Angelika Braun und Louise Probst

Fear as an Affective Trait

Abstract: Dieser Beitrag beschäftigt sich mit dem Phänomen Angst als akut empfundener Emotion und nicht als allgemeiner Stimmungsgrad. Obwohl Angst als eine der sog. Basisemotionen recht gut erforscht ist und eine Reihe von Untersuchungen zur phonetischen Kodierung von Angst sowie zu der Fähigkeit von Hörern vorliegen, diese zu erkennen, gibt es nach wie vor offene Fragen. Zwei davon werden in diesem Beitrag behandelt. Zum einen geht es um unterschiedliche Grade von Angst: Werden sie durch identische Merkmale, wenn gleich in unterschiedlich starker Ausprägung, kodiert, oder durch kategorial unterschiedliche Merkmale? Zum anderen wird die interkulturelle Dimension der Emotionskodierung und -wahrnehmung thematisiert. Es wird untersucht, inwieweit Hörer in der Lage sind, angstbehaftete Äußerungen aus unvertrauten Sprachen und Kulturen allein anhand stimmlicher Merkmale zu dekodieren. Als Materialien dienen hierbei amerikanisch-englische, deutsche und japanische Stimuli, die einer TV-Serie entnommen wurden. Abschließend geht es um den Stellenwert visueller Informationen als Ergänzung auditiver Hinweise. Die konkrete Frage lautet hier, ob die visuelle Information eine korrekte Wahrnehmung von Angst erleichtert.

This contribution addresses the issue of fear not so much as a general sentiment or mood but as a momentarily experienced emotion. Although – being one of the so-called basic emotions – fear is rather well-researched, and there is a large number of studies covering the phonetic cues to fear and listener ability to recognize this emotion based on vocal cues alone, some questions still remain open. Two of them will be covered in the present contribution: One is the issue of different intensities of fear – will they be coded by differing degrees of identical cues or possibly by different cues altogether? The other concerns the intercultural dimension of encoding and decoding emotions. Are listeners able to decode fearful utterances across languages and cultures? This aspect is studied based on American English, German, and Japanese samples taken from the same TV series. Finally, the role of visual information complementing the audio is studied.
1 Introduction

Anxiety or fear can be looked at from many different angles. Specifically, they may be regarded as quasi-permanent mindsets or states which form part of the personality or as temporary affect bursts involving various physiological and cognitive processes, i.e. traits (Pekrun 2000). While the former is in the center of attention of the project as it stands, this contribution aims to add a different perspective.

Intuitively, we all have an idea of what “emotion“ entails, but when it comes to defining the term, things become more complicated. Kleinginna and Kleinginna (1981) collected close to one hundred definitions of “emotion“, reflecting different approaches to the concept. Generally speaking, three main approaches can be distinguished: basic emotions, dimensional approaches, and emotions as response patterns. These are not necessarily to be considered mutually exclusive; they just cover different aspects of categorizing emotions.

1.1 Basic Emotions

The basic emotions approach goes back to Charles Darwin (1872). Paul Ekman is one of the researchers who currently advocate the concept of basic emotions (1992a, 1992b). It relies on the visual presentation of photographs to viewers from different cultural backgrounds. A minimum of five emotions are thus identified, one of them fear. The others are anger, sadness, joy, and disgust. However, basic emotions were established from visual cues alone (Ekman 1992a, 1992b). They are assumed to be “biologically hardwired“ and universal (Kitayama & Markus 1994, 6). In the context of the present study, acoustic cues are the ones to be taken into account. It remains to be demonstrated that the concept of basic emotions extends to the auditory domain.

1.2 Dimensional Approaches

To some researchers, the basic emotion approach has been too simplistic since it is categorical. Therefore, dimensional models were constructed, characterizing emotions on continuous scales in a multi-dimensional space. These attempts go back to Wundt (1902), who proposed the dimensions pleasantness – unpleasantness, arousal – depression, and tension – relaxation. More recent representatives are Schlosberg (1954), Osgood (1966), and Russell (1980). However, the dimensions assumed are similar but not identical. For instance,
Schlosberg (1954) lists pleasantness – unpleasantness, attention – rejection, and sleep – tension (activation). Russell (1980), on the other hand, suggests that “affective states are, in fact, best represented as a circle in a two-dimensional bipolar space“ (p. 1161–1162) with the dimensions pleasure – displeasure and degree of arousal. Kienast (2002) and Paeschke (2003) propose a different model still: according to them, any (degree of) emotion can be placed in a three-dimensional space comprising valence (is the stimulus pleasant or unpleasant), arousal (high or low activation), and potency (weak/passive or strong/active) (Kienast 2002: 16). Irrespective of the exact model which one subscribes to, the dimensional approaches stress the point that emotions are multifaceted and that there may be no constant one-to-one relationship between the production and acoustics of emotionalized speech.

1.3 Appraisal Theories

In the context of the present study, the concept of regarding emotions as a physical reaction to events which are potentially posing a threat to the individual and whose significance must be assessed in order to react appropriately has to be discussed. This type of approach was published in the 1980s and 1990s by various researchers (Zentner & Scherer 2000). Probably the most elaborate concept of emotions as an appraisal is tied to the name of Klaus Scherer, who is among the most experienced scholars in the field of vocal emotion research (Scherer 1985, 1986, 1987, 1989, 1990, 2000, 2003, Banse & Scherer 1996, Scherer & Bergmann 1984, Scherer & Wallbott 1990, 2001 Scherer et al. 1991, Scherer et al. 2001).

