Application of Expressive TTS Synthesis in an Advanced ECA System

Jan Romportl[1], *Enrico Zovato*[2], *Raúl Santos*[3], *Pavel Ircing*[1], *José Relaño Gil*[3], *Morena Danieli*[2]

[1]Department of Cybernetics, University of West Bohemia, Pilsen, Czech Republic
[2]Loquendo, S.p.A., Turin, Italy
[3]Telefónica I+D, Madrid, Spain

[1]{rompi,ircing}@kky.zcu.cz, [2]{enrico.zovato,morena.danieli}@loquendo.com, [3]{e.rsai, joserg}@tid.es

Abstract

The research project COMPANIONS aims at developing an advanced embodied conversational agent (ECA). This ECA is used in two scenarios and two languages (English and Czech), and it requires a TTS system being able to generate very natural expressive and emotional speech output. This paper describes application issues of two such systems within the ECA, introduces approaches to expressive speech handling in unit selection methods, and discusses similarities and differences between these systems and approaches.

Index Terms: expressive speech synthesis, TTS, unit selection, embodied conversational agent, dialogue system

1. Introduction

Automatic spoken dialogue systems are currently in focus of many research teams across the world, mostly due to the fact that the technologies that are necessary for the development of a practically usable system have made considerable progress in the recent years. However, the performance of the "core" modules that would be required for a truly natural dialogue system still leaves a lot to be desired — this concerns not only natural language understanding (NLU), natural language generation (NLG), dialogue management (DM), automatic speech recognition (ASR), but also text-to-speech synthesis (TTS).

This is exactly the reason why not only the commercially available, but also state-of-the-art laboratory dialogue systems are nowadays still restricted to a limited domain, such as train timetable inquiries [1] or only a slightly more general tourist information service [2]. These systems can usually make use of commonly available TTS systems delivering intelligible and convenient neutral speech.

The research project COMPANIONS (www.companions-project.org) aims at developing a dialogue system that would go somewhat beyond those domain limitations and, even more importantly, allow the user to develop some "relationship" with the system. This ambitious goal will be achieved by the system's adaptability to the specific user needs and also by the ability to perceive and express emotions to a higher degree. For the sake of maximum emotional coverage, it was chosen that our system would be conceived as an embodied conversational agent (ECA) with high quality expressive and emotional speech and avatar visualisation able to use elements of emotional communication such as gesturing and facial expressions. Therefore, there are much higher requirements posed on a TTS system used in such an ECA.

The project investigates the aforementioned man-machine companionship on the basis of two scenarios: a) 'How was your day?' scenario; b) 'Senior Companion' scenario.

Since multilinguality is also a very important aspect of modern communication systems, the 'How was your day?' scenario is in English, whereas the 'Senior Companion' scenario is in Czech.

This paper discusses a role of TTS in such a dialogue system, introduces requirements on TTS synthesis in the 'How was your day?' and 'Senior Companion' scenarios, and presents implementation details of both Czech and English expressive TTS systems together with their integration in the whole framework of the ECA.

1.1. 'How was your day?' scenario

The scenario for the English prototype is based around the idea of a user who freely discusses his/hers working day at a typical office environment and the avatar providing comfort and advice in a natural 'social' dialogue situation. It is called 'How was your day?' (henceforth HWYD). In this scenario, the research challenges are mainly: a) to study the usage of longer utterances from both the user and the system to express complex opinions, b) to use emotions also at both sides, and c) to explore several advanced dialogue features, such as interruption handling and replanning.

To tackle (a), the system allows and even encourages the users to freely express their opinions. This often results in users providing elaborate inputs, often in the range of 50–100 words per turn. Along with this, we use emotion detectors (b) at acoustic (based on EmoVoice) and semantic levels, the output of which is fused with the speech recognition, so the dialogue management engine has the most accurate picture of the state of dialogue. These long utterances induce several side effects (c) that occur in natural dialogue, such as interruptions. When the system is providing the user with a long opinion, we can expect the user to disagree with it while it is being uttered. We handle these events by stopping the system and reassessing the dialogue status by discarding the remaining utterances or repairing them at some level so they fit in the new context. The internals of these, including the decision making process, are further explained by Crook et al. [3].

