# Emotions in Speech: Juristic Implications

Erik J. Eriksson[1,*], Robert D. Rodman[2,**], and Robert C. Hubal[3]

[1] Dept. Philosophy and Linguistics, Umeå University, Sweden
[2] Dept. Computer Science, North Carolina State University, USA
rodman@ncsu.edu
[3] Technology Assisted Learning Ctr., RTI International, USA

**Abstract.** This chapter focuses on the detection of emotion in speech and the impact that using technology to automate emotion detection would have within the legal system. The current states of the art for studies of perception and acoustics are described, and a number of implications for legal contexts are provided. We discuss, inter alia, assessment of emotion in others, witness credibility, forensic investigation, and training of law enforcement officers.

**Keywords:** acoustic parameters, affect, emotion, emotional categories, forensic, juristic, speech.

## 1 Introduction

Current natural language computational systems are able to infer much information from a person's spoken input based both on gross acoustic features and on lexical, syntactic, and semantic analyses. Information that can be inferred includes meaning, style, and certain speaker characteristics (Cole, et al., 1997 [1] Frank, et al., 2002 [2]). However, current systems are only able to draw inferences from a subset of features in the acoustic signal including intonational and stress patterns, overall loudness, peculiarities of phonation, and other distinctive properties of speech. Human listeners, on the other hand, have access to a full set of features and are able to integrate them in such a way as to acquire a wealth of information from them and apply this knowledge to identifying and classifying speakers, apprehending subtle shades of meaning, inferring implicatures and other pragmatic factors and, most relevant to the present work, perceiving affect and interpreting the speaker's emotional state.

Automated detection of emotion in speech holds considerable promise in many areas, including deception detection during interviews (Fuller, et al., 2006 [3]),

[*] Erik Eriksson's contribution to this paper was funded by a grant from the Bank of Swedish Tercentenary Foundation Dnr K2002-1121:1-4 to Umeå University for the project 'Imitated voices: A research project with applications for security and the law'.

[**] Corresponding author. Address all correspondence to: Robert Rodman Department of Computer Science, Box 8206 North Carolina State University Raleigh, NC 27695-8206, Fax: 919-515.7480, 919-515.7896.

C. Müller (Ed.): Speaker Classification I, LNAI 4343, pp. 152–173, 2007.

separating cognitive from affective experiences in a clinical setting (Susca, 2006 [4]), and call center applications (Yacoub, et al., 2003 [5]). Here, we focus on legal settings. The entirety of legal actors (Maroney, 2006 [6]) – attorneys, defendants, executive officials, expert witnesses, judges, jurors (grand and otherwise), law enforcement officers, legislators, plaintiffs, prosecutors, regulators, suspects, victims, and witnesses - experience emotion and may display those emotions, the knowledge of which may have far-reaching ramifications for other actors. We first identify several areas where emotions in speech can affect legal judgment and decision making. Broadly, these areas involve assessment of emotion in others, emotions and memory (concerning witness credibility), emotions and culture (including effects on forensic investigation), and emotions in legal scholarship. Later, we describe some implications of emotion detection on the training and assessment of law enforcement officers, attorneys, and other actors in juristic processes. We also describe perceptual and acoustic studies supporting an optimistic outlook for the automated detection of emotion in speech.

## 2 Effects in Legal Contexts

### 2.1 Assessment of Emotion in Others

Detection of a particular emotion and the degree to which it is felt based on measurable acoustic parameters of speech could prove useful to many actors in the juristic system. Knowledge of emotional state would surely have powerful implications in the following situations and many more similar to them:

- Law enforcement officers would benefit by knowing what emotions a suspect is experiencing during interrogations, perhaps to gauge credibility or extent of knowledge.
- Attorneys could gauge credibility of a client or the client's level of knowledge or understanding by discerning the client's emotional state or state changes during discussions of the client's past behavior.
- Prosecution attorneys could profit by knowing what emotions potential jurors are experiencing during jury selection questioning so as to gauge their likely reactions to evidence.
- Defense attorneys could benefit by knowing what emotions the prosecution attorneys are experiencing as they present a case, perhaps to gauge the effects of alternate defense strategies.
- Jurors could benefit by knowing what emotions a witness is experiencing during testimony, to gauge credibility and remorse.
- Defendants could benefit by knowing what emotions a judge is experiencing when handing down a decision, perhaps to gauge the possibility of suspension of the sentence or parole.

