Speech synthesis and emotions: a compromise between flexibility and believability

Enrico Zovato[1], **Jan Romportl**[2]

Abstract. The synthesis of emotional speech is still an open question. The principal issue is how to introduce expressivity without compromising the naturalness of the synthetic speech provided by the state-of-the-art technology. In this paper two concatenative synthesis systems are described and some approaches to address this topic are proposed. For example, considering the intrinsic expressivity of certain speech acts, by exploiting the correlation between affective states and communicative functions, has proven an effective solution. This implies a different approach in the design of the speech databases as well as in the labelling and selection of the "expressive" units. In fact, beyond phonetic and prosodic criteria, linguistic and pragmatic aspects should also be considered. The management of units of different type (neutral vs expressive) is also an important issue.

1 INTRODUCTION

In the last few years, affective computing has become a key aspect in many technological environments. In particular, the inclusion of emotions in computer interfaces like embodied conversational agents is considered a key point to significantly improve human-computer interactions. Moreover, in multimodal interfaces, the voice plays an important role for conveying emotions. For example, rhythm and intonation of the voice seem to be important features for the expression of emotions [1].

The synthesis of emotional speech has therefore to deal with these aspects. It should be capable to reproduce the acoustic patterns typical of certain expressive styles. Nonetheless many state-of-the-art synthesizers fall short of being able to generate the variability typical of human vocal expressivity [2]. It is therefore likely that for obtaining believable expressive synthetic speech, different strategies should be followed. In terms of naturalness, concatenative TTS systems can significantly benefit from working on the linguistic level beyond the acoustic one. This is particularly evident in limited domain speech synthesis systems in which the design of the corpus strongly depends on the application context. In the same way, interesting results are achieved by linking the affective states with the intrinsic acoustic characteristics of certain expressions that convey emotions.

This work reports about the research activities concerning speech synthesis and emotions, undertaken within the EU funded COMPANIONS project[3] [3]. In this project, two languages are considered: English and Czech and two synthesis (or Text-To-Speech) systems are exploited in the speech interface. The first one, used in the English prototypes, is provided by Loquendo and the second one is developed by the University of West Bohemia. Although the two systems are different in their architecture, they are both based on the concatenative technology that yields a satisfactory degree of naturalness at the cost of less flexibility in terms of prosodic and voice quality modifications. This is a limitation when the objective is the simulation of the variability typical of the human voice expressivity.

The parallel research activities undertaken by Loquendo and the University of West Bohemia are reported in the next sections. It is worth noticing that both research teams are trying to link the expression of emotions in synthetic speech with the domain specific pragmatic functions of certain speech acts.

2 EMOTIONAL STATES AND COMMUNICATIVE FUNCTIONS

Many methods of emotional (affective) state classification have been proposed. Very briefly and in simplicity – the basic distinction is whether a particular classification system is categorical, or dimensional. Among many we can name a categorical classification system by P. Ekman [4] which distinguishes emotional states such as *anger, excitement, disgust, fear, relief, sadness, satisfaction*, etc. In a dimensional model, emotions are defined as positions (or coordinates) in a multidimensional space where each dimension stands for one property of an emotional state. Various dimensions have been proposed out of which a widely accepted set is the one by J.A. Russell [5] with two axes: valence (positive vs. negative) and arousal (high vs. low activation). Other models also consider a third dimension that is power or dominance and some even a fourth dimension: unpredictability.

It is quite difficult to classify human speech according to one of these models (both categorical and dimensional) with the perspective of finding out acoustic correlates useful for generation purposes. Instead, we have settled for the assumption that a relevant affective state (of the conversational agent) goes implicitly together with a *communicative function* of a speech act (or utterance) which is more controllable than the affective state itself. It means that we do not need to think of modelling an emotion such as "guilt" per se – we expect it to be implicitly in an utterance like *"I am so sorry about that"* with a communicative function "affective apology".

Moreover, there are objective difficulties with the acquisition of affective speech data. We can basically divide affective states into three classes according to the way they have been produced:

1) *Spontaneous affective states* can be acquired from some recorded emotionally extreme situations, such as

[1] Loquendo SpA, Turin, Italy. Email: enrico.zovato@loquendo.com

[2] Dept. of Cybernetics, Faculty of Applied Sciences, Univ. of West Bohemia, Pilsen, Czech Republic. Email: rompi@kky.zcu.cz.

[3] This work has been partially supported by the EU's 6th framework project "COMPANIONS" (www.companions-project.org), IST 034434.

various TV competitions, therapeutic sessions, dangerous flight situations, customers' complaints, etc. Due to their very nature they are very difficult to retrieve and, most importantly, definitely not suitable for speech synthesis. However, they can serve as a good source of information for hypothesis formulation.