Without rejecting the dimensional approach to emotions, Scherer introduces the concept of stimulus evaluation checks (SECs) some of which are split up into a number of subchecks as part of the reaction to sudden events (Scherer 1986). They are the novelty check, the intrinsic pleasantness check, the goal/need significance check, the coping potential check, and the norm/self-compatibility check. The outcomes of these checks are assumed to have somatic consequences for the speech production process.

The novelty check decides whether an event is novel or was to be expected. In the former case, the reaction will involve changes in the speech production process.

The intrinsic pleasantness check determines whether a stimulus is pleasant and therefore to be approached or unpleasant and thus to be avoided. The outcome of this check will result in very different reactions of the organism.
The *goal/need significance check*, including its various subchecks, determines the degree to which the organism gets involved. The higher the significance of a stimulus, the higher the ergotropic arousal which in turn has an impact on the speech production process.

The *coping potential check* involves an assessment of the organism’s potential to respond to the stimulus in question. A higher coping potential leads to high ergotropic arousal, low chances of coping with the challenge will result in trophotropic dominance.

The *norm/self-compatibility check* is assumed to occur late ontogenetically and phylogenetically. Therefore, Scherer does not assign specific responses of the vocal mechanism to this check but rather postulates an interaction with the reaction to other checks (Scherer 1986: 156).

The reactions to these checks are somatic and they help to explain the physiological processes within the vocal tract, which in turn evoke certain acoustic characteristics of speech. Due to the rapid succession of the SECs, the processes going on in the speech organs may be the result of a combination of SECs (and their subchecks) as opposed to separate reactions to each single one.

Vocalization, like all motor action, is achieved by the action of the striated musculature. It is assumed that each outcome of a SEC will have a specific effect on the somatic nervous system (SNS), producing a distinctive change in general muscle tension or local muscle action (which in turn will be modified by the effect of the following SEC outcome) (Scherer 1986: 148; 156).

Based on these observations, Scherer narrows down the concept to speech production and attempts to predict physiological processes within the speech organs as a reaction to the stimuli. These processes cue a certain and well-defined vocal behavior, which according to Scherer is universal by definition.

The results of the SECs for *fear* are hypothesized to be as follows: Fear involves a high degree of *novelty*, and a low *pleasantness* rating. Its *relevance* is high while it does not meet the expectations within the *goal-plan sequence*. It clearly obstructs reaching one’s goals, and it demands very urgent reaction. At the same time, the *control* over the event is low, as are the *coping power* and – to a lesser degree – the potential for internal *adjustment*. The *norm/self-compatibility* check is assumed not to be applicable to this emotion (Scherer 1986: 147).

According to this approach, cultural differences in emotional phenomena can be explained by differences in the perception and appraisal of events. Culturally different interpretations of an event may result in different emotions being felt.

No matter which of these approaches one subscribes to, the phonetic perspective is on how a speaker implements various emotions and if and how a
listener correctly interprets those emotions based on vocal cues alone, i.e., without facial expression or gestures being available. Further questions are whether those cues remain the same across varying degrees of an emotion, and whether they will work across languages and cultures. According to Ellsworth (1994: 30), “[s]everal dimensions of emotional appraisal are consistent across cultures.” Still, the stimulus evaluation may also reveal cultural differences.

1.4 Methodological Issues

There have been a number of studies trying to confirm empirically what was predicted by the theoretical considerations outlined above. However, any empirical research on emotion faces some fundamental methodological issues for which no generally agreed solution has been found until now:

1. It is very difficult, if not impossible to find recordings of naturally occurring emotions which are of a technical quality permitting acoustic analysis. An alternative is to evoke emotions in the laboratory, e.g. by having subjects play games or by presenting them with emotionalized movies. However, it seems ethically questionable to evoke “negative” emotions like sadness and fear in this way. If that is done, the resulting emotions are likely not to be very strong; in fact they may be too weak to render valid results. Therefore, most empirical studies have relied on actors’ renditions of simulated emotions instead.\(^1\)

In studies which use naturally occurring emotions, the question is whether there is a sufficient degree of emotion to be coded in terms of discrete vocal cues (Scherer 1986: 144). In studies which use actors simulating emotions, the question is whether they are in fact sufficiently natural. As is often observed in impersonation, salient cues may be overemphasized whereas subtle ones may be neglected (Scherer 1986: 144). Thus, the full range of cues serving to signal emotions may not be exploited by the actors and will then not show up in these production studies\(^2\). A compromise between evoked

---

1 This is the situation which listeners are confronted with when trying to decode emotions while talking over the telephone.

2 A rare exception is the study by Williams & Stevens (1972), which includes an analysis of the live radio coverage of the Hindenburg disaster on May 6, 1937.

3 It has also been argued that the dividing line between real-life and acted emotions may not be as sharp as it appears since emotions form part of human interaction which automatically involves some enactment (Banse & Scherer 1996: 618). The authors use this as an argument in favor of using actors as speakers.
and laboratory speech consists in what can be called authentic speech, i. e. speech produced by actors not with the intention of portraying an emotion for research purposes but with the intention of sounding natural. The continuum between natural speech and laboratory speech can be summarized as follows: Naturally occurring emotions – emotions evoked by movies or games – emotions produced by actors with the intention of rendering natural emotions (authentic speech) – emotions produced by actors with the intention of portraying emotions for scientific analysis (laboratory speech).