In all these situations, the legal actors naturally exhibit emotion through their behaviors, including their gestures, facial expressions, body language, and, in particular, their speech. A device of some sort with the capability of detecting vocal patterns influenced by emotion could augment information gained through lexical, syntactic, and semantic analyses. The device need not be hidden or unobtrusive; the point would be to gather additional potentially useful information in a noninvasive manner.

To gather such information, automated emotion detection relies on exposure to a database of speech segments tagged with their emotional category (neutral, sad, angry, etc.). Each segment is also classified in accordance with its acoustic properties.

Cowie, et al. (2005) [7] present a number of such databases, generic in the sense that a general model, applicable to all speakers of the particular language, is to be constructed. Such databases may be thought to underlie *speaker independent* emotion detection. (Further detail is provided in the appendix.)

Speaker specificity of feature sets in emotion encoding was shown by Hozjan and Kačič (2006) [8]. *Speaker dependent* emotion detection may turn out to be far more effective, since one is modeling a single individual for the purpose of predicting that individual's emotional state, but it is formidable from the data collection point of view. This option is open only to experimenters who work with the same participants over a prolonged period and therefore have sufficient time to collect the amount of data needed for the modeling process. One such environment occurs in the context of vocal computer aided instruction (CAI). After some weeks of interacting with the same students, an automated CAI system, working in parallel with an intelligent tutoring system, would observe their frustrations, anxieties, achievements, and glowing successes, and from these observations build individualized databases (Burns & Capps, 1988 [9]). With such a database, a model of vocal affect/emotion individualized to the student could be produced. From that point forward, whenever the tutoring system would identify emotion in the student's voice, it could respond appropriately.

In most forensic and other legal settings, one would use a generic model. For instance, during jury selection, there are scarcely any data available regarding individual members in a jury pool. Similarly, most interrogations are relatively short, hardly enough to enable the speaker-specific learning that would need to occur for overcoming a generic model. However, in situations where witnesses testify at great length (e.g., Slobodan Miloševic at his trial in The Hague), there may be ample time to assemble and put to use a speaker dependent emotion detection system.

Finally, it remains to find conclusive results as to whether emotion detection is language dependent or independent. That is, if, say, the lowering of pitch is a mark of sadness in an English-speaking environment, would it be so in a Serbian-speaking environment, or a Zulu-speaking environment? Would the situation differ between tonal languages such as Mandarin Chinese, and the mostly non-tonal languages of Europe? These and other interesting questions are open to basic research, hence their effects on legal processes are yet unknown.

## 2.2 Emotions and Memory

Perception studies have shown that humans correctly recognize emotions only 60 to 70 percent of the time (Picard, 1997 [10]; Scherer, 2003 [11]). Courts often mistakenly put too much trust in eye and ear witnesses (Solan & Tiersma, 2003 [12]). Solan and Tiersma argue that courts might look further into mathematical and/or computer aided analysis of witnesses' testimonies to gauge their reliability, and this may include an assessment of the emotional state of the witness.

Witness credibility is so important that judgment of credibility of witnesses is included in the jury instructions, since eyewitness testimony is viewed as direct evidence and adds to the prosecution's circumstantial evidence. But aside from being able to accurately assess the emotions that a witness is experiencing during testimony, emotions play a crucial role in memory that needs to be better understood.

Cognitive psychologists commonly distinguish among memory formation or encoding, association, and reconstruction. All of these processes can be affected by emotion (Forgas, 2001 [13]). For instance, emotional events are thought to receive some preferential processing (Christianson, 1992 [14]; Taylor & Fragopanagos, 2005 [15]) and thus, like all stimuli that receive attentive processing, lead to more stable and perhaps more accurate memory traces. By the same token, surrounding stimuli associated with the event that are not attended to are not encoded, hence are not retrievable later. Similarly, when events are reconstructed during eyewitness testimony, salient stimuli are better recalled than less salient stimuli. Salience can be related to three factors: first, to the witness's prevailing emotion (the closer to the emotional stress of the experience, the more accurate the memory is considered to be) (Jackson, 1995 [16]; see also the encoding specificity principle in Tulving & Thomson, 1973 [17]); second, to the witness's confidence (which, however, is largely uncorrelated with the accuracy in memory of an event; Olsson, 2000 [18]); and third, to suggestions provided by others (Loftus & Ketcham, 1991 [19]; Loftus, 2003 [20]). Emotionally encoded stimuli can also alter attention; such stimuli can divert attention to themselves and away from lesser emotionally laden stimuli and thus render the emotionally encoded stimuli more salient in the context (Taylor & Fragopanagos, 2005 [15]).

