# emotional speech

Advanced Signal Processing
Winter Term 2003

franz zotter

# contents

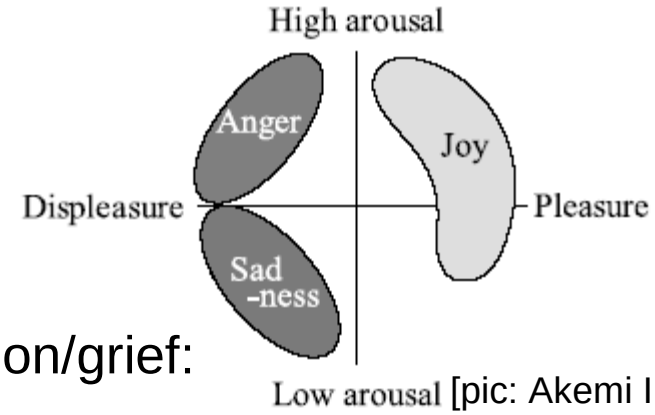
- emotion psychology
- articulation of emotion
  - physical, facial
  - speech
    - acoustic measures
    - features, recognition
    - affect bursts
- emotional speech detection/ synthesis
  - applications
  - synthetic feature generation methods
  - HMM (hidden markov models)
  - neural network models
  - available systems

# emotion psychology

- C. Darwin: archetypes of emotion, biological survival reasons (anger, disgust, fear, sadness, surprise, happiness)
- W. James: biological reasons, feel bodily changes: *“we are afraid because we tremble”*
- Cognitive: (M. Arnold) emotion determined by *appraisal* (consider: novelty, pleasantness, responsibility, effort)
- Social Constructivist: (J. Averill) cultural based emotional behaviour, social rules and moral values

# physical features of emotion

autonome nervous systems
(sympathetic/parasympathetic)



[pic: Akemi Ilda]

fear/anger:

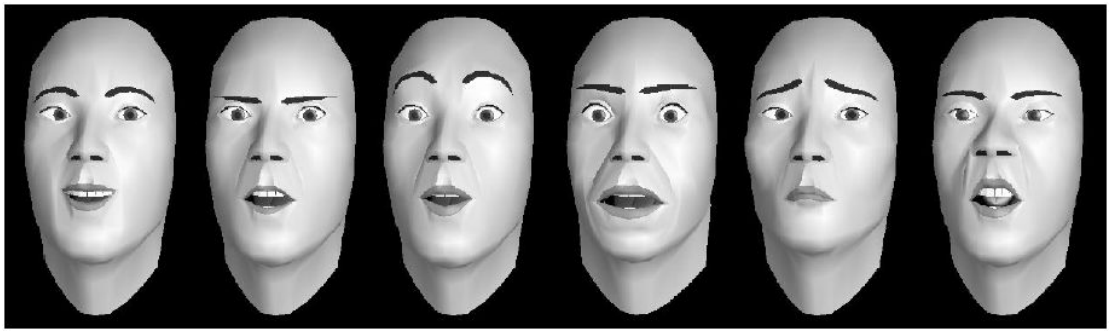
- short respiration cycles
- respiration rhythm irregular
- high subglottal pressure
- dry mouth
- muscle tremor
- high blood pressure and heart rate

relaxation/grief:

- smooth respiration cycles and rhythm
- low subglottal pressure
- Increased salivation
- low blood pressure and heart rate

[Janet E. Cahn]

# facial features of emotion



happiness, anger, surprise, fear, sadness, disgust

- facial expression is very accurate in recognition
- carries conscious (controlled) and unconscious information about emotion

## gestures / postures

- gestures mostly with hands and motion
- posture: e.g. turn sb. so‘s back, crossing arms on the chest, etc.

# emotion in speech (2)

impacts on speech:

[Janet E. Cahn]

fear/anger:

- increased speed and loudness
- higher pitch
- expanded pitch range
- disturbed speech rhythm
- precise articulation
- increased higher frequency energy

