ECIRCUS: Building Voices for Autonomous Speaking Agents

Christian Weiss, Luis C. Oliveira, Sergio Paulo, Carlos Mendes, Luis Figueira [1], *Marco Vala, Pedro Sequeira, Ana Paiva*[2], *Thurid Vogt, Elisabeth Andre*[3]

[1] INESC-ID/IST, Spoken Language Systems Laboratory, Lisbon, Portugal
[2] INESC-ID/IST, GAIPS, Lisbon, Portugal
[3] Institute of Computer Science, University of Augsburg, Germany
{christian.weiss, lco}@l2f.inesc-id.pt

Abstract

This paper describes our work integrating automatic speech generation into a virtual environment where autonomous agents are enabled to interact by natural spoken language. The application intents to address bullying problems for children aged 9-12 in the UK and Germany by presenting improvised dramas and by asking the user to act as an "invisible friend" of the victimised character. As we are addressing an elementary school environment one specification of the resulting voice was building age-corresponding young school kids voices. The second specification addresses building a low-resource speech generation system which is capable to run on older school computers but is still fast enough in response time to guaranty a fluent conversation between the agents. Third requirement was integrating the speech-module with the agents. We focus on the speech generation system itself, pointing out possible implementation issues in building non-controlled speech interaction in virtual environments Furthermore we describe the problems arising in building unit-selection based child's' voice TTS and shows alternative methods to child's voice recording by deploying voice transformation methods.

Index Terms: Social learning and education, speech synthesis, spoken interaction

1. Introduction

Virtual animated characters in dramatized scenarios are no longer used only on computer games. Learning and educative environments can benefit from the ever growing familiarity of users with virtual environments.

The eCircus (Education through Characters with Interactive Role-playing Capabilities that Understand Social interaction) project is an ongoing interdisciplinary EU-project focusing on novel conceptual models and innovative technology to support social and emotional learning through role-play and affective engagement for Personal and Social Education. Main aspects are to create a virtual environment for emotional and social learning focusing on the domains of bullying and refugee integration in school [1]. This paper describes our work in integrating an automatic speech generation module into the first showcase of the technology developed in this project, a virtual learning environment on anti-bullying education, called FearNot!. In this application autonomous agents need to communicate with each other in a away understandable by the user. The inter-agent communication using speech acts is converted into either English or German by a language generator engine that is translated into speech using a speech synthesizer. Figure 1 shows a screenshot of a bullying scenario in FearNot!. Although the 3D animated synthetic characters are cartoon like figures, our previous work showed that the users expect them to have naturally sounding voices [2].

As we are addressing an elementary school environment with students at the age between 9 and 12 years old, one specification of the generated voice was building an age-corresponding young school kids voice. The second specification addresses building a low-resource speech generation system which is capable to run on older school computers but is still fast enough in response time to guaranty a fluent conversation between the agents and the user. Third requirement was including audio-visual synchronization with the agents' actions.

This paper is organized as follows. In section 2 we address the problems arising while building a unit-selection based child voice and point out the difficulties and show our solution. In section 3 we describe our implementation of the voice building software and focus on the integration of the various modules usually needed by speech synthesis systems. The next section describes the experiment that was conducted to evaluate the system and its results. The final section presents the conclusions and the planned future work.

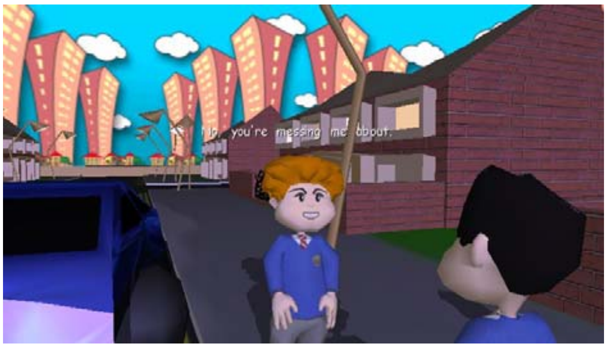

Figure 1: Screenshot of a FearNot! scenario

2. Child Voices

When trying to produce voices for child like characters the first approach that comes to mind is to record real children voices. We started by recording a set of 100 English sentences by a 9 year old girl and a boy of age 10. Although these recordings were very useful for our analysis of the acoustics of children's speech it soon became obvious that the recording of a larger set of sentences would be impractical. Children require shorter recording sessions and at slower pace than an adult speaker. It is also more difficult to assure the same speaking style among recording sessions since it depends on the child mood in that specific day. Given this difficulties it was decided to record carefully selected adults and modify their voices to make them sound as children's voices. To select the voice talents and to understand what type of modifications were required, we analysed our own recordings (table 1) and in general confirmed the results published in [3].