The user interface used for the ECA in this scenario is shown in Figure 1. For a video demonstration of this system in action, see http://www.companions-project.org/demonstrators/english/. The video does not show the final version of the ECA and TTS, but it helps to make a clearer idea of what the HWYD scenario means.

1.2. 'Senior Companion' scenario

The Czech scenario is aimed at research and development of a system which is able to conduct a natural dialogue with elderly

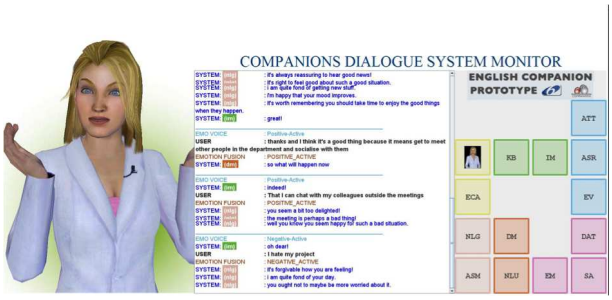

Figure 1: *User interface of the ECA in the 'How was your day?' scenario.*

users, mostly to keep them company and letting them to stay mentally active. Such a dialogue domain is still very broad to be reasonably tackled by a machine, and therefore it was decided to narrow the task further to the reminiscing about the family photographs. The 'Senior Companion' (henceforth SC) is thus supposed to create companionship with elderly users by discussing their lives over their family photo albums which help the dialogs keep going.

At least a reasonably large sample of data reflecting the nature of the dialogues that are likely to take place during the interaction with an automatic dialogue system is necessary, not only for DM training, but also to see what the emotional responses of the users are and what emotions and expressiveness should be rendered by the TTS system of the ECA. We have therefore employed the Wizard of Oz (WoZ) technique [4] to collect a corpus of 65 hours of sample dialogues between the seniors and a WoZ-controlled ECA.

This data acquisition was carried out in such a way that human subjects were placed in front of a computer screen and were told that the software they were interacting with was using "artificial intelligence" techniques to conduct a natural dialogue with them. In reality, ASR, NLU and response generation was simulated by a human operator (the "wizard"). Only the speech of the "artificial companion" was genuinely produced from the wizard's texts using a non-expressive TTS system [5] coupled with a 3D avatar head, which further reinforced the subjects' belief that they were truly interacting with a virtual person. More technical details can be found in [6]. Figure 2 shows the user interface for the ECA involved in the SC scenario.

Figure 2: *User interface of the ECA in the 'Senior Companion' scenario.*

2. Requirements on TTS in ECA

Both scenarios strongly rely on the ECA's ability to rightly express its emotional involvement in the user's situation as well as its compassionateness and emotional understanding of the user. It is very reasonable to suppose that there could be perfect DM and NLU modules in the dialogue system forming the ECA, but it would not be sufficient for the scenarios at all if the TTS output produced by the ECA did not show any natural expressive phenomena corresponding to the communicative and emotional function of the situation, or the suprasegmental features of the speech were in conflict with this situation, such as in case of a common news-reading-like synthetic voice.

This means that the ECA requires a TTS system to be able to produce naturally sounding synthetic speech in an adequate range of expressive and emotional styles. The adequateness here means that we do not need to synthesise absolutely all the voice-presented emotions which a human is capable of — only those relevant for the HWYD and SC scenarios are sufficient enough.

