Emotional Speech

Franz Zotter

December, 2003

1 Four Theories about Emotion

(Randolph R Cornelius)

2 Darwin

(Charles Darwin, 1872 "The Expression of Emotion in Man and Animals") Emotions are phenomena with important survival functions for a species. An emotion occurs if one faces the associated problem, in order to help us solving it. The assignment emotion-problem has developed during evolution. Thus emotional expression also serves the function of survival, i.e. helps to solve a problem. The most noticeable expression of emotion is the facial expression. Darwin described those emotion expressions in detail. In the 80s and 90s contemporary psychologists reduced the expressions of emotion to a few universal (cross-cultural) and recognizable set of emotion archetypes: (see also 1)

- happiness
- sadness
- fear
- disgust
- anger
- suprise

3 James-Lange

(William James 1884, Carl Lange) Carl Lange and William James both claimed that emotions occur after bodily changes. James controversively stated:

- the nervous system is a bundle of predispositions to react in a particular way upon particular features of the environment
- emotions occur due to bodily changes:
  - "we feel sorry because we cry, ..."
  - "angry because we strike, ..."
  - "afraid because we tremble, ..."
  - "afraid because we run"
  - if I had no body, i would be "excluded from the life of the affections"

An addition to the last statement: people with spinal cord injuries indeed show a decrease in intensity of experienced emotion, due to lack of feedback from the body. (Hohmann 1966)

It might sound strange that emotional experience comes after making faces. Therefore several theorists proposed the existence of "affect programs" which activates expression and experience.

4 Cognitive Approach

(Magda Arnold 1960, Smith) In the *appraisal* events are judged as good or bad for oneself, this occurs unreflected and automatically. Bodily changes and emotions are then the results of the *appraisal*. Emotions are associated to a characteristic pattern of *appraisal*.

Following judgements are supposed to be done in the *appraisal* concerning the situation one is exposed to:

- novelty
- pleasantness
- responsibility
- effort
- certainty
- control
- ...

4.1 Social Constructivist

(James Averill 1980, Rom Harré) In this theory emotion is not seen as biologically determined, but as cultural product that arise from learned social rules. Thus they are social constructions and they can only be fully understood on a social level of analysis. In this view anger is a very sophisticated feeling which is based on a moral judgement, and it comes up, if someone violates some standard of behaviour. Even the intention of other persons play an important role in becoming angry. Also "losing control" during oneself's aggression is not a subjective but a social judgement. It is possible to find explanation for every basic emotion as a social construction.

5 Emotion Articulation

In the following sections a few aspects of expression shall be shown, concerning the features of each the basic emotion. Tough emotional content in speech acoustics is here the main topic, some side-informations are interesting, some correlations exist.

5.1 Physical Aspects

Due to arousal of either the *sympathetic* or *parasympathetic* nervous system the arousal in the speaker's body is controlled. Those two systems have a opposing regulative effect on mostly internal organs. Some emotions and proprietary physical affects:

- fear/anger:
  - increased heart rate and blood pressure
  - changes in depth and pattern of respiratory movements
  - increase in respiration rate, causing increased subglottal pressure
  - drying of the mouth
  - occasional muscle tremor
- relaxation or grief
  - decreased heart rate and blood pressure
  - increased salivation
  - decrease in respiration rate, low subglottal pressure
  - relaxed respiration

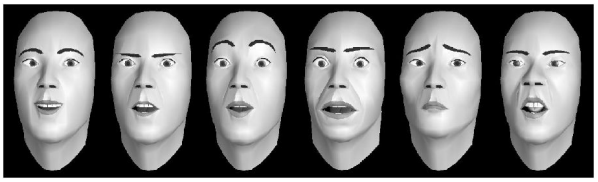

Figure 1: archetypes of facial expressions [Dominic Massaro]: happiness, anger, suprise, fear, sadness, disgust

5.2 Body Language

Body language consist of gestures and posture. Posture is mainly a static position of the body from which one can derive certain attitudes and emotions (e.g.: crossed legs against someone could mean 'keeping distance', crossing arms could mean 'rejecting something', turning one's back to someone could also mean rejection, sitting crooked, ...). Gestures are movements, mostly with the hands (e.g.: reaching an open hand, grabbing, making a fist, ...) that are very expressive and are consciously used as for symbolic communication, but they are also made unconsciously showing one's attitudes and constitutions. I try to give some examples:

- anger: making fists and contracting the arm's muscles, turning down the head
- sadness: crooked posture, loosely hanging arms/ hands on the chest, looking down
- happiness: standing tall, reaching out both hands, looking up
- ...

