Specifying Affect and Emotion for Expressive Speech Synthesis

Nick Campbell

ATR Human Information Science Laboratories, Kyoto, Japan.
nick@atr.co.jp

Abstract. Speech synthesis is not necessarily synonymous with text-to-speech. This paper describes a prototype talking machine that produces synthesised speech from a combination of speaker, language, speaking-style, and content information, using icon-based input. The paper addresses the problems of specifying the text-content and output realisation of a conversational utterance from a combination of conceptual icons, in conjunction with language and speaker information. It concludes that in order to specify the speech content (i.e., both text details and speaking-style) adequately, selection options for speaker-commitment and speaker-listener relations will be required. The paper closes with a description of a constraint-based method for selection of affect-marked speech samples for concatenative speech synthesis.

1 Introduction

For unrestricted text-to-speech conversion, the problems of text anomaly resolution and given/new or focus determination can be profound. They can require a level of world-knowledge and discourse modelling that is still beyond the capability of most text-to-speech synthesis systems. One implication of this is that the prosody component of the speech synthesiser can only be provided with a default specification of the intentions of the speaker or of the underlying discourse-related meanings and intentions of the utterance, resulting in a flat rendering of the text into speech. This is not a problem for the majority of synthesis applications, such as news-reading or information announcement services, but if the synthesiser is to be used in place of a human voice for interactive spoken dialogue, or conversation, then the speech will be perceived as lacking in illocutionary force, or worse, it will give the listener a false impression of the intentions of the utterance and of the speaker-listener relationships, leading to potentially severe misunderstandings.

When a synthesiser is to be used in place of a human voice in conversational situations, such as in a communication aid for the vocally impaired, in speech translation systems, or in call-centre operations, then there is a clear need for the vocal expression of more than just the semantic and syntactic linguistic content of the utterance. Paralinguistic information related to dialogue turns, and speaker interest is signalled along with the syntactic structure of the speech by means of prosody and voice quality [1].