There is some concern, though, that emotion or stress adversely affects eyewitness memory (Deffenbacher, et al., 2004 [21]). Understanding a witness's emotion may have interview and courtroom implications. For instance, during an interview, a defense attorney may note how an adversarial witness is becoming agitated or aroused and may justifiably claim bias in the witness's recall of events. Similarly, during questioning of a witness in the courtroom, a judge may notice the witness becoming stressed or emotionally involved by the questioning and may call for a different line of questioning or a recess to calm the witness. As stated, the credibility of a witness depends in large part on the witness's level of emotion.

### 2.3 Emotions and Culture

Cultural differences in emotions might impose serious problems in a forensic investigation. For instance, foreign language interpretations in police interviews have been shown to generate problems, especially if the interpreter is not properly trained, or if a police officer acts as the interpreter (Berk-Seligson, 2002 [22]). Russell (2002) [23] argued that even highly trained interpreters who are not serving dual positions such as interpreter and police officer during interviews (see Berk-Seligson, 2002 [22]), may affect interview outcomes or even verdicts in court proceedings. Russell argued that literal and correct translations of foreign languages should be emphasized. However, this might not be possible as there are numerous translation difficulties and ambiguities between languages (Wierzbicka, 1999 [24]) and between cultures (Semin, et al., 2002 [25]). Wierzbicka reported, as an example of problematic translations, the word marah, which is the closest translation of angry in Malaysian. However, the Malaysian word is incompatible with aggression and is closer in meaning to resentful than angry. So if a translator interprets the word marah as angry for a person under investigation for a crime involving aggressive behavior, the person could very well appear aggressive, even though the exact meaning of the word does not imply aggression.

Similarly, cultural differences can play a role in the expression of emotion during an investigation or during courtroom proceedings. Jackson's (1995) [16] descriptions of emotional expression, for instance, are based on the Anglo-Saxon culture which encourages hiding of emotions (Tsai & Chentsova-Dutton, 2003 [26]; Wierzbicka, 1999 [24]), a norm that is not as true in other cultures. A number of studies (e.g., Ekman, et al., 1987 [27]; Ekman & Keltner, 1997 [28]; Markus & Kitayama, 1991 [29]; Mesquita & Markus, 2004 [30]) have investigated cultural similarities and differences in facial expression and interpretation of facial expression, finding that a number of expressions represent universal emotional displays, but that cultural influences on the context of the expression or on self-concept can affect judgments of expression.

Further, during courtroom proceedings, as well as other negotiations, emotions provide means to conclude a positive result (Kopelman, et al., 2006 [31]). Kopelman, et al. showed that participants in negotiations of varying kinds (immediate outcomes, time limited ultimatum, and prolonged negotiations) displaying positive emotions throughout the discussion were more likely to arrive at a subjective interest-based agreement (e.g., a win). By making the legal actors aware of their own display of emotions, as well as letting a mediating judge be aware of these, a more neutral, and even perhaps less biased, ruling might be achieved.

### 2.4 Emotions in Legal Scholarship

The judicial system already acknowledges emotions as an integral component. The system itself is based on social morality norms which, in turn, are based on emotional values and views of the society (Karstedt, 2002 [32]). For instance, hate crimes are described by the culprit's attitude towards the victim and the punishment of such a crime is controlled, partly, by the culprit's emotions surrounding the event, the impact of the event on the victim(s), and the judge's perception of the social implications of the event (i.e., the need for statutory ruling) (Karstedt, 2002 [32]). Thus, emotions are inherently intertwined with the law (Vidmar, 2002 [33]).

