

Expresion of emotions in speech synthesis

--

A bird's eye snapshot

HUMAINE Plenary Meeting, 4-6 June 2007, Paris
Marc Schröder, DFKI
schroed@dfki.de

Overview

- ◆ Modelling speech expressivity for synthesis
- ◆ State of the art in expressive speech synthesis
- ◆ Challenges ahead

1. Modelling speech expressivity for synthesis

- ◆ Knowing what to express
 - ➔ How to measure perceptually relevant acoustic correlates of emotional expressivity?

1. Modelling speech expressivity for synthesis

- ◆ Global prosodic settings
 - ➔ pitch level, range
 - ➔ energy level, range
 - ➔ speaking rate
- ◆ This is what everybody measures, because it is easy to measure
- ◆ It is clearly insufficient

1. Modelling speech expressivity for synthesis

◆ Voice quality

→ model-based: LF model of glottal flow

- direct measurement through (manually corrected) inverse filtering of the speech signal (Gobl & Ní Chasaide, 2003)
- “trial-and-error” modelling, minimising prediction error (Vincent, Rosec & Chonavel, 2005)
- easier to compute approximation: NAQ (Alku et al., 2002)

→ spectral descriptions

- implicit: MFCCs
- explicit: energy in spectral bands, spectral tilt, glottal formant, ...

1. Modelling speech expressivity for synthesis

◆ Contours

- ➔ evolution over time
- ➔ interaction with linguistic structure
 - e.g., prominence of accented syllables may depend on emotion
 - same intonation contour may have different meaning depending on sentence type (Scherer, Ladd, Silverman, 1984)

1. Modelling speech expressivity for synthesis

- ◆ Interjections / affect bursts
 - word-like local phenomena
 - densely convey emotional meaning (Schröder, 2003)

1. Modelling speech expressivity for synthesis

- ◆ Measuring perceptual relevance
 - ➔ Brunswikian lens model (Bänziger, 2004)
 - ➔ controlled synthesis (Audibert et al., 2006)
- ◆ Fully automatic methods preferred for data-driven synthesis

2. State of the art in expressive speech synthesis

◆ Explicit modelling approaches

- emotional prosody/voice quality rules
- realise through parameterisable synthesis method (formant, diphone synthesis)



Burkhardt (2000)
anger



Schröder (2003)
active-negative-dominant

- sounds unnatural

2. State of the art in expressive speech synthesis

◆ “Playback” approaches

- unit selection: speaking style defined by recordings
- no explicit model needed
- no control possible



Schröder (~2005)
soccer comments



Loquendo (2004)
pre-recorded affect bursts

- can sound natural, but only in the recorded style

2. State of the art in expressive speech synthesis

◆ Implicit modelling approaches

- statistical models trained on / adapted to expressive databases
- HMM-based speech synthesis
- interpolation possible

	trained from corpus:		trained from corpus:	
Miyanaga et al. (2004)	<u>neutral</u>	<u>0.5 joyful</u>	<u>joyful</u>	<u>1.5 joyful</u>
		interpolated		interpolated

- has potential

3. Challenges ahead

◆ long-term

- ➔ explicit models with fully controllable synthesis (e.g., articulatory synthesis)
- ➔ will need efficient ways of training models on large databases
- ➔ the general solution is not in sight

3. Challenges ahead

◆ short-term

- ➔ make playback approach more parameterisable
 - emotion-related selection criteria
 - interpolation between expressive voices (Schröder, submitted)
- ➔ make implicit models more natural-sounding
 - actively pursued by Tokuda et al.
- ➔ acceptable solutions for targeted applications are in reach