Emotions in Speech: Tagset and Acoustic Correlates

Sofia Gustafson-Capková
Department of Linguistics
Stockholm University
sofia@ling.su.se

ABSTRACT

In recent years the interest has grown, for automatically on the one hand detect and interpret emotions in speech, and on the other hand to generate certain emotions in speech synthesis. Areas where such knowledge might improve a system is e.g. dialogue/expert systems but also applications for disabled people.

This report is a short survey of research within the field of emotions in speech. Special attention is paid to what categories are used for tagging speech corpora for emotions, and to what acoustic correlates these categories might be connected.

0 Introduction

Much research has been done on the topic emotions in speech. Already in the early 20th century attempts were done to connect certain forms of speech to certain emotions (Armstrong & Ward, 1926). From that time and onward these attempts were repeated. The early research was rather carried out in the field of psychology, but with the introduction of speech synthesis and automatic speaker recognition (ASR), the more psychological branch is accompanied by a more technical application based approach.

Many researchers in the field of speech technology have during the last decade worked on different aspects of emotions in speech. One of the goals is to make speech synthesis sound more natural, another goal is to be able to recognise the emotive state of a speaker in e.g. a dialogue system.

One motivation for the claim that emotions are signalled in speech prosody, is experiments which have shown that subjects can recognise the emotive content in a speech sample, also when all word meaning is filtered out (e.g. Banse & Scherer, 1996, Brown, 1980, Mozziconacci, 1998, Pereira, 2000, Scherer, 1981, Soskin & Kauffman, 1961). I.e. with the sole information of intonation, subjects are able to recognise the emotion behind the utterance.

Emotions colour the language, and can make meaning more complex. As listeners we also react to the speakers emotive state and adapt our behaviour depending on what kind of emotions the speaker transmit, e.g. we may try to show empathy to sad people, or if someone hesitates we try to make the person clarify what s/he means or wants. To classify the emotive state by a speaker on basis of the prosody and voice quality, we have to classify acoustic features in the speech as connected to certain emotions. This also implies the assumption that voice alone really carries full information about emotive state by the speaker. This assumption is often taken for granted, but e.g. Stibbard (2001) takes this assumption for questionable.

Research in emotional speech has a long tradition, but in the recent years the need for applicable results in this field has become more important in more and more sophisticated automatic spoken language systems, like e.g. SmartKom (Batliner, et al. 2001). Access to the emotive information in speech could gain certain applications, i.e. it would be useful to be able to take into account whether the speaker in e.g. a dialogue system is frustrated, irritated or content. With recognition of emotions on basis of features in the speech signal, a system might be able to detect the emotive state by a person, and respond accordingly, as well as being able to speak in a way that people can feel comfortable with. Experiment results has shown that an agent who can signal empathy is significantly more effective than the control conditions, in helping relieve frustration levels (Klein, 1999).

However, the findings in the area of acoustic correlates to emotions are not always encouraging, and also not too homogenous. Results point sometimes in contradictory directions, and it is hard to define what is valid data concerning emotive speech.

In this paper I am going to give a brief survey over research in the field of emotive speech. After a survey of some earlier investigations I will focus specifically on three aspects: i) Speech data, ii) emotive categories and iii) tagsets for emotions. The outline of the paper is as following: In section *1. Choice of corpora* I will examine what kind of data that have been used in different corpora and investigations. In section *2. Categories of emotion* I will give account for psychological motivations to categorisation of emotions. I will discuss along what dimensions emotions can be classified, and also the distinction between emotions and attitudes. Section *3. Acoustic correlates* consist of an overview of results from studies in the field of emotional speech, regarding the acoustic correlates of emotions. In section *4. Evaluations of tagsets* I will give account for an evaluation of an existing tagset,

developed on basis of spontaneous speech, and in section *5. Synthesis and recognition of emotions in speech* I will give account for the results for some experiments with emotional speech synthesis. I will finish with section *6. Summary and Discussion*.

In the very end of the paper I have listed some links and resources that one can find on the www. The list is not meant to be exhaustive, but rather to give the interested reader a small sample of web pages of interest (please try the demos, they are really VERY funny!).

1 Choice of corpora

In this section I will describe what kind of data that have been used in this field. I will call each data collection a corpus, even though the material might differ from each other regarding both characteristics and size and degree of analysis.

The first problem one faces in starting to investigate emotions in speech is the problem of choosing valid data. Generally materials from three categories are used in investigating emotional speech:

- Spontaneous speech
- Acted speech
- Elicited speech.

All three groups have both pros and cons, and no of the groups can be pointed out as generally optimal.

In this chapter I will review what kind of data was used in some earlier studies, and also describe strong and weak sides of these different kinds of data.

1.1 Spontaneous speech

Spontaneous speech is often argued to contain the most direct and authentic emotions, but the difficulties in collecting this kind of speech are also extensive. In the ideal condition speakers should be recorded without knowing about it, so that they behave completely naturally, but this kind of data collection rises difficulties, since such a routine is ethically problematic (e.g. Campbell, 2000, 2001). Problems with spontaneous speech may in other cases, e.g. such as clips from the television, might instead give problems with copyright (Stibbard, 2001).