2. Another problem is that – whether when working with actors’ simulations including evoking emotions in the laboratory or with naturally occurring material – there are varying intensities of emotional states. For instance, there is a vast difference between being a little sad because the dress one wanted to buy was not available in the right size on the one hand and the heartbroken state of someone whose spouse has just walked out on them on the other. Little, if any, attention has been given to this aspect in the past. Again, to the present authors’ knowledge, Scherer and colleagues are the only ones to have addressed this issue in any detail (cf. e.g. Scherer 1986, Banse & Scherer 1996). In these authors’ view, varying degrees of emotion will help explain some discrepancies and contradictions in previous findings on the encoding stage as well as variation in listener ability to identify certain emotions. Their predictions on the effect of different degrees of emotion are derived from various studies with varying research designs (Banse & Scherer 1996: 617). Then they proceed to test a small number of acoustic phonetic parameters against their predictions and find many but by no means all of their hypotheses confirmed.

3. Even if one were to get hold of authentic recordings, the problem of mixed emotions would arise. For analytic purposes, it is desirable to study emotions in isolation, but in reality, there will probably be mixed emotions most of the time, i. e. fear mixed with sadness, sadness mixed with anger, etc.

4. When it comes to intercultural studies, another aspect has to be considered: As was indicated earlier on (cf. 1.3 above), there seems to be some kind of cultural filter which is applied to the emotion which one actually feels, i. e. so-called feeling rules and display rules (Matsumotu 1990: 1993). These establish limitations on the degree of emotionality which may be externalized or even perceived. So the sensation of an emotion as well as the degree to which it is shown may vary according to cultural rules. It may e.g. be considered improper to make them known to other people, let alone foreigners. A widely known example of this type of behavior is the reaction by the locals to the Fukushima disaster in Japan – people were smiling when
interviewed about it. Western viewers had to be educated about the interpretation: it was not that the Japanese were happy, they had just been trained not to show their anxiety and fears.

5. There is still a striking paucity of intercultural studies on emotions which involve both the production and the perception sides of all languages involved. Most of the time, the perception of emotions encoded in one language is studied in listeners belonging to a different speech community. Exceptions are Abelin 2004; van Bezooijen et al. 1983; Braun & Heilmann 2012, Shimoda et al. 1978, and Tickle 2000. Even though they all find differences between different languages/cultures, the numbers are still too small to draw firm conclusions.

The present research tries to address two of these issues, attempting to shed light on different degrees of emotions and intercultural aspects. The research questions asked are as follows:

1. Are different degrees of emotions in general and fear in particular encoded in a linear way, i.e., essentially using the same mechanisms, but to a different extent?
2. Are there any intercultural differences in the encoding and decoding of fear?
3. If so, do the differences in the encoding and the decoding of fear increase with linguistic and/or cultural distance?
4. Will the presence of visual information aid the decoding process?

2 Experiment 1 – Degrees of Emotion

There are a few studies, notably Banse & Scherer (1996) which aim at covering different facets of emotionality and do so by asking actors to portray members of the same emotion family (e.g. hot and cold anger, panic and anxiety, elation and happiness). The approach taken here is somewhat different: speakers are expressly asked to portray varying degrees of one and the same emotion. The difference between these approaches is not trivial: We have chosen this approach in order to avoid varying interpretations of concepts like anxiety and elation as well as the danger of eliciting mixed emotions. Furthermore, the present authors do not share the view held by Scherer (e.g. 1986: 157) who regards
cold anger as a lesser degree of hot anger. We consider both to be variants of anger, which can each be expressed in different degrees.  

2.1 Materials and Methods

The aim of this study was to include the production and perception of three degrees of the basic emotions hot anger, cold anger, joy, sadness, disgust, and fear. In view of the complexity of the encoding task, it was decided to work with professional actors. The six professional speakers selected (native speakers of German; 3 female, 3 male) each had many years of experience in radio, television, and stage work. For the purpose of this study, five nonsense sentences were constructed which consists of words forming part of the German lexicon and adhere to German syntax but are nonsensical like e.g. Die Maserung der Wand geht ins Wasser ‘The pattern on the wall disperses into the water’. Speakers were asked to portray these sentences in six different emotions and in a neutral speaking style. Specifically, their task was to express three degrees of each emotion: low, medium and extreme. No further instructions were given to the speakers, and none of them signaled any trouble or questions concerning the task.

Recordings were made in a professional studio of one of the largest German radio stations (Westdeutscher Rundfunk) using a condenser microphone Neumann U 87 and a DHD-RM4200 preamplifier. The sampling rate was set to 48 kHz with a 16 bit quantization. Speakers were encouraged to repeat the stimuli until they were satisfied with the results, and they actually made use of this option.

This procedure led to a total of 90 items for fear (6 speakers x 3 degrees of emotionality x 5 utterances), and 30 more which were considered neutral and served as a reference. This contribution will focus on the production and perception of fear.

2.2 Analysis

Several acoustical parameters were analyzed in order to establish how speakers portray an emotion. They include measurements of fundamental frequency and

---

4 For similar reasons, we have reservations against translating cold anger as Ärger and much prefer kalte Wut.
speaking tempo, but also an assessment of voice quality. With respect to fundamental frequency (f0), the mean, standard deviation and range (max f0 - min f0) were measured. In order to avoid errors due to the automatic pitch extraction, all measurements were manually checked. Since male and female voices were to be compared, all results were converted to semitones.

Speaking tempo was analyzed in terms of articulation rate (AR in terms of syllables per second excluding pauses), because this measure best represents the actual velocity of articulator movement. The number of pauses was analyzed separately.

Voice quality was analyzed auditorily according to the classification by Laver (1980). Only laryngeal voice qualities were included in the analysis, i.e. creaky voice, creak, breathy voice, whispery voice, whisper, tense voice, harsh voice, falsetto as well as combinations of the above which are physiologically possible.