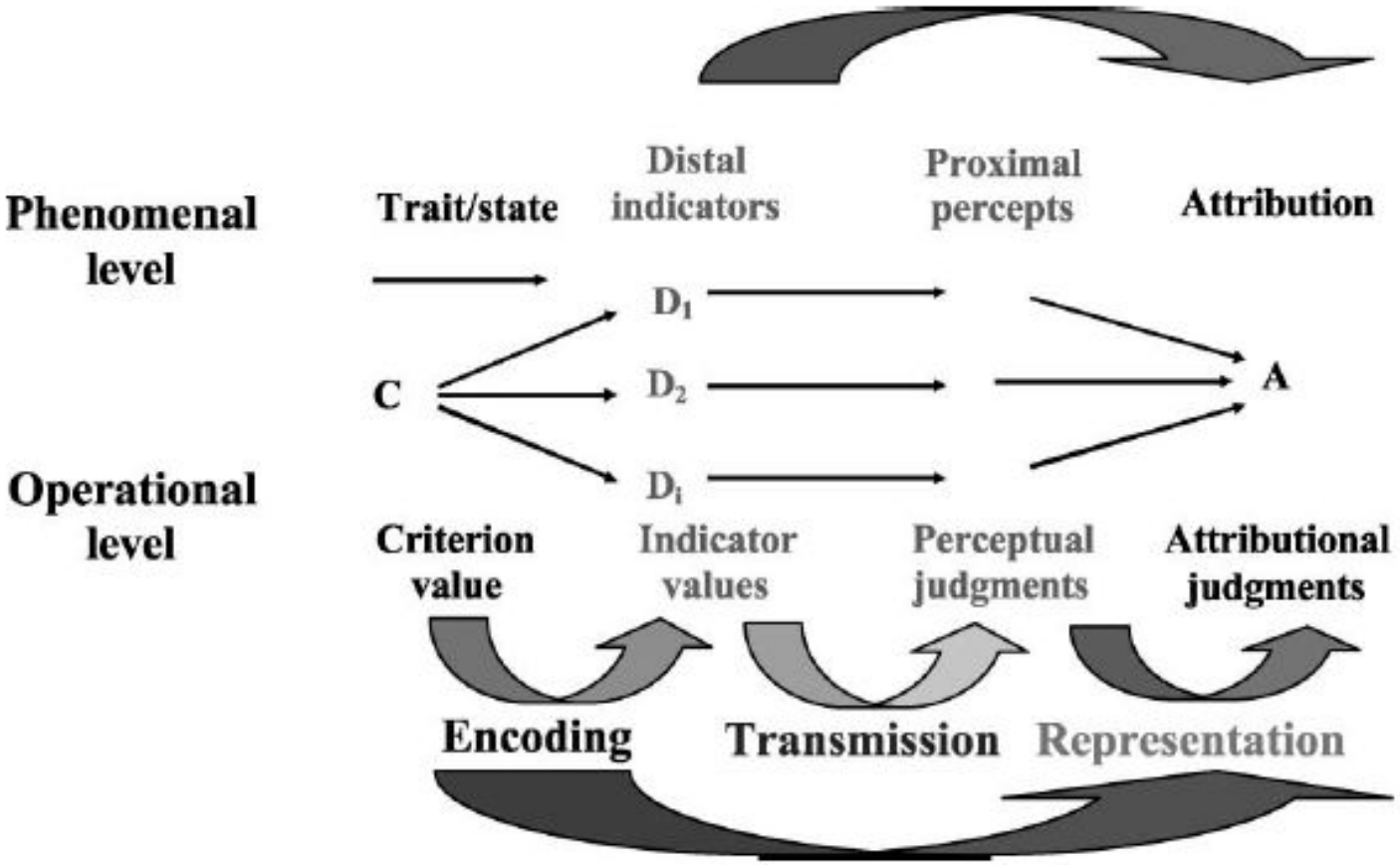
relaxation/grief:

- low speed and loudness
- low pitch
- smaller pitch range
- smooth speech rhythm, fluent speech
- imprecise articulation: formant change towards schwa
- decreased higher frequency energy

# emotion in speech (1)

[Klaus Scherer,

Brunswikian
Lens model]



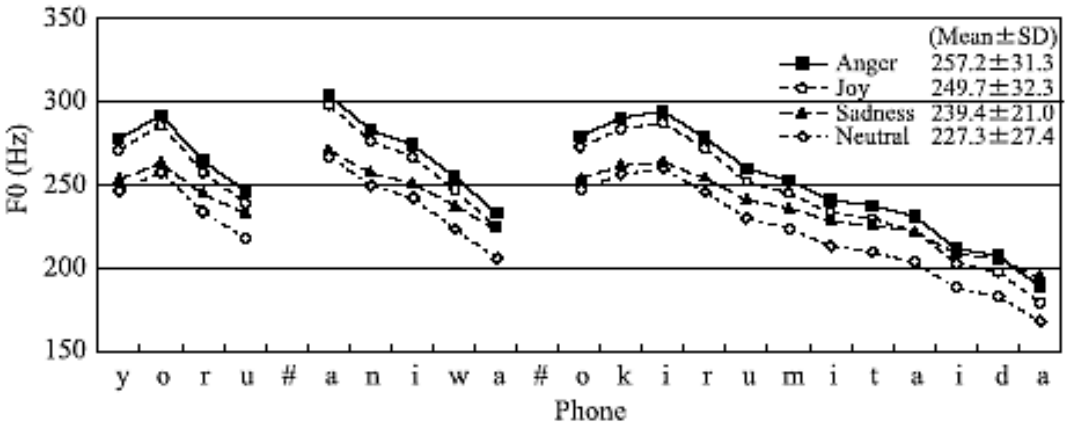
# acoustic measures

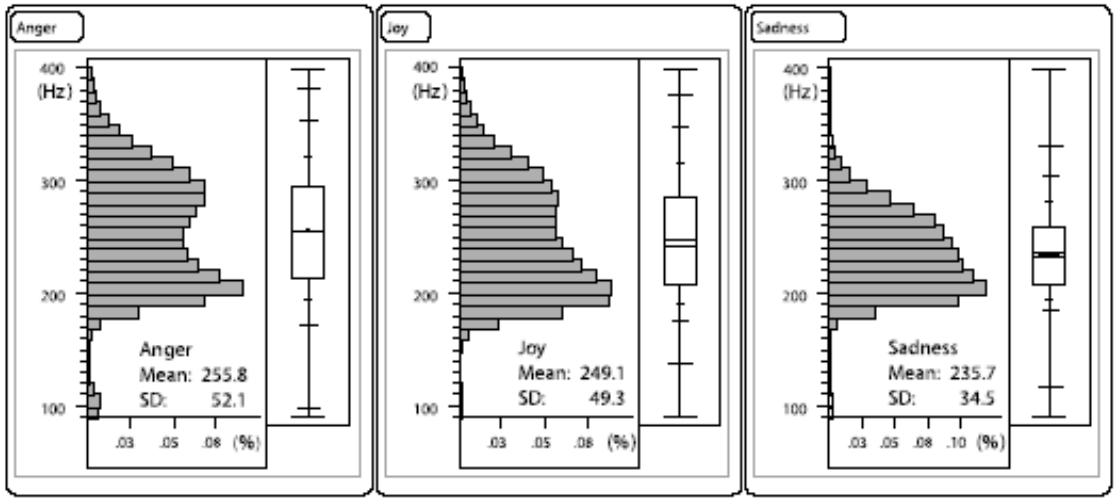
- pitch (F0): [Janet E. Cahn]
  - pitch range
  - pitch average
  - contour slope (up/down)
  - accent shape and range
- harmonicity:
  - breathiness: amount of respiration noise
  - Laryngealisation: due to small subglottal pressure (narrow pulse shape, irregular period)
  - tremor / jitter: irregular pitch period T
- brilliance (energy ratio between high and low frequencies)
- loudness (psycho-acoustic weighing)
- timing:
  - intensity contour (pauses, hesitation)
  - word duration
  - vowel / consonant duration
  - intensity of plosive bursts
- spectral information:
  - formant positions, bandwidths
  - articulation precision