Although difficult to collect, there exist corpora of spontaneous speech, but they are generally not distributed, e.g. The Belfast database (Douglas-Cowie et al. 2000) and The Leeds-Reading Emotion in Speech Corpus (e.g. Greasley et al., 1995), both consisting of clips from different television programs. Other examples are the JST database (e.g. Campbell 2001), also consisting of recordings of natural speech in natural situations and the SUSAS corpus (Hansen, 1999) which consists of air force pilots conversation i the cockpit, which of course also is a “natural situation”, but still less common than many everyday situations.

Aviation data, i.e. crew conversations in cases where the aircraft is crashing, has also been used (by e.g. Brenner et al., 1983, Williams & Stevens, 1969) as well as the radio recordings of the reporting of the Hindenburgh catastrophe (e.g. used by Stibbard 2001, Williams & Stevens, 1969).

Other researchers using spontaneous speech is e.g. Scherer and Ceshi (1997, 2000), who used recordings of airline passengers waiting in vain for their luggage, and Huttar (1968) who used recordings of 30 utterances of one American male.

1.2 Acted speech

Acted speech does not have the same ethical problems that are present by collecting spontaneous speech, but the degree of naturalness is often questioned. The differences regarding the quality of acting also has to be taken into account. Some collections of acted speech consist of recordings with professional actors (e.g. Banse & Scherer, 1996, Mozziconacci, 1995, Scherer et al. 1996), in other cases amateur actors have been used (e.g. Fairbanks & Hoaglin, 1941) or students of drama (e.g. Green & Cliff 1975), or any students (e.g. Levy, 1964). Of course the quality of the acting could be suspected to differ between those recordings.

In the first place the quality of acted speech is a function of the quality of the acting performed, which might affect the manifestations of the emotions. However, there are further unclear parts of using acted speech; one is whether acted speech really can be said to mirror authentic emotions. It is clear that acted speech has acted emotions, and how relevant are they? Acted speech is easier to control, but on the expense of naturalness. Stibbard (2001) mentions that the acted speech is merely conforming to stereotypes of how people believe that emotions should be expressed in speech, not to how emotions actually are expressed. This indicates that acted speech is more stereotypical, and that the expression of emotions is more extreme than in spontaneous speech. For a speech synthesis

application this might not be a problem, perhaps it is rather an advance to use stereotypical emotional expressions. It is just suitable if synthesised speech gives the most prototypical and easily interpretable emotive correlates, instead of real, but complex and hard to interpret variants (of course, the degree of stereotypicity cannot be too high, then it is just sounding ridiculous). However, in speech recognition this mismatch between ideal and reality gives rise to problems. In recognising speech we have to cope with the complexity of reality.

1.3 Elicited speech

In elicited speech the idea is that certain emotions are induced. The procedure can be that subjects watch a film, which should evoke specific emotions, and then they have to retell the film to the experimenter. Here the idea is that the speech shall be coloured by the emotion induced. It is also possible to put a subject into a situation meant to evoke a specific emotion, and then record her/his speech. However, this method suffers from ethical problems, i.e. it is not fully ethical to scare someone, and then record her/his speech (perhaps it is even more unethical to do this, than to just record someone who is already scared.). As a result of this problem the induced or elicited emotions are often too mild; when they are not too mild they are instead too unethical.

The induction method has the positive feature that it gives control over the stimulus, on the other hand, different subjects may react differently on the same stimulus. The validity of such elicited, or induced, speech depends to a large extent on how successful the induction process was (Stibbard, 2001).

Studies, which have used induced emotional speech, are e.g. Skinner (1935), Friedhoff et al. (1962), Hecker et al. (1968), Iida et al. (1998).

1.4 Summary

So, what kind of speech is best to use? The question is not easily answered. On the one hand, spontaneous speech is the most genuine source, but, as Stibbard (2001) points out, also spontaneous speech is in a way acted and constrained by e.g. social constraints. “Pure” outbursts of emotions, as for instance those we can see in children, is not a social accepted way to express anger, sorrow, happiness etc.; as we grow up we are expected to learn to control and verbalise intensive emotions. However, we are still interested in the subtle traces that those emotions might leave in voice.

Acted speech gives extensive possibilities to control the speech, but it is not genuinely natural, but, it seems that researchers are not fully convinced whether they should judge this as good or bad.

One answer to the question posed, what kind of data is best might be that the “best” data is something task-dependent. I.e. in an aviation application the very limited data from aviation is probably the best, while such data would be disastrous in a system, intended to perform some kind of general conversation. For such a system one would certainly need more general speech data.

2 Categories of emotion

What categories of emotions are relevant in trying to establish the correspondences between emotions and speech? Of course the answer is to a certain extent depending on the task, i.e. different applications may gain from different categorising. But still what kind of set would be a generally good set to start with? The question does not seem to generate just one answer. Researchers today are not agreeing on some specific general tagset, and the categorising is often done on a relatively subjective basis. This gives difficulties when one wants to compare the results from different studies.

