What voice do we expect from a synthetic character?

João Cabral, Luís Oliveira, Guilherme Raimundo, Ana Paiva

INESC-ID/IST
Rua Alves Redol 9, 1000-029 Lisbon, Portugal
{jpcabral,lco}@l2f.inesc-id.pt,{guilherme.raimundo,ana.paiva}@gaips.inesc-id.pt
http://www.inesc-id.pt

Abstract

The emerging new applications of synthetic characters, as a way to achieve more natural interactions, puts new demands on the synthetic voices, in order to fulfill the expectations of the user. The work presented in this paper evaluates a synthetic voice used by a synthetic character in a storytelling situation. To allow for a better comparison, a real actor was filmed telling a children's story. The pitch, duration and energy of the recorded speech were copied to the synthetic speech generated with a FESTIVAL-based LPC-diphone synthesizer. At the same time, the synthetic character was also animated with the gestures, emotions and facial expressions used by the actor.

Using different conditions combining the synthetic voice, synthetic character with the real voice, and the real character, the voice was evaluated regarding the comprehension of the storyteller, the expression of emotions, its credibility and the user satisfaction.

1. Introduction

Several storytelling applications have been developed, e.g. [1], [2] and [3], but very little attention has been paid to the impact of a virtual storyteller when compared with a human. This work intends to study a multimodal system in which the embodied agent acts a story as if it was a person on the stage. It must use the adequate gestures and speech intonation so that the user understands the narrative, perceives the emotional meaning, considers it is credible, and feels satisfaction when interacting with the system [4]. To evaluate the acceptance by a person of such a system, we filmed an actor telling a story and then we tried to reproduce the movie with a synthetic character. In this paper we only report the results achieved in the performance of the synthetic voice.

In storytelling situations, speech is typically rich in emotional content and variability. For this reason, and to understand the differences and subtleties needed for the generation of synthetic speech for storytelling, we copied the rhythm and the intonation from the real speech to the synthetic speech. These perceptual effects are important in the expression of emotions, although there are other relevant aspects that were not considered such as voice quality [5], [6]. The conducted experiment allowed us to conclude about the perception of emotions from the synthetic voice in this particular context. The interaction of the storyteller with the voice as well as the quality of the synthetic voice were also investigated.

This paper is organized as follows. Section 2 describes the filming of the story with a real actor and the preparation of the speech corpus to obtain the orthographic transcription and the phonetic labels of the original speech. These transcriptions were necessary for the speech synthesis and to permit the synchronization of speech with facial movements. Section 3 explains how the synthetic speech was generated with the same pitch, duration and energy as the original. Section 4 describes the evaluation carried out. The results are presented in section 5 and discussed in section 6. Finally, Section 7 summarizes the obtained results and anticipates the future work.

2. Corpus

The corpus consisted of the video and audio recordings of an actor telling a children's story. The gestures and emotions of the facial expression were annotated and the prosodic parameters of the speech were analyzed to be used in the expression of the synthetic agent. This section only describes the preparation of the speech corpus.

2.1. Filming of the story

A male speaker with experience of amateur acting was asked to interpret freely a children's story. The actor spoke European Portuguese, his native language. The story was approximately 8 minutes long. The actor performed in the stage of a room with good acoustic conditions and he used an adequate headset microphone.

2.2. Corpus preparation

The recorded audio signal was down-sampled to 16 kHz and divided into short segments to facilitate the signal processing.

In order to command the facial movements of the synthetic character synchronously with the acoustic realization of the actor it was necessary to phonetically annotate the original speech segments. This annotation was processed semi-automatically. Since the actor did not have to follow strictly the script, after the filming it was necessary to produce the orthographic transcription of the story as told by the actor. From the text, the necessary analysis levels were performed automatically to obtain the phone sequence. Then, this sequence was aligned automatically with the original speech signal using a DTW-based phonetic aligner [7],[8]. Finally, the resulting phonetic labels and phone boundaries were manually verified and corrected when necessary. The amount of manual correction was higher than expected because the actor used a falsetto voice in the filming while the automatic alignment system was trained and adapted for neutral voices. Another consequence of the use of such voice quality was that the original speech signal presented a significant higher mean F_0 than the usual for a male speaker. Later, it will be explained how this affected the quality of the synthesized speech.

