

## **Inferring and Expressing Prosody in Text-to-Speech Systems**

**Marc Fabiani. Nuance**

Text-to-Speech systems have achieved a high degree of naturalness and can be indistinguishable from recorded speech. But even in the best systems, truly consistent quality obtains only in contexts that are narrow either lexically (e.g., natural numbers, date, times) or prosodically (e.g., name and address strings). The result is that the use of Text-to-Speech tends to be restricted to such contexts or where automation is not possible by any other means. It is clear that the ability to infer prosody more intelligently from text is a key to broader adoption of Text-to-Speech, but one of the hazards of commercial deployment is that it is often more costly to make an error than it is to maintain a neutral consistency: as in medicine, a 'first do no harm ethic must be balanced with innovation.

In this talk I will review the current state of Text-to-Speech and the degree to which the best systems can accurately infer prosody and produce natural sounding speech output accordingly. I will show that the challenges are formidable and have largely gone unmet until recently. I will conclude with a discussion of some promising areas of research, and provide audio samples of some early results.