Voice transformation and speech synthesis for video games

Snorre Farner[1], Christophe Veaux[1], Gregory Beller[1], Xavier Rodet[1], and Laurent Ach[2]
[1] IRCAM, 1 place Igor Stravinsky, 75004 Paris
[2] Cantoche, 68 rue d'Hauteville, 75010 Paris

May 26, 2008

Abstract

Voice and expressivity transformation as well as text-to-speech synthesis with high degree of naturalness are now available. A set of tools permitting a large range of voices to be made from a single voice, whose speech may be produced from text and given a certain expressivity, is proposed. In the context of multiplayer video games, for instance, this technology allows for creation of the speech of non-player characters as well as for transforming the player's voice into the voice of her character. The technology behind these tools will be presented. A demonstration using cartoon characters will also be provided.

1 Introduction

Speech is already widely used in video games and is gradually replacing the use of written text. Speech material is traditionally simple studio recordings that are stored and replayed without modification when asked for in the game. This limits the use to the narrator's comments and to simple, predetermined utterances by non-player characters (NPC). Furthermore, it requires the recording of a number of actors to represent the voices of the different NPCs in the game. In modern multiplayer role-playing games, the players create their own characters by choosing their appearance, abilities, and personality. These characters communicate with the other participants in the game by writing messages in a console, including the use of emoticons like smileys, and by making their characters perform gestures to express emotions and other attitudes, intentions, etc., all of which is called *expressivity*.

In this presentation, we propose a set of tools to facilitate and develop the use of speech in video games. These tools provide the following facilities:

- to transform the voice of a given actor into several different characters,
- to design the voice of a character based on another voice or that of the player,
- to modify speech to express a certain emotion or affect, and
- to produce arbitrary sentences based on Text-To-Speech synthesis (TTS), and apply the mentioned transformations to these sentences.

All this can be done in real time, allowing the developer to take into account the instantaneous context and the course of the game as well as to render the player's voice into that of the character on the fly.

Without going into technical depth, we present some background for these facilities and show their usage by a few demonstrations, in particular in connection with the *Living Actor*™ technology developed by Cantoche.

2 Transformation of type and nature of the voice

Given a person's voice, it is now possible to convert it to sound like another person with a certain size, sex, and age with a high degree of naturalness. Two important factors for such transformations are the *pitch range* and the *timbre* of the voice. The pitch range is related to the length and thickness of the vocal folds and distinguishes, for instance, low-pitched male voices from high-pitched child voices. The timbre, on the other hand, is related to how the vocal folds move and the physics of the vocal tract, including the nasal cavity. Differences in timbre, described by adjectives like dark, bright, nasal, hollow, etc., make us able to separate the voices of two men from each other, a low-pitched woman's voice from that of a tenor male voice, or a mature adult from a teenager even though the pitch might be similar. There are also other aspects of vocal quality, such as the degree of breathyness, hoarseness, tenseness, etc.

After studying the characteristics of different types and natures of voices, we have developed the *IrcamVoiceTrans* software library of transformations. IrcamVoiceTrans is capable of changing the voice type between those of men, women, children, old persons, etc., as well as rendering the voice hoarse (as that of a smoker), creaky, whispering, etc.

Apart from a short presentation of some of the mechanisms used in the transformations, we will demonstrate how the recording of a single person's voice can give a conversation between different persons, and how our technology can be used to modify the voice produced by our Text-To-Speech synthesis system (see below). We will also apply a couple of transformations in real-time on the speaker's voice while he gives the talk.

3 Transformation of expressivity

In video games as in real life, emotions, attitudes, and intentions conveyed by the voice are an important part of communication. Examples of such *expressivities* are *anger, happiness, fear, and sadness*, each of which may be directed inwards or expressed extrovertedly, as well as positive and negative surprise, disgust, discretion, excitation, and confusion. The use of *emoticons* such as smileys is an example of the need for expressing non-verbal information in textual interfaces.

For expressive speech to be available in TTS synthesis, the database must be recorded with this expressivity. This requires costy preparation of many databases, so our research has rather been centered at another way to render a voice expressive: transformation of neutral speech, synthesised or recorded, to expressive speech. A talented professional actor is able to speak with different degrees of the above-mentioned expressivities. Our approach has thus consisted in recording actors speaking with various expressivities, and in building models of how they affect the voice in terms of pitch, timing, and timbre by means of automatic machine learning. These models may then be used to modify the neutral voice to exhibit the same characteristics as expressive speech to a given degree. We will demonstrate this type of transformation by some sound examples.

4 Text-to-speech synthesis (TTS)

While the above subjects concern transformation of the voice, application of speech synthesis in video games extends the use of the voice to allow intelligent NPCs to speak with the players and the generation of various narrator's comments to be spoken depending on the game evolution.

Corpus-based or unit-selection synthesis, which is the latest generation of speech synthesis methods, now offers a very high degree of naturalness and intelligibility. In this approach,

speech is synthesized by intelligent selection and concatenation of small units of continuous speech that are stored in a database. The database contains a large number of units in order to deal with various phonetic contexts and different prosodies (speech melody and rhythm). The text input is first converted into a phonetic sequences with prosodic specifications. Then the optimal sequence of units is searched in order to fit the phonetic and the prosodic specifications while minimizing artifacts when joining units.

We have developed *IrcamTTS*, a corpus-based synthesizer, which uses a database of nine hours of speech recording in an anechoic room. The synthesis can also be done on the fly and offers a high level of naturalness. Moreover, an entirely automatic segmentation has been developed to create the database of units from recorded speech. Therefore the procedure to build a new voice is greatly simplified and can complement the use of transformation. A demonstration of synthesised speech transformed into different voices and expressions will also be given.

5 Conclusion

We summarise the presentation by a motion-picture cartoon by the Living Actor$^{\text{TM}}$ system of Cantoche where the characters' voices have been created by transformation of a single person's voice. The high quality of the result is expected to have a great impact both on human-computer interaction and on communication between the characters in video games. Other possible applications are within educational games, e-learning, and "serious" games, for instance.