fowl: A Wizard of Oz evaluation of dialogue strategies in the restaurant domain., in: Language Resources and Evaluation Conference, Gran Canaria, Spain, 2002.

[19] M. Hajdinjak, F. Mihelič, The Wizard of Oz system for weather information retrieval, in: V. Matoušek, P. Mautner (Eds.), Text, Speech and Dialogue, proceedings of the 6th International Conference TSD, Vol. 2807 of Lecture Notes in Computer Science, Springer, Berlin-Heidelberg, Germany, 2003, pp. 400–405.
https://doi.org/10.1007/978-3-540-39398-6_57

[20] J. A. Russell, A circumplex model of affect, Journal of Personality and Social Psychology 39 (1980) 1161–1178.

[21] A. Mehrabian, Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament, Current Psychology 14 (1996) 261–292.
https://doi.org/10.1007/BF02686918

[22] R. R. Cornelius, The science of emotion: Research and tradition in the psychology of emotions, Prentice-Hall, Englewood Cliffs, NJ, USA, 1996.

[23] A. K. Syrdal, A. Conkie, Y.-J. Kim, M. Beutnagel, Speech acts and dialog TTS, in: Proceedings of the 7th ISCA Speech Synthesis Workshop – SSW7, Kyoto, Japan, 2010, pp. 179–183.

[24] E. Zovato, A. Pacchiotti, S. Quazza, S. Sandri, Towards emotional speech synthesis: A rule based approach, in: Proceedings of the 5th ISCA Speech Synthesis Workshop – SSW5, Pittsburgh, PA, USA, 2004, pp. 219–220.

[25] J. M. Montero, J. Gutiérrez-Ariola, S. Palazuelos, E. Enríquez, S. Aguilera, J. M. Pardo, Emotional speech synthesis: From speech database to TTS, in: Proceedings of the 5th International Conference on Spoken Language Processing – ICSLP, Vol. 3, Sydney, Australia, 1998, pp. 923–926.

[26] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, M. A. Picheny, The IBM expressive text-to-speech synthesis system for American English, IEEE Transactions on Audio, Speech, and Language Processing 14 (4) (2006) 1099–1108.
https://doi.org/10.1109/tasl.2006.876123

[27] A. J. Hunt, A. W. Black, Unit selection in a concatenative speech synthesis system using a large speech database, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, 1996, pp. 373–376.
https://doi.org/10.1109/ICASSP.1996.541110

[28] H. Zen, K. Tokuda, A. W. Black, Statistical parametric speech synthesis, Speech Communication 51 (2009) 1039–1064.
https://doi.org/10.1016/j.specom.2009.04.004

[29] H. Zen, A. Senior, M. Schuster, Statistical parametric speech synthesis using deep neural networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 7962–7966.
https://doi.org/10.1109/ICASSP.2013.6639215

[30] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: A generative model for raw audio, in: Arxiv, 2016. arXiv:1609.03499v2.

[31] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al., Tacotron: Towards end-to-end speech synthesis, arXiv preprint arXiv:1703.10135
https://doi.org/10.21437/interspeech.2017-1452

[32] A. Kain, M. W. Macon, Spectral voice conversion for text-to-speech synthesis, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, 1998, pp. 285–288.
https://doi.org/10.1109/icassp.1998.674423

[33] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, K. Shikano, GMM-based voice conversion applied to emotional speech synthesis, IEEE Tranactions on Speech and Audio Processing 7 (1999) 2401–2404.

[34] J. Parker, Y. Stylianou, R. Cipolla, Adaptation of an expressive single speaker deep neural network speech synthesis system, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5309–5313.
https://doi.org/10.1109/ICASSP.2018.8461888

[35] J. Matoušek, D. Tihelka, J. Romportl, Current state of Czech text-to-speech system ARTIC, in: Text, Speech and Dialogue, proceedings of the 9th International Conference TSD, Vol. 4188 of Lecture Notes in Computer Science, Springer, Berlin-Heidelberg, Germany, 2006, pp. 439–446.
https://doi.org/10.1007/11846406_55

[36] D. Tihelka, J. Kala, J. Matoušek, Enhancements of Viterbi search for fast unit selection synthesis, in: Proceedings of Interspeech, Makuhari, Japan, 2010, pp. 174–177.

[37] M. Grůber, M. Legát, P. Ircing, J. Romportl, J. Psutka, Czech Senior COMPANION: Wizard of Oz data collection and expressive speech corpus recording and annotation, in: Z. Vetulani (Ed.), Human Language Technology. Challenges for Computer Science and Linguistics, Vol. 6562 of Lecture Notes in Computer Science, Springer, Berlin-Heidelberg, Germany, 2011, pp. 280–290.
https://doi.org/10.1007/978-3-642-20095-3_26

[38] R. Cowie, Describing the emotional states expressed in speech, in: ISCA Workshop on Speech and Emotion, Newcastle, uk, 2000, pp. 11–18.