A perception experiment was carried out to study the perceptual relevance of the acoustic findings. In order to avoid listener exhaustion, only the low and extreme degrees of emotionality were used. 61 students, all native speakers of German, served as listeners. They were 18-30 years of age, 40 being female and 21 male. Their task was to listen closely and mark the perceived emotion as a forced choice decision on an answering sheet listing five emotions (joy, fear, sadness, anger, and disgust) as well as neutral.

Results will be presented in two steps: First, the relation between the three degrees of emotionality will be discussed. In a second step, the emotional renditions will be compared to the neutral utterances serving as a reference.

### 2.3 Results

#### 2.3.1 Fundamental Frequency (f0)

As is shown in Figure 1, the difference in mean f0 between low and medium degrees of fear is not very sizable. However, the extreme degree of fear sticks out, showing a considerably higher f0 than the other two degrees. Thus, the relationship among the various degrees of emotionality is incremental, but by no means linear. This corresponds perfectly well to the expectations as e.g.
expressed by Banse & Scherer (1996). Fear definitely and reliably causes F0 to rise, and the higher the degree of fear, the stronger the increase.

With respect to the relation to neutral, most speakers signal a low and even medium degree of fear by a slightly lower than neutral mean f0, whereas the values for the extreme degree of fear are always well above those for neutral. Thus mean F0 will not serve as a cue which can readily be exploited by listeners to recognize a low degree of this emotion.

![Fig. 1: Mean f0 for the various degrees of fear](image)

The standard deviation (SD) of the fundamental frequency corresponds to the impression of melodiousness in a voice. All speakers exhibit a decrease in SD for the low and medium degrees of fear, making it more monotonous. Five out of six speakers mark the extreme degree by a larger f0 variability. Once again, only the highest degree of fear presents a reliable cue for listeners to exploit in the identification of this emotion.

The f0 range shows a slight increase from low to medium degree. For the low degree the f0 range is comparable to that of the neutral stimuli for most speakers. With the exception of two speakers, the extreme degree of fear is marked by a substantially wider range (up to 5.2 semitones wider than neutral). Figure 2 shows the differences of f0 range for the various degrees of fear in relation to neutral.

Two speakers show a behavior differing sharply from that of the other four. They either exhibit little change in their F0 range or even narrow it further
across the three conditions. For these two speakers, listeners cannot rely on this parameter for the recognition of this emotion.

2.3.2 Speaking tempo

_Fear_ is the only emotion contained in our study which shows an increase in rate of articulation with the degree of emotionality, if one averages over speakers. However, this does not hold if individual speakers are considered. There is considerable individual variation, the only constant being that five out of six subjects show a faster than neutral rate of articulation in the portrayal of an extreme degree of _fear_. The group of speakers is split in half for the lower degrees, three of them articulating faster and three lower than neutral. Certainly, the individual data do not support the notion of a linear increase with degree of emotionality. However, differences from _neutral_ are not very large in absolute terms (about .5 syllables/second), and the actual speaking tempo shows overlap with not only the _neutral_ stimuli but also with low degrees of _joy_ and _sadness_.

Articulation rate in itself thus does not seem to be used as a marker for this emotion by our speakers in any consistent way except for extreme degrees of _fear_ for which it does increase. This raises the question whether pausing may have played a role in the portrayal of _fear_. A look at the data shows that the
number of pauses increases compared to *neutral* but remains essentially unchanged with degree of *fear* (cf. Table 1).

| Degrees of fear | low | medium | extreme | total |
|-----------------|-----|--------|---------|-------|
| speaker         |     |        |         |       |
| F1              | 0   |        |         | 0     |
| F2              | 2   | 2      |         | 4     |
| F3              | 1   | 1      | 1       | 3     |
| M1              | 1   | 1      | 4       | 6     |
| M2              | 3   | 1      | 4       | 8     |
| M3              | 2   | 3      | 1       | 6     |
| **total**       | 6   | 6      | 7       | 19    |

On an individual level, and considering the portrayal of various degrees of *fear*, a pattern does emerge. Speakers tend to either speed up or use pauses to mark a certain degree of *fear*. For instance, speaker M3 uses pauses, but not speech rate to portray low and medium degrees of *fear*, whereas he uses the latter for an extreme degree. Speaker F1, on the other hand, does not pause at all, but uses speech rate to mark medium and high degrees of fear. Thus in our material there are indications to the effect that something like a trade-off exists between these two parameters for the emotion of *fear*.

### 2.3.3 Voice quality

In order to establish voice quality, the material was analyzed auditorily, and the presence of qualities differing from modal voice was noted. There may be more than one voice quality other than modal voice present in a single utterance. For instance, a speaker may start out with modal voice, then switching to tense voice and finally to whispery voice. Therefore, the number of observed voice qualities exceeds the number of utterances.
Whereas all speakers stick to modal voice in their neutral reference utterances, they use a wide range of voice qualities for the representation of fear: modal voice, breathy voice, tense voice, whispery voice and whisper.  

The number of different voice qualities per item increases when speakers move from the low emotional degree to the medium or extreme level of emotionality. The low degree of fear is displayed mostly by modal voice and breathy voice, one of these two often prevailing throughout the whole utterance. In the medium degree, combinations of voice qualities become more frequent. Whispers and – to some extent – also tense voice occur. It is interesting to note that there is a gender effect on whispery voice in the sense that it is only used by male speakers. The extreme degree of fear is mostly externalized by tense voice (either throughout the utterance or for most parts of it), often accompanied by whispery voice, occasionally ending in (voiceless) whisper. No speaker utters the extreme degree using modal voice only. In some cases breathy voice is added to modal and tense voice, while speakers rarely combine modal voice or tense voice with whispery voice.