Maroney (2006) [6], therefore, points out that emotions can and should be studied owing to their undisputed relevance to the law. A six-pronged approach is recommended for study within the law-and-emotion rubric:

1. Focus on a particular emotion such as disgust, fear, or shame, and pertinent legal considerations. For example, in the legal defense of battered women who strike back, legal recognition of the experience of the fear emotion, and the behaviors it may engender, could and should be taken into account in the course of legal proceedings.
2. Focus on causes of emotional states, or "affective forecasting". For example, a litigant who imagines winning a certain level of damages in a civil case may experience projected happiness, and that experience may induce the litigant to make important legal decisions such as rejecting a low, but perhaps appropriate, offer of settlement because it doesn't produce the projected level of happiness. The reverse, projecting sadness in the event of losing, may persuade a litigant to accept an inappropriately low offer of settlement.
3. Focus on theories of emotion - both methodological and within disciplinary categories - and how current law reflects any one particular theory, and whether current law favors one theory over another.
4. Focus on legal doctrine. Whereas the first three items focus initially on emotion, this item examines how a particular area of the law - most obviously *criminal law* - is subject to understanding emotion. Indeed, emotions such as passion are encoded in the legal system wherein a crime committed "in the heat of *passion*" may be regarded differently than a crime committed "in cold blood". But needless to say the entire panoply of legal taxonomy may be affected by the emotional state of the legal actors that are involved.
5. Focus on the theory of law, as contrasted with focusing on the theory of emotion. Here, the starting point is a particular theoretical approach to law followed by an analysis of theories of emotion from that particular point of view. For example, one might examine the emotional dimensions of "restorative justice".
6. Focus on how a particular legal actor's behavior is influenced by his or her emotions, and the emotional state of those with whom he or she interacts. This, in fact, is the primary focus of the present chapter and of most research in this area, the first five foci being as yet relatively unexplored.

Clearly, then, knowledge of emotions can be useful for the judicial system. However, means of collecting signals that carry emotional content are needed. One apparent source of such signals is the voice. It is readily available and can be collected non-invasively. Key here is to separate salient features for emotion recognition. The next section presents means of collecting, classifying, and analyzing emotion in speech.

# 3 Emotions in Speech

In interactions between individuals the voice is a major tool and often (e.g., during telephone conversations) the only tool of communication. Individuals not only convey the explicit meaning of their utterance via vocal quality when they are interacting with other humans, but they also present the receiver with information of a more complex nature through vocal affect, which is one of the surface manifestations of the emotions that the speaker is experiencing.

Studies of emotions in speech can be divided into two major fields, perceptual and acoustic. Perceptual studies use human listeners to assess the emotional content in a speech segment. These studies are often used for cross-culture comparisons, or to test the relation of specific acoustic cues to particular emotions (Yang & Campbell, 2001 [34]) (For example, when a change in pitch is heard, how frequently will a human listener perceive anger, or confusion, or finality?).

The other type, acoustic studies, uses speech data to extract salient features that are linked to specific emotions. Often the emotional content of speech is first perceptually evaluated and tagged before acoustic feature extraction is employed. (For example, when a human listener perceives anger, how frequently is, say, a rise in overall pitch detected acoustically?)

A more detailed description of how emotions and acoustic features are investigated may be found in the Appendix. The following sections present findings from studies in perceptual analysis and acoustic analysis. Legal implications are also presented.

## 3.1 Perceptual Studies

A number of perception studies have been undertaken (e.g., Breitenstein, van Lancker, & Daum, 2001 [35]; Cowie, et al., 2000 [36]; Douglas-Cowie, et al., 2000 [37]; Laukka, et al., 2005 [38]; Mullennix, et al., 2002 [39]; Scherer, et al., 2001 [40]; Thompson & Balkwill, 2006 [41]; Wurm, et al., 2001 [42]). Human classification rates have been found to lie near 70% for unknown voices (Picard, 1997 [10]; Polzin & Waibel, 2000 [43]). However, human listeners have been shown to use other effects, such as cultural or linguistic influences on the processing of emotions (Scherer, et al., 2001 [40]; Thompson & Balkwill,2006 [41]; Tickle, 2000 [45]), to assist in deriving additional information from speech.

At least four forms of speech have been used as emotional stimuli to present to human listeners. One form, speech collected from actors or amateurs, achieves its emotional content from emotions elicited by either instruction or by self-induced emoting. That is, the actors are asked to try to feel the emotion while recording speech. A second form, speech collected from actors or amateurs who are unaware of the purpose of the recording, comes from emotion elicited by a mood induction technique. That is, the emotion is induced by the nature or content of what is recorded. A third form is speech collected from real life television or radio shows. A fourth form is synthetic speech constructed by altering one or several acoustic features to (purportedly) reflect a particular emotion (Iida, et al., 2003 [46]).