There are philosophical, psychological and biological motivations in the approaches to the categorising of emotions, all of these combined in different ways. In addition to these basic approaches, we can also find differences in the dimensions of the categorisation of the emotions; discrete categories or scalar values.

In the PHYSTA research report on human emotions, the concept of emotion is described as consisting of two main ideas. The first is the idea of basic emotions, originating from Descartes, meaning that humans have a number of universal emotions. These basic emotions may also blend into complex emotions. The second is the idea of biologically motivated emotions, originating in work of Darwin, which means that the basic emotions are motivated by an evolutionary need. Plutchiks (1980) psychoevolutionary theory of basic emotions can be taken as a central standpoint for this view on emotions.

The classification of emotions can be viewed from two perspectives: i) The view on emotions as discrete, ii) The view of emotions as scalar (or dimensional), i.e. as points in a n-dimensional space. However, as Murray and Arnott (1993) points out, the both views are not mutually exclusive, since the

discrete values also can be plotted in the scalar space. The description of emotions as scalar has its root in the work of Wundt (1897), who describes emotions as values along a number of axes.

Just as there are different ways to classify the space of emotions, there are different ways to define the concept of emotions. Sometimes the concept of emotion is including all kinds of emotions, attitudes and beliefs, but sometimes a sharper distinction is done. Wichmann (2000), for instance, draws a distinction between attitudes and emotions, attitudes being less governed by affect and more based on a belief. An emotion should then be more purely a state of affect. Wichmann argues that only emotion is likely to be reflected in the speech signal. One example on the difference between attitude and emotions could be e.g. "trust" (attitude) and "fear" (emotion). It is interesting to note that attitude, in terms of evidentiality, in some languages, e.g. Makah, is coded into the morphological system (Saeed, 1995), but, to the knowledge of the author, no language codes an emotion as e.g. fear morphologically.

However, this fine-grained distinction between emotions and attitudes is not always made, but instead attitudes and emotions are often integrated into one bag labelled emotions. The need to distinguish between attitudes and emotions appears just when the set of categories is large, i.e. as a function of a wish to make a full description of everything a human might perceive. However, since most researchers do not attempt to make this very fine-grained categorising but instead choose a limited set of basic emotions, the need to distinguish between attitudes and emotions is in most cases not crucial.

In the following sections I will describe and give examples of discrete and scalar approaches to emotive categories.

2.1 Discrete categorising of emotions

In this class each emotion is regarded as a separate unit. The main problem, however, is to define a basic set of those discrete emotions. A glance into the literature shows very disparate sets of emotions, ranging from two (e.g. Weiner & Graham) up to 73 (EISP, e.g. Stibbard, 2001). What are then the motivations for some given sets? In Table 1 (other side), adapted from Ortony and Turner (1990) an overview of different motivations for different sets of categories is given.

It is clear, that also researchers, which have the same basis or motivation ("basis for inclusion") for the emotions (as e.g. Gray, Izard, Pankzepp and Watson, who all postulate hard-wired, i.e. innate, emotions), suggest quite different emotive categories. Also the motivations are differing (e.g. such as "hard-wired", "universal facial expressions", "relation to action tendencies") between researchers, even though they are all share some kind of psychobiological basis for them.

2.2 Scalar categorising of emotions

In the scalar view, emotions are not regarded as discrete categories. Instead emotions are regarded as consisting of "different mixings of the same ingredients", i.e. ingredients from *n* dimensions, but for each emotion the proportional mixture is different. This describes emotions as values on *n* number of axes. According to Stibbard (2001) this approach can be traced back to Spencer (1890) and Wundt (1897).

Generally dimensional systems contain three dimensions: Strength (corresponds to attention – rejection), Valence (corresponds to positive – negative) and Activity (corresponds to sleep – tension) Stibbard (2001), Murray & Arnott (1993). Other three-dimensional systems may consist of the dimensions arousal level, valence and activity, or another example is the dimensions pleasure, arousal and power (e.g. Pereira, 2000). Dimensional systems have been used by e.g. Schlosberg (1954), this system was later modified by e.g. Tomkins (1964), Scherer (1979a) and Huttar (1968).

Also two-dimensional systems are in use, e.g. with the dimensions active-passive and positive negative. The response tracing system FEELTRACE (Cowie et al., 2000) is an example of a system using a two-dimensional system for emotions. FEELTRACE has been developed as a tool to let observers track the emotional content of a stimulus as they perceive it over time (Cowie et al., 2000).

A dimensional categorising of emotions is claimed to allow a systematic investigation of relations between emotions, an important possibility that the discrete systems are lacking.