For the difference between whispery voice and whisper cf. Laver (1980: 121).
2.3.4. Perception

Perception rates differ for the five emotions and neutral speaking style, and also for the varying degrees of emotionality. Fear is correctly identified in 40% of the items (for comparison: neutral is correctly identified 84% of the time). However, sorting the items for degree of emotionality shows that the extreme degree of fear is much more often recognized correctly than the low degree, as is evident from Figure 4.

![Figure 4: Perception rates for low and extreme degrees of fear](image)

The confusion patterns also merit attention: The low degree of fear was not even recognized as such at chance level. By far the most frequent response was sadness, followed by neutral. This is remarkable because sadness and fear are located at opposite ends of the activation dimension and thus should not be prone to be confused. For the extreme degree of fear the picture is different. The recognition rate is quite high, but the confusion pattern remains the same (predominantly sad and neutral).

These results demonstrate that listeners cannot be expected to correctly identify low levels of emotionality at a level above chance. This may in part explain different recognition rates in previous studies. It also reflects the difference between the three intensities of fear on the production side. Specifically, the production data for the low degree of fear were often very close to neutral, but also to those for sadness. This applies in particular to the F0 measures.
(mean, SD, and range). In view of this similarity, it does not come as a surprise that low degree fear was more often mistaken for sadness and neutral than correctly identified.

## 2.4 Discussion

It can be concluded from the present study that speakers handle the portrayal of fear in part conventionally and in part individually, making use of changes in fundamental frequency, speaking tempo and voice quality. As a general observation, fear is represented by a higher pitch (increased mean f0), and less modulation for the low and medium degrees and more for the extreme degree, and generally a wider f0 range than neutral. These findings are in line with those reported by Scherer (1986), Banse & Scherer (1996) and Paeschke (2003) as far as mean f0 and f0 range are concerned. These studies also report greater variability (greater SD) for fear, but this expectation is not met by our data.

Speaking tempo seems to be a highly individual parameter, but all speakers make use of it to distinguish emotional speech from neutral. For the portrayal of fear, the general observation of increased tempo confirms the assumptions based on the results of previous studies (Banse & Scherer 1996, Kienast 2002, Paeschke 2003). Pauses tend to be a suitable method in portraying emotions, however they also appear to be strongly dependent on individual speaker preferences. Speakers seem to link these two parameters: the less they increase their articulation rate, the more they make use of pauses. This trade-off between the mechanisms of speech tempo regulation suggests different strategies of encoding.

There is hardly any previous research on voice quality in emotional speech. Kienast (2002: 89) reports the presence of whisper and breathy voice for fearful speech. In the present study, speakers prefer to assign breathy voice to the low degree of fear, whispery voice to the medium degree and tense voice to mark the extreme degree, the latter being a setting not studied by Kienast. Our results indicate that voice quality is applied variably to different degrees of emotion and cannot be expected to be tied to any single emotion in a uniform way. As Kienast (2002: 126) already notes, there is no single phonatory setting solely characteristic to fearful speech. The expectation by Klasmeyer & Sendlmeier (2000) that falsetto might play this role, does not hold.

An important question in the context of this study is whether vocal cues remain the same or change with the varying degrees of emotion, in other words, whether the portrayals are linear or categorical. It emerges from the measurements that the encoding of the various degrees of emotion is exponential rather
than linear. The acoustic representations are incremental concerning mean f0, pauses and – more or less – articulation rate, where individual externalizations differ. The encoding of fear with respect to voice quality is neither linear nor exponential but can be termed categorical, meaning that there is a change in preferred voice quality.

Running counter to the concept of basic emotions as being universal and therefore generally easy to recognize, fear is getting the lowest recognition rates of all emotions tested in this study. The low degree seems substantially more difficult to perceive than the extreme, and listener performance in detecting it is at less than chance level. This does not come as a surprise considering that findings on recognition rates for fear differ greatly between previous studies. While Tickle (2000) finds rates of 40% (Japanese speakers and listeners) and 49% (English speakers and listeners), Kienast reports 85% correct recognition for her German subjects. Braun & Heilmann found correct recognition of fear in 60% (Japanese speakers and listeners) to 91% (German speakers and listeners) of all cases. Banse & Scherer (1996) report 42% correct identification for anxiety and 36% for panic. Some of these results suggest a cultural difference in recognition rates. This question will be addressed in Experiment 2.

3 Experiment 2 – Intercultural aspects

An important issue in emotion research concerns the notion of universality vs. cultural determination of emotions. The concept of basic emotions presupposes universality, but it heavily relies on the visual domain. Mesquita and Frijda (1992) present a review of earlier studies and find evidence for and against universality both with respect to photographs and speech. Specifically, Japanese viewers/listeners are cited as exhibiting an encoding and decoding behavior which is quite different from that of American subjects. For instance, recognition rates were lower for Japanese listeners than for Americans. We therefore carried out an experiment involving American, German, and Japanese subjects.

3.1 Materials and Methods

In an attempt to use authentic speech rather than recordings made by actors for the purpose of studying emotions, excerpts from the leading male and female characters in the TV series Ally McBeal were used as stimuli. The American original as well as the German and Japanese dubbings were studied. We chose these
languages because we intended to study one language which is linguistically and culturally close to the American English original as well as a one that is distant from it. The results reported here form part of a study covering four basic emotions (anger, sadness, fear, and joy) as well as neutral stimuli. We had aimed at studying stimuli from the leading male and female characters, but fearful utterances were found for the female character only. Four short scenes of about 5-8 sec. duration were selected for analysis based on the assessment of the German dubbed version by three female researchers. Special care was taken to choose utterances whose wording did not convey a particular emotion. The stimuli were analyzed auditorily and acoustically. In order to study the perception side in addition to the production data, three listener groups (American, German, and Japanese) were asked to assign different emotions to the stimuli.