Westermann, et al. (1996) [47] conducted a meta-analysis, investigating some 250 studies and comparing eleven methods of emotion induction. They found that pictorial and movie elicitation caused the strongest effect on listeners. They further found that the effect is shifted in favor of negative emotions, particularly in self-imposed methods. Finally, they found that the elicitation effect is raised if the listener is aware of the purpose of the experiment. One implication is that emotion-inducing images (e.g., a bloody stairwell) presented, say, as evidence in a courtroom may have strong effects on observers such as jurors, particularly for negative emotions (of possible benefit to the prosecution), while making jurors aware of the strong effects may help them understand their reactions (of possible benefit to the defense).

It has been argued that by recording actors eliciting emotions, full-blown and unambiguous emotions can be collected (Liscombe, et al., 2003 [48]; Scherer, et al., 2001 [40]). However, it has also been argued that acted speech is different from genuine emotions (Bachorowski & Owren, 2003 [49]; Batliner, et al., 2003 [50]) but may contain a core of truth as it often is reliably decoded by listeners (Scherer, 2003 [11]). Actors are thought to over-characterize emotions when producing them and tend to elicit emotions primarily via pitch and prosody (Batliner, et al., 2003 [50]). In fact, it has been argued that the material used in emotion research always should be collected in the real world (e.g., Cowie, et al., 2000 [36]; Cowie & Cornelius, 2003 [51]; Douglas-Cowie, et al., 2000 [37]; Picard, et al., 2001 [52]) and thus be authentic and genuine. However, this generally means that the emotions elicited will be less strong (Batliner, et al., 2003 [50]; Cowie & Cornelius, 2003 [51]; Douglas-Cowie, et al., 2000 [37]). Douglas-Cowie, et al. also point out that in using real-life material the emotions may be hidden by social or other factors or can be expressed in a degraded or mixed version. This implies that when the effects of emotional stimuli need to be made explicit to observers or recipients, particularly for real-world stimuli, the social or environmental context in which the stimuli occur should be recreated as best as possible. Witness memory and credibility may depend on these ideas.

During perceptual evaluations (also applicable to acoustic analysis), the emotional content of the utterance is related to a specific class of emotion. Cowie and Cornelius (2003) [51] argued that the size of the chosen set of emotions to identify would impact the level of categorization. That is, the more emotion classes to distinguish between, the more insecure the categorization will be. Hence, Cowie, et al. (2000) [36] and Douglas-Cowie, et al. (2000) [37] argued that the use of two (or three) continuous dimensions was a more succinct way of representing emotions. The dimensions here were valence and arousal (and power). Cowie, et al. (2000) [36] showed low inter-subject variation in mapping emotional content onto these dimensions. Laukka, et al. (2005) [38], however, argued that three dimensions are insufficient to represent differences between certain emotion classes, but increasing the number of dimensions to represent the emotional space not only makes data analysis difficult (hard to train listeners, etc.), it also limits the amount of explanation each dimension can give

(Cowie, et al., 2005 [7]; Scherer, 2003 [11]). Further, listeners often agree on the emotional content of stimuli in sets of limited number of categories (Scherer, 2003 [11]) promoting the more “classical” emotion categories (anger, happiness, etc.). The implication for legal actors, then, may be to limit their interpretations of vocal affect to these classical or universal (Ekman, et al., 1987 [27]) emotions, on the basis that other actors will tend to agree with those interpretations.

These perceptual evaluations can be done using either experts (e.g., Banse & Scherer, 1996 [53]) or laymen (e.g., Cowie, et al., 2000 [36]; Douglas-Cowie, et al., 2000 [37]; Tickle, 2000 [45]). However, a pre-processing by experts to filter out poor examples has been argued to be inappropriate as emotional elicitation may differ greatly between participants and any elicitation is an exhibition of the emotions, regardless of whether it is strong or not (Bachorowski & Owren, 2003 [49]). Hence, the expertise or experience of the observer or recipient may affect how emotional stimuli are interpreted. For example, judgments or responses to different cultural emotional stimuli may depend on knowledge of the other culture.