3 Acoustic correlates

In focusing on spoken language, intonation is taken as the predominant marker of emotions. The acoustic correlates of emotions traditionally investigated are pitch (fundamental frequency, both average and range), duration, intensity and voice quality (Murray & Arnott, 1993). If we just consider

the acoustics of speech, the emotional state of the speaker affects predominantly his intonation, but not exclusively. This is so because the physiological effects that emotions have on the larynx are reflected

	Basic Emotions	Basis for Inclusion
Plutchik	**Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise**	**Relation to adaptive biological processes**
Arnold	**Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness**	**Relation to action tendencies**
Ekman, Friesen, and Ellsworth	**Anger, disgust, fear, joy, sadness, surprise**	**Universal facial expressions**
Frijda	**Desire, happiness, interest, surprise, wonder, sorrow**	**Forms of action readiness**
Gray	**Rage and terror, anxiety, joy**	**Hardwired**
Izard	**Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise**	**Hardwired**
James	**Fear, grief, love, rage**	**Bodily involvement**
McDougall	**Anger, disgust, elation, fear, subjection, tender-emotion, wonder**	**Relation to instincts**
Mowrer	**Pain, pleasure**	**Unlearned emotional states**
Oatley and Johnson-Laird	**Anger, disgust, anxiety, happiness, sadness**	**Do not require propositional content**
Panksepp	**Expectancy, fear, rage, panic**	**Hardwired**
Tomkins	**Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise**	**Density of neural firing**
Watson	**Fear, love, rage**	**Hardwired**
Weiner and Graham	**Happiness, sadness**	**Attribution independent**

Table 1. Table over emotions and their motivation as defined by different researches, taken from Ortony and Turner, (1990).

---

in substantial acoustic variations. E.g. Williams & Steven (1972) points out that physiological correlates of increased subglottal pressure, excessive dryness of the mouth or salivation, and decreased smoothness of motor control can influence the waveform.

If we in addition to the speech signal also consider facial expression, it is questionable whether intonation is the predominant marker of emotions, since emotions can probably be recognised more reliable by seeing the face of the speaker. This might be one reason, why findings concerning correlates between specific emotions and specific acoustic features sometimes have been quite disparate. E.g. in taking speech data from a documentary video clip, the speaker certainly also makes use of the visual channel, while in speech data from acted speech, recorded in a studio, the actor can be assumed to make less use of the visual channel.

Briefly, the results from investigations searching for acoustic correlates to emotions have been both uplifting and pretty discouraging. The findings are not fully consistent across different kinds of data, instead the answers are sometimes pointing in opposite directions. This might well be a function of the differences in i) the data used and ii) the categories of emotion used.

Few studies using spontaneous speech are reported, but e.g. Stibbard (2001) found nearly no correspondences at all between emotive categories and acoustic correlates. Quite discouraging!
However, in studies using acted or elicited speech, the results are more encouraging. Here many researchers report quite good correspondences.

The rest of this section is a report over findings of acoustic correlates for specific emotions. It shall be pointed out that a very small part of the data, which the findings in this section are based on, is spontaneous natural speech. In the division of categories I have made a synthesis of categories used by

Murray & Arnott (1993) and Stibbard (2001). The whole section is based on what is reported in Murray & Arnott (1993) and Stibbard (2001).

3.1 Primary emotions

Anger

According to Stibbard (2001) this is the emotional category where findings from both spontaneous and elicited material consistently report features such as high mean F0, wide pitch range, high energy and fast tempo. In detail Fairbanks and Provonost (1939, neutral phrase, acted speech, subjects recognised emotions, English) report that anger generally is characterised by high pitch and a wide pitch range. Fónagy & Magdics (1963, recordings of spoken Hungarian, subjective analysing,) reports that anger is characterised by mid pitch and a straight rigid melody. Öster and Risberg (1986, conversation, perception experiment with hearing impaired subjects, Swedish) reports high tempo, normal pitch range and normal pitch level.
None of those investigations have however distinguished between different kinds of anger, such as “hot” anger and “cold” anger (Stibbard, 2001).

Joy, happiness, humour

Generally researchers has focused on more extreme forms of happiness, but still, here the findings reported are quite consistent between researchers. However, Stibbard (2001) points out that concerning voice quality the findings are contradictory.
In detail Skinner (1935, emotions induced with music, subjects vocalised emotions through "ah", new subjects recognised the emotions), Cowan (1936, acted speech, subjective enelysis) and Öster & Risberg (1986), report that happiness gives an increase in pitch and pitch range. Öster and Risberg noted a slow tempo, while Fónagy & Magdics (1963, conversations, acted speech, music, subjective analysis, Hungarian) described it as “lively”. Davitz (1964, subjects rated stimuli with 14 emotive adjectives) reported an increase in speech rate along with an increase in intensity, this intensity also being noted by Skinner (1935).

Sadness

General findings describe sadness as exhibiting normal or lower pitch, narrow pitch range and slow tempo (Skinner (1935), Davitz (1964), Fónagy (1981, Murray and Arnott (1993) describe the material as being "a wide range of analysis and experimental techniques"), Öster and Risberg, (1986).
However, there are contradictory findings concerning voice quality (Stibbard, 2001).

Fear/Anxiety

Generally reported features are increased mean F0, increased F0 range, increased F0
Fairbanks and Pronovost (1939, neutral phrase spoken with different emotive expression by non-actors, subjects recognised emotions) reported the relatively highest pitch and the widest pitch range and the highest pitch median. Fairbanks and Hoaglin (1941, neutral phrase spoken with different emotive expression by non-actors, subjects recognised emotions) noted high speech rate. Williams and Steven (1972, acted speech, emotions induced by content in specially written play with control clusters) reported low F0, but with occasional F0 peaks, low speech rate.