Apart from the expressiveness of the synthesised speech, there are other requirements which the TTS system must fulfil:

- Integration issues: the whole dialogue system consists of many specialised modules developed by different partners in the project. The TTS system, therefore, must be fully compatible with them, sharing their integration framework, so that it is possible to plug it into the system as another module.
- Response time: the dialogue system has an indispensable time overhead due to very complicated data processing at every level. In the same time, the dialogue naturalness and quality significantly decreases with higher latency of the system's responses. Therefore, it is very important to use any measure to minimise the time needed for the TTS synthesis, which is not a trivial task to accomplish in unit selection TTS systems searching through vast speech segment databases.
- Visualisation: synthesised speech is reproduced by a visualised avatar, thus the TTS system must handle all the related issues, such as lip and gesture syncing, phoneme to viseme conversion, etc. The TTS emotion parameterisation must also be compatible with the avatar's gesture and facial expression capabilities.

In previous paragraphs, we have discussed the importance of the emotionality in a TTS system for fulfilling the goals posed by the HWYD and SC scenarios. However, here we must present a note which is, quite surprisingly, contrary to the aforementioned statement. We have found out that at least in the SC scenario the actual quality, naturalness, emotionality and expressiveness of the synthetic speech play only a very little role in how the senior users get involved in the dialogue. We have analysed the WoZ recordings of the dialogues with the senior users (65 one-hour dialogues, 37 female users, 28 male users, the average age 69.3 years) where only a common news-reading-like Czech unit selection speech synthesis without any expressiveness was used, and we must conclude that the majority of the users got emotionally involved in the dialogue just as if they were really talking to a human [7], although the voice was clearly synthetic and emotionless (yet of a good quality).

The goal of creating a natural emotional speech synthesis, nevertheless, remains intact — only its motivation has shifted somehow more towards curiosity and challenging the limits of state-of-the-art TTS systems. However, there might be one risk

hidden in such an attitude: what if the users enjoyed the dialog mainly because they *felt* by all their senses that they were talking to a machine, and thus their experience with such a kind of machine was "balanced". In other words: adding high quality and natural speech emotions into an otherwise still somehow limited machine dialogue system may lead to the *uncanny valley* [8], derogating from the users' positive experience with the system.

3. English TTS

The speech synthesis system used in the English HWYD scenario is Loquendo TTS (henceforth LTTS) with some extensions developed to fulfil the requirements of the project.

3.1. System overview

LTTS is a multilingual speech synthesiser based on a unit selection synthesis technique, which provides high intelligibility and good acoustic characteristics [9].

The architecture of this system is very modular and is conceptually composed of a text analyser whose output feeds a waveform generation algorithm that selects the acoustic units from the speech database and then joins them together. LTTS is organised as a set of separate software modules that can be used jointly, depending on the application. The basic system is composed of an engine together with voice and language components.

Language specific text analysis includes the conversion of any input text into a suitable expanded graphemic form, a morphological analyser and a syntactic parser. Text analysis also detects the basic phrase structure of the input sentence by performing a rule-based syntactic chunking. The syntactic phrases (i.e. NP, PP, VP, and AdvP) are then processed on the basis of rhythmical rules in order to get a set of candidate syntactic-prosodic boundaries which are used to generate silent pauses of different durations [10].

English grapheme to phoneme conversion is based on machine learning techniques [11] which make use of large phonetic lexicons. Phonetic exceptions can, however, be managed through the use of user lexicons.

Given the conversion of the input text into a phonetic and prosodic target, waveform generation is based on the selection of the best matching sequence of acoustic units according to similarity and transition constraints. A concatenation module is used to join the speech units. It also smoothes down pitch and spectral discontinuities.

The TTS speech database of the voice used in the Companions prototype consists of a set of high quality audio samples, recorded by a female professional speaker whose mother tongue is British English.

3.2. Expressive TTS features

The emotional model adopted in the HWYD scenario considers two time constants associated with two distinct communication loops. The short term loop is aimed at reacting immediately to the user input. This reaction has to be immediate and consistent with the user's affective state (recognised by the emotion detectors mentioned in Section 1.1). We have therefore exploited the LTTS capability of producing small expressive prompts with specific communicative functions (i.e. expressive speech acts). This repertoire consists of a set of phrases widely used in human communication, especially with pragmatic intentions. These sentences, obtained through the linguistic analysis of different representative corpora, are classified into 17 categories: *Refuse*, *Approval*, *Disapproval*, *Recall*, *Announce*, *Request of Confirmation*, *Request of Information*, *Request of Action*, *Prohibition*, *Contrast*, *Disbelief*, *Surprise*, *Regret*, *Thanks*, *Greetings*, *Apologies* and *Compliments*.