# acoustic measures (1/4)

[Janet E. Cahn]

- pitch (F0):
  - pitch range
  - pitch average
  - contour slope (up/down)
  - accent shape and range

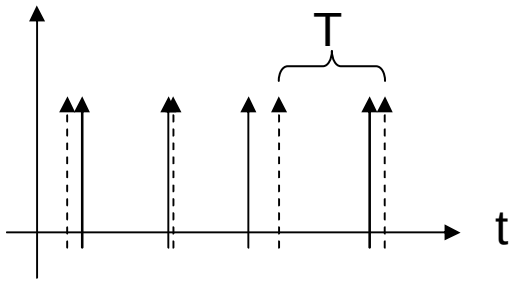



[Akemi Iida, pitch contour, histogram]

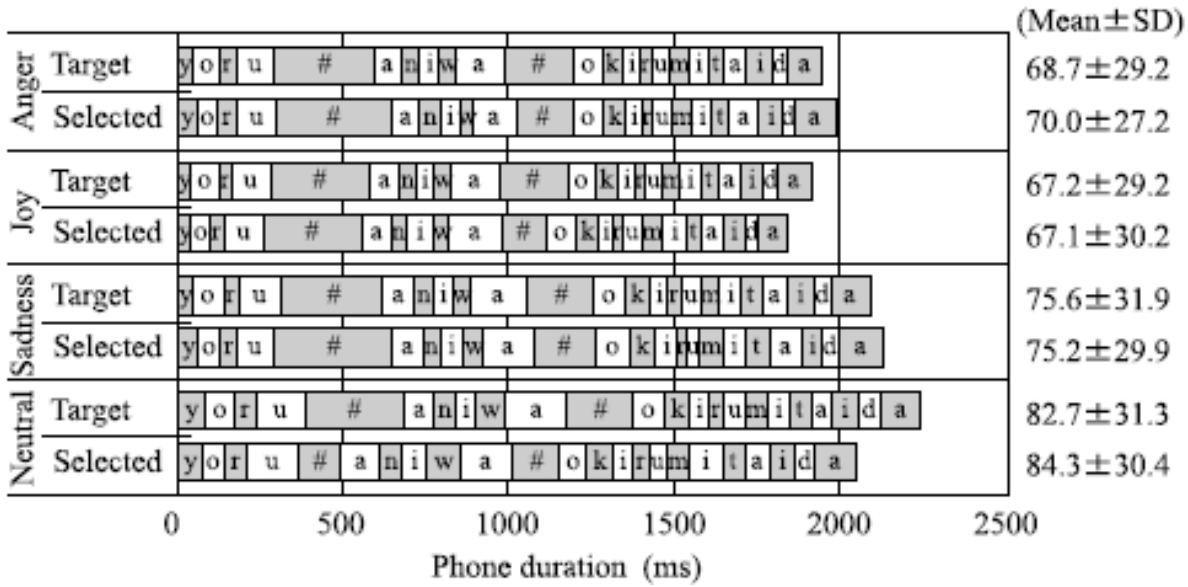
# acoustic measures (2/4)

- harmonicity:
  - breathiness: amount of respiration noise
  - Laryngealisation: due to small subglottal pressure (narrow pulse shape, irregular period)
  - tremor / jitter: irregular pitch period T



# acoustic measures (3/4)

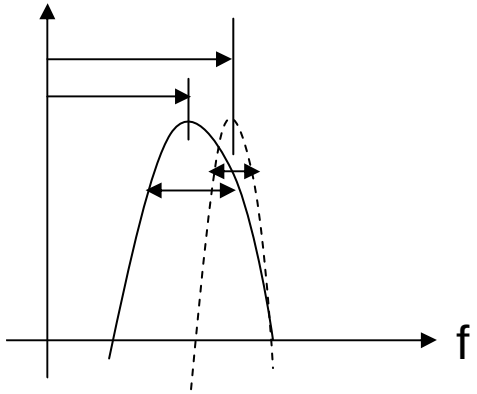
- brilliance (energy ratio between high and low frequencies)
- loudness (psycho-acoustic weighing)
- timing:
  - intensity contour (pauses, hesitation)
  - word duration
  - vowel / consonant duration
  - intensity of plosive bursts

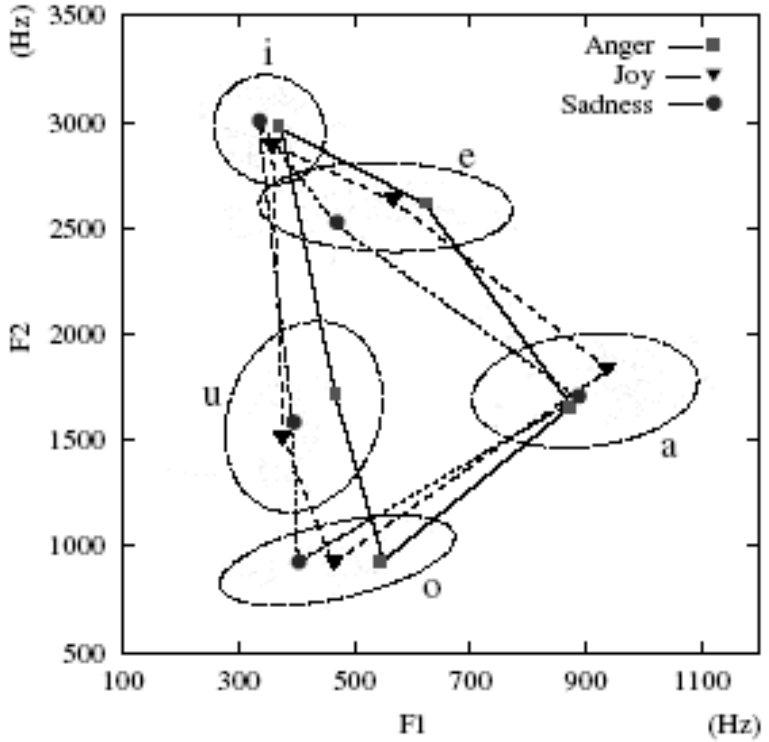



[Akemi Iida, phone duration]