2) *Affective states in experimental conditions*. Many affective states can be stimulated in a "laboratory environment". Stress can be evoked by forcing the subject to solve difficult tasks in time pressure. Other emotions are evoked by audiovisual stimulations (emotively strong pictures or videos) or experimenteror's behaviour [1]. Such data are, however, again hardly usable as resources for speech synthesis.
3) *Simulated affective states* are actually the only option for speech synthesis data acquisition. They are based on the ability of a speaker to pretend various emotions. As [1] again notes, it has been shown that non-professionals usually are not able to simulate emotions well enough to achieve good results. It is often important to support even the professional actors by a good script and this poses further technical requirements on a speech corpus preparation.

Our decision to simulate the communicative functions during the corpus recording instead of simulating affective states is further based on the assumption that it is more feasible in our conditions to prepare scripts leading to various communicative functions rather than emotions.

In the parallel research undertaken by Loquendo and the University of West Bohemia we have decided to test two slightly different approaches (separately for the English and Czech synthesisers) which will be described in the following sections.

3 ENGLISH EXPRESSIVE SPEECH SYNTHESIS

The English TTS developed by Loquendo [6], is a multilingual system and it is mainly composed of a text analyzer and a speech synthesizer. The first part converts the input text into a stream of phonemes, each associated with a prosodic label and with values of duration and fundamental frequency (F0 or *pitch*). The second module actually converts this stream into signal samples, by searching the best fitting units in the speech database and by smoothly concatenating them.

As above mentioned, with this kind of technology, it is quite difficult to control acoustic variables like prosodic or spectral features. In fact, the achieved naturalness of the synthesized speech is the result of the recombination of segments of human voice, accurately selected from the speech database by means of a cost function. Prosodic variations are allowed, even if they should be set within certain ranges, otherwise distortion artifacts could occur. For example, they are exploited in the synthesis of interrogative phrases, in which a target intonation contour has to be reproduced. In order to synthesize expressive speech trough unit selection, huge databases should be considered, covering the the phonetic and acoustic domain. However, in many applications only a subset of expressive data, related to frequently used phrases with a pragmatic function, could be inserted into the system database and then appropriately labeled for the selection process.

In our approach, these expressive units (phrases) can be inserted in the speech flow, providing different styles depending on the selected items. In fact, these particular units differ from the neutral data, characterized by an almost standard reading style.

In the Loquendo TTS, we designed an inventory of discourse markers and speech acts that constitute contextually relevant part-of-speech that are combined with neutral speech [7]. At the basis of this approach is the design of 17 speech act categories (*Refuse, Approval , Disapproval , Recall , Announce , Request of Confirmation , Request of Information , Request of Action, Prohibition, Contrast, Disbelief, Surprise, Regret, Thanks, Greetings, Apologies, Compliments*). For each of them we have made a linguistic inquiry in order to select the most frequent expressions used in the English spoken language. These expressions, that yield a communicative and pragmatic task, are recorded with an expressive attitude with respect to the sentences recorded for the "neutral" databases.

3.1 Speech acts analysis

Within the Companions projects, we have analysed the transcripts of the English dialogues relative to the two application scenarios: a virtual companion that helps elderly people organize and browse photographs, and the companion that keeps track of the user's diet, monitors and suggests physical exercise. We have extracted the most frequent expressions that were adopted by the interviewers in the recorded dialogues. We have then classified them according to their pragmatic functions into six major categories (see Table 1), that are already available in the TTS system.

GREETINGS	*Good Morning!* *Hi!*
THANKS	*Thank you for talking to me!* *Many thanks!*
REQUEST OF ACTION	*Let's continue!* *Let's see another photo.*
REQUEST OF INFORMATION	*What's this?* *Please, tell me about this photo.*
APPROVAL	*That's very interesting!* *That's great!*
APOLOGIES	*Sorry.* *My apologies*

Table 1. Speech acts categories in the Companions data

We have also conducted an experiment in which the same selected expressions used by the agent in the Companions dialogues have been recorded. The speakers were asked to interpret the texts in a natural way, taking into account their semantics and context. In a second session, they recorded the same material adopting a neutral style (a 'synthesis' like style). From the analysis of the aggregated data, it is possible to notice that each category has its own specific prosodic characteristics. Moreover, all of them show significant distances from the neutral style (see Figure 1). This is a further proof that there is a link between the linguistic/pragmatic level and the acoustic one that justifies the adopted approach.