5.3 Facial Expression

The facial expression is very essential reading someone's emotions. Even unwanted emotional expression can often be derived from it. As a part of the body language there are again movements and static expressions due to positioning or movement of the face muscles.

- anger: contracting lips, lowering the eye brows
- happiness: smiling
- ...

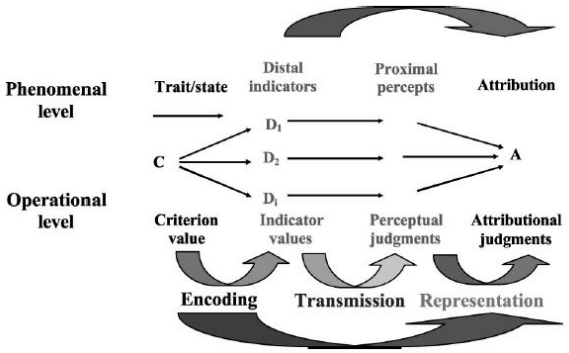

Figure 2: Brunswikian Lens Model for emotion transmission [Klaus Scherer]

5.4 Expression in Speech

5.4.1 Acoustic Measures

In Figure 2 (Klaus Scherer) a schematic model of emotional communication is shown. In this model the states of the speaker are encoded into the distal cues, which are distant the observer, but can be (partially) perceived as proximal cues and afterwards decoded. In our concern the distal cues are transmitted together with the speech signal During transmission information can either get lost or not be perceived by the observer. The decoding process then is the judgement of the observer in which his emotional state and attitude interfere with his internalized representations of proximal cues and their meaning. The acoustic features of the speech signal

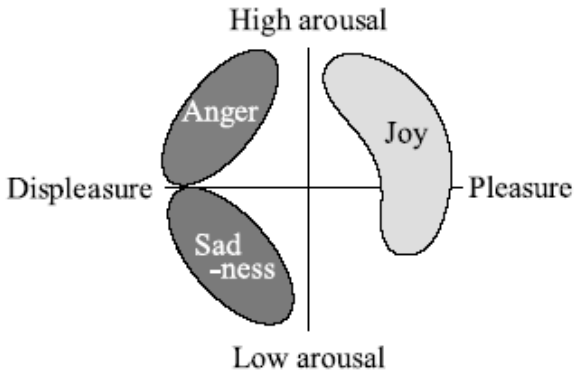

Figure 3: Emotions and Arousal [Akemi Iida]

can be measured, but do not necessarily correspond to the proximal perception of an observer. acoustic features that show correlation to emotional states are:

- pitch (fundamental frequency):
  - average pitch
  - contour slope
  - pitch range
  - pitch discontinuity
  - final lowering at the end of sentences
  - accent shape
- harmonicity
  - breathiness: amount of respiration noise in speech
  - Laryngealization: creaky voice, irregular glottal excitation
  - Tremor: irregular pitch periods (*jitter*)
- brilliance (relation between low and high frequency energies)
- loudness (measure of perceived intensity)
- timing:
  - intensity slopes (pauses, hesitations)
  - speech rate (word duration)
  - vowel duration
  - consonant duration
  - intensity of plosive bursts
- spectral information
  - formant positions
  - articulation precision

5.4.2 Emotion Impact on Speech

In general there can be seen basic impacts on speech that are mainly ruled by physical changes in an affected speaker

- fear/anger:
  - greater speed and loudness
  - higher pitch average
  - expanded pitch range
  - increased higher frequency energy
  - disturbed speech rhythm due to short respiration cycles
  - increased precision of articulation

