FUTURE DIRECTIONS FOR SPEECH SYNTHESIS – A PERSONAL VIEW

Nick Campbell
Project Leader, ATR Human Information Science Laboratories
Research Director, JST/CREST ESP Project.

ABSTRACT

Keywords: speech synthesis, unit-selection, prosody, voice-quality, contexts, relationships

This paper presents an overview of the needs for future speech synthesis, based on an analysis of trends up to the present day. Whereas "contexts" can be considered the keyword of unit-selection synthesis, this paper will argue that "relationships" also need to be considered if we are to progress to the next level of speech synthesis quality. Lexical, syntactic and discoursal contexts have been shown to affect the acoustic characteristics of the speech waveform, and consequent consideration of the prosodic environment as a selection criterion has resulted in significant improvements to the quality of corpus-based synthesised speech.

From a preliminary analysis of the acoustic characteristics in a large conversational-speech corpus of spontaneous Japanese, it is clear that speaker-listener relationships, and speaker-commitment relationships also have similar effects on the speech. This paper will summarise the types of meaningful variation that arise when the talker is addressing a different interlocutor, and when the talker expresses different degrees of commitment towards the content of an utterance. The necessity for synthesised speech to mimic these characteristics of a human speaker will be discussed.

Due to the rapid growth of information-access technology, resulting mainly from developments in telephony and the internet, our early expectations of speech synthesis as a "reading machine" have been replaced by the vision of a "talking machine" instead. The main difference being that whereas the former was primarily concerned with only the message, the latter is also concerned with the ways in which that message can be presented. Applications such as customer-care and remote marketing, for example, now require that speech synthesisers be able to express personality in addition to utterance content, so it is speculated that future research will focus on modelling the different levels of the information content of speech (linguistic, paralinguistic, and extralinguistic) in a multi-dimensional manner, in order to reproduce speech which is not only meaningful, but also appropriate to the context of the discourse and to the expectations of the listener.

Two such dimensions might be proposed: the dimension of "commitment", or content-relationship, which governs the expressed sincerity of the speaker, including expression of emotion and revelation of the speaker's attitudinal bias, and the dimension of "friendliness", or listener-relationship, which governs the formality and the degree of familiarity that can be expressed in the speech. Samples of naturally-occurring human speech will be presented, in order to illustrate the multi-dimensionality of information carried by the human voice in conversational interactions, and suggestions will be offered for the categorisation and parameterisation of these variables.

The challenge to speech coding and synthesis will be to recognise such paralinguistic features of the speech signal and to reproduce them identifiably with a small number of parameters. A grammar of spoken language in context will then be needed in order to map between the different levels of information carried by the speech signal. Work in progress, part of the JST/CREST Expressive Speech Processing project, to prepare such data for statistical modelling, will be reported.

0-7803-7549-1/02/$17.00