Disgust

For the category disgust induced data and acted data has shown directly contradictory results Studies using induced data reports an increase in mean F0, while studies using acted data found a decrease (Stibbard, 2001).

Hatred/contempt/scorn

Fairbanks & Pronovost (1939) noted low pitch median, wide pitch range and very wide downward inflections at phrase endings. Fairbanks & Hoaglin (1941) noted lowest observed speech rate, caused by prolonged speech rate and increased pause length. Fónagy & Magdics (1963) noted that scorn is reflected by a descending melodic line, intoned to a very low level. However, Stibbard (2001) points out, that the only parameter analysed in studies more than one is tempo, why it is difficult to evaluate the findings.

3.2 Secondary emotions

Grief/Sorrow

Fairbanks and Pronovost (1939) noted low pitch median, the narrowest observed pitch range, the narrowest observed mean inflectional range. Fairbanks and Hoaglin (1941): noted low speech rate, and high ratio of pausing. Slow speech rate was caused by long pauses. Low pitch level and narrow range was also noted by Fónagy (1978), Cowan (1936), and Williams an Stevens (1972) although the latter authors noted that the low speech rate was due to both long pauses and prolonged vowels.

Affection /tenderness

For describing this emotion a variety of expressive phrases are used. Fónagy and Magdics (1963) reported that tenderness was "expressed on a higher pitch level " which "does not fluctuate in this case" with "the melody of the phrase". Fónagy (1981) noted an "absence of aggressiveness". Davitz (1964) Characterised affection by "soft" loudness, "low" pitch, "resonant" timbre, "slow" rate, "steady and slightly upward" inflection, "regular" rhythm and "slurred" enunciation.

Sarcasm/irony

Gibbs (1986) noted that both irony and sarcasm is characterised by a contradiction between the verbal and the non-verbal level. Murray & Arnott cites comments as nasalisation, slow speech rate.
However, I don't think sarcasm is an emotion at all!

Surprise/astonishment

This emotion is not extensively studied (Stibbard, 2001), but Murray and Arnott (1993) describe some findings in their survey paper. Fónagy and Magdics (1963) noted that in surprise "the voice suddenly glides up (or up-down), falls to a mid level (joyful surprise) or to a lower level (stupefaction). Öster and Risberg also noted a very wide pitch range for surprise, with tempo and pitch median normal or higher

As a short summary I give a table over acoustic characteristics for the different emotions, as summarised in Murray & Arnott (1993).

	Anger	**Happiness**	**Sadness**	**Fear**	**Disgust**
Speech rate	**Slightly faster**	**Faster or slower**	**Slightly slower**	**Much faster**	**Very much slower**
Pitch average	**Very much higher**	**Much higher**	**Slightly slower**	**Very much higher**	**Very much lower**
Pitch range	**Much wider**	**Much wider**	**Slightly narrower**	**Much wider**	**Slightly wider**
Intensity	**Higher**	**Higher**	**Lower**	**Normal**	**Lower**
Voice quality	**Breathy, chest tone**	**Breathy, blaring**	**Resonant**	**Irregular voicing**	**Grumbled chest tone**
Pitch changes	**Abrupt, on stressed syllables**	**Smooth, upward inflections**	**Downward inflections**	**Normal**	**Wide downward terminal inflections**
Articulation	**Tense**	**Normal**	**Slurring**	**Precise**	**Normal**

Table 2. Summary of most general correlates to emotions in speech. Table from Murray & Arnott (1993).

Thus, in fact there seem to exist some general tendencies for certain acoustic correlates by certain emotions. It is, however, unclear what aspect of the emotions the findings capture. Stibbard (2001:64) concludes as following: "While correspondences have been found between the strength of emotional arousal and phonetic features resulting from the physiological effects of arousal, no such features have been discovered which facilitate the more interesting goal of distinguishing the quality and valence of emotions". This rather pessimistic view might perhaps be a fruit of his work with developing tag categories for emotions based on spontaneous speech.

4 Evaluations of tagsets

We have seen that emotional categories are quite fuzzy in their definitions, and different researchers use different sets. However, examining the list over correspondences between emotions and acoustic features, there seem to be a connection between them. In this section I will give account for an evaluation of a tagset developed on basis of spontaneous speech, i.e. the EISP corpus (Greasley et al., 1995).

Systematic and careful evaluations of tagsets used for labelling emotions are generally lacking. One out of few is Richard Stibbards Ph D thesis (Stibbard 2001), which contains analysis and evaluation of two tagsets used in the EISP project. The EISP corpus consists of recordings of tv and radio documentaries, hand annotated on nine tiers which are time-aligned with the speech signal. The nine tiers contained information about perceived emotional force (four tiers), paralinguistic effects (one tier), intonation (one tier), word juncture (one tier) the verbal content (one tier) and miscellaneous noises (one tier).

In the EISP project perception experiments were carried out, with 158 subjects, in aim to get a valid classification of the emotions in a subset (89 speech samples) of the corpus. Two systems for emotion annotation in the perception tests were used.