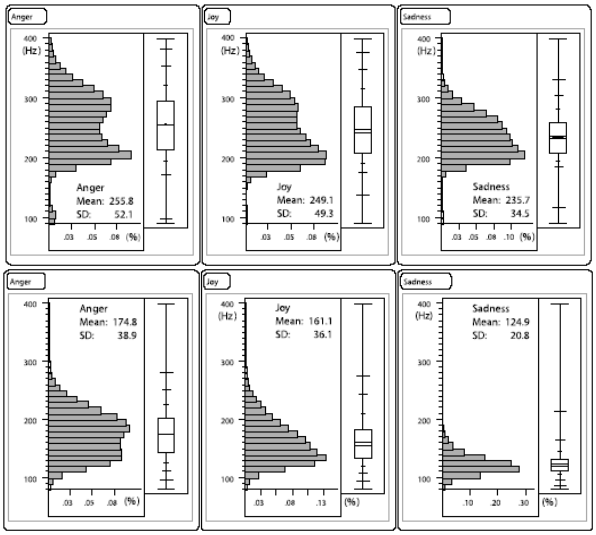

Figure 4: Pitch histograms in female and male emotional speech [Akemi Iida]

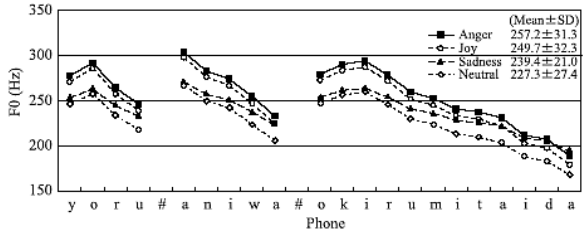

Figure 5: Pitch contours of a sentence with different emotional expression

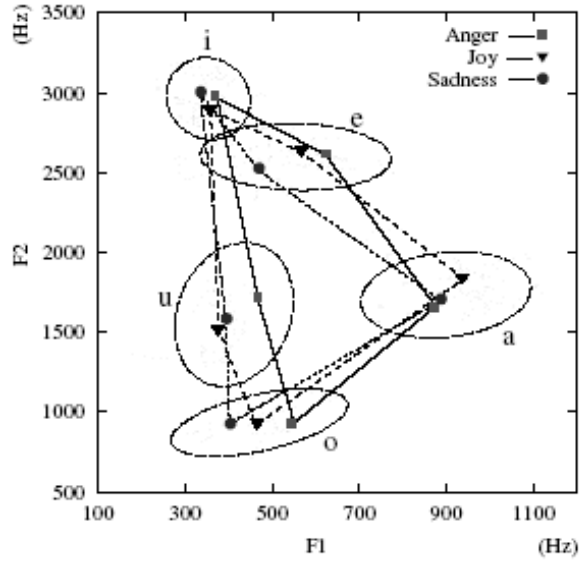

Figure 6: Formant Positions in different emotional state [Akemi IIda]

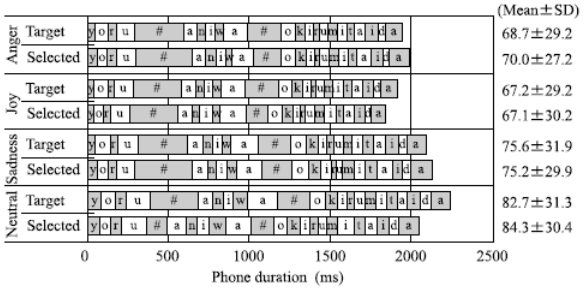

Figure 7: Vowel duration [Akemi Iida]

- relaxation/grief:
  - low speech rate and loudness
  - decreased higher frequency energy
  - lower pitch average
  - smaller pitch range
  - fluent speech
  - reduced pitch range
  - decreased precision of articulation, in general
  - formant position change (F1, F2, F3) because of decreased articulator movements

Experiments show more specific classifications of features for several emotional states (Scherer) depicted in the following pictures.

Unfortunately an experiment on the accuracy of emotional speech communication in comparison to facial expression features conducted in western and non-western countries shows that it is hard to communicate emotions only by speech.