Table 1: Parameters from our own recordings.

Boy avg. F0 (Hz)	Boy avg. Formant values	Girl avg. F0 (Hz)	Girl avg. Formant values (Hz)
270	570 1400 2700 3900	280	570 1800 3000 4100

The main characteristic that distinguishes children's voices from adult voices results from the smaller size of their vocal tract. This results in higher pitched voices due to shorter vocal folds and in the scaling of the formants as a result of a shorter vocal tract.

The most significant changes in f0 occur for male speaker from age 12 to 15 resulting in f0 dropping from an average value of 226 Hz at age 12 to a value of 127 Hz at age 15. This drop is much smaller in female speakers with no significant pitch changes after age 12, with an average f0 of 231 Hz. For our target age of 10, the average f0 for boys is around 260 Hz while girls have an average value of around 270 Hz. This suggested the use of female adult voices as a base for the voice of children of both genders.

The analysis of formant frequencies shows a clear linear scaling trend as a result of the axial growth of the vocal tract [3]. The main gender difference is that the scaling factors of male speakers are approximately the same for all formants while each formant of the female speakers evolves differently as a function of age. Since the formant scaling factor from an adult male voice to an adult female voice is, on average, 30%, female voices are also in this respect better for being transformed into children’s voices. This way, using the data in [3], the average scaling factor from an adult female voice to a voice of a 10 years old boy, would be of around 10% for all formants. The average scaling values for a voice of a girl of age 12 would be 20% for F1, 15% for F2 and 10% for F3.

Taking into account these results it was decided to search for voice donors with the following characteristics: females of small stature, corresponding to a small vocal tract, with experience in interacting with children of the target age, without strong social or regional accent and with the ability to produce the required intonation in a regular way. The selected speakers were two English teachers of children of age 10. The recording tests showed that they were able to produce the required intonation patterns and that their voices could be modified by both the PSOLA technique [4] and spectral scaling with little distortion. By applying different small scaling factors to both f0 and formant frequencies, we could produce voices for the different synthetic characters. For the German version a female and a male voice were recorded. As expected, the pitch of the male voice could not be changed to the values usually observed in children’s voices but informal tests showed that the modified voices were acceptable for cartoon like characters.

3. Voice Building Process

The speech corpus for the recordings was built based on the language engine that converts into English or German the speech acts used for the communication between agents. The input text of the synthesizer is thus limited by the variability of the text generated by the language engine. This suggests the use of a limited domain speech synthesizer [5][6].

To create the inventory required to synthesize the utterances spoken by the characters we started by modifying the language engine to generate all the possible sentences. This resulted in a total 7496 sentences, with 1206 distinct words. A greedy algorithm was used to select a subset of these sentences with full word coverage, distinguishing words in the middle of intonational phrases and words close to prosodic boundaries. The greedy algorithm selected 552 sentences for the English inventory. A similar procedure was applied to the German language engine generating a total of 4690 sentences, with 1557 distinct words, from which the greedy algorithm selected an inventory of 622 sentences.

The two selected voice donors for each language recorded all the sentences of the English inventory. The recordings were conducted in the sound proof booth of INESC-ID and the speakers were asked to read the prompts with some, but not excessive, expressiveness. The recordings required four sessions of 2 hours for each speaker.

3.1. Integrated Voice Building

Building a voice for a TTS is a non-trivial task as needs a lot of pre-processing steps. In order to remove errors and repetitions from the utterances' orthographic transcriptions, the text prompts were manually verified. Then, they were automatically split into prosodic phrases by using the MuLAS system [7] so that every single file contains only one prosodic phrase. The resulting 552 phrases were then automatically segmented by our own phonetic segmenter [8] that was specifically adapted for British English. Gender dependent models were trained using the British English WSJ corpus, which reached 85% and 84% of accuracy at 20 milliseconds for female and male speakers, respectively. A speaker adaptation procedure was performed 2 times, by using the canonical word pronunciations for the segmentation stage. At the 3rd iteration, the segmenter was provided with a pronunciation graph accounting for the canonical pronunciations together with some alternative pronunciation raised by the post-lexical rules.

Using a multi-level unit inventory we are able to generated new words which are not occurring in the recorded speech corpus. We call this approach a semi-limited domain synthesis while not all words existing in on language can be reproduced.

Our voice building software is capable of building voice inventories using only the label-files which include the segment start time, the word and syllable boundaries as well as syllable stress information. Furthermore we need the according utterance-files and the recorded audio-files. Once all files are gathered an automatic process starts and builds a context depended voice inventory stored as a XML based representation of each label, utterance and audio- file. Please see section 3.1.1 for a detailed description of the XML representation.