3.2 Results

3.2.1 Encoding

The production parameters studied included those covered by Experiment 1, complemented by perceived loudness, fluency, and breathing patterns. Table 2 summarizes the results. They are depicted with reference to the results for neutral stimuli. At first glance, there are no big between-speaker differences. All three speakers show a greatly increased mean F0 as compared to neutral utterances, amounting to about 45 Hz, a slight increase in perceived loudness as well as disfluencies and dyspnea. These findings are prototypical for the emotion of fear and are predicted by Scherer’s model (Banse & Scherer 1996: 617).
On the other hand, differences between the speakers emerge with respect to voice modulation as expressed by F0 standard deviation, voice quality, and speaking tempo. Whereas the American and Japanese speakers show hardly any deviation from neutral, a strong increase in F0 modulation and range is observed for the German speaker. She once again differs from the other two with regard to speaking tempo: while the American and the Japanese speakers speed up when displaying fear, the German speaker shows a tendency to slow down. Moreover, the German speaker is the only one to show creaky voice, i.e. a voice quality which is usually not encountered in fearful utterances (cf. also Experiment 1). The Japanese speaker, on the other hand, shows fewer disfluencies than the other two. This is the only parameter which suggests an intercultural difference between Japanese on the one hand and English and German on the other. As for the rest, the Japanese speaker sticks out much less than the German speaker does.

To sum up, the German speaker in particular deviates from the other two in several respects, which might make it difficult for Japanese and American listeners to recognize the intended emotion.

---

7 It has to be kept in mind, though, that we analyzed what can be considered a standard set of parameters, and that the perceptually relevant feature which distinguishes Japanese encoding of fear has yet to be discovered.
3.2.2 Decoding

What do the said differences between the German speaker and the other to mean for the perception side of things? Table 3 shows the results.

**Tab. 3: Recognition rates for fear by the different listener groups**

| Speakers         | Listeners          |              |                |              |
|------------------|--------------------|--------------|----------------|--------------|
|                  | American Engl.     | German       | Japanese       | mean         |
| American English | 82%                | 79%          | 49%            | 72%          |
| German           | 41%                | 91%          | 44%            | 56%          |
| Japanese         | 62%                | 58%          | 60%            | 61%          |
| Mean             | 64%                | 74%          | 52%            | 66%          |

The table shows that each listener group performed best with stimuli in their own language, the margin being larger for the American listeners than for the other two groups. This confirms previous findings, e.g. by van Bezooijen et al. (1983). Generally speaking, German listeners perform best. On the other hand, German stimuli are recognized worst – 56% as opposed to 62% for the Japanese and 72% for the American speakers. This may be due to the fact that the German speaker used an encoding strategy which was somewhat different from the other two (cf. 3.2.1 above).

As far as the intercultural perspective is concerned, the German and the Japanese listeners behave exactly as expected: They achieve the best results with stimuli in their own language and do worst with those in the structurally distant language. The latter does not apply to the American listener group, though. They do much worse with the German stimuli than with the Japanese, mistaking them for sadness much more frequently (57%). We can only speculate about the reasons for this behavior: these results may be attributable to the differences in encoding or possibly German sounded even more foreign than Japanese to the American listeners, none of whom had ever been exposed to German.

If one takes a look at confusion matrices, i.e. at the emotions with which *fear* was most often confused, the result is quite straightforward and also very similar to that of Experiment 1: *fear* is by far most often mistaken for *sadness* (cf. Table 4). The reverse is also true – *sadness* is most frequently confused with fear. This is surprising since *fear* and *sadness* are at opposite ends of the activation dimension, and Banse & Scherer (1996: 616) generally predict opposite
acoustic manifestations. This, however, is not confirmed by our data (Braun & Heilmann 2012). Specifically, we found correspondences between sadness and fear on the production side with respect to voice quality, dyspnea, and fluency.

**Tab. 4:** Confusion matrices for intended fear and sadness according to listener language

| Listeners perceived | All listeners | AE listeners | G listeners | J listeners |
|---------------------|--------------|--------------|------------|------------|
| Fear                | 66%          | 64%          | 74%        | 52%        |
| Sadness             | 20%          | 25%          | 11%        | 30%        |
| Neutral             | 7%           | 5%           | 8%         | 9%         |

Still another reason for the poor performance and the confusion may be that the only channel available to the listeners was the audio. We therefore checked whether performance increased with video only or with both audio and video information present. Table 5 shows the results. Somewhat surprisingly, listener performance does not increase substantially if both audio and video information are available.8

**Tab. 5:** Recognition rates for fear: audio, video, and audio-video combined

| Channel          | American English | German | Japanese | mean  |
|------------------|------------------|--------|----------|-------|
| audio            | 64%              | 74%    | 52%      | 63%   |
| video            | 56%              | 53%    | 34%      | 48%   |
| audio-video      | 64%              | 85%    | 54%      | 68%   |
| mean             | 61%              | 71%    | 47%      |       |

The largest gain from the added video information is exhibited by the German listeners who show an improvement in recognition rate by 11 percentage points compared to audio alone. The fact that the video signal alone was always recognized much worse than the audio alone comes as a surprise given that most emotion theory is based on observations in the visual domain.