5.4.3 Affect Bursts

The term *affect burst* is referred to in the work of Mark Schröder and is defined as short emotional non-speech expression. Actually *affect bursts* sometimes allow a little more accurate emotion recognition. Figure 16 shows experiment results.

6 Synthetic Emotion

6.1 Emotional Speech Synthesis Applications

Synthetic emotion is highly interesting not only in terms of man-machine communication but also as aid for impeded people (e. g. mute people, . . .). Several target applications can be found:

- emotion/trouble recognition in automatic dialog systems
- emotion recognition for psychic analysis
- lie detection in forensic investigations
- speech driven facial animations
- text-to-speech synthesis with emotional content
- voice xml
- speech manipulation (emotion conversion)

6.2 Emotional Feature Generation

As we see in section 5 several parameters have to be controlled to get an appropriate representation of emotion in synthetic speech.

What methods can be used to apply emotional parameters on synthesized speech:

- *affect burst* insertion
- excitation/residual signal processing:
  - PSOLA, LP-PSOLA, RP-PSOLA, ...:
    - timing (duration of syllables, words, vowels, pauses, hesitation, ...)
    - pitch manipulation (final lowering (melody), stresses, pitch range, pitch average)
    - F0 interpolation (fluent speech, transitions)
  - Jitter processing
  - additive noise: breathiness
  - Ring modulation (AM): harmonicity/breathiness (spectral shift)
  - linear filtering: loudness, brilliance, glottal excitation shape
- envelope/intensity slopes: transitions (plosives, pauses, stressed words, loudness, ...)
- spectral modifications (LPC, MFCC, ...):
  - formant repositioning/bandwidth change: articulation precision
  - emphasis: brilliance, loudness
  - frame rearrangement: phoneme duration
  - reflection coefficient interpolation: phone transitions, articulation precision

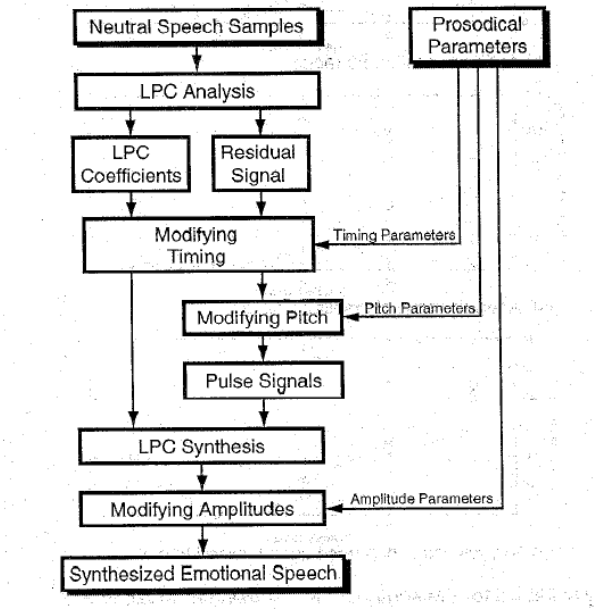

Figure 8: Emotional LPC resynthesis model [Jun Sato]

6.3 Parameter derivation in TTS

In text-to-speech systems following methods seem to be necessary to derive emotion-driven prosodic parameters:

- Grammar, sentence analysis
- key-words (*affect bursts*)
- context analysis
- mark-ups

In concatenative text-to-speech synthesis RP-PSOLA is recommended. A Reference Pitch-PSOLA unit database consists of words/phonemes that are spoken at reference pitch, so that naturalness for wide range pitch manipulation can be preserved by an extended database size.

6.3.1 Emotion in TTS-Systems

- HAMLET (DECTalk, formant synthesis, Iain Murray)
- LAERTES (British Telecom Laureate concatenative TTS, Iain Murray)
  - anger: increased pitch range, raised pitch, increased speech rate, laryngealisation

  - happiness: increased pitch range, further raised pitch, slightly increased speech rate
  - sadness: lowered pitch contour, decreased speech rate, laryngealisation
  - fear: further raised pitch, much increased speech rate

- AffectEditor (DECTalk, Janet E. Cahn, MIT, Formant Synthesizer): A *very* large number of parameters and presets are used to synthesize speech with emotional information. Even a sentence analysis can be done. In Figure 14 the whole lot of used speech parameters can be seen.

