Abstract. In the present paper, several experiments on text-to-speech system personification are described. The personification enables TTS system to produce new voices by employing voice conversion methods. The baseline speech synthesizer is a concatenative corpus-based TTS system which utilizes the unit selection method. The voice identity change is performed by the transformation of spectral envelope, spectral detail and pitch. Two different personification approaches are compared in this paper. The former is based on the transformation of the original speech corpus, the latter transforms the output of the synthesizer. Specific advantages and disadvantages of both approaches are discussed and their performance is compared in listening tests.

1 Introduction

Within the concatenative corpus-based speech synthesis framework, a new voice can be obtained by recording a new large speech corpus by the demanded speaker. From that corpus, containing several thousands of utterances, a new unit inventory is created and used within the synthesis process [1]. However, recording of such a great amount of speech data is a difficult task. Usually, a professional speaker is required.

Alternatively, text-to-speech system personification [2] enables this system to produce new voices by employing voice conversion methods. Much fewer speech data are necessary. Our voice conversion system [3] converts spectral envelope and pitch by probabilistic transformation functions; moreover, spectral detail is transformed by employing residual prediction method.

Two different personification approaches are described and compared in this paper. The former is based on the original speech corpus transformation, the latter transforms the output of the synthesizer. Specific advantages and disadvantages of both approaches are discussed and the performance is compared by using preference listening tests.

The paper is organised as follows. In Section 2 the baseline TTS system planned to be personified is described. In Section 3 the voice conversion methods are specified. Section 4 deals with the TTS system personification task. Section 5 describes our first personification experiments. In Section 6 the results are discussed and future work is outlined.
2 Baseline TTS System

The text-to-speech system ARTIC employed in our personification experiments was in detail described in [1]. It has been built on the principles of concatenative speech synthesis. Primarily, it consists of three main modules: acoustic unit inventory, text processing module and speech production module. It is a corpus-based system, i.e. large and carefully prepared speech corpora are used as the ground for the automatic definition of speech synthesis units and the determination of their boundaries as well as for unit selection technique.

Our TTS system was designed for the Czech language, nevertheless many of its parts are language-independent. For our personification experiments, a female speech corpus containing 5,000 sentences (about 13 hours of speech) was employed. The block diagram of our TTS system is shown in Fig. 1.

![Fig. 1. A scheme of our TTS-system ARTIC including the both personification approaches – see dashed and dotted blocks](image)

3 Voice Conversion System

The voice conversion system utilized for the aforementioned system personification was introduced in [3]. A simplified version of that system is described in this section. For the training of transformation functions, parallel utterances (i.e. pairs of source and target speakers’ utterances) are employed. Voiced speech is analysed pitch synchronously; each segment is three pitch periods long and the shift of analysis window is one pitch