---

8 This result is even more surprising given the fact that the AV testing was done about a week after the audio testing, so that listeners already knew both the video and the audio. If anything, recognition rates should have gone up due to that circumstance.
3.3 Discussion

On the production side, there is no specifically “Japanese encoding“ of fear. The Japanese and the American speakers make use of very similar mechanisms to signal this emotion. If there is one of the speakers who differs from the other two, it is the German one. This comes unexpected given previous results on the restrictions on showing fear in the Japanese culture. For instance, Tickle (2000) found that her Japanese listeners did worse with Japanese stimuli than with English ones at identifying fear and attributes this finding to the fact that Japanese speakers are hesitant to externalize fear in the first place. We did not find evidence to support this notion. It is possible, though, that the Japanese speaker modelled her speech behavior after the American original and expressed fear more freely than usual.

In decoding, the Japanese listeners generally did worse than Germans and Americans. It thus seems as if there may be a cultural effect in the sense that Japanese listeners have considerably more problems to decode fear than e.g. Germans do, probably they are not used to fear being uttered freely. This may also be one reason why Japanese listeners mistake fear for sadness more frequently than the other two.

Even though it sounds implausible from a dimensional perspective, we are not the only ones to have found this confusion pattern. Abelin & Allwood (2000), Scherer et al. (2001), and Tickle (2000) report the same pattern, the latter even both ways, i.e. sadness was also mistaken for fear. This reciprocal confusion also occurred in our study.

In our data, the role of the video was not as prominent as expected. Recognition rates were lower than for the audio and audio-video mode alone, and the latter was only mildly superior to the audio alone. This casts some doubt on putting an emphasis on the visual channel in emotion theory and strongly suggests including the auditory information to a larger extent.

4 General Discussion

Experiment 1 is the only one that we are aware of which aimed at studying three intensities or degrees of emotions. The only comparable study is that by Banse & Scherer (1996), but those authors used two different emotional labels as opposed to different degrees of one and the same emotion, e.g. anxiety and panic for the emotion family of fear. What makes results more difficult to compare is that they do not list their findings with reference to neutral recordings but to the
overall mean across all emotion recordings. Their predictions are made with reference to a neutral state of mind, though. Banse & Scherer basically assume a sharp increase in all parameters studied here for high intensity *fear* (*panic*) and a moderate increase for low intensity *fear* (*anxiety*). Not all of these predictions are met by their own study. For instance, mean F0 for *anxiety* was found to be lower than predicted.

In our Experiment 1, the three degrees of emotion were distinguishable and marked by different speech production features. The acoustic differences between them were rather not linear but exponential for mean F0 and its standard deviation, i.e. the differences between low level *fear* and medium fear were much less pronounced than those between medium *fear* and extreme *fear*. In the case of speaking tempo, the signaling is more complex, i.e. we observe an interplay between tempo and pausing which seems to be individual.

Our results with respect to mean F0 are in agreement with Banse & Scherer’s (1996) findings but not with their predictions: low degree *fear* generally exhibited lower values than *neutral*. This may indicate that the prediction needs to be revised including the implications for emotion theory. We found speaking tempo for low intensity *fear* quite close to *neutral* and medium degrees of *fear* to be lower than *neutral*, whereas Banse & Scherer predict an increase. It seems as if there is a lot of variability in everything but a high degree of emotion, an aspect to which much more attention should be devoted.

As far as voice quality is concerned, the changes are categorical, and the results of our two experiments coincide. Scherer has pointed to the importance of voice quality analysis early on (1986: 145) but we are not aware of extensive studies including a qualitative analysis except for Braun & Heilmann (2012). The reason for this is obvious – the analysis is tedious, requires extensive training, and the results do not lend themselves to statistical analyses. Still, the results reached so far strongly suggest to include a more detailed analysis of voice quality parameters. The acoustic analysis of assumed correlates to selected voice quality features like e.g. the difference in intensity between the first and second harmonic may be too crude to capture the fine detail of emotion expression.

On the perceptual strand, low degree of *fear* fails to be recognized by listeners above chance level even though the actors were satisfied to have produced the emotion in question. Thus statements like “listener-judges are able to recognize reliably different emotions on the basis of vocal cues alone” (Banse &

---

9 We think it fair to assume that the samples from Ally McBeal represent a medium intensity of *fear*. 
Scherer 1996: 615) clearly require some qualification. This finding may also help to explain why previous researchers found differences in acoustic cues for fear and also why the recognition rates vary to the extent that they do (cf. 2.4 above). A further study would have to look into the impact of added video information on low degrees of verbal emotion. It would probably be much larger than it was in our study.

Furthermore, the perceptual fear-sadness conjunction is striking. This means that fear is often mistaken for sadness and vice versa. This conjunction applies to both the stimuli created by professional actors in Experiment 1 and the authentic speech from the TV series in Experiment 2, which means that it cannot easily be attributed to a flawed elicitation procedure. Furthermore, this confusion pattern applies to all listener groups alike, i.e. it does not follow a cultural pattern. Other authors have observed similar confusions in their perception experiments (cf. 3.3 above).

This result is surprising because dimensional models of emotion will generally place fear and sadness at opposite ends of the arousal dimension (cf. Russell 1980). However, there is evidence for the similarities between these two emotions in the production data. Mean F0 and related parameters as well as voice quality were found to be very similar in both experiments. This casts some doubt on the placement of both these emotions on the arousal dimension which is part of basically all dimensional models of emotion (cf. Russell 1980) both from a production as well as a perception perspective.

