Whence and Whither: The Automatic Recognition of Emotions in Speech

PIT 2008, 17.6.2008

Anton Batliner

Lehrstuhl für Mustererkennung (Informatik 5)

Friedrich-Alexander-Universität Erlangen-Nürnberg

Overview

- whence: a short history
- terminology: what "is" *emotion*
- *corpus engineering*
  - scenario
  - annotation
  - segmentation
- features & feature selection
- classification, evaluation
- whither: progress and applications
- the future

Whence: A Short History

- basic studies on emotion in speech from the 20ies onwards
- bulk of basic research from the 80ies onwards
- first studies on automatic recognition of acted emotions
  mid 90ies
- first studies using "realistic" data
  end of the 90ies & this century $\Rightarrow$ tenth anniversary
- still, too many studies using acted data

What "is" Emotion: Terminological Remedies

- {despair, fear, joy, ...} vs. {stress, tiredness, interest, ...}
  = (full-blown) emotion vs. mood, ...
- cover terms used:
  - emotions ("emotional intelligence")
  - emotion-related states
  - affective states
  - user states (changing over time) vs. user traits (stable)
  - **pervasive emotions** (ex negativo: *whatever present in most of life but absent when people are emotionless*)
- term not used:
  - (speech) register (*subset of a language used for a particular purpose or in a particular social setting*)

Corpus Engineering: Scenario, Design

- sub-optimal but common breakdown:
  $\pm$ {acted, induced, natural(istic)}
- vs. $\pm$ {natural, spontaneous}
- representative non-acted/non-prompted scenarios
  - mother-child interaction
  - call-center interaction
  - stress detection in driving scenario
  - human-machine (computer/robot) interaction
  - tutoring dialogues
  - appointment scheduling dialogues
  - human-human multi-party interaction

Corpus Engineering: Annotation

- basics: orthographic transcription plus lexicon
- labellers: experts or naïve, how many
- catalogue of terms
  - deduced (catalogue of terms) or data-driven
  - categories or dimensions
  - mixed ("early or late mixture") or pure
- assessment
  - reliability (kappa etc. vs. classifier performance)
  - external ground truth
  - ecological validity

Corpus Engineering: Segmentation

- normally, **units of analysis** not addressed
  - either given trivially (acting using semantically void sentences or short dialogue moves)
  - or defined on an intuitive basis
- beneficial/necessary in case of longer utterances
- emotion units correlate with
  - prosodic units (intonation units and/or pauses)
  - syntactic-semantic units & dialogue acts
  - size depending on register (e.g. in *FAU-Aibo*, 2.7 words/unit)
- better performance & time constraints

  within end-to-end system (time-alignment)

Corpus Engineering: Selection of Cases

- almost never:
  - random selection
  - using all cases
- choosing more prototypical cases (majority voting)
- up- or down-sampling
- usually no rejection class

Features

- extraction
  - frame-based (ASR), segment-based, global
  - manual or automatic
- low-level (or high-level: sub-optimal)
- raw or normalized (implicitly or explicitly)
- types
  - Low Level Descriptors LLDs: acoustic, linguistic
  - functionals

Features

- acoustic LLDs
  - voice quality (jitter/shimmer, HNR)
  - pitch
  - spectrum and formants
  - cepstrum (MFCC features from ASR)
  - energy
  - duration
  - and: Teager operator, dynamic features for HMM, ...
- linguistics LLDs
  - non-verbals, disfluencies
  - stemming, e.g. part of speech (POS)
  - bag of words (from document retrieval tasks)
  - n-grams: (mostly uni-grams)

Functionals

- percentiles (e.g. quartiles)
- specific functions (e.g. regressional)
- extremes: (e.g. min/max)
- higher statistical moments: (e.g. std. dev.)
- means
- sequential and combinatorial (e.g. two functionals applied, e.g. mean of max)

⇒ CEICES feature encoding scheme

Feature Selection

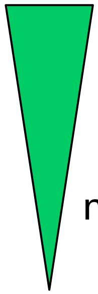
- # features: some ten ⇒ some thousands
  - small number of pattern + high number of features
    ⇒ reduction + selection necessary
- feature reduction: PCA, LDA, ICA, ...
- feature selection
  - wrapper, filter methods such as IGR
  - sequential forward SFS or backward selection SBS, ...,
    Sequential Floating Forward Selection SFFS

The "Best" Feature Vector?