# acoustic measures (4/4)

- spectral information:
  - formant positions, bandwidths due to small lip's opening or
  - articulation precision differences due to speaker's arousal




[Akemi Iida, vowel position]

# features of emotional speech (1)

[Klaus Scherer]

Synthetic compilation of the review of empirical data on acoustic patterning of basic emotions (based on Johnstone and Scherer, 2000)

| | Stress | Anger/rage | Fear/panic | Sadness | Joy/elation | Boredom |
|---|---|---|---|---|---|---|
| Intensity | ↗ | ↗ | ↗ | ↘ | ↗ | |
| F0 floor/mean | ↗ | ↗ | ↗ | ↘ | ↗ | |
| F0 variability | | ↗ | | ↘ | ↗ | ↘ |
| F0 range | | ↗ | ↗(↘) | ↘ | ↗ | ↘ |
| Sentence contours | | ↘ | | ↘ | | |
| High frequency energy | | ↗ | ↗ | ↘ | (↗) | |
| Speech and articulation rate | | ↗ | ↗ | ↘ | (↗) | ↘ |

Accuracy (in %) of facial and vocal emotion recognition in studies in Western and Non-Western countries (reproduced from Scherer, 2001)

| | Neutral | Anger | Fear | Joy | Sadness | Disgust | Surprise | *Mean* |
|---|---|---|---|---|---|---|---|---|
| Facial/Western/20 | | 78 | 77 | 95 | 79 | 80 | 88 | 78 |
| Vocal/Recent Western/11 | 74 | 77 | 61 | 57 | 71 | 31 | | 62 |
| Facial/Non-Western/11 | | 59 | 62 | 88 | 74 | 67 | 77 | 65 |
| Vocal/Non-Western/1 | 70 | 64 | 38 | 28 | 58 | | | 52 |

*Note:* Empty cells indicate that the respective emotions have not been studied in these regions. Numbers following the slash in column 1 indicate the number of countries studied.

# features of emotional speech (2)

[Klaus Scherer]

Predictions for emotion effects on selected acoustic parameters (based on Table 4 and appraisal profiles; adapted from Scherer, 1986)

| | ENJ/ HAP | ELA/ JOY | DISP/ DISG | CON/ SCO | SAD/ DEJ | GRI/ DES | ANX/ WOR | FEAR/ TER | IRR/ COA | RAG/ HOA | BOR/ IND | SHA/ GUI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *F0* | | | | | | | | | | | | |
| Perturbation | <= | > | | | > | > | | > | | > | | |
| Mean | < ✔ | > ✔ | > | <> | <> ✔ | > ✔ | >? | >> ✔ | <> ✔ | <> | < ✔ | >? |
| Range | <= | > | | | < | > | | >> | < | >> | | |
| Variability | < | > | | | < | >? | | >>? | < | >> ✔ | | |
| Contour | < | > | | | < | > | > | >> | < | = | | > |
| Shift regularity | = | < | | | | | | < | | < | > | |
| *Formants* | | | | | | | | | | | | |
| F1 Mean | < | < | > | > | > | > | > | > | > | > | > | > |
| F2 Mean | | | < | < | < | < | < | < | < | < | < | < |
| F1 Bandwidth | > | <> | << | < | <> | << | < | << | << | << | < | < |
| Formant precision | | > | > | > | < | > | > | > | > | > | | > |
| *Intensity* | | | | | | | | | | | | |
| Mean | < ✔ | > ✔ | >? | >>? | << ✔ | > ✔ | | > ✔ | > ✔ | >> ✔ | <> | |
| Range | <= | > | | | < | | | > | > | > | | |
| Variability | < | > | | | < | | | > | | > | | |
| *Spectral parameters* | | | | | | | | | | | | |
| Frequency range | > | > | > | >> | > | >> | | >> | > | > | > | |
| High-frequency energy | < | <> ✔ | > | > | <> | >> ✔ | >? | >> | >> | >> ✔ | <> | > |
| Spectral noise | | | | | > | | | | | | | |
| *Duration* | | | | | | | | | | | | |
| Speech rate | <? | > ✔ | | | < ✔ | > | | >> ✔ | | > ✔ | | |
| Transition time | > | < | | | > | < | | < | | < | | |

*Note:* ANX/WOR: anxiety/worry; BOR/IND: boredom/indifference; CON/SCO: contempt/scorn; DISP/DISG: displeasure/disgust; ELA/JOY: elation/joy; ENJ/HAP: enjoyment/happiness; FEAR/TER: fear/terror; GRI/DES: grief/desperation; IRR/COA: irritation/cold anger; RAGE/HOA: rage/hot anger; SAD/DEJ: sadness/dejection; SHA/GUI: shame/guilt; F0: fundamental frequency; F1: first formant; F2: second formant; >: increase; <: decrease. Double symbols indicate increased predicted strength of the change. Two symbols pointing in opposite directions refer to cases in which antecedent voice types exert opposing influence. (✔) prediction supported, (?) prediction contradicted by results in (Banse and Scherer, 1996).