1. Free choice. Subjects were free to label a stretch of speech with any emotion they found appropriate.
2. Forced choice. Subjects were instructed to choose one label among 5 specific emotions in a predefined set, consisting of anger, disgust, fear, happiness and sadness.

The free choice alternative resulted in 73 categories, and the forced choice alternative resulted in pure (significant majority) categorising of 57 out of the 89 samples. In the further analysis the data from forced choice was used.

Very few results indicating a correspondence between emotion and acoustic features were found. In short Stibbard (2001) reports findings supporting the category sad (low pitch and vocal qualification). Further he reports that angry speech had a significantly higher speaking rate than sad speech, but just for males, since female angry speech was so sparse in the data so no reliable conclusions could be drawn.

Since so few correspondences between the emotional categories and the acoustic features could be found, Stibbard argues that it is not possible to generalise from experiments with acted or synthesised emotional speech to natural speech, even though high recognition rates often can be found in those former experiments.

Stibbard (2001) concludes that a tagset with five basic categories is too coarse to model the spectrum of emotions in spontaneous speech, and a free choice classification gives an unsystematic categorisation, which is impossible to use. He further states that since the complexity of the manifestation of emotions in natural speech, the phonetics-only approach to a classification of emotions in speech is not possible. An approach, which in addition takes notice of other features and modalities, such as facial expressions, is according to Stibbard a better approach.

5 Synthesis and recognition of emotions in speech

Many attempts have been done in aim to incorporate emotions in synthesised speech. To evaluate synthesised speech is a different situation than evaluating spontaneous speech. In spontaneous speech all kind of emotions may appear, and the task is to model the emotive space. In evaluating synthesised speech, the task is simplified; the categories are already existing, and the task is to make the synthesis as prototypical as possible given the emotive categories.

Evaluations have shown quite good figures, and this is not surprising. Bear in mind that an argument against acted speech was that the emotions signalled were too stereotypical. This could be gained by the synthesis, i.e. in the synthesis one might try to model the most prototype manifestation of one emotion, and then one could guess that a human listener could recognise the emotion quite well.

5.1 Synthesis of emotions in speech

The reported results are often quite optimistic. In this subsection I will give examples of some experiments with synthesised speech.

Cahn (1989) reports findings from evaluation of results of a system for emotional speech, Affect Editor, which is based on formant synthesis. Affect Editor contains six emotional categories: Angry, disgusted, glad, sad, scared and surprised. The performance was evaluated with a forced choice perception test, and the results reported in a χ-square test gave highly significant figures. The emotions which was easiest to identify was sadness, witch was identified in 96% of the cases. The identification rate for the other emotions all lied around 50%.

In CHATAKO (Iida et al., 2000), a system based on concatenative speech synthesis, three emotional categories were evaluated. The categories were: Joy, Anger and Sadness. Also here a forced choice evaluation was conducted. Again, sadness was the most correctly identified category (82%)[1], followed by anger (60%) and joy (52%).
Similar results are reported by Carlson et al. (1992), who report that in an experiment with synthesised speech sadness was recognised in 95% of the cases, anger in 80%, and happiness in 42%.

5.2 Automatic recognition of emotions in speech

Automatic detection of emotions is, to my understanding, the most difficult part in this field. The reason is that we have to handle spontaneous speech as input, and that we i) do not have a reliable set of categories of emotions, and ii) do not have defined reliable acoustic correlates for emotions in spontaneous speech.
However, there are still attempts to attack this task. One is the experiment reported by McGilloway et al. (2000). In this setting they used speech with elicited emotions, i.e. subjects had to read texts designed to evoke strong specific emotions. The set of emotions consisted of : Afraid, Happy, Neutral, Angry and Sad. The data was processed in the ASSESS system, and after evaluation of the results McGilloway et al. report that 50% correct for a set of five emotional states is an attainable goal, based on the findings in their work.

6 Summary and Discussion

Concerning a robust and valid categorisation of emotions we are far from home. One main reason is, that we do not yet know exactly on what criteria humans judge emotions in natural speech. Work with more stereotypical varieties of speech, such as acted or synthesised gives better results than spontaneous speech. However, Schröder (2001) points out, that in evaluation tasks with forced choice, it is not a matter of *identifying* an emotion, rather to *discriminate* among the emotions in a given set, which is a much simpler task. We can get a taste of that difference in comparing the recognition rates for e.g. CHATAKO with the free choice figures given in Stibbards (2001) study. However, these two studies also gives a picture of the difference in working with spontaneous speech (as Stibbard) and working with synthesized speech (as in CHATAKO).
The successfully synthesised emotions are generally prototypical and strong. Such emotions does not occur in natural dialogues that often, a fact that might be due to sociocultural factors. Cowie (2000) points out this, and stresses the need for more research on milder varieties of emotions. A task, which is, I guess, hard to carry out before we have more detailed information of transmission of emotions in general.