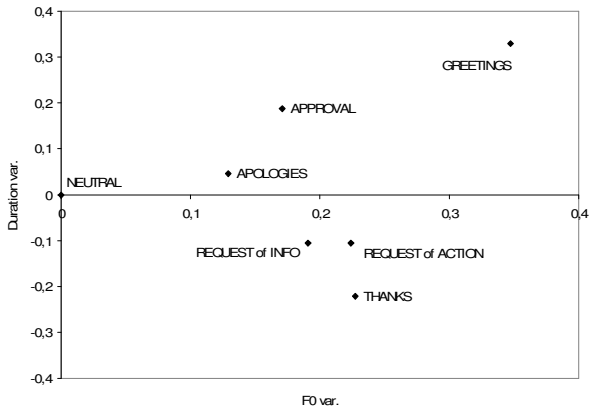

Figure 1. Scatter plot of Mean F0 variations vs duration variations for different speech act categories.

The expressive cues can be effectively used by the language generation module of the dialogue manager, once the list of available units is a priori known. In fact, depending on the message to be generated, the system could choose one of these expressions from a particular category set and then the TTS engine will select from its database the corresponding acoustic realization that intrinsically brings expressivity. Of course, this solution is limited to certain parts of the phrases that have to be generated, typically short initial expressions. The question is how to treat those parts of the phrases that do not constitute a speech act. For example in the sentence *"Sorry, I did not mean to interrupt"*, the first part (*Sorry*) could be synthesised by selecting the corresponding expressive phrase unit, while the remaining part should be synthesised with the standard neutral data.

3.2 Analysis of expressive data

The risk of introducing an unnatural contrast between the two types of speech segments is concrete, when adopting the above described solution. This is due to the different acoustic characteristics of the involved units.

To improve this aspect, we are investigating the possibility of slightly modifying the prosodic characteristics of the variable part of the generated phrases while trying to avoid significant reduction of quality and naturalness. However, we have initially decided to restrict the problem by identifying the typical acoustic patterns of two basic emotional styles with opposite valence: one with negative valence and low activation and the other with positive valence and possibly high activation. To this end a small speech database has been recorded in a quiet environment. The scripts were composed of sentences extracted from novels and classified into two categories: those conveying negative messages and those conveying positive messages. Up to date, we have recorded data from two American speakers (one male and one female) and one British English speaker (male). Also in this case, two versions of speech data were acquired, one yielding clear emotional intentions and the other in a neutral reading style. Analysis regarded prosodic features like F0 (min, average, max, range), intensity and duration. Further classifications have been made to account for the presence of lexical stress within the units, and the relative position within the sentence. In this way we could make some comparisons and study how the prosodic features vary, depending on one of the two styles.

This analysis provided clear correlations as concerns the intonation contour. The "negative" style is characterised by lower values of the average F0 and F0 range, while the "positive" one is characterised by a significant increase of the range of F0 with respect to the neutral style. Moreover, the length of the acoustic units tends to decrease in the first case. (see figure 2) .

These results can be used to reproduce the prosodic characteristics of these two basic emotional styles. In fact, clear information on how to modify pitch and speech rate can be assumed. Preliminary experiments regarding the manipulation of neutral data, according to these analysis results, have shown that this solution is feasible.

The same material was also used to study the phrasing strategies adopted by the speakers in these two expressive styles. In particular, we have analysed the phrase boundaries by observing the occurrences and positions of strong pauses as well as of prosodic pauses, i.e. those pauses associated with a significant change in the slope of the intonation contour. This analysis shows that in the "positive" style there is the tendency to increase the number of prosodic pauses, mainly after a long subject and before an objective complement. This is probably correlated with the increase in the speech rate.

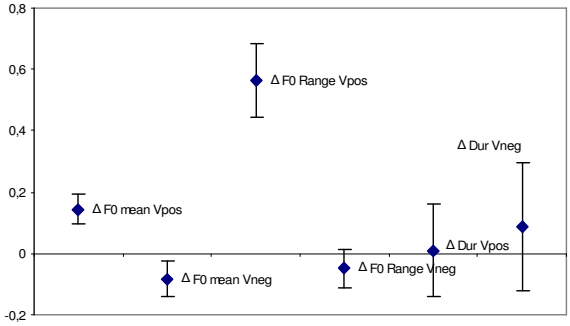

Figure 2. Average variations of prosodic parameters for two emotional styles with opposite valence.

4 CZECH EXPRESSIVE SPEECH SYNTHESIS

The Czech TTS developed by the University of West Bohemia is also based on the concatenative technique.