3. Generation of the synthetic speech

For the speech synthesis it was also necessary to guarantee the synchronism between the actor's video and the synthetic speech. Synthesis was performed by imposing the original's phone durations. In addition to the duration parameter, the original pitch and energy contours were also imposed on the synthetic speech to produce the same prosodic expression as the original.

Figure 1 shows the schematic procedure to generate the synthetic speech. The text-to-speech synthesis was performed using the Portuguese LPC diphone-based FESTIVAL synthesizer with male voice, developed at INESC-ID. In this process the diphone-units were concatenated without modification of the original pitch so that the resulting signals were approximately monotone with frequency 130 Hz. The pitch and energy of the synthetic signals were then modified using the Pitch-Synchronous Time-Scaling (PSTS) method [9] according to the prosodic parameters computed from the recorded speech. Since the actor used falsetto voice quality, the mean F_0 of the speech corpus was significantly higher than the voice used by the speech synthesizer. Consequently, the required pitch-scale factors were often higher than 1.5 which typically produces audible distortion. To avoid this effect, the target pitch contour was adjusted by lowering its mean value by 50 Hz and decreasing the pitch range in 40% so that the minimum pitch frequency was acceptable. Although we could use an unit-selection speech synthesizer, a diphone synthesizer was used because this permits more flexibility in speech generation.

Based on informal experiments and the authors' quality judgements, our approach generated speech with higher quality than using only the FESTIVAL synthesizer to produce the same expressive speech. This can be explained by the fact that the FESTIVAL system generates speech by performing the necessary prosody transformations on the speech units before concatenating them. This introduces higher acoustic discontinuities when joining the diphone units which results in speech quality degradation. In our approach, we used the FESTIVAL framework to concatenate speech units with constant F_0 reducing the pitch mismatches in the concatenation points and performed the prosodic transformations to the synthesized speech segments. This was only possible because we did not need to produce speech in real-time. Another factor was that the PSTS method performs better for large pitch-scale transformations than the PSOLA method which is used by the FESTIVAL text-to-speech system. The PSTS technique also permits to transform glottal source parameters related to voice quality which are relevant for production of speech with emotions. The transformation of these parameters to simulate emotion in speech is a current topic of research [10] so the emotional content on the synthetic speech was only expressed by variations in the intonation and rhythm.

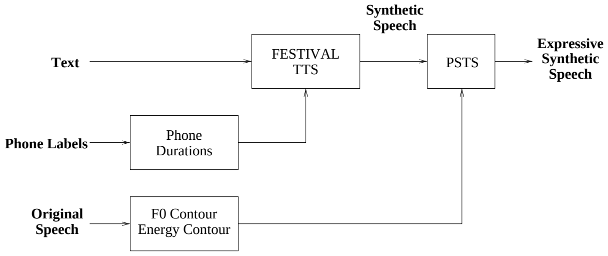

Figure 1: Schematic representation of the speech generation framework.

4. Perceptual experiment

A perceptual test was conducted to evaluate the performance of the storytelling character when compared with the human actor.

4.1. Stimuli

For the experiment four videos sequences were produced, corresponding to the situations in which the character telling the story was either real or virtual and the voice was either the original or synthetic.

These combinations are represented by the four different groups in Table 1. Figures 2 and 3 show images from the movies with the filmed actor and the 3D synthetic character, respectively. The virtual character represents a mix of the "old man" figure typical in storytelling and the famous characters "The Tweenies" who participate in a television programme aimed for children, broadcasted on a BBC channel.

		Character	
		Real	Synthetic
Voice	Real	CRVR	CRVS
	Synthetic	CRVS	CSVS

Table 1: The four types of movies used in the perceptual experiment.

Figure 2: Actor telling the story.

4.2. Subjects

The subject's panel was composed of 108 students from the Instituto Superior Técnico (Technical University of Lisbon), 89 were male and 19 were female, with ages ranging from 18 to 28. The participants did not have previous knowledge about the test.

Figure 3: Synthetic character telling the story.