Concerning the correspondences between categories of emotions and acoustic features, at least one category seems to be quite stable; Sadness. This emotion always got high recognition rate in the synthesis experiments, and it was also identified in the Stibbard (2001) study. Otherwise, spontaneous speech seems to be extremely hard to approach in the "right" way, since there is so much contextual information involved. In acted speech, there seem to be more correspondences, at least between "intended" emotions and acoustic features. This means that humans are able to produce stereotype emotive manifestations, a fact which in turn implies that humans have some kind of prototype categories for how emotions are realised in speech. However, humans do not seem to use those stereotypical forms in spontaneous speech. Perhaps because emotions are more complex in real life, and also because they might to a certain extent be masked, due to social convention. Just think about the saying "boys don't cry". Sayings like this imply that we learn actively from an early age to mask certain emotions.

Emotions are, however, also transmitted via the visual channel in face to face conversation. Stibbard (2001) stresses that in the field of emotions in speech, multimodality is not an option, it is a condition. The performance improves drastically when adding e.g. a visual face to the speech signal. One thing is for sure, there is much left to explore in this field!

7 Future Directions

In the EISP project spontaneous speech was use, and it was taken from documentary television programs. This seems to me a quite good way to overcome the ethical problems with spontaneous

[1] The figures are given for female speech, for male speech the figures are slightly worse.

speech. It might however be argued that the speech then is not fully naturally, but I think it is still more natural than acted speech. One thing to do might be to use documentary soap operas. These programs contain non-scripted speech, and are aimed to contain a lot of emotions. The participants are also discussing their feelings, which might give a hint of what they felt in a certain situation. An other gain with using documentary soap operas would also be access to multi-party dialogue. (Actually I heard (source:gossip) that Nordtalk is going to use exactly such data!).
The multimodal approach might also be fruitful. With regard to new annotation tolls, which allow alignment of speech and visual image, this work is getting better and better data collection resources. One drawback, however, is the drastic increase of data amount, which calls for knowledge about relevant features.
We should also remember to pose the question: what kinds of emotions are important to detect? What kinds of emotions are crucial for a smooth dialogue, and in what kind of applications do we want to use the emotional speech.

Acknowledgements

Many thanks to both of my reviewers, Loredana Cerrato and Sara Rydin. I am glad for all your comments that gave me the opportunity make a more clear and correct term paper! I also want to thank for all comments and suggestions made by people at the closing seminar, of the Speech technology course, at CTT.

8 References

Please note that the references marked with a star are references cited in some other work. Of course this should have been marked out in the text (or even better, the original reference should have been read!), but now it is not the case. I just decided to keep it this way, even though I know it is not fully fair.

*Banse, R. & Scherer, K. R. , (1996): Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70 (3).

*Brenner, M., Shipp, T., Doherty, E., Morrissey, P. (1983): Voice Measures of Physiological stress – Laboratory field data. In Titze & Scherer (Eds.): Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control. Denver, Colorado, USA.

*Brown, B. L. (1980): The detection of emotion in vocal qualities. In Giles, H., Robinson, W. & Smith, P. (Eds.): Language: Social Psychological Perspectives. Selected Papers from the First International Conference on Social Psychology and Language, held at the University of Bristol, England, July 1979. Oxford, UK: Pergamon.

Campbell, N. (2000): Databases of Emotional Speech. In Cowie, R., Douglas-Cowie, E. & Schröder, M. (Eds.) Proceedings of the ISCA Workshop on Speech and Emotion. Belfast, Ireland.

Campbell, N (2001): Building a Corpus of Natural Speech – and Tools for the Processing of Expressive Speech – the JST CREST ESP Project. In Proceedings of Eurospeech 2001, Aalborg, Denmark.

Carlson, R., Granström, B. & Nord, L. (1992): Experiments with emotive speech - Acted utterances and synthesized replicas. Proceedings of the International Congres on Spoken Language Processing.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000): Feeltrace: An Instrument for Recording Perceived Emotion in Real Time. In Cowie, R., Douglas-Cowie, E. & Schröder, M. (Eds.) Proceedings of the ISCA Workshop on Speech and Emotion. Belfast, Ireland.

*Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Kollias, S., Fellenz, W. & Taylor, J. G. (2001): Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine.

Douglas-Cowie, E. Cowie, R., & Schröder, M. (2000): A New Emotion Database: Considerations, Sources and Scope. In Cowie, R., Douglas-Cowie, E. & Schröder, M. (Eds.) Proceedings of the ISCA Workshop on Speech and Emotion. Belfast, Ireland

Cruttenden, A. (1986): Intonation. Cambridge University Press.

Greasley, P., Setter, J., Waterman, M., Sherrard, C., Roach, P., Arnfield, S., and Horton, D. (1995): Representation of prosodic and emotional features in a spoken language database. Proceedings of the 13th International Congress of Phonetic Sciences. Stockholm. 242-245.

*Friedhoff, A. J., Alpert, M., & Kurtzberg, R. L. (1962): An effect of emotion on voice. Nature, 193.

Hansen, J. (1999): Speech Under Simulated and Actual Stress (SUSAS). LDC 99S78.

*Huttar, G. L. (1968): Relations between prosodic variables and emotions in normal American English utterances. Journal of Speech and Hearing Research, 11.

