# EMOTION IN SPEECH

Nick Campbell
ATR Human Information Science Labs
Keihanna Science City, Kyoto, Japan
nick@atr.jp

# Towards an Advanced Media Society

There is a strong need to produce speech technology that is of use to the normal citizen in addition to that designed for the business or military user.  As illustrated by the boom in portable telephone ownership and usage, speech is the first medium of choice for the majority of day-to-day social and business interactions, yet current speech technology is incapable of processing even the full range of linguistic information in speech.

When we consider that approximately half of speech information is supra-linguistic (expressing affect, social relationships, and discourse roles) then we begin to understand how limited is the current machine processing of human speech.

# Text versus speech processing

- The main needs of business applications can perhaps be met by processing just the linguistic information in speech, but the needs of social interaction are to express and interpret states and relationships that are not well described by such linguistic information alone, and that are signalled alongside it as a fundamental component of conversational speech.
- Much of the transcribed text of a spoken conversation would be unintelligible or meaningless without a knowledge of the way it has been spoken, and of the contextual and interpersonal variables that are conveyed by its prosody and tone-of-voice.

# Is it ‘emotion’ that is missing?

- Current technological trends are very disappointing in this respect; both speech synthesis and recognition are trained and evaluated at the lingustic level alone, using text (often well-formed and complex grammatical constructions that bear little relation to the clipped and repetitive units of conversational speech) as the intermediate representation.
- There is a growing realisation within the speech processing communities that this is insufficient, but it is explained as being due to a lack of 'emotional' content in speech processing.

# Information versus Affect

- Our analyses of the 1000-hr ESP speech corpus have shown that the overt expression of 'emotion' (in the generally-accepted sad/happy/angry/fearful dimensions) is extremely rare. Social norms and societal roles discourage the expression of such raw emotions in daily conversation, but we found that much more subtle levels of information were expressed instead.
- Speakers speak not just to transmit information, but also to convey feelings, nuances, and to display relationships. Conversational interactions are at least as much social as informational or transactional.

# Social Communication Devices

- ‘Social Robots', ‘Customer-Care' applications, or ‘Talking Machines‘ have the potential for monitoring or taking a part in a conversation without actually 'understanding' it --- in much the same way that domestic pets can intepret their owners' tone-of-voice and speech patterns without (we presume) being able to understand the linguistic content of their speech.
- They offer an avenue for phatic communication, for using speech sounds to reassure a human that contact is being made and that understanding is taking place at a social level if not a linguistic one.

# Supra-segmental Speech Info

- The development of an advanced media society should take into consideration these levels of supra-linguistic information in speech, both for monitoring and interpreting as well as for generation of communicative content.
- With the development of ubiquitous computing and pervasive devices, there is great potential for such non-linguistic processing of speech.
- Our pre-human ancestors apparently managed to communicate and to form advanced societies well before the development of formal language, by using grunts and non-verbal speech sounds. It may be that we first 'communicate' with domestic machines in the same way, much as we do with domestic animals, by using sounds that are natural to us and come 'instinctively', yet which may bear little linguistic content.

# Sensitive Technology

- We should aim to produce a technology that is at least sensitive to the multiple levels of information carried by speech, so that people with different levels of educational development, technical ability, and understanding can gain access to machine-mediated information and services through the use of voice.
- This goal will not be achieved if we limit our view of speech to only the information that can be reproduced in a written transcription of its content. Nor to simply 'emotion'!
- Instead, we need to develop a model of the multiple layers of interpersonal, social, discoursal, and affective information that is carried by the voice, and a means for integrating these different streams of information so that a more global interpretation of the speech content can be achieved.