4.3. Experiment

Subjects were distributed by four groups corresponding to the four types of videos. Thus, each group was composed of 27 elements. Subjects watched the video associated with their group and then they had to answer a questionnaire to classify the effectiveness of the gestures, facial expression and voice used by the character. These dimensions were evaluated in terms of the contribution to the story understanding, emotional expression, credibility of the character and user satisfaction. The questionnaire contained twelve different statements that the subjects had to classify from 1 to 7 according to their degree of agreement. The following translated sentences were used to classify the speech:

- "I understood everything the character told";
- "The voice of the character expressed the appropriate emotions";
- "The voice of the character was credible";
- "I liked the voice";

5. Results

As we said earlier, this section reports only speech-related results. The study of the gestures and facial expressions will be reported elsewhere. Tables 2 to 5 present the results of the classifications for the three statements related to the voice in the questionnaire. They show the rates of the negative values (disagreement), neutral values (neither agreement or disagreement) and the positive values (agreement). The percentage values were calculated for the four videos CRVR, CSVR, CRVS, and CSVS.

6. Discussion of the results

In general, the values obtained with different characters but the same speech signal are very similar, indicating that the type of character has little influence on the evaluation of the different characteristics of the voice. When we compare the classifications between the videos with the same type of character but different voices, there is a clear disadvantage in using the synthetic voice for all cases.

"I liked the voice"

	Video			
	CRVR	CSVR	CRVS	CSVS
Negative	0	7,4	59,3	59,3
Neutral	18,5	18,5	14,8	11,1
Positive	81,5	74,1	**25,9**	**29,6**

Table 2: Classifications for the satisfaction of the user with the voice, in percentage.

"I understood everything the character told"

	Video			
	CRVR	CSVR	CRVS	CSVS
Negative	0	7,4	18,5	33,3
Neutral	3,7	0	3,7	3,7
Positive	96,3	92,6	**77,8**	**63**

Table 3: Classifications for the intelligibility of speech, in percentage.

"The voice of the character expressed the appropriate emotion"

	Video			
	CRVR	CSVR	CRVS	CSVS
Negative	0	3,7	22,2	11,1
Neutral	7,4	0	7,4	18,5
Positive	92,6	96,3	**70,4**	**70,4**

Table 4: Classifications for the emotional content of speech, in percentage.

"The voice of the character was credible"

	Video			
	CRVR	CSVR	CRVS	CSVS
Negative	0	0	40,7	**33,3**
Neutral	7,4	14,8	**14,8**	3,7
Positive	92,6	**85,2**	**44,5**	63

Table 5: Classifications for the credibility of the voice, in percentage.

Table 2 shows clearly that most of the subjects preferred the original to the synthetic. This was most probably due to the fact that speech produced using LPC-diphone synthesizers typically sounds artificial. We could use an unit selection speech synthesizer to obtain a more natural sounding speech but it would not allow us to have such a flexibility in matching the actors performance.

A considerably high percentage of subjects understood everything the storytellers with synthetic voice told, which reflects the good intelligibility characteristic of LPC speech synthesizers. In the case of the storytellers with human voice that result was clearly better, which was expected given the lack of naturalness in the synthetic speech. In subjective tests like this, most subjects cannot separate naturalness from understanding. It is also interesting to note that in Table 3 there were more subjects who understood the story told by the real character with synthetic voice (CRVS) than by the synthetic character (CSVS). We think this may be related with differences in lip movements by the synthetic character.

According to Table 4, most subjects considered that the synthetic voice expressed appropriate emotions (positive rate of about 70% for CRVS and CSVS). This confirms that the variations in rhythm and intonation are an important factor for conveying emotions. However, the performance achieved by the synthetic voice is far from the rates obtained by the video sequences with the original voice (CRVR and CSVR). In our view, one of the reasons for this difference is that there are other acoustic cues for emotion that were not taken into account. It is expected that imposing more speech parameters related to voice quality on the synthetic speech from the original speech signal the positive rates would increase. The rates obtained for the real and synthetic characters for a given type of voice are very similar, which indicates that the body expression of the character had very little influence in the perceived emotional content of the voice.

