Towards More Reality In The Recognition Of Emotional Speech

Björn Schuller[1], Dino Seppi[2], Anton Batliner[3], Andreas Maier[3], Stefan Steidl[3]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2]ITC-irst, Trento, Italy [3]Institute for Pattern Recognition, Universität Erlangen-Nürnberg, Germany

Introduction

Facing **new challenges** in automatic emotion recognition based on speech:

- speaker independence
- spontaneous speech with naturalistic emotions
- difficult noise and microphone conditions

Databases

Acted Data

1. Danish Emotional Speech Database (DES)
- 4 emotions: anger, joy, sadness, and surprise plus neutral
- 4 professional Danish actors (2 m, 2 f)
- words "yes" and "no", 9 sentences, 2 text passages
- perception test (20 persons): 67.3 % accuracy

2. Berlin Emotional Speech Database (EMO)
- 6 emotions: anger, disgust, fear, joy, sadness, boredom plus neutral
- 10 professional German actors (5 m, 5 f)
- 10 sentences of emotionally undefined content
- selection of 494 phrases: more than 60 % natural, at least 80 % clearly assignable in perception tests
- perception test (20 persons): 84.3 % accuracy

Spontaneous Data

3. AIBO Emotion Corpus
- 51 children (21 m, 30 f) communicating with Sony's pet dog Aibo
- spontaneous speech
- 11 user state labels, majority voting of 5 labelers on word level
- selection of 3990 turns, 4-class problem: motherese, neutral, emphatic, anger (cover class for angry, touchy/irritated, reprimanding)
- classroom recordings with a wireless head-set microphone
- additional audio stream of the video camera

Noise and Microphone Conditions

1. Acted Data:
- studio recordings + additive noise overlay at different SNR levels

2. Spontaneous Data:
- close-talk microphone (CT)
- artificial reverberation: CT data convoluted with different impulse responses (CTRV)
- audio data of the video camera: real noise and reverberation (RM)

Features and Classification

Two Different Feature Sets:
- Feature set '**Set 1**'
  - broad feature set (≈ 4000) for subsequent feature selection
  - covering prosodic, articulatory, and voice quality aspects
  - calculated on turn level by applying functionals (Table 2) to the base contours (Table 1)
- Feature set '**Set 2**'
  - compact knowledge-based prosodic set: 26 features
  - supra-segmental prosodic features
  - calculated at different levels:
    - ⋆ **word level:** segmentation by manual annotation, automatic forced alignment of the transliteration, or automatic speech recognition
    - ⋆ **chunk level:** chunks of variable length

contour	Set 1	Set 2
log-energy	√	√
pitch	√	√
duration	√	√
harmonics-to-noise ratio	√	-
pos., bandwidth & ampl. of formants	√	-
jitter and shimmer	√	-
16 MFCCs	√	-
spectral flux, centroid, 95%-roll-off	√	-

Table 1: Extracted acoustic base-contours.

functional	Set 1	Set 2
mean & standard deviation	√	√
centroid	√	-
skewness & kurtosis	√	-
quartiles	√	-
ranges	√	-
extremes & relative positions	√	√
zero-crossing-rate	√	-
roll-off-points	√	-
lin. regr. coefficients & error	√	√
quadratic regr. coefficients	√	-

Table 2: Applied functionals for acoustic feature calculation.

Classification:
- random forests
- 2-fold speaker-independent cross-validation

Experiments

Acted Data: DES and EMO, features 'Set 1'

[%]	∞ dB	20 dB	10 dB	0 dB	-5 dB	-10 dB
	Danish Emo. DB (5 classes)					
RR	53.5	51.3	46.6	44.3	43.7	41.6
CL	54.3	51.2	46.5	43.8	43.3	41.5
	Berlin Emo. DB (7 classes)					
RR	72.3	71.7	67.6	64.5	64.3	62.9
CL	67.4	65.6	61.9	58.7	58.5	56.5

Table 3: Accuracies at selected SNR levels using all features. RR: recognition rate, CL: mean class-wise RR.

Acc. [%]	Danish Emo. DB		Berlin Emo. DB	
feat. sel.	all	*n* best	all	*n* best
∞ dB	53.5	57.1	72.3	72.5
-10 dB	41.6	49.4	62.9	66.8

Table 4: Accuracies with all features and a selection of the *n* best features at two selected SNR levels.

Spontaneous Data: AIBO, 'Set 1' and 'Set 2'

[%]	C1	C2	C3	C4	C5	C6
Feature Set	Set 1	Set 2	Set 2	Set 2	Set 2	Set 2
Segmentation	TL	MA	VL	VL	FA	AR
Transcription	-	MA	MA	-	MA	AR
	close-talk (CT)					
RR	51.3	53.5	51.7	49.6	49.2	50.0
CL	46.2	51.0	51.0	47.9	46.7	47.1
	close-talk reverberated (CTRV)					
RR	46.6	52.8	50.9	48.9	49.8	49.5
CL	43.1	50.6	50.5	48.7	47.3	48.3
	room microphone (RM)					
RR	40.0	52.0	50.3	48.6	49.3	47.0
CL	35.0	49.4	49.7	47.2	48.9	45.7

Table 5: Accuracies under different noise and microphone conditions, diverse feature combinations C1-C6, MA manual annotation, VL variable length, TL turn-level, FA forced alignment, and AR recognizer output. 'Set 1' features are reduced to 105 (CT), 90 (CTRV), 94 (RM).

Summary of the Results

1. Acted Data
- additive noise overlay: significant decrease in accuracy for each step (cf. Table 3)
- reduction of the feature set helps to improve performance (cf. Table 4)
- but: reduced feature sets differ largely at various noise levels

2. Spontaneous Data
- only minor influence of noise and reverberation on feature set 'Set 2' (C2-C6, Table 5)
- under bad conditions, the word-based feature set 'Set 2' clearly outperforms the turn-based feature set 'Set 1' (C2-C6 vs. C1)
- in contrast to speech recognition, (word-based) emotion recognition is robust against noise
- only little influence of the segmentation
- best results with manual transliteration and manually corrected segmentation (C2)