Indeed, perceptual studies have been undertaken to investigate effects such as cross-cultural similarities in emotion decoding (Scherer, et al., 2001 [40]; Thompson & Balkwill, 2006 [41]) or both emotion encoding and decoding (Tickle, 2000 [45]). The results of these studies suggest that there are similarities in both encoding and decoding, but differences also exist. Scherer, et al. found that individuals from nonwestern cultures, specifically Asian, were less successful in decoding the emotional content encoded by German actors, than were, for instance, Germans, French, and Americans. Note that Scherer, et al. used carefully constructed nonsense stimuli to lower the impact of language, though that may actually adversely affect individuals from cultures that demand larger contexts in which to assess emotional display (Mesquita & Markus, 2004 [30]). Tickle (2000) [45] found similar effects using English and Japanese encoders and decoders. Thompson and Balkwill suggested that there are both universal cues and culture-specific ones. They found in-group advantages to the extent of significance, suggesting that there are cultural-specific cues that other cultures overlook during decoding. The results were based on English listeners’ judgments of English, German, Chinese, Tagalog, and Japanese utterances with prosodically encoded emotions.

Perceptual investigations are needed to test the salience of acoustic features to specific emotions, or to find unambiguous emotional content in speech samples. That is, either a hypothesized feature (e.g., pitch or fundamental frequency) is tested for emotional salience, in which case the feature is manipulated synthetically (Mozziconacci, 2002 [54]), or collected material needs to be emotionally labeled and confirmed (Batliner, et al., 2003 [50]; Douglas-Cowie, et al., 2000 [37]). Most of the studies presented in the next section used data perceptually evaluated before extraction of features.

Further details regarding emotion categories may be found in the Appendix.

### 3.2 Acoustic Studies

The main goal of acoustic studies has been to link any particular acoustic feature (or set of features) in a speech sample to the emotional state expressed (given by perceptual investigations) in that sample. Further, mathematical algorithms for classification have been used to correlate acoustic features with emotion exemplars, and therefore support a method of classification of emotions based on acoustic features.

There are two ways of investigating the emotional salience of acoustic features in speech. Mozziconacci (2002) [54] argued that the best way of finding acoustic correlates to specific emotions was to employ an analysis / re-synthesis method. A specific acoustic feature is determined a priori to be correlated with some emotion(s). The feature is then manipulated by voice synthesis while keeping other features constant. If listeners to the synthetic voice perceive the emotion in the presence of the feature, the feature can be said to correlate to some degree with the emotion(s).

The other way of investigating the emotional salience of acoustic features in speech is to use collected material and use data driven methods of extracting multitudes of features and measure the emotional salience for each of these features.

Numerous acoustic features have been investigated over the years. In one of the most inclusive studies (Batliner, et al., 2003 [50]), the task was to find acoustic correlates of user frustration. Features such as fundamental frequency (F0) and statistics for F0 (mean, standard deviation, overall range, minimum and maximum), temporal durations (length of pauses, etc.) with various reference points, speech rates, and spectral energy and tilt were all examined.

Oudeyer (2003) [55] included a plethora of features, but reduced the original number (exceeding 200) to a succinct few using a feature selection algorithm. Based on that algorithm, Oudeyer found the most salient content to be localized in the first part of the spectrum (0 - 250 Hz). However, only three of the commonly used features (mean, minimum, and maximum) for F0 were found to be among Oudeyer's top 20 features.

Features based on models of the spectrum have also been used, focusing on the first ten (Oudeyer, 2003 [55]) or sixteen (Polzin & Waibel, 2000 [43]) coefficients of a mel-frequency cepstrum and a twelve feature log frequency power coefficient vector (Nwe, et al., 2003 [56]). In sum, it appears that pitch (fundamental frequency) (e.g., Banse & Scherer, 1996 [53]; Bänziger & Scherer, 2005 [57]; Batliner, et al., 2003 [50]; Mozziconacci, 2002 [54]) or spectral information below 250 Hz (McGilloway, et al., 2000 [58]; Oudeyer, 2003 [55]) have high impact for emotion classification purposes. This is in accordance with findings by, inter alia, Williams and Stevens (1972) [59]. However, Toivanen, et al. (2004) [60] found, in conjunction with Oudeyer, that the commonly used measurements of pitch (fundamental frequency) such as mean and median did not show great significance for emotion classification. Toivanen, et al. used spoken Finnish as their language of choice and Oudeyer used French, whereas many of those that found fundamental frequency mean and median to have an impact used English as a language of choice. Therefore it can be argued that mean and median of fundamental frequency might be language specific cues.

Batliner, et al. (2003) [50] also followed McGilloway, et al. (2000) [58] in that they divided the speech into sections of interest. McGilloway, et al. called these sections "tunes" and defined them to be sections of arbitrary length between specified events (such as pauses of approximately 180 milliseconds). Hence, both Batliner, et al. and McGilloway, et al. could specify prosodic events based on the fundamental frequency curve during any particular tune and thus use these as features.