Affect burst classes within each intended emotion

| Intended emotion | Affect burst class | expert transcription | | Listening test | | | Written perception test | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Segments, voice qual. | Intonation | No. of stimuli | Emotion recognised | Recogn. rate | Orthographic transcription | Transcr. variability | Emotion recognised | Recogn. rate |
| Admiration | Wow | [ɪːə̯] | 3–1 | 4 | ✓ | 91% | wow | 3 | ✓ | 90% |
| | Boah | [bo̤a̤ː] | 1 | 4 | ✓ | 90% | boah | 2 | ✓ | 90% |
| Threat | Hey | [hɛɪ] | 3–2 | 5 | ✓ | 81% | ej | 3 | ✓ | 65% |
| | Growl | [m̩ː] | 1 | 2 | Anger | 80% | mrr | 8 | Anger | 50% |
| Disgust | Buäh | [bʊæ̯ː] | 3–2 | 6 | ✓ | 92% | uäh | 1 | ✓ | 63% |
| | Igitt | [iːgɪtʰ] | 3 | 1 | ✓ | 100% | igitt | 1 | ✓ | 100% |
| | Ih | [ɪːə̯] | 3–2 | 1 | ✓ | 95% | irgh | 29 | ✓ | 84% |
| Elation | Ja | [jaː] | 3 | 4 | ✓ | 69% | jaaa | 2 | ✓ | 47% |
| | Yippie | [jɪpiː] | 4–3 | 2 | ✓ | 100% | jippii | 1 | ✓ | 100% |
| | Hurra | [huʁaː] | 4–3 | 2 | ✓ | 80% | hurra | 0 | ✓ | 95% |
| Boredom | Yawn | | 3–1 | 4 | ✓ | 81% | uuahh | 20 | Startle | 53% |
| | Sigh | [ə̯ː] | 2–1 | 2 | ✓ | 45% | hmm | 12 | ✓ | 63% |
| | Hmm | [mː] | 1–2 | 2 | ✓ | 83% | mmh | 7 | ✓ | 60% |
| Relief | Sigh | [a̤ː] | 2–1 | 3 | ✓ | 85% | ahh | 5 | ✓ | 50% |
| | Uff | [ʊfː] | 2–1 | 3 | ✓ | 98% | uff | 6 | ✓ | 80% |
| | Puh | [pʰu̟ɸː] | 3–1 | 2 | ✓ | 95% | puh | 1 | ✓ | 85% |
| Startle | Int. breath | | 3 | 6 | ✓ | 92% | he | 8 | Threat | 40% |
| | Ah | [a] | 3 | 2 | ✓ | 80% | a | 8 | Relief | 37% |
| Worry | Oje | [o̤je̤ː] | 2–1 | 4 | ✓ | 96% | ujeh | 6 | ✓ | 75% |
| | Oh–Oh | [ʔoʔːoː] | 3–2 | 2 | ✓ | 85% | o-oh | 6 | ✓ | 67% |
| | Oweh | [o̤ːβe̤ː] | 3–1 | 1 | ✓ | 50% | oh jee | 14 | ✓ | 85% |
| | Hmm | [m̩m] | 2–1 | 1 | ✓ | 70% | hmm | 9 | Boredom | 63% |
| Contempt | Laughter | [ħ̩ʱħ] | 1 | 5 | ✓ | 77% | hähä | 10 | ✓ | 74% |
| | Pha | [phaʔ] | 1 | 2 | ✓ | 95% | pah | 4 | ✓ | 95% |
| | Tse | [tsʰə] | 3 | 1 | ✓ | 100% | tse | 5 | ✓ | 85% |
| Anger | Growl | [m̩ː] | 2–1 | 4 | ✓ | 69% | ahr | 2 | ✓ | 39% |
| | Breath out | [ħː] | 1 | 3 | ✓ | 55% | chrr | 8 | ✓ | 39% |
| | Oh | [ɔ̤ː] | 2–1 | 1 | ✓ | 45% | ooh | 4 | Admiration | 53% |

The 'emotion recognised' columns indicate the most frequent answer for that affect burst in the respective test ('✓' = intended emotion). Recognition rates are given for that most frequent answer. 'int. breath' designates a rapid intake of breath.

# *affect bursts*

[Marc Schröder]

definition:

short emotional non-speech expressions.