In this system, unit selection speech synthesis with emotions is based on the assumption that the presence of a reasonable number of sentences representing particular affective states in a source corpus of the synthesis allows the generation of emotionally rendered sentences during the synthesis process. This assumption is based rather on practical results achieved so far (e.g. [8]) than on rigorous theoretical analysis. The results are, nevertheless, encouraging at least for partially limited domain speech synthesis, we have therefore decided to acquire speech data of emotionally marked utterances with strong link to the intended domain of usage of the system (i.e. a conversational agent discussing photographs with seniors).

4.1 Classification of emotional states and communicative functions

As discussed in Section 2, it is very difficult and unconvincing to classify isolated sentences from a speech synthesis corpus according to either categorical or dimensional emotion models. Moreover, we have actually uncovered that such a thing is from the point of view of our goal pointless and rather introduces more complications (e.g. low inter-judgement agreement in manual classification of already recorded utterances, serious difficulties in the recording process where a speaker tries to mimic a given emotion, uncertainty whether a higher level dialogue manager of the conversational agent would be able to manage such states reasonably and whether they make sense in the given application domain, etc.). We have therefore given up our effort to label the corpus utterances in terms of the aforementioned emotions and instead of it we have decided to record new sentences with the most needed *communicative functions*. We have borrowed a basic list of communicative functions from [8] and we have further modified it for our target domain. The list is reported in Table 2. The communicative functions express only such features which are not represented in our framework of prosodic structures [9].

Imperative	
DIRECTIVE	*Please call the technical assistance.*
REPEAT	*Could you repeat that once more, please? I'm sorry, I didn't get it.*
REQUEST	*Let's go on to the next photo.*
WAIT	*Please wait a minute.*
WARNING	*The picture will be definitely deleted.*
Affective	
APOLOGY	*Sorry, I wasn't able to do that. I don't know.*
EXCLAMATION-NEGATIVE	*Oh! Oops!*
EXCLAMATION-POSITIVE	*Great! Got it! Wow! Fantastic!*
EMPATHY	*I'm so sorry to hear that.*
APPROVAL	*Hmm, that's very interesting!*
ENCOURAGEMENT	*Go on, it's very interesting!*
CHEER	*That was much fun! You've got great children!*
GOODBYE	*Good bye. Have a nice day.*
GREETINGS	*Hi, how are you?*
Other	
CONFIRMATION	*OK. Yes. I see!*
DISCONFIRMATION	*No, I don't see it.*
FILLED PAUSE	*Hmmm.*

Table 2. Communicative functions in the Czech corpus (examples are obviously literal translations or close English equivalents).

The corpus for our speech synthesis consists of a standard set of neutral sentences selected from newspaper texts. This set has been substantially extended by sentences acquired through statistical analysis of transcripts of the interviews with seniors carried out using the Wizard of Oz method [10]. In this way we have ensured the relevancy of the corpus towards the target domain. A communicative function is then manually assigned to each newly recorded sentence in our corpus and it can be later used as one parameter of the target cost during the unit selection process.

The speech corpus obtained by the WoZ method consists of 56 dialogs between seniors and our audio-visual TTS system – i.e. a 3D talking head [11]. The head has uttered 7681 sentence tokens out of which there are 3681 unique sentences.

Since Czech is a highly flective language, a significant decrease of the number of the unique sentences has been reached by their clustering using automatic lemmatisation. Each sentence cluster is then represented by such a sentence which is richest in terms of the phoneme variability. All these cluster representatives are manually annotated with the tags of the communicative functions from Table 2.

Due to various constraints (mainly the technical demands on affective speech recording, resulting from the facts described in Section 2) we have decided to set the limit of the newly recorded sentences to 1000. However, even after the lemmatisation their number was still significantly higher, hence the sentence selection algorithm similar to the one described in [12] has been utilised. This algorithm has selected such a sentence subset which best satisfied the criterion given by the diphone, word and communicative function coverage.

During the process of studio recording of the selected sentences the speaker is introduced into the communicative situations by seeing the discussed photographs and reading aloud whole pieces of conversation (from the talking head's side) while listening to the recorded seniors' replicas.

Post-processing of the recorded sentences must then involve together with precise phonetic annotation also the verification of the communicative function labels – the way in which a sentence has been actually pronounced does not necessarily implies the same communicative function that had been intended in the sentence pre-annotation.

4.2 Annotation of semantic accents

We assume that a speaker may emphasise any number of words by acoustic means to express their prominence in comparison with other words. The acoustic prominence of a word can deliver various kinds of information – from this point of view we can observe the acoustic prominence even on words where a phrase end is acoustically realised. However, such a prominence has a different function: we want to find words whose acoustical prominence has an emphasising function in terms of semantics or pragmatics. We call such a phenomenon a semantic accent.