The present study supports the argument for culture specificity of emotions. We implemented a research paradigm suggested by those authors (Scherer et al. 2001: 88), i.e. working with both “encoders and decoders from several different countries [...] to test whether decoders from the countries involved would recognize emotion portrayals by encoders from their own countries most accurately.” In Experiment 2, we actually found this to be the case. All three listener groups were best at recognizing fear that was encoded in their own language and performed much worse with linguistically and culturally more distant samples.

In past research on emotion and culture, it has been taken for granted that cultural distance coincided with linguistic distance. What remains to be studied in this context is whether a cultural effect can be distinguished from a language effect. This would involve studying culturally close but typologically different languages such as Norwegian and Finnish and vice versa, e.g. Chinese and Japanese. Through experiments separating the two, effects created by the linguistic structure, e.g. the prosody of a language, could be separated from effects which are really created by cultural mechanisms.
References

Abelin, Åsa (2004): Cross-Cultural Multimodal Interpretation of Emotional Expressions. An Experimental Study on Spanish and Swedish. Proceedings of the ISCA supported International Conference on Speech Prosody, Nara, 647–650.

Abelin, Åsa & Jens Allwood (2000): Cross-linguistic interpretation of emotional prosody. Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, 110–113.

Banse, Rainer & Klaus R. Scherer (1996): Acoustic Profiles in Vocal Emotion Expression. Journal of Personality and Social Psychology 70 (3), 614–636.

Bezooijen, Renée van, Stanley A. Otto & Thomas A. Heenan (1983): Recognition of Vocal Expressions of Emotion: A Three-Nation Study to Identify Universal Characteristics. Journal of Cross-Cultural Psychology, 14, 387–406.

Braun, Angelika & Christa M. Heilmann (2012): SynchronEmotion. Frankfurt am Main: Lang.

Darwin, Charles (1872: The Expression of Emotions in Man and Animals. Chicago: University of Chicago Press. [Reprinted 1965].

Ekman, Paul (1992a): An Argument for Basic Emotions. Cognition and Emotion 6, 169–200.

Ekman, Paul (1992b): Are There Basic Emotions? Psychological Review 99 (3), 550–553.

Ellsworth Phoebe C. (1994): Sense, Culture, and Sensibility. In: Shinobu Kitayama & Hazel Rose Markus (Hrsg.), Emotion and Culture. Empirical studies of mutual influence, 23–50. Washington/D.C: American Psychological Association.

Kienast, Miriam (2002): Phonetische Veränderungen in emotionaler Sprechweise. Aachen: Shaker.

Klasmeyer, Gudrun & Walter F. Sendlmeier (2000): Voice and Emotional States. In: Raymond D. Kent & Martin J. Ball (Hrsg.), Voice quality measurement, 339–357. San Diego: Singular Publishing Group.

Kleinginna, Paul R. Jr. & Anne M. Kleinginna (1981): A Categorized List of Emotion Definitions, with Suggestions for a Consensual Definition. Motivation and Emotion 14 (3), 195–214.

Matsumoto, David (1990): Cultural Similarities and Differences in Display Rules. Motivation and Emotion 14 (3), 195–214.

Matsumoto, David (1993): Ethnic Differences in Affect Intensity, Emotion Judgments, Display Rule Attitudes, and Self-Reported Emotional Expression in an American Sample. Motivation and Emotion 17 (2), 107–123.

Mesquita Batja & Nico H. Frijda (1992): Cultural variations in emotions: A review. Psychological Bulletin 112 (2), 179–204.

Osgood, Charles E. (1966): Dimensionality of the semantic space for communication via facial expressions. Scandinavian Journal of Psychology 7 (1), 1–30.

Paeschke, Astrid (2003): Prosodische Analyse emotionaler Sprechweise. Berlin: Logos. Reihe Mündliche Kommunikation. Bd. 1.
Pekrun, Reinhard (2000): Persönlichkeit und Emotion. In: Jürgen H. Otto, Harald A. Euler & Heinz Mandel (Hrsg.), Emotionspsychologie. Ein Handbuch, 334–348. München: Beltz.
Russell, James A. (1980): A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39 (6), 1161–1178.
Scherer, Klaus R. (1985): Vocal Affect Signaling: A Comparative Approach. In: Jay S. Rosenblatt, Colin Beer, Marie-Claire Busnel & Peter J. B. Slater (Hrsg.), *Advances in the Study of Behaviour* 15, 189–244. New York: Academic Press.
Scherer, Klaus R. (1986): Vocal affect expression: A review and a model for future research. *Psychological Bulletin* 99 (2), 143–165.
Scherer, Klaus R. (1988): Criteria for emotion-antecedent appraisal: A review. In: Vernon Hamilton, Gordon H. Bower & Nico H. Frijda (Hrsg.), *Cognitive Perspectives on Emotion and Motivation*, 89–126. Dordrecht: Nijhoff.
Scherer, Klaus R. (1989): Vocal correlates of emotional arousal and affective disturbance. In: Hugh Wagner & Antony Manstead (Hrsg.), *Handbook of social psychophysiology*, 165–197. Oxford: John Wiley.
Scherer, Klaus R. (1990): Theorien und aktuelle Probleme der Emotionspsychologie. In: Klaus R. Scherer (Hrsg.), *Psychologie der Emotion*, 1–38. Göttingen: Hogrefe.
Scherer, Klaus R. (2000): Partikuläre und integrative Ansätze. In: Jürgen H. Otto, Harald A. Euler & Heinz Mandel (Hrsg.), Emotionspsychologie. Ein Handbuch, 151–164. München: Beltz.