assigned emotions are often easily recognized

# emotion detection and synthesis

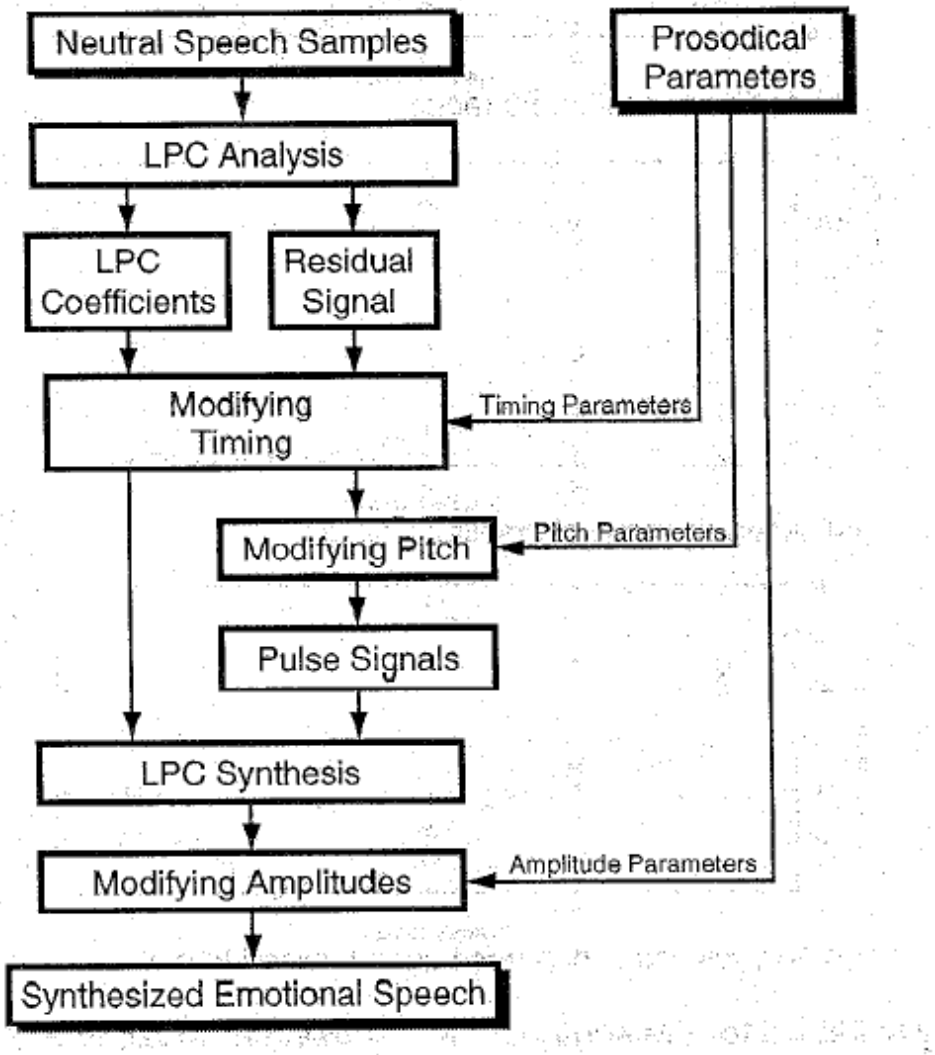
applications:

- automatic dialog systems (trouble recognition)
- emotion analysis:
  - pathologic purposes (schizophrenia, parkinson, …)
  - forensic purposes (lie detection)
- speech driven facial animations
- TTS (text-to-speech) synthesis with emotion (context, xml)
- speech manipulation (conversion)

# synthetic feature generation

- *affect burst* insertion
- residual excitation manipulations (source-filter models: LPC, …):
  - pitch manipulation (MBROLA, PSOLA, RP-PSOLA, …)
    - timing
    - accents, pitch slope, F0 interpolation
    - pitch shift
  - jitter processing
  - additive noise (breathiness)
  - ring modulation (spectral shift -> harmonicity)
  - linear filtering (emphasis)
  - wave-shaping (exciter: higher harmonics, non-linear)
- envelope modulation: (pauses, hesitation, plosive bursts, stressed words)
- spectral modification:
  - formant positions and bandwidth rearrangement
  - emphasis (brilliance)
  - frame rearrangement (timing, diphone transitions)
  - reflection coefficient interpolation (LPC, articulation precision)

# synthetic feature generation




neutral to emotional speech synthesizer

[Jun Sato, residual signal driven emotional speech synthesizer]

# HMM (hidden markov model)

[Hansen, Ghazale]

$$P(Q|O,\lambda)$$

lambda: model parameters (propabilities for observed states with respect to their history)

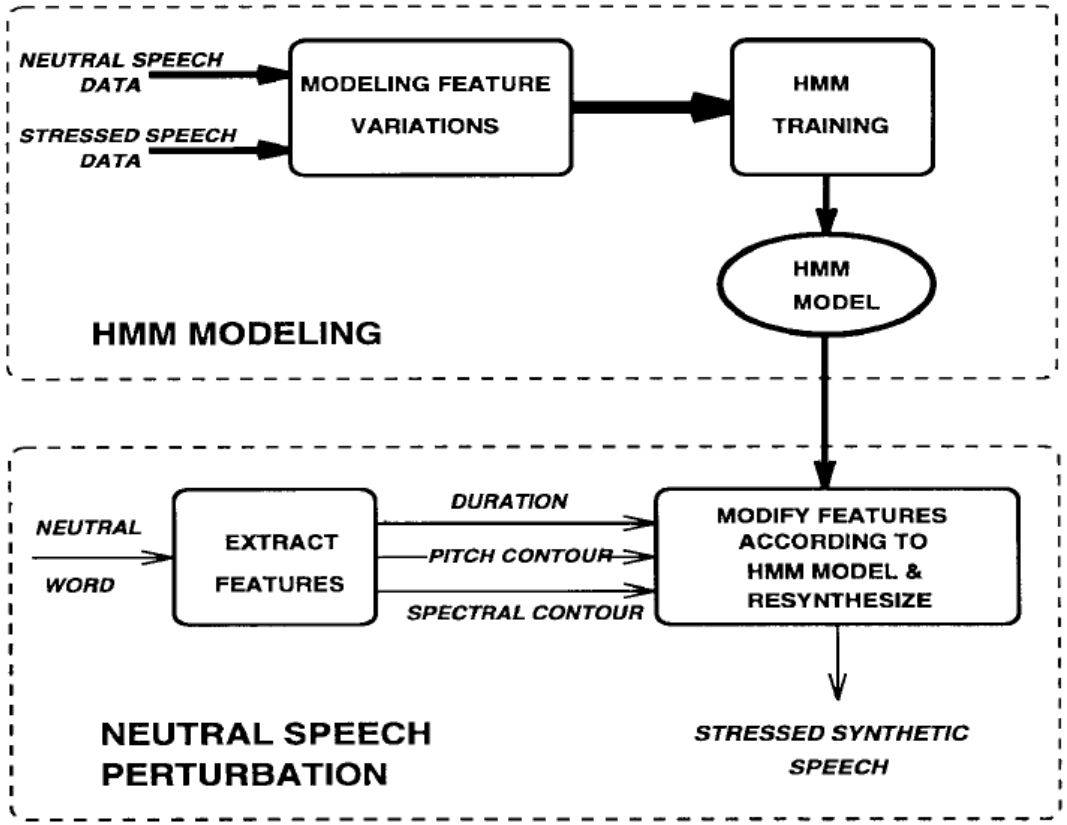
O: observed prosodic parameter sequence

Q: emotional state of the listener

The HMM training estimates all model parameters lamda

In the end you can:

- detect emotions by propability measures
- create prosodic features with the viterbi algorithm



# neural network models

[Jun Sato]

task of each node: arousal of output nodes (next layer) due to the input arousal (previous layer)

emotion space (3rd layer output):

2 nodes: 2dimensional emotion space:

- emotion intensity
- emotion type

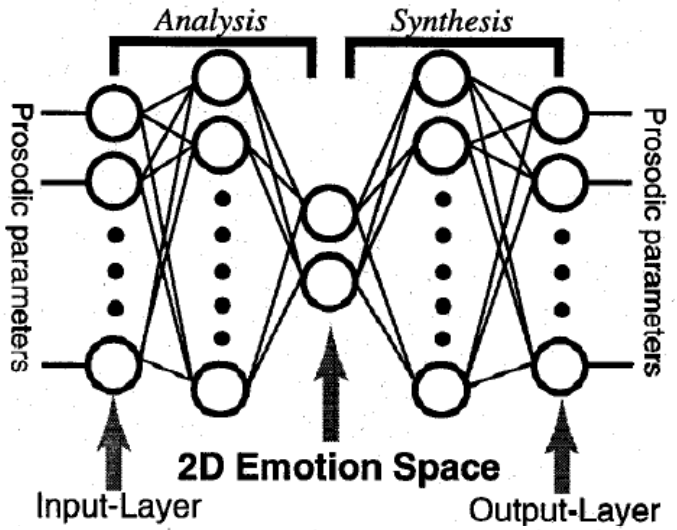
The node‘s in- and output behavior has to be estimated in a training process.

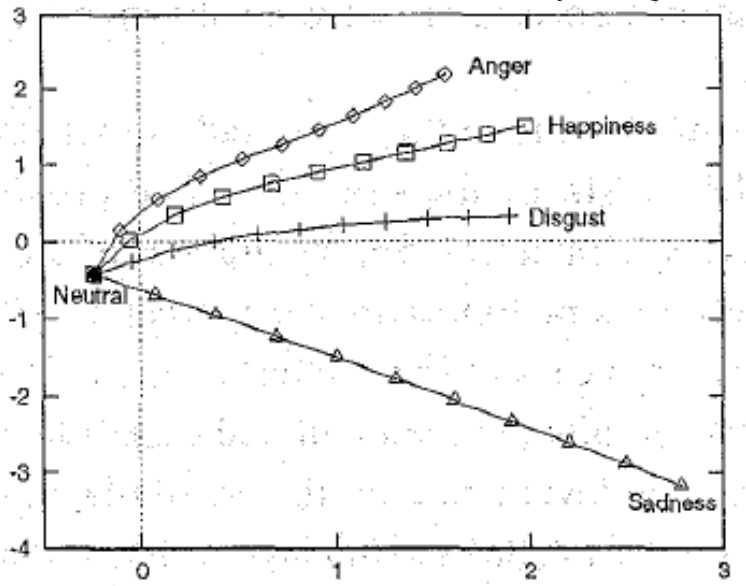
tasks:

- emotion detection from prosodic params
- given emotion ->prosodic parameter generation

[picture: example for one sentence]

Problem: context dependent



# available systems

- HAMLET (DECTalk, formant synthesis, Iain Murray)
- LAERTES (BT Laureate, concatenative synthesis, Iain Murray)
- CHATAKO (CHATR, unit selection, Akemi Iida)
- AffectEditor (DECTalk, formant synthesis, Janet E. Cahn, MIT)
- VieCtoS (OFAI, concatenative, Erhard Rank)
- emosyn (MBROLA, TU-Berlin, Felix Burkhard)
- neural networking: Jun Sato

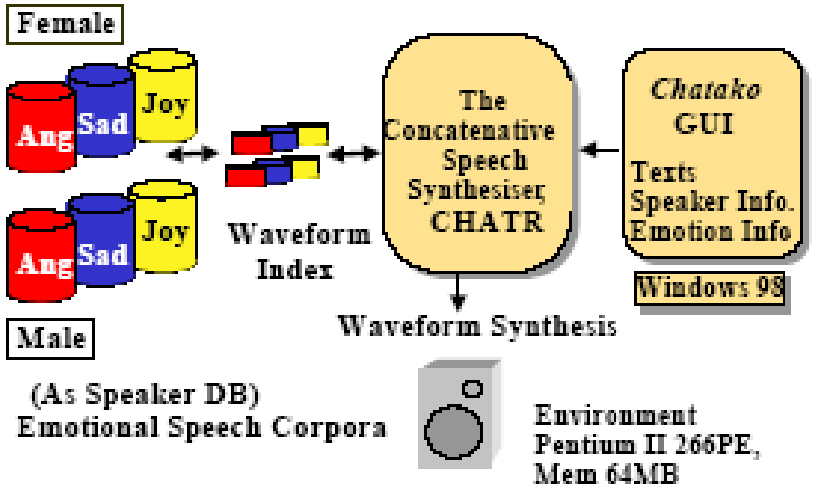
# screenshots, examples (1)