*Hecker M., Stevens, K., von Bismarck, G. & Williams, C. E. (1968): Manifestations of task-induced stress in the acoustic speech signal. Journal of the Acoustical Society of America.

Iida, A., Campbell, N., & Yasamura, M. (1998): Design and Evaluation of Synthesised Speech with Emotion. Journal of Information Processing Society of Japan, 40 (2).

Klein, J. (1999): Computer Response to User Frustration. Technical report # 480, MIT Media Laboratory Vision and Modelling Group.

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Machiel, W., & Sybert, S. (2000): Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark. In Cowie, R., Douglas-Cowie, E. & Schröder, M. (Eds.) Proceedings of the ISCA Workshop on Speech and Emotion. Belfast, Ireland.

*Mozziconacci, S. (1998): Speech Variability and Emotion: Production and Perception. Eindhoven, Netherlands: Technische Universiteit Eindhoven.

Mozziconacci, S. (2000): The expression of emotion considered in the framework of an intonational model. In Cowie, R., Douglas-Cowie, E. & Schröder, M. (Eds.) Proceedings of the ISCA Workshop on Speech and Emotion. Belfast, Ireland.

Murray, I. R. & Arnott, J. L. (1993): Towards the Simulation of Emotion in Synthetic Speech: A Review of the Literature of Human Vocal Emotion. Journal of Acourstic Society of America 93 (2), 1097-1198.

*Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? Psychological Review, 97, 315-331.
*Pereira, C. (2000): Perception and expression of emotions in speech. Macquarie University: PhD thesis.
Pereira, C. (2000): Dimensions of Emotional Meaning in Speech. In Cowie, R., Douglas-Cowie, E. & Schröder, M. (Eds.) Proceedings of the ISCA Workshop on Speech and Emotion. Belfast, Ireland
PHYSTA. http://www.image.ntua.gr/physta/
*Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), Emotion: Theory, research, and experience: Vol. 1. Theories of emotion (pp. 3-33). New York: Academic.
Saeed, J. (1995): Semantics. Blackwell publishers.
*Scherer, K. R. (1981): Speech and Emotional States. In Darby, J. K. (Ed.) Speech Evaluation in Psychiatry. New York, Grune and Stratton.
*Scherer, K. & Ceshi, G. (1997): Lost luggage emotion: A field study of emotion-antecedent appraisal. Motivation and Emotion, 21.
*Scherer, K. & Ceshi, G. (2000): Studying affective communication in the airport: The case of lost baggage claims. Personality and Social Psychology Bulletin, 26 (3).
*Skinner, E. R. (1935): A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness. Speech Monographs.
*Soskin W. F. & Kauffman, P. E. (1961): Judgements of Emotions in Word-free Voice Samples. Journal of Communication.
Stibbard, R. M. (2001): Vocal Expression of Emotions in Non-laboratory Speech: An Investigation of the Reading/Leeds Emotion in Speech Project Annotation Data. Unpublished PhD thesis. University of Reading, UK.
*Weiner & Graham
Wichmann, A. (2000): The Attitudinal Effects of Prosody, and How They Relate to Emotion. In Cowie, R., Douglas-Cowie, E. & Schröder, M. (Eds.) Proceedings of the ISCA Workshop on Speech and Emotion. Belfast, Ireland
*Williams, C. E. & Stevens, K. N. (1969): On determining the emotional state of pilots during flight: An exploratory study. Aerospace Medicine, 40.
Williams, C. E, & Stevens, K. N. (1972): Emotions and Speech: Some Acoustic Correlates. The Journal of the Acoustic Society of America, vol 52 (4) pp. 1238-1250.

9 Some Selected Links and Resources

9.1 Demos

Felix Burkhardt, Technische Universität, Berlin. Simulations of emotion with speech synthesis.
http://www.kgw.tu-berlin.de/~felixbur/speechEmotions.html
Simulation of emotions, synthesized music+ singing. HAMLET.
http://www.computing.dundee.ac.uk/staff/irmurray/hamlet.asp

9.2 Corpora

Speech Laboratory, University of Reading, "Emotion in Speech Corpus". Not available for distribution.
SUSA, Speech Under Simulated and Actual Stress
http://www.ldc.upenn.edu/ldc/news/release/SUSAS.html

9.3 Projects

Affective computing, MIT Media Laboratory.
http://www.media.mit.edu/affect/
Expressive Speech Processing, Crest ATR
http://www.his.atr.co.jp/esp/
HAMLET, the Helpful Automatic Machine for Language and Emotional Talk. Applied Computer Science, Dundee University.
http://www.computing.dundee.ac.uk/staff/irmurray/hamlet.asp
MIMIC, Voice, Accent and Emotion Synthesis. Communications & MultiMedia Signal Processing, Brunel University.
http://www.brunel.ac.uk/departments/ee/Research_Programme/COM/Proj_Mimic/home.html

PHYSTA, Principled Hybrid Systems: Theory and Applications
http://www.image.ece.ntua.gr/physta/
The Reading/Leeds Emotion in Speech Project. Speech Laboratory, University of Reading.
http://midwich.reading.ac.uk/research/speechlab/emotion/