Emotional Speech Synthesis: a comparison of different methods

Melanie Kroes
Judith Kessens
Mark Neerincx
Outline

- Introduction
- Emotional speech synthesis
- Methods
- Results
- Conclusions
Introduction

• Internship TNO Human Factors
• Literature study
• Experiment
  • Research question: to what extent is it possible to identify intended emotions in synthetic speech, which are produced by altering synthesis parameters (duration and F0)?
Emotional speech synthesis (1)

Emotions

- How to describe emotions?
  - Categories
  - Dimensions
    - Arousal
    - Valence

- Which emotions?
  - Four basic emotions: anger, fear, happy and sad
  - Other emotions: relaxed and neutral
Emotional speech synthesis (2)

Synthesis techniques

• Diphone synthesis
  • Concatenation of diphones
  • Emotion modeling possible with open source systems
  • Most relevant for TNO research
Emotional speech synthesis (3)

Parameter settings

- EmoFilt (Felix Burkhardt)
  - Standard settings for 9 emotions (categories)
  - Converts durations and F0

- EmoSpeak (Marc Schröder)
  - Parameters (duration and F0) are changed according to the dimensions arousal, valence and power

- Copy synthesis
  - F0 and durations copied from naturalistic emotional speech
Methods (1)

Data selection

• Emotional speech database
  • Belfast Naturalistic Database (Queens University Belfast)
  • Dimensional and categorical annotations

• Semantics
  • Neutral vs. emotional

• Language
  • English
Methods (2)

Speech generation

- With EmoFilt & EmoSpeak
  - Creating durations and F0 pattern with MARY
  - Manipulation of durations and F0

- With copy synthesis
  - Extraction of F0 with Praat
  - Extraction of durations with TNO speech recognizer

- Speech generation with MBROLA diphones
Methods (3)

Conditions

- 35 unique conditions:
  - 6 emotions
  - 2 types of semantics
  - 3 different settings, plus neutral synthesis
- 2 sentences per condition
- 10% (7 sentences) presented twice
- each participant annotated 77 trials categorically and dimensionally
- 20 participants
Methods (4)

Setup

• For both dimensional and categorical annotations:
  • Training with natural and synthetic speech
  • Experiment:
    Dimensional
    Categorical
Results (1) \textbf{Categorical annotations}

\textbf{Emotions}
- Percent of correct recognition significantly better than chance
- Confusion between:
  - Relaxed/sad and neutral (and vice versa)
  - Fear and anger (but \textbf{not} vice versa)

\begin{table}
\centering
\begin{tabular}{|l|c|c|c|c|c|c|}
\hline
 & Anger & Fear & Happy & Neutral & Relaxed & Sad \\
\hline
Intended & 49.0 & 9.6 & 17.3 & 10.4 & 4.4 & 9.4 \\
Fear & 21.3 & 31.7 & 13.3 & 18.8 & 5.8 & 9.2 \\
Happy & 18.8 & 13.8 & 37.4 & 15.7 & 5.5 & 8.8 \\
Neutral & 2.9 & 5.4 & 3.2 & 34.1 & 22.5 & 32.0 \\
Relaxed & 2.3 & 2.6 & 6.3 & 28.2 & 42.0 & 18.5 \\
Sad & 1.3 & 3.5 & 4.0 & 27.1 & 18.3 & 45.8 \\
\hline
Average & 15.9 & 11.1 & 13.6 & 22.4 & 16.4 & 20.6 \\
\hline
\end{tabular}
\end{table}
Results (2) Categorical annotations

Settings
• Settings of EmoFilt and EmoSpeak only differ on the emotion fear, copy synthesis had the worst recognition

Comparison with neutral synthesis
• Modification of parameters was better than neutral synthesis

Semantics
• Semantically emotional sentences were better recognized than semantically neutral sentences
Results (3) **Dimensional annotations**

**Comparison with natural speech**
- Annotations made in the experiment are less extreme than in the naturalistic database

**Settings**
- Fear synthesized with EmoSpeak less extreme values than EmoFilt and copy synthesis
- Sad synthesized with copy synthesis less extreme values than EmoFilt and EmoSpeak
Results (4) **Dimensional annotations**

**Comparison with neutral synthesis**
- Modification of parameters results in more extreme arousal values than no modification
- Valence values lie (with and without settings) around zero

**Semantics**
- Semantics do not influence the dimensional annotations