Prosodic events have also been used to separate emotional events into categories (e.g., Mozziconacci, 2002 [54]; Paeschke, et al., 1999 [61]; Polzin & Waibel, 2000 [43]; Schröder, et al., 2001 [44]). Batliner, et al. (2003) [50] found that prosodic cues alone were insufficient to achieve high classification rates, however Bänziger and Scherer (2005) [57] found successful discrimination between four emotions using "... simple F0 contours - such as F0 mean or F0 range ..." (p.265). Mozziconacci and Hermes (1999) [62] successfully correlated some intonational patterns to a subset of emotions, but found only partial correlation during a perceptual evaluation functioning as validation of the findings. Other studies have used fundamental frequency data to separate emotional content (Burkhardt & Sendlmeier, 2000 [63]; Dellaert, et al., 1996 [64]; Lee, et al., 2001 [65]; McGilloway, et al., 2000 [58]; Oudeyer, 2003 [55]; Paeschke, et al., 1999 [61]; Polzin & Waibel, 2000 [43]; Roy & Pentland, 1996 [66]). These studies are relevant because they imply that observers will rely on not just one acoustic feature of an emotion-inducing stimulus to categorize the effects. Legal actors ranging from law enforcement officers to interviewers need to learn to assess the breadth of behaviors exhibited by a subject, including the variety of vocal characteristics, before determining his or her emotional or psychological state (see Hubal, et al., 2004 [67]; Link, et al., 2006 [68]).

Voice quality (e.g., formant distributions of particular vowels, or phonation types such as creaky voice) has been studied by a few researchers (e.g., Burkhardt & Sendlmeier, 2000 [63]; Gobl & Chasaide, 2003 [69]). Burkhardt and Sendlmeier found some correlation to emotion involving voice qualities but this was not the case for Gobl and Chasaide (see also Janniro & Cestaro, 1996 [70]) who argued that the correlation between voice quality and expressed emotion is uncertain. Clearly, further research is needed in this area.

In order both to assess the multivariate description of emotions by a set of features and to use features to classify new utterances, multivariate tools and classification algorithms are needed. Oudeyer (2002, 2003) [71] [55] performed a comparison between several classification algorithms. These included several different types of neural networks, decision trees, *k*-star, kernel density, linear regression, several support vector machines, and AdaBoost. Oudeyer found that the most successful algorithm for his data was the AdaBoostM1/C4.5 method, which applies a machine learning technique to refine and stabilize the output of a decision tree method. This produced results as high as 96.1% classification rates with speaker dependent data and optimal feature sets. Other popular classification algorithms include a maximum likelihood Bayes classifier (Dellaert, et al., 1996 [64]; Polzin & Waibel, 2000 [43]), kernel regression (Dellaert, et al., 1996 [64]), k-nearest neighbor (Dellaert, et al., 1996 [64]; Toivanen, et al., 2004 [60]) and hidden Markov models (Nwe, et al., 2003 [56]).

Cues in which emotion is conveyed can also be found in higher linguistic levels. Batliner, et al. (2003) [50] used conversational artifacts, syntactic structure, and dialogue acts to find trouble in communication. They found that repetitions of statements, especially when an utterance is repeated word for word, are cues to severe problems and therefore annoyance in the observer or recipient.

Similarly, Hozjan & Kačič (2006) [8] used various durations (e.g., sentence, syllable, specific sounds) in conjunction with fundamental frequency and amplitude measurements. They found that the mean energy of segmented speech (i.e., energy means taken over segments of speech) had the highest significance for emotion classification. This measurement was closely followed by the durations of affricates, plosives, sonorants, and fricatives in that order. (See also Petrushin & Makarova, 2006 [72], for this effect in Russian.) They also found that any cue on its own was insufficient to separate any emotion pair, but combinations of different cues did. However, the combinations differed for each emotion pair and each speaker and they suggested that speakers might have personal preferences when selecting available cues to depict a specific emotion. It should be noted that their speakers had a mixed language background, which could indicate cultural or language dependent differences, although speakers of the same language showed no more similarities in cue setup than mixed-background speakers.