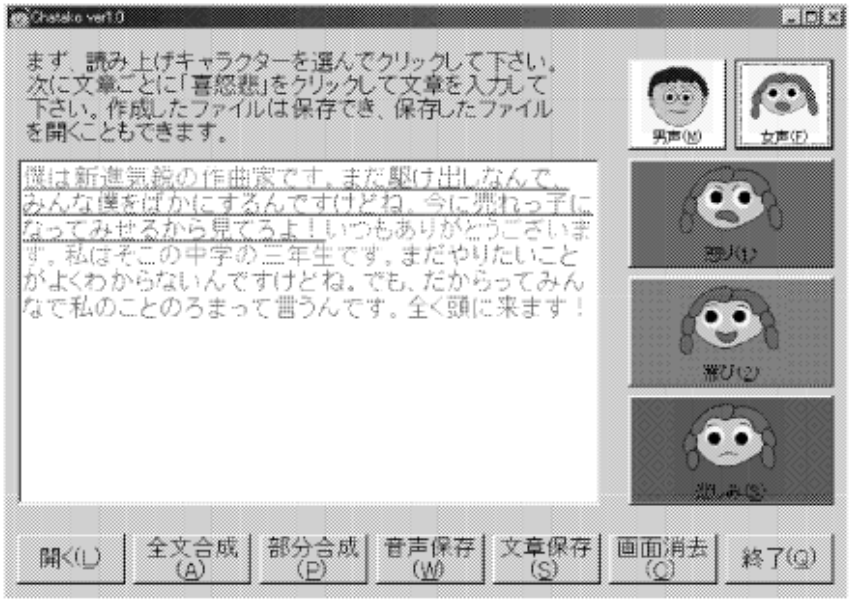
CHATAKO (unit selection)

anger

happiness

sadness



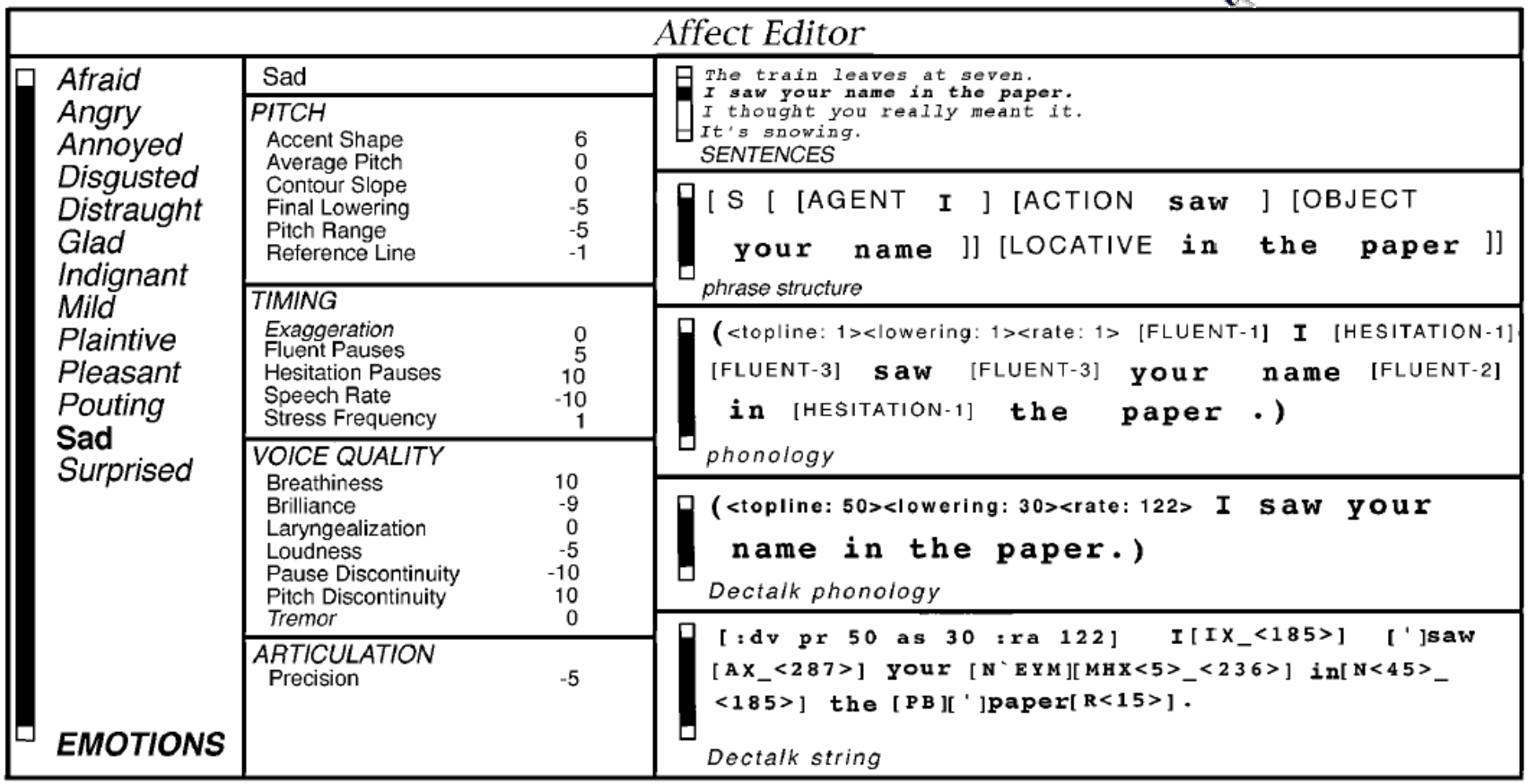
# screenshots, examples (2)

AffectEditor (formant synthesis)

anger

happiness

sadness



# examples

- Emofilt (rule based prosody)

anger

happiness

sadness

- VieCtoS

anger

happiness

sadness

# papers

- [Akemi Iida, ]
- [Janet E. Cahn]
- [Iain Murray]
- [Klaus Scherer, Speech Communication 40, 2003]
- [Mark Schröder, Speech Communication 40, 2003]
- [Jun Sato, IEEE Robot and Human Communication 1996]
- [Randolph Cornelius, Speech Communication 40, 2003]
- [Sahar Bou-Ghazale, John Hansen, IEEE Transaction on Speech and Audio Processing, 1998]
- ….