We have organised extensive listening tests and in this way acquired 99 parallel annotations from different listeners of 250 corpus utterances. These annotations involve labels for subjective perception of phrase boundaries and semantic accents. A reference model of inter-subjective perception of these phenomena has been estimated by an EM algorithm [13].

The idea of the whole process is the following: prosodic phrase boundaries and semantic accents are manually designated in a reasonable sub-part of the whole real speech database so that there is agreement as high as possible among many independent listeners. The phrase boundaries and semantic accents (their model respectively) obtained in this way are considered to be the "real" ones in the sense of "objectiveness", no matter our subjective opinion. In the second, a machine classifier trained on these data can automatically extend the phrase boundary and

semantic accent designation to the rest of the speech database, without being "confused" by inconsistencies in training data subjectively annotated by a single person.

Our current research focuses on designing a machine classifier which is able to extend automatically the reference model labels from these 250 sentences (training data) to the whole corpus so that we can synthesize emphasized words by introducing one more component of the target cost – a flag whether a particular word is or is not a semantic accent.

In addition to this, from the results presented in [13] we can make one more important conclusion: it proves the hypothesis that there can be at the most one semantic accent in a phrase. The test participants had no prior information about expected or allowed numbers of the semantic accents in the phrases. Still the resulting statistically underlain deployment places no more than one semantic accent into any phrase, no matter whether the semantic accents have been assessed concurrently with the phrase boundaries or separately afterwards.

CONCLUSIONS

The synthesis of emotional speech is still a matter of research and experimentation. In fact, the state of the art Text-To-Speech technology has reached a satisfactory degree of naturalness but lacks in flexibility in terms of vocal expressivity. A realistic approach for synthesizing emotional speech within the Unit Selection technique is the topic of this work. In the two considered synthesis systems, the key idea is relying on the implicit affective states that can be conveyed by expressions that have a communicative function. Beyond their own peculiarities, two similar solutions for the treatment of expressive phrase units have been described. The results obtained so far show that the proposed solutions are effective and believable for domain specific applications.

Acknowledgements

This work has been partially supported by the EU's 6th framework project "COMPANIONS" (contract IST 034434) and partially by the Ministry of Education of the Czech Republic, project LC536.

REFERENCES

[1] Scherer, K.R., Vocal communication of emotion: A review of research paradigms, Speech Communication, n. 40, (2003).

[2] Schröder, M. (2001). Emotional Speech Synthesis: A Review, In Proceedings of EUROSPEECH 2001, pp. 561 – 564, Aalborg, (2001).

[3] http://www.companions-project.org

[4] Ekman, P., Basic Emotions. In T. Dalgleish and T. Power (Eds.) The Handbook of Cognition and Emotion, pp. 45 – 60. Sussex, U.K. (1999).

[5] Russell, J.A., A circumplex model of affect. In Journal of personality and social psychology, pp. 1161 – 1178 (1980).

[6] Balestri M., A. Pacchiotti., S. Quazza, P. Salza and S. Sandri, Choose the Best to Modify the Least: a New Generation Concatenative Synthesis System. Proceedings of EUROSPEECH 1999: 2291-2294, Budapest (1999).

[7] Zovato, E., Tini Brunozzi, F., Danieli, M., Interplay between pragmatic and acoustic level to embody expressive cues in a Text to Speech system, Proc.AISB 2008, Aberdeen, UK, (2008)

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

[9] Romportl, J., Kala, J. Prosody modelling in Czech text-to-speech synthesis . In Proceedings of the 6th ISCA Workshop on Speech Synthesis, pp. 200–205, Rheinische Friedrich-Wilhelms-Universität, Bonn (2007).

[10] Legát, M., Grůber, M., Ircing, P., Wizard of Oz data collection for the Czech senior Companion dialogue system. Submitted to the Fourth International Workshop on Human-Computer Conversation, Bellagio, Italy.

[11] Krňoul, Z., Železný, M., Realistic face animation for a Czech Talking Head. Lecture Notes in Artificial Intelligence, no. 3206, pp. 603–610 , Springer, Berlin, Heidelberg (2004).

[12] Matoušek, J., Romportl, J., On building phonetically and prosodically rich speech corpus for text-to-speech synthesis. Proceedings of the second IASTED international conference on Computational intelligence, pp. 442–447, ACTA Press, San Francisco (2006).

[13] Romportl, J., Prosodic phrases and semantic accents in speech corpus for Czech TTS synthesis. In Lecture Notes in Computer Science, Berlin, Heidelberg : Springer, in print (2008).