Schröder (2000) [73] investigated the impact of acted German affect bursts on perceived emotion. Affect bursts are short utterances produced appropriately for a specific emotion. For example, growling was used to convey threat. Schröder had actors choose whichever burst they saw fit for a specific emotion and found correlations between emotion and choice of sound. However, in a follow-up listening experiment the correlation was much weaker and confusion rates were greater.

## 4 Implications for Training and Assessment of Legal Actors

As was suggested, detection of emotion in others has implications for witness credibility and forensic investigation. Looking forward, automated detection of emotion using tools based on the research just described may have further implications for legal training and assessment.

### 4.1 Interaction Skills Training

Law enforcement officers and others in the legal system who regularly encounter suspects and witnesses need training on learning to assess those persons' emotions. As an example, there is a need for training law enforcement officers in managing encounters with the mentally ill (Engel & Silver, 2002 [74]). Frank, et al. (2002) [2] describe a system for that form of training, where law enforcement officers encounter a synthetic character (i.e., a computerized agent) and the officers must learn, using interaction skills alone, to de-escalate the situation. Along with gestural and facial expressions given by the character, emotion expressed in speech is critical and informative for these officers.

De-escalation is just one procedure that persons in the legal system perform. Other procedures include interrogation, negotiation, and crowd control, and emotion comes into play for all of these procedures. For instance, law enforcement interrogations done incorrectly can be suggestive and can lead witnesses to confident, emotionally laden, detailed mistaken memories (Loftus, 2003 [20]). All of these procedures also require, at some point, assessment of emotion as part of determination of intent. Training systems using technology similar to that of Frank, et al. that incorporate emotional characters offer great advantages in reliability, safety, and ultimately success in performance on the job.

### 4.2 Situated Assessment

Not only must the emotional state of individuals sometimes be assessed by an observer, but also the individual's responses to emotional stimuli must sometimes be assessed. This might be true, for instance, to gauge a defendant's behavior when emotional evidence is introduced. The closer the social and environmental context is to that which is on trial, the more realistic the response can be expected to be. That is, whereas practitioners of situated learning strive to have learners gain knowledge and acquire skills in the contexts that reflect how knowledge and skills are applied in everyday situations (e.g., Anderson, Reder, & Simon, 1996 [75]), a new line of research aims to place the individual within a simulated environment that closely mirrors the real environment, and measures the individual's assessment of the situation (e.g., Paschall, et al., 2005 [76]). The situation might measure physical behavior, but also verbal behavior (i.e., speech) exhibited by the individual. Paschall, et al. showed that a simulation is capable of differentiating between groups of participants, such as individuals diagnosed or not diagnosed with conduct disorder. The ability to detect a person's state through behavior exhibited in response to emotional stimuli holds promise for interrogation, for identifying remorse or feelings of guilt, for judging the effects of culture, and for judging credibility.

### 4.3 Admissibility of Machine-Detected Emotion as Evidence

Like all new technologies (e.g., fingerprints or DNA testing, at different times), admissibility as evidence may depend on the court's perception of the technology's reliability as well as its appropriateness in the particular kind of juristic process (e.g., criminal vs. civil) in question. As a precursor to the chain of judicial rulings that will undoubtedly come about in the future, a widely accepted principle of the admissibility of novel scientific evidence, called the "Frye Test" (from *Frye v. The United States in 1923*), is likely to be invoked. The criteria are that the technology would be first subjected to rigorous analysis by the scientific community during its experimental stage, and only after this community arrived at a consensus that the technique was valid would evidence of its use be admissible in court.

## 5 Summary

Automated detection of emotion in speech may improve legal decision making in areas that involve assessment of emotion in others, emotions and memory, emotions and culture, and training of participants in the legal process. Though current natural language systems are not yet fully able to interpret a person's emotion, ongoing perceptual and acoustic studies paint a promising picture for automated detection of the wealth of information available in the acoustic signal of speech. The advent of this technology will spur research into its effect on all aspects of the juristic system.

## References

1. Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A., Zue, V., Varile, G., Zampolli, A. (eds.): Survey of the State of the Art in Human Language Technology. Cambridge Studies In Natural Language Processing Series. Cambridge University Press, Cambridge (1997)
2. Frank, G., Guinn, C., Hubal, R., Pope, P., Stanford, M., Lamm-Weisel, D.: JUST-TALK: An Application of Responsive Virtual Human Technology. In: Proceedings of the Interservice/Industry Training, Simulation and Education Conference, pp. 773–779