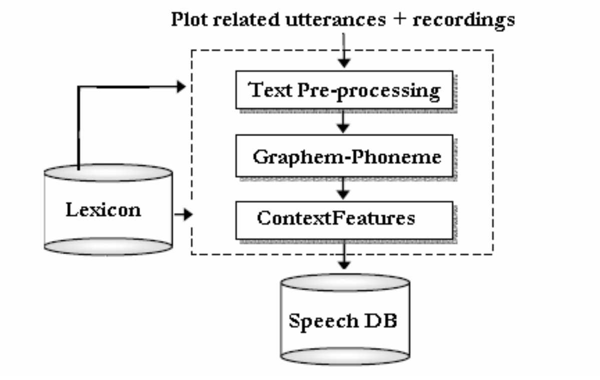

Figure 2: Diagram of the voice building system.

Figure2 shows a schematic flow-chart of the steps which were applied during the voice building process. These steps are:

- Text normalization
- Acoustic and spectral parameter extraction; Duration, F0, MFCC
- Extracting phonological and quantitative features.
- Grapheme-Phoneme conversion

For F0 and spectral feature extraction we use standard tools like the Snack-library and HTK. For dynamic feature prediction we use conditional log-linear models, please see section 3.2.

3.1.1. XML –Representation

The XML based structure consists of the features as listed below:

Table 2: Overview of features

Unit	Feature
Word	Preceding, following word Sentence type Distance left/right in sentence POS Duration, log duration Average F0, log F0 First/last frame MFCCs
Syllable	Preceding, following syllable Distance left/right word, sentence Stress Duration, log duration Average F0, log F0 First/last Frame MFCCs
Phone	preceding, following phone distance left/right word, syllable Duration, log duration average F0, log F0 First/last Frame MFCCs

Once we extracted all features which are describing the segments we build a temporarily XML-based left-right context representation of the utterance and store this information in a voice inventory database.

3.2. Conditional Log-Linear Models for Dynamic Feature Prediction

For Grapheme-Phoneme conversion, Part-of-Speech Tagging syllable boundary detection, as well as for duration and F0 predicting we applied conditional log-linear models also known as Maximum-Entropy models [9], [10].

The conditional log-linear model framework is a well known approach for ambiguities resolution in natural language processing where many problems can be reformulated as a classification problem. The task of such a reformulation is to include a context and to predict a correct class. The objective is to estimate a function $X \to Y$, which predicts an object $x \in X$ to its class $y \in Y$. Y represents the predefined classes for either each task of our prediction problem.

In the field of stress prediction we are dealing with a binary classification where the class is true for stressed syllables and false for non-stressed.

The same binary classification task has to be solved in the domain of syllabification where we have a syllable boundary or not.

X consists of quantitative and phonological features where we include the context and the resulting input for the classification. The classifier $X \to Y$ can be seen as a conditional probability model in the sense of

$$C(x) = \arg\max_{y} p(y|x) \qquad (1)$$

where x is the object to be classified and y is the class. Including the context we get a more complex classifier

$$C(x_1, x_2, ..., x_n) = \arg\max_{y_1 ... y_n} \prod_{i=1}^{n} p(y_i | x_1 ... x_n, y_1 ... y_{i-1}) (2)$$

where $x_1 ... x_n, y_1 ... y_{i-1}$ is the context at the i^{th} decision and y_i is the outcome.

This model we use in all our dynamic feature prediction tasks during the offline voice building process as well as during runtime.

3.3. Acoustic Synthesis with F0 Smoothing

The acoustic synthesis module follows the variable-size unit selection algorithm. We apply a pre-selection strategy while the algorithm tries to find a segment that matches the predefined target structure in a left-right context. If this does not result in any found segment we simplify the structure matching but keep the left-right context. When no segment is found at the word-level, the algorithm searches for syllable segments and, as a last alternative, a phoneme-level segment selection is performed.

Using a predefined structure matching for segment selection we save computational resources in target and join-cost distance calculation. The target distance calculation is done by summing the differences between the values of the features of the selected and of the target segment. Some kind of normalization is needed given the different ranges of the feature values (for example, the log F0 and the duration values). This normalization is done using the following equation:

$$normcost = \frac{x^2}{1 + x^2} \qquad (3).$$

where *x* is the difference between the values of the feature of the selected and of the target segment. The join cost calculation is done by a Euclidian distance measure between the successive frames MFCC's of the segments.

4. Experimental Evaluation

The experimental evaluation was conducted only on the English version of the synthesizer. The evaluation followed a procedure very similar to the one described in [11]. Given that we are not using real children's voices, one of the objectives was to check if the modified voices were acceptable for the FearNot! characters. Two types of tests were conducted: half of the subjects could only listen to the characters voices, while the other half watched movie clips with different animated characters (Figure 2). Although the lip movements were random, they were synchronized with the duration of the utterance making an acceptable illusion of lip synchronization given the small size of the characters mouth.