- CHATAKO (ATR CHATR, Akemi Iida, Japan, concatenative synthesis): In the CHATAKO concatenative speech synthesis there are three database corpuses, each representing an emotional state. In Figure 12 the principle, I didn't get information about pitch parameters. The cute graphical user interface can be seen in Figure 13.

- VieCtoS (Austrian Institute for Artificial Intelligence, Erhard Rank, concatenative synthesizer)

6.4 Intelligent Systems (automatic feature detection/reproduction)

The main advantage of the systems described later is the ability to take into account a huge amount of input data to derive an estimate of the speaker's emotional state. It is especially exciting that training of the system behaviour could be done by usage of broadcasted or recorded speech data. The remaining problem is building system models concerning input parameters (prosody of words, sentences, ...) that provide reliable indicators for automatic emotion recognition or resynthesis.

6.4.1 Hidden Markov Modelling (HMM)

To detect the probability for a certain emotional state of the speaker, one can use hidden markov models. Hidden Markov Models are based on the probability that a certain (perceivable) state $\mathcal{Q}$ of the speaker is indicated with respect to the observed parameters $\boldsymbol{O}$ and the estimated model parameters $\boldsymbol{\lambda}$:

$$P(\mathcal{Q}|\boldsymbol{O},\boldsymbol{\lambda}) \qquad (1)$$

In reverse the viterbi algorithm can be used to produce a prosodic parameter sequence $\mathcal{Q}$ to feed an emotional speech synthesizer. The main task here is to find appropriate indicators for emotional states to build up a model. Due to the complexity in emotion encoding in speech a HMM has to be able to deal with several input parameters appropriate for emotion indication. For a stressed/relaxed state decision the **TEO** (Teager Energy Operator) for utterances seems to deliver appropriate results [Hansen]:

$$TEO = x^2[n] + x[n+1]x[n-1] \qquad (2)$$

A general model of a speech resynthesis process could look like this:

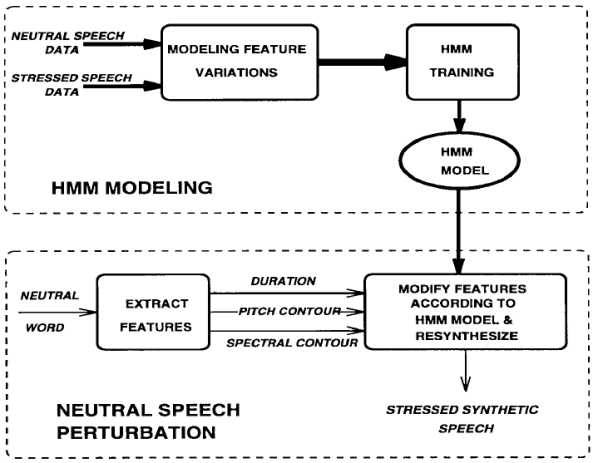

Figure 9: Generalized Model for HMM with prosodic parameters

6.4.2 Neural Network Modelling

A neural network consists of nodes in different layers. Each of this nodes has connections to nodes in the previous layer as input and connections to nodes of the next layer on the output. The task of each node is to derive the stimulation of the following nodes due to the previous node's stimulation pattern. Like in Figure 11 the first layer's nodes are driven by prosodic features, and the third layer shall represent a two-dimensional emotion space, at which the intensity and kind of emotion should be represented. The behaviour of each node is trained by test-sequences, so that the network shall be able to accurately detect emotion kind and intensity. As the reverse process prosodic features can be produced out of a desired point in the emotional space

(intensity, kind). A problem in the work of Jun Sato is, that the emotion space is context dependent. Therefore also appropriate input data and models are required.

