Predicting the prosody of speech for synthesis

Nick Campbell
ATR Human Information Sciences Research Labs
Keihanna Science City, Kyoto, Japan.

When this paper was proposed, I was asked to write about the prediction of prosody for speech synthesis, but in view of the multiplicity of problems involved, I would prefer to talk instead about why we *cannot* predict prosody, and about what alternatives there might be for the control of speaking style in synthetic speech.

The reasons for the difficulty of prosody prediction are many, but they can be reduced to two main factors: sparcity of input information, and inadequacy of the prediction models. Prosody in speech adds several layers of information over-and-above that signaled by the word string alone; it shows how the speaker relates to the hearer and to the content of the message, it includes para-linguistic and extra-linguistic details signalling the intentions and state of the speaker, and how the word string should be interpreted. Very little of this information can be reliably estimated from an analysis of the text alone.

Statistical models exist for the mapping between input features and output parameters, but they are limited in that they can predict only averages; and unless the input is very finely specified, the output can only be an estimation of the most likely average for any given combination of input settings. But the variety of expression observed in speech is not random; *all* variants can be interpreted as carrying some form of meaning, so the likelihood of a mistaken interpretation is high unless the prediction is limited to the most basic of levels. Currently, all that can be obtained automatically from an analysis of the input text is an estimation of the syntactic and semantic relationships between the words; not their intended interpretation.

Speech synthesis was originally conceived as a form of media conversion, mapping between text and speech in the form of a reading machine. However, there are few uses for read speech; and many more for the transfer of information via the oral medium. In other words, we need talking machines, not reading machines! Speech-translation, customer-care, communication aids, weather-forecasts, stock-price announcements, car-navigation, even speaking clocks, can better be envisaged as talking machines, expressing information for human consumption. Computers may not be expected to laugh or cry, but speaking machines may have to express sadness or joy on behalf of human sources, mediated via machine.

We should try to consider the range of uses of the human voice, and the flexibility that we use in expressing facts and feelings, if we are to design a prosody component for future speech synthesis. This talk will present examples of the types of speaking style encountered in everyday human speech, and describe the labels that we are currently using for the description of these features. It will conclude with a proposal that future prosody prediction may avoid the statistical modeling between features and parameters, and go directly from the feature-based specification to unit selection for synthesis.