The subjects were asked to rate the utterances in terms of 6 factors: (1) overall sound quality (2) naturalness of the intonation (3&4) extent to which the utterance sounded like a boy or a girl (5&6) extent to which the utterance sounded like it was pronounced by the bully or the victim.

The stimuli were produced in 8 different versions: the original recordings of both speakers, synthesized speech using unmodified inventories of both speakers, one modified version for each speaker original recordings and synthesized speech using inventories of modified voices. Each subject was asked to rate a total of 48 stimuli. Like in [11] the ratings were on a Likert scale with 1 for very bad and 5 for very good. The test was conducted over the internet and the subjects used headphones. The results showed that the presence of video result in a better rating on the overall perceive quality: 3.42 (with a significance of $p<0.005$) vs 3.70 ($p<0.00001$). Without the video the overall rating of boy, girl, victim and bully was not significant ($p>0.05$). The presence of the animated character made the voices believable especially for the victim (3.68, $p<0.00001$). The modified voices had the same rating in overall quality as the unmodified voices for the audio only test (3.42, $p<0.04$) but were better rated when played in video clips (3.82, $p<0.00001$ vs 3.59, $p<0.009$). The results for the overall quality of both the modified and unmodified recording were above 4 (4.45, $p<0.00001$). The ratings for synthesized speech were not significant and the analysis of the results showed that although the evaluators agreed on some sentences (usually with score above 4) they did not agree on the rating to assign to sentences with noticeable concatenation discontinuities.

Figure 3: Image of one of the video clips used in the audio and video evaluation task.

5. Conclusion and Future Work

Limited domain synthesis allowed us to produce voices for 3D animated characters with almost natural speech quality as expected by the users of virtual learning environments. In order to minimize concatenation mismatches we asked the adult voice donors to refrain their expressiveness during the recordings. This affected mostly the bully character's voice that was found less credible, but with a sufficiently good rating. Although there was no story context in our evaluation, the video of the animated characters influenced positively the perceived overall quality and intonation.

Using the results of this study, we will now generate additional modified voices for the remaining characters of the FearNot! application. We also plan to correct some segmentation and concatenation problems detected during this evaluation, and to improve the voice modification algorithm by using a more robust epoch detector. The German language version of the system is also being developed. The effectiveness of the FearNot! application against bullying in schools will soon be fully investigated when the final version of the system is placed in schools in the UK and Germany for a large scale longitudinal evaluation.

6. Acknowledgements

This work was partially supported by European Community (EC) and is currently funded by the eCIRCUS project IST-4-027656-STP. The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

7. References

[1] Zoll, C., Enz, S., Schaub, H., Aylett, R., Paiva, A., "Fighting Bullying with the Help of Autonomous Agents in a Virtual School Environment", *7th International Conference on Cognitive Modelling*, Trieste, Italy, 2006

[2] Cabral, J. and Oliveira, L.C., Guilherme Coelho Barreira Raimundo, Ana Paiva, "What voice do we expect from a synthetic character?", *SPECOM*, pages 536-539, 2006.

[3] Lee, S., Potamianos, A. and Narayanan, S., "Acoustics of children's speech: Developmental changes of temporal and spectral parameters", *JASA*, 105:1455–1468, 1999.

[4] Charpentier, F., Moulines, E., "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Eurospeech*, Paris, 1989.

[5] Black, A. and Lenzo, K., "Limited Domain Synthesis", *ICSLP*, Beijing, China, 2000.

[6] Schweitzer, A., Braunschweiler, N., Klankert, T., Möbius, B., Sauberlich, B., "Restricted unlimited domain synthesis", *Eurospeech*, 1321-1324, 2003.

[7] Paulo, S.G., and Oliveira, L.C., "MuLAS: A Framework For Automatically Building Multi-Tier Corpora", *Interspeech*, Antwerpen, 2007.

[8] Paulo, S.G. and Oliveira, L.C., "Generation of Word Alternative Pronunciations Using Weighted Finite State Transducers", *Interspeech*, pages 1157-1160, 2005.

[9] Berger, A., Della Pietra, S.A., Della Pietra, V.J., "A Maximum Entropy Approach to Natural Language Processing", *Computational Linguistics*, 22(1), 1996.

[10] Ratnarparkhi, A., *Maximum Entropy Models for Natural Language Ambiguity Resolution*, PhD Dissertation, University of Pennsylvania, 1998.

[11] Johnson, W.L. Narayanan, S. Whitney, R. Das, R. Bulut, M. LaBore, C., "Limited domain synthesis of expressive military speech for animated characters", *IEEE Workshop on Speech Synthesis*, September 2002.