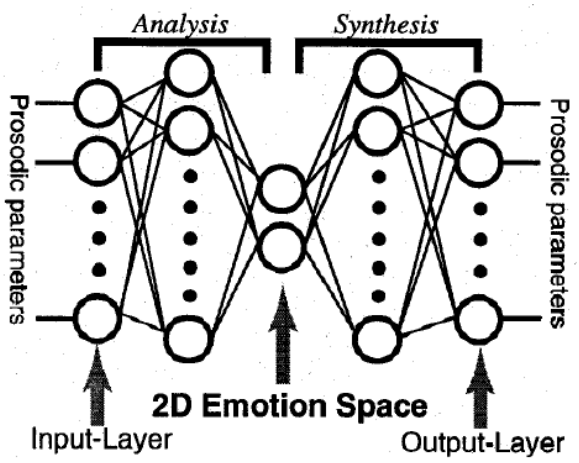

Figure 10: Neural network for prosodic parameters [Jun Sato]

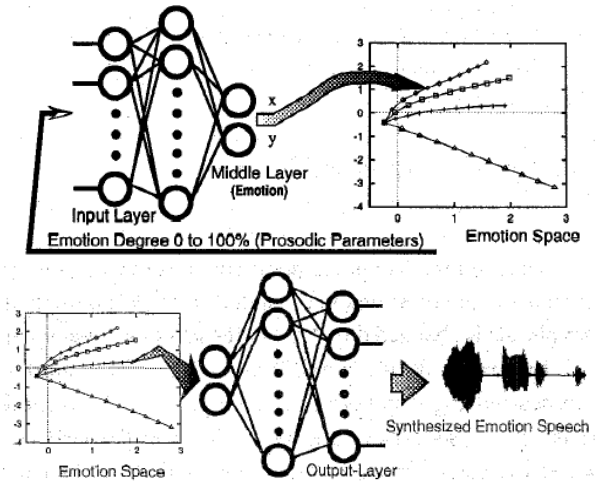

Figure 11: Neural network prosodic emotion recognition and resynthesis [Jun Sato]

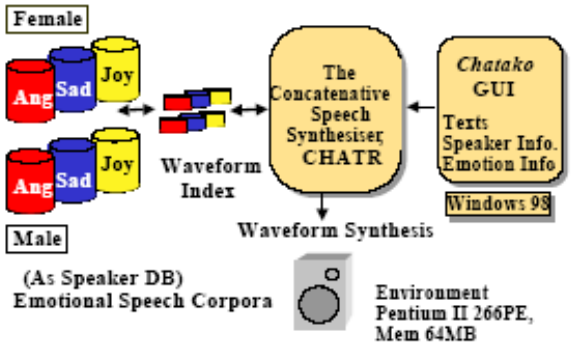

Figure 12: CHATAKO block diagram [Akemi Iida, CHATR]

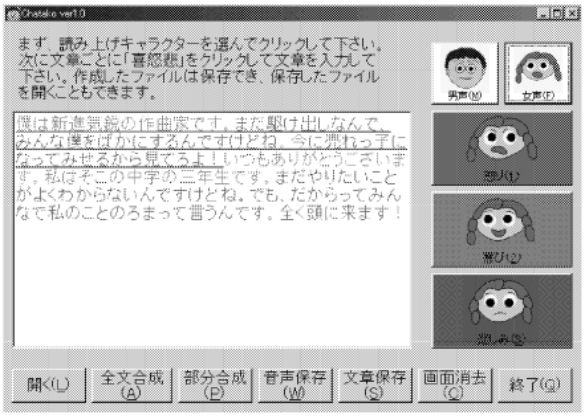

Figure 13: CHATAKO graphical user interface

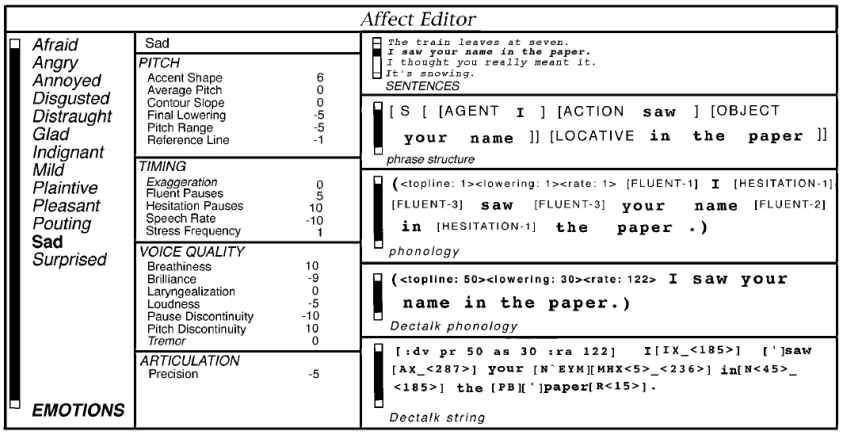

Figure 14: AffectEditor user interface [Janet E. Cahn, MIT]