[39] A. K. Syrdal, Y.-J. Kim, Dialog speech acts and prosody: Considerations for TTS, in: Proceedings of Speech Prosody, Campinas, Brazil, 2008, pp. 661–665.

[40] M. G. Core, J. F. Allen, Coding dialogs with the DAMSL annotation scheme, in: Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines, Cambridge, MA, USA, 1997, pp. 28–35.

[41] J. Allen, M. Core, Draft of DAMSL: Dialog act markup in several layers, WWW page, [online] (1997).

[42] D. Jurafsky, L. Shrilberg, D. Biasca, Switchboard-DAMSL labeling project coder's manual, Tech. Rep. 97–02, University of Colorado, Institute of Cognitive Science, Boulder, Colorado, USA (1997).

[43] S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, J. J. Quantz, Dialogue acts in VERBMOBIL, Tech. rep., German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany (1995).

[44] J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, M. Siegel, Dialogue acts in VERBMOBIL-2 - second edition, Tech. rep., German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany (1998).

[45] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Statist. Soc. Ser. B 39 (1) (1977) 1–38, with discussion.

[46] 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, Vol. 5246 of Lecture Notes in Artificial Intelligence, Springer, Berlin–Heidelberg, Germany, 2008, pp. 493–500. https://doi.org/10.1007/978-3-540-87391-4_63

[47] J. L. Fleiss, Measuring nominal scale agreement among many raters, Psychological Bulletin 76 (5) (1971) 378–382. https://doi.org/10.1037/h0031619

[48] J. L. Fleiss, J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, Educational and Psychological Measurement 33 (3) (1973) 613–619. https://doi.org/10.1177/001316447303300309

[49] J. A. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1) (1960) 37–46. https://doi.org/10.1177/001316446002000104

[50] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data., Biometrics 33 (1) (1977) 159–174. https://doi.org/10.2307/2529310

[51] D. Tihelka, J. Matoušek, Unit selection and its relation to symbolic prosody: a new approach, INTERSPEECH 2006 – ICSLP, proceedings of 9th International Conference on Spoken Language Procesing 1 (2006) 2042–2045.

[52] M. Grůber, Enumerating differences between various communicative functions for purposes of Czech expressive speech synthesis in limited domain, in: Proceedings of Interspeech, Portland, Oregon, USA, 2012, pp. 650–653.

[53] 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, IEEE, 345 E 47TH ST, NEW YORK, NY 10017, USA, 2011, pp. 267–272. https://doi.org/10.1109/isspit.2011.6151576

[54] L. Latacz, W. Mattheyses, W. Verhelst, Joint target and join cost weight training for unit selection synthesis, in: Proceedings of Interspeech, ISCA, Florence, Italy, 2011, pp. 321–324.

[55] X. L. F. Alias, Evolutionary weight tuning for unit selection based on diphone pairs, in: Proceedings of Eurospeech, Vol. 2, Geneve, Switzerland, 2003, pp. 1333–1336.

[56] Z. Hanzlíček, Czech HMM-based speech synthesis, in: Text, Speech and Dialogue, proceedings of the 13th International Conference TSD, Vol. 6231 of Lecture Notes in Computer Science, Springer, Berlin-Heidelberg, Germany, 2010, pp. 291–298. https://doi.org/10.1007/978-3-642-15760-8_37

[57] J. Nouza, J. Psutka, J. Uhlíř, Phonetic alphabet for speech recognition of czech, Radioengineering 6 (4) (1997) 16–20.

[58] J. Romportl, J. Matoušek, D. Tihelka, Advanced prosody modelling, in: Text, Speech and Dialogue, proceedings of the 7th International Conference TSD, Vol. 3206 of Lecture Notes in Artificial Intelligence, Springer, Berlin-Heidelberg, Germany, 2004, pp. 441–447. https://doi.org/10.1007/978-3-540-30120-2_56

[59] J. Yamagishi, K. Onishi, T. Masuko, T. Kobayashi, Modeling of various speaking styles and emotions for HMM-based speech synthesis, in: Proceedings of Eurospeech, Geneva, Switzerland, 2003, pp. 2461–2464.

[60] K. Miyanaga, T. Masuko, T. Kobayashi, A style control technique for HMM-based speech synthesis, in: Proceedings of Interspeech, 2004, pp. 1437–1440.

[61] T. Nose, Y. Kato, T. Kobayashi, A speaker adaptation technique for MRHSMM-based style control of synthetic speech, in: Proceedings of ICASSP, 2007, pp. 833–836. https://doi.org/10.1109/icassp.2007.367042

[62] M. Grůber, Z. Hanzlíček, Czech expressive speech synthesis in limited domain: Comparison of unit selection and HMM-based approaches, in: Text, Speech and Dialogue, Vol. 7499 of Lecture Notes in Computer Science, Springer, Berlin-Heidelberg, Germany, 2012, pp. 656–664. https://doi.org/10.1007/978-3-642-32790-2_80

[63] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. W. Black, The HMM-based speech synthesis system (HTS), [online].