Regarding the credibility of the voice, there are some intriguing results. From Table 5, the rates of positive and neutral responses for the credibility of a real character talking with synthetic voice (CRVS) were higher than expected because it is obvious that an actor cannot have that voice. Another interesting result is that, in general, the synthetic character talking with human voice (CSVR) was more credible than the synthetic character talking with synthetic voice (CSVS). It seems that a user finds more natural that a synthetic character speaks with a human voice even if it does not have the appearance of a person. We think that this is due to a long exposure to animated characters that use recorded speech in day-to-day events such as in movies, children's cartoons, toys, computer games, etc.

Finally, the falsetto voice quality used by the actor could have negatively influenced the results obtained for the characters with synthetic voice. Since it resulted in a higher F_0 than the voice of the speech synthesizer, the mean F_0 and F_0 range were scaled with factors obtained empirically. Thus, we think that the quality of the synthetic speech would be higher if it was not necessary to adapt the pitch of the original speech to the voice of the synthesizer.

7. Conclusion and outlook

This work reports the results of an experiment to evaluate the performance of a synthetic character telling a story when compared with an actor telling the same story.

The speech synthesis framework was based in a LPC diphone synthesizer and used a pitch-synchronous time-scaling method to avoid distortion caused by the pitch-scale transformations. The pitch, duration and energy were imposed in the synthetic speech to produce the same rhythm an intonation as the real speech.

The results of the perceptual test showed that the voice was significantly relevant for the good understanding of the story, expression of appropriate emotional content, credibility of the voice, and satisfaction with the voice. Although the synthetic voice only obtained a negative classification for satisfaction, it was clearly outperformed by the real voice in all aspects. In general, the gestures and facial expressions presented little influence on the classifications of the voice.

The participants in the test expected more appropriate emotional content from the synthetic voice. This suggests that it is necessary to consider more speech parameters related to emotions such as voice quality.

The synthetic character was significantly more credible when it spoke with a real voice than with a synthetic voice. It seems that the figurative representation of the character is not important for its credibility: it seems to be enough to have a synchronized movement of the lips to induce the idea that the speech signal was produced by human vocal tract.

Research on the speech quality and stylization of the synthetic voice that meet the expectations of the user are certainly good topics for future work.

8. Acknowledgements

This work was partially funded by the Portuguese Foundation for Science and Technology (FCT/FEDER/POSC).

9. References

[1] Theune, M., Faas, S., Nijholt, A. and Heylen, D., “The Virtual Storyteller”, in ACM SIGGROUP Bulletin, Vol. 23, Issue 2, ACM Press, pp. 20–21, 2002.

[2] Schell, J., *Understanding Entertainment: Story And Gameplay Are One*, in the Human-Computer Interaction Handbook, J.A. Jacko and A. Sears (Eds.), Lawrance Erlbaum Associates, 2003.

[3] Swartout, W. and Lent, M., “Making a Game of System Design”, in Communications of the ACM, 46(7):pp.32–39, 2003.

[4] Sawyer, R., *The way of the Storyteller*, New York: Penguin Books, 1942.

[5] Cowie, R., Douglas-Cowie, Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J., “Emotion Recognition in Human-Computer Interaction”, in Proc. of the IEEE Signal Processing Magazine, 18(1), pp. 32-80, 2001.

[6] Cahn, J. E., *Generating Expression in Synthesized Speech*, Master's Thesis, MIT, 1989.

[7] Paulo, S. and Oliveira, L., “Improving the Accuracy of the Speech Synthesis Based Phonetic Alignement Using Multiple Acoustic Feautures”, Computational Processing of the Portuguese Language - Proc. of the 6th Intl. Workshop, PROPOR 2003, Faro, Portugal, pp. 31–39, June 2003.

[8] Paulo, S. and Oliveira, L., “DTW-based Phonetic Alignment Using Multiple Acoustic Features”, in Proc. of the EUROSPEECH'2003 - 8th European Conference on Speech Communication and Technology, Genéve, Switzerland, 2003.

[9] Cabral, J. P. and Oliveira, L., “Pitch-Synchronous Time-Scaling for Prosodic and Voice Quality Transformations”, in Proc. of the Eurospeech-Interspeech'2005, Lisbon, Portugal, 2005.

[10] Cabral, J. P., *Transforming Prosody and Voice Quality to Generate Emotions in Speech*, MSc Thesis, Instituto Superior Técnico, IST/INESC-ID, Lisbon, Portugal, 2006.