7 Literature

- "Theoretical Approaches To Emotion", Randolph R. Cornelius, ISCA Workshop on Speech and Emotion, Belfast 2000
- "Vocal communication of emotion: A review of research paradigms", Klaus R. Scherer, Speech Communication 40 (2003)
- "Experimental study of affect bursts": Mark Schr"oder, Speech Communication 40 (2003)
- "A corpus based speech synthesis system with emotion", Akemi Iida, Nick Campbell, Fumito Higuchi, Michiaki Yasumura, Speech Communication 40 (2003)
- "HMM-Based Stressed Speech Modeling ...", Sahar E. Bou-Ghazale, John H. L. Hansen, IEEE Transactions on Speech and Audio Processing, Vol. 6, Nr. 3, May 1998
- "Emotion Modeling in Speech Production using Emotion Space", Jun Sato, Shigeo Morishima, IEEE International Workshop on Robot and Human Communication, 9/96
- "Generating Expression in Synthesized Speech", Janet E. Cahn, MIT Media Laboratory, diploma thesis, 1990
- http://www.kgw.tu-berlin.de/ felixbur/speechEmotions.html by Felix Burkhardt
- "Erzeugung emotional gefärbter Sprache mit dem VieCToS Synthesizer", Erhard Rank, "OFAI 1998
- "Rule Based Emotional Speech Synthesis Using Concatenated Speech", Iain Murray, ISCA Workshop on Speech and Emotion, Belfast 2000
- "Real-Time Speech-Driven Face Animation With Expressions Using Neural Networks", Pengyu Hong, Zhen Wen, Thomas S. Huang, IEEE Transaction on Neural Networks, VOL. 13, NO. 4, 2002
- http://www.qub.ac.uk/en/isca/proceedings/ ISCA workshop 2000
- ...

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).

Figure 15: Emotional Features in Prosody [Klaus Scherer]

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	[ɪːə]	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	[m̰ː]	1	2	Anger	80%	mrr	8	Anger	50%
Disgust	Buäh	[bʊ̯æː]	3–2	6	✓	92%	uäh	1	✓	63%
	Igitt	[iːgɪtʰ]	3	1	✓	100%	igitt	1	✓	100%
	Ih	[ɪːə]	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	[əː]	2–1	2	✓	45%	hmm	12	✓	63%
	Hmm	[mː]	1–2	2	✓	83%	mmh	7	✓	60%
Relief	Sigh	[aː]	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	[o̞je̞ː]	2–1	4	✓	96%	ujeh	6	✓	75%
	Oh–Oh	[ʔoʔːoː]	3–2	2	✓	85%	o-oh	6	✓	67%
	Oweh	[o̞ːβe̞ː]	3–1	1	✓	50%	oh jee	14	✓	85%
	Hmm	[m̩m]	2–1	1	✓	70%	hmm	9	Boredom	63%
Contempt	Laughter	[ɦe̞ɦ]	1	5	✓	77%	hähä	10	✓	74%
	Pha	[phaʔ]	1	2	✓	95%	pah	4	✓	95%
	Tse	[tsʰə]	3	1	✓	100%	tse	5	✓	85%
Anger	Growl	[m̰ː]	2–1	4	✓	69%	ahr	2	✓	39%
	Breath out	[ɦː]	1	3	✓	55%	chrr	8	✓	39%
	Oh	[ə̹ː]	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.

Figure 16: Experiment on the recognition of *affect bursts* [Mark Schröder]

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.

Figure 17: Experiment on emotional prosody recognition in western and non-western countries [Klaus Scherer]