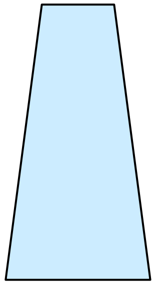
- a *holy grail* - there is no "best" feature vector
- depends on what's available
  - spoken word chain or ASR result
  - manual and/or automatic extraction of feature values
  - ...
- (high) correlation between (types of) features
  $\Rightarrow$ some (any?) combination will be adequate

The "Best" Feature Vector: a Suggestion

- types of acoustic features
  - energy
  - MFCC
  - duration
  - pitch
  - voice quality, spectrum
- any linguistic information
- types of functionals
  - means (robustness)
  - extremes
  - higher statistical moments (regression)
  - layered (combining smaller and larger units of analysis)



varied data
speaker-independent
no full-blown emotions



uniform data
(synthesis)
personalized
sadness vs.
anger

Classification: Methods

- pattern recognition & data mining
- availability of tools such as WEKA and HTK
- from changing standard methods to multiplicity
  - Linear Discriminant Classifiers, Nearest Neighbour, Decision Trees (Random Forests) , Artificial Neural Networks
  - Naive Bayes, Support Vector Machines, Hidden Markov Models, Ensembles, ...
- Regression
- Fusion: early or late

Classification: Evaluation

- train vs. test
  - leave-one-case out
  - leave-one-part out (10-fold cross-validation in WEKA)
  - leave-one-speaker-out
  - stratified cross-validation
  - train + validation + test
- measures
  - recognition rate RR
  - class-wise computed recognition rate CL
  - ...
  - basis: confusion matrix

Classification: Assessment

- a simple ranking of classifier performance is not very enlightening
- worse classifier = less classifier tuning
- *any classification method is as good provided a good feature vector*
- *there is no free lunch*
- out-of-the-box procedures are competitive
- standards nowadays e.g.: SVM and ensemble methods

Whither: an Epistemological Loop?

- basic research
  - clash of cultures
  - if based on acted data, no transfer to analysis/recognition of real data
- real data necessary
- from real data to real(istic) applications
  - proof of the pudding
  - not generic, too focused?
  - recall & false alarm rate can/should be different for different applications

Applications: a Taxonomy

System Design

online	system reacts (immediately/delayed) while interacting with user
offline	no system reaction, or delayed reaction after actual interaction
mirroring	user gets feedback as for his/her emotional expression
non-mirroring	system does not give any explicit feedback
emotional	system reacts itself in an emotional way
non-emotional	system does not behave emotionally but “neutral”

Meta-assessment

critical	application's aims are impaired if emotion is processed erroneously
non-critical	erroneous emotion processing does not impair application's aims

Emotional Monitoring (1)

Main scenario: Call Center

Anger detection with discontent users
Quality of service

Interaction with user
Quality control of agent

Fully automatic system
Handing over to agent

online
offline

mirroring
non-mirroring

emotional
non-emotional

critical
non-critical

⇒ necessary: <u>very</u> low false alarm rate

Emotional Monitoring (2)

Main scenario: Call Center

Anger detection with discontent users
Quality of service

Interaction with user
Quality control of agent

Fully automatic system
Handing over to agent

⇒ beneficial: high recall (and not too many false alarms)

online
offline

mirroring
non-mirroring

emotional
non-emotional

critical
non-critical

Emotional Monitoring (3)

Main scenario: Call Center

Anger detection with discontent users
Quality of service

Interaction with user
Quality control of agent

Fully automatic system
Handing over to agent

⇒ screening, i.e. average recall over time should and can be close to human performance - but don't forget the trade unions!

online
offline

mirroring
non-mirroring

emotional
non-emotional

critical
non-critical

Nothing Said About yet

- performance - that's what you remember
  - state-of-the-art
    - > 60% for 4 classes, ~80% for 2 classes
    - ~~up to > 80% for 4 classes and > 90% for 2 classes~~
  - worse with ASR, esp. linguistic features
  - what about rejection classes?
- generation and synthesis
  - ECAs are speaker-dependent and want to *convey* emotions
  - our field is mostly speaker-independent and we want to *recognize* emotions
  - bridging not possible yet

The Future

- more natural data needed
- taking into account realistic applications
- fine-tuning classification will help - but not much
- examples of promising applications
  - call centers - millions of calls, pays off
  - autismus therapy - adequate treatment
  - tamagotchi - "meaningless" and successful
  - MP3-player - yet another feature
  - the sex industry - real money
- the rubicon: higher performance than now, lower than needed for dictation systems
- the challenge: to bring together
  - analysis & synthesis
  - basic research & applications

Some Take-Away Messages

- acted data are not useful
- segmentation into emotion units is necessary
- the "best" feature vector might consist of some combination
- there is no best classifier
- let's use ASR and all cases

and two more caveats: it’s no panacea

- to employ voice quality features: they are speaker-dependent and multi-functional
- to employ multi-modality (= facial gestures and speech): sequential multi-modality is ok, but simultaneous multi-modality will mostly not pay off because ground truth is vague and partly antagonistic

Acknowledgments

many thanks to my colleagues in the CEICES initiative

⇒

The Automatic Recognition of Emotions in Speech

Anton Batliner, Björn Schuller, Dino Seppi, Stefan Steidl,
Laurence Devillers, Laurence Vidrascu, Thurid Vogt,
Vered Aharonson, Noam Amir

in

HUMAINE Handbook on Emotion. ed. Paolo Petta et al., Springer, 2009

Thank you for your attention