The voice talent who recorded these prompts (about 500 for the English language) was asked to adopt the most suitable speaking style depending on the speech acts.

The hypothesis at the base of this solution is that the expression of emotions in human voice is often linked to semantic and pragmatic contexts, and it is likely that the speech acts involve acoustic modifications, lexical choices, and extra-linguistic phenomena, such as back-channelling sounds [12].

The selection mechanism applied to these units is different from the traditional one, although very simple. These expressive utterances are in fact treated as "phrase" units which cannot be selected partially and concatenated with standard neutral units, but are indivisible and have to be selected as a whole. They are stored in special databases and are assigned labels which beyond phonetic information include their text and typology. The selection is performed by matching the text, phonetics and punctuation. These expressive cues are used as short expressive verbal feedbacks, but could also be inserted in the synthesized message without compromising the naturalness and the fluency of the speech provided by the unit selection algorithm. The acoustic, prosodic and lexical structure of these phrases actually improves the expressiveness of the speech flow, therefore reinforcing the pragmatic intention of the message, that in the case of the HWYD scenario is empathising with the user according to his emotional status.

Also human sounds without any linguistic/semantic content, such as laughs, sighs, breaths, coughs, etc., have been considered for the interaction with the user in the emotional short loop. They can be used as short feedbacks or intros generated by the ECA depending on the context. For example, a throat sound could communicate embarrassment, distancing, request of attention, or it could be used as a back-channel sound. The inventory available in LTTS comprises different kinds of hesitations, interjections, coughs, sighs, bitter and hearty laughs.

In case of the the emotional long term loop, we have experimented with another feature which can be applied to any input text. It comprises synthesising speech in two different expressive styles according to the polarity of the emotional model of the application. This is achieved through prosodic and spectral processing aimed at obtaining speech with characteristics similar to the target patterns of two emotional speaking styles with the opposite valence. In particular, two presets were implemented within the HWYD scenario: a "negative" style (sad) and a "positive" style (happy or joyful).

These speaking styles are characterised by prototypical acoustic patterns. In the sad style, the values of the fundamental frequency (F0) are lower, and limited in range; speech rate tends to be slower. As for the happy speaking style, F0 is higher, with more variability, and speech rate is increased. The processing applied to the speech units is based on an analysis-synthesis scheme. In the analysis phase, acoustic and prosodic patterns are calculated, while during the synthesis phase the patterns modified according to the expressive targets are applied. Post-filtering is also executed so as to slightly change the spectral shape. A significant amount of off-line manual tuning was experimented to get the best configuration of the synthesis parameters and to avoid the introduction of signal processing artifacts.

4. Czech TTS

The speech synthesis used in the Czech SC scenario is based on the TTS system ARTIC, developed at the University of West Bohemia, Pilsen [5].

4.1. System overview

The system in its unit-selection version is structurally and technically very similar to LTTS. It employs the unit selection algorithm, but here without any signal modification of concatenated segments. Prosody is controlled using the diphone target-cost features derived from the prosodic structure of synthesised sentences, such as position in a prosodic phrase, prosodeme type, etc. [5]. Czech grapheme to phoneme conversion is rule-based because of very regular pronunciation in Czech.

The baseline version generates neutral speech (a female voice in case of the SC scenario) using a speech segment database created from a corpus of 10,000 neutrally recorded declarative sentences. To achieve the synthesised speech quality as high as possible, we use the whole corpus, which on the other hand increases time demands of the unit selection algorithm. As a result of this, the baseline version struggles with the real-time constraints, synthesising only about 1.5 times faster than the duration of the synthesised utterances. Since this is far too slow for the real-time dialogue system with the advanced ECA capabilities (cf. Section 2), we have introduced an enhancement of the unit selection algorithm which is able to speed up the unit selection process at least 30 times. It uses several stopping and pruning schemes in the Viterbi search algorithm and it is described by Tihelka et al. [13]. This way we are able to significantly decrease the latency in the ECA's responses, which is also supported by caching and reusing the synthesised clauses.

4.2. Expressive TTS features

Unlike the HWYD ECA, the SC ECA does not support any short loop at this stage of development, and therefore there are no databases of common expressive prompts in Czech. Instead of this, we have focused on expressive and emotional enhancement of longer ECA's turns in the long loop.

Since it is often difficult to classify human speech according to categorical or dimensional emotion models, we have settled for the assumption that a relevant affective state of the ECA 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 present in an utterance like "I am so sorry about that" with a communicative function "(affective) apology". This is very similar to the approach used in English LTTS, but here we have introduced a different set of communicative functions and we have used them in the long loop instead of the "positive"/"negative" emotion styles (cf. Section 3.2).

This approach, however, needs also a different speech data acquisition method than the one used with LTTS. We have selected all the sentences from the WoZ senior dialogue corpus uttered by the WoZ-controlled ECA, i.e. those sentences actually written by the WoZ operators and synthesised by the non-emotional TTS during the WoZ data acquisition (cf. Section 1.2), and we have used them as the source data for a new emotional TTS corpus recording scenario: the same female voice talent who had already recorded the speech corpus of 10,000 sentences in the neutral style, was listening to the recorded WoZ senior dialogues with omitted ECAs turns (this means she was listening only to the turns of the seniors) and her task was to put herself into the real dialogue and record her turns (i.e. the original WoZ-controlled ECAs turns) with proper expressiveness. This way we acquired the corpus of approximately 6,000 emotionally and expressively recorded sentences suitable for the SC scenario.

Table 1: *The communicative function inventory for the 'Senior Companion' ECA.*

Comm. function	Example
Directive	Tell me that.
Request	Lets 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	It's really terrible.
Happy empathy	It's nice. Great.
Showing interest	Can you tell me more about it?
Confirmation	Yes. Yeah. Well.
Disconfirmation	No. I don't understand.
Encouragement	Well. For example? And what about you?
Not specified	Do you hear me well?

After thorough analysis of this corpus we postulated a set of expected communicative functions relevant for the SC ECA (see Table 1) — this means that this set definitely does not comprise all possible communicative functions, but only those which were demonstrably present in the WoZ data. Every utterance in this corpus was then objectively annotated by a tag representing its real communicative function (considering both its meaning and the actual acoustic and prosodic forms). Such an objective annotation was achieved by large-scale parallel listening tests evaluated by a maximum likelihood model, following the same procedure as we have proposed for prosodic phrase annotation [14].

As a result, every diphone in the corpus received one more descriptive feature: a label representing the communicative function of the utterance the diphone appears in. This feature was also added to the unit selection algorithm of the baseline system — namely as one parameter of the target cost function. The synthesis process thus prioritises units with the same communicative function tag as the sentence being synthesised, which results in a significantly increased probability that the generated sentence is perceived as having the desired communicative function. Such an approach is indeed "domain-connected": it can synthesise any sentence (i.e. it is not "domain-specific") but sentences closer to the source domain are more likely to sound almost naturally in terms of affectiveness and voice quality. However, this limitation is definitely not a problem in a dialogue system such as the SC ECA because the NLG and DM modules hold complete control over the input of the TTS module, and therefore the domains of almost all sentences to be synthesised are known in advance.

5. Integration

Building modern multimodal spoken dialogue systems needs the integration of a large number of components, and therefore it often leads to a complex architecture. To minimise the impact of this in our work, the decision was taken to use a common underlying infrastructure. We used Inamode, a loosely coupled multi-hub framework which facilitates a loose, non-hierarchical connection between any number of components. The system features XML message passing, in which every component in

the system is connected to a repeating hub which broadcasts all messages sent to it to all connected components. The hub and the components connected to it form a single domain. Facilitators are used to forward messages between different domains according to filtering rules. A lightweight communication protocol is used to support components implemented in various programming languages. A common XML message "envelope" specifies the basic format of messages, including commonly used parameters such as routing, unique identifiers and turn numbers. The message has then a provision for a customised payload section, whose format is not dictated by the framework but left to define among the endpoints of the connection.

In addition to an underlying framework, both our prototypes share a number of components almost seamlessly, most significantly the avatar system and the module that adapts the graphical avatar rendering to the dialogue platform, known as the Multimodal Fission Manager (MFM henceforth). This has allowed us to speed up the development.

For both Czech and English prototypes we use the Haptek^{TM} commercial avatar engine, providing a human-like torso along with a low level API to control its movements and expressions. It is natively able to connect with SAPI-compliant TTS engines.

The MFM drives both the Avatar rendering and the TTS engine (see Figure 3), offering the possibility to construct complex communicative acts that chain series of utterances and gestures to better transmit the system outputs to the user. To provide a communication channel with the rest of the dialogue system we use a scheme based on a subset of the FML capabilities proposed by the SAIBA initiative. This methodology, which is adapted from the work described by Hernández et al. [15], allows us to discriminate between the actual textual surface forms that the system speaks (e.g., a sentence "That's good") and the communicative intent of them (e.g., irony or confirmation in "That's good"). This way we can provide an expressive, nuanced dialogue description, so that the emotional TTS modules can have proper input to select the details of their output.

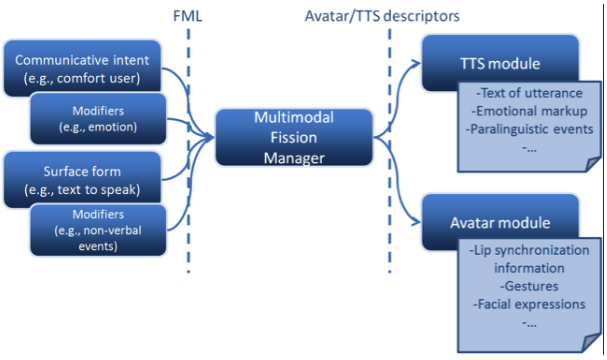

Figure 3: *Multimodal Fission Manager data flow.*

The system features also a template-based input mode, in which a module can call the ECA to perform actions without having to build a full FML-based XML message. This is intended to be used in feedback loops, such as nodding the head and saying 'That's good!' just when the user finishes saying a turn that has been detected as containing positive emotions by the acoustic analyser. In these actions we produce most of our content leveraging on the purely affective aspects of the communication because a proper semantic analysis and language generation would be far too slow for the ECA to be natural.

Due to the above, emotional TTS becomes essential in these feedback loops. In the subsections that follow we will explain the details of the integration of both emotional TTS subsystems.

5.1. Integration of English TTS

The integration of the TTS subsystem in the English HWYD prototype was technically more straightforward than in the Czech prototype because LTTS is fully SAPI compliant and the avatar natively supports visualisation and synchronisation of English articulation. Yet, since SAPI does not implement emotional TTS tagging out of the box, a number of specific solutions had to be implemented.

In particular, controls for the expressive "positive" and "negative" speaking styles had to be integrated in the SAPI layer of LTTS. The implemented solution allows for prompt switching among styles, once specific control tags are indicated in the input textual string:

```
\voice=Kate_positive I had an exciting working day.
\voice=Kate_negative I had problems in the office.
\voice=Kate I will complete the report by 6 p.m.
```

The last control tag is used to resume the standard neutral style. The SAPI interface was therefore modified in order to allow the use of these control tags and to be fully compatible with the MFM.

The mapping between the subset of FML affective labels considered in the Companions Project and facial expressions and body movements could also be extended to these speaking styles for better control of the avatar expressiveness.

On the contrary, the selection of expressive speech acts and human sounds did not require any changes in the SAPI layer and no particular tags were needed in the input text. In fact, selection is based on parsing of the input text. The TTS engine compares it, including punctuation, with the strings available in the system's inventory. Once matching conditions are encountered, the engine retrieves the corresponding expressive speech acts from the appropriate database. Punctuation plays an important role: in fact, when an exclamation mark is used in the input text, the corresponding expressive prompt, if available, is reproduced. For example, if the TTS input text is "*That's great news!*" and there is an expressive phrase unit in the database with the corresponding label, it is selected and the acoustic output is performed expressively. On the contrary, if the punctuation mark is different, the output is produced by selecting and concatenating units from the "neutral" speech database.

Regarding non-verbal human sounds, a more complex mechanism had to be developed in order to fulfil the avatar synchronisation requirements. These sounds, due to their non-verbal nature, are not phonetically labeled in the LTTS system, but this information is generally used to produce a temporal sequence of visemes that are used by the avatar to determine which movements have to be produced, such as mouth opening, height, width and protrusion. In order to overcome this limitation, all these audio events available in the database have been assigned a particular pseudo-phonetic labelling. Audio samples were then manually annotated by inserting labels and time markers so as to get the best approximation in terms of viseme generation. A specific syntax and a lexicon were introduced to handle the selection of these units. This solution was necessary in order to select these units from the appropriate database and not, for example, from the neutral speech database. Vice-versa, it avoids the selection of these non-verbal units during the synthesis of verbal messages.

In the TTS version implemented in the HWYD scenario, human sounds are activated by following a specific syntax. For example the string extra_Laugh02! activates the production of the corresponding laugh sound. A scheme of the different selection strategies for expressive and neutral speech is shown in Figure 4.

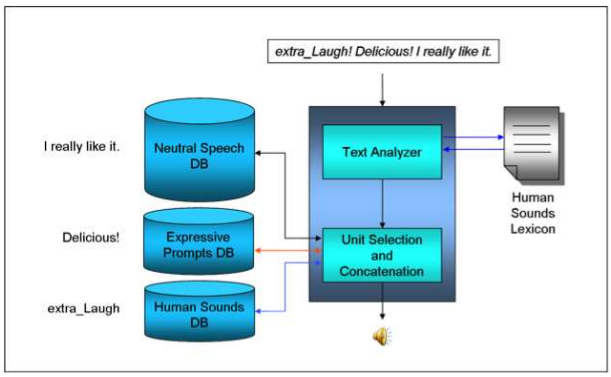

Figure 4: *An example of different selections depending on the input text formalism.*

5.2. Integration of Czech TTS

The Czech TTS system ARTIC is connected to Inamode in the same fashion as LTTS; however, its integration asked for several extra issues to be solved. These issues are determined by the following facts: a) the Czech SC ECA utilises a different set of communicative functions than the English HWYD ECA and the shared MFM (cf. Sections 3.2 and 4.2); b) the shared avatar does not support Czech natively; c) this version of ARTIC is not fully SAPI compliant.

In order to cope with (a), we designed an intermediate proxy module between the MFM, TTS and avatar modules. This proxy module accepts incoming messages from the MFM and on their basis it coordinates parallel activities of the TTS and avatar modules. Moreover, it comprises its own rule-based method for translation of the MFM's native communicative functions into those from Table 1.

The issues (b) and (c) were tackled in the TTS module itself: ARTIC comprises an algorithm for approximation of Czech phonemes into Spanish visemes (the avatar module supports English and Spanish). Every synthesised utterance is converted into the Ogg Vorbis format which supports an accompanying textual track and this textual track is filled by time notifications of Spanish visemes corresponding to the Czech phonemes of the synthesised utterance. The resulting Ogg Vorbis file is then streamed (using Inamode) to the avatar module which plays its acoustic content and performs lip syncing according to the textual content.

6. Conclusions

The integration of emotional TTS synthesis in real-world systems is a very difficult task to accomplish. Human perception is extremely sensitive about such high level language and cognitive phenomena, and therefore users are very discriminating when judging emotional speech synthesis. The described application of two emotional TTS systems in the Companion ECA proved successful in informal tests and preliminary subjective evaluations during the course of the project that are yet to be published.

The system was built on the premise that emotional speech synthesis is more acceptable to users when integrated in an advanced ECA. A systematic and objective real user evaluation of the system including the emotional responses that addresses the validity of this assumption remains a complex problem and will be the main focus of the project from now on.

7. Acknowledgements

This work was funded by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

8. References

[1] F. Jurčíček, J. Švec, and L. Müller, "Extension of HVS semantic parser by allowing left right branching," in *IEEE International conference on acoustics, speech, and signal processing*, Las Vegas, USA, 2008.

[2] S. Young, M. Gasic, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu, "The hidden information state model: A practical framework for POMDP-based spoken dialogue management," *Computer Speech & Language*, vol. 24, no. 2, pp. 150–174, 2010.

[3] N. Crook, C. Smith, M. Cavazza, S. Pulman, R. Moore, and J. Boye, "Handling user interruptions in an embodied conversational agent," in *Proceedings of AAMAS*, 2010.

[4] S. Whittaker, M. Walker, and J. Moore, "Fish or fowl: A Wizard of Oz evaluation of dialogue strategies in the restaurant domain," in *Proceedings of LREC 2002*, 2002.

[5] D. Tihelka and J. Matoušek, "Unit selection and its relation to symbolic prosody: a new approach," in *Proceedings of Interspeech 2006*, vol. 1. Bonn: ISCA, 2006, pp. 2042–2045.

[6] M. Grůber, M. Legát, P. Ircing, J. Romportl, and J. Psutka, "Czech senior COMPANION: Wizard of Oz data collection and expressive speech corpus recording," in *Proceedings of LTC 09*, Poznan, Poland, 2009, pp. 266–269.

[7] M. Otradovcová, "Multimodální komunikace člověk-stroj (Multimodal man-machine communication)," Master's thesis, University of West Bohemia, Pilsen, 2010.

[8] M. Mori, "On the uncanny valley," in *Proceedings of the Humanoids-2005 workshop: Views of the Uncanny Valley*, Tsukuba, Japan, 2005.

[9] S. Quazza, L. Donetti, L. Moisa, and P. L. Salza, "ACTOR®: a multilingual unit-selection speech synthesis system," in *Proceedings of SSW4*, Pitlochry, UK, 2001.

[10] E. Selkirk, "The interaction of constraints on prosodic phrasing," in *Prosody, theory and experiment: studies presented to Gösta Bruce*, M. Home, Ed. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2000, pp. 231–261.

[11] F. Mana, P. Massimino, and A. Pacchiotti, "Using machine learning techniques for grapheme to phoneme transcription," in *Proceedings of Eurospeech 2001*, Aalborg, Denmark, 2001.

[12] E. Zovato, F. Tini-Brunozzi, and M. Danieli, "Interplay between pragmatic and acoustic level to embody expressive cues in a text to speech system," in *Proceedings of AISB 2008*, Aberdeen, UK, 2008.

[13] D. Tihelka, J. Kala, and J. Matoušek, "Enhancements of Viterbi search for fast unit selection synthesis," in *Proceedings of Interspeech 2010*, Makuhari, Japan, 2010.

[14] J. Romportl, "On the objectivity of prosodic phrases," *The Phonetician*, vol. 96, pp. 7–19, 2010.

[15] A. Hernández, B. López, D. Pardo, R. Santos, L. Hernández, J. R. Gil, and M. C. Rodríguez, "Modular definition of multimodal ECA communication acts to improve dialogue robustness and depth of intention," in *The First Functional Markup Language Workshop Workshop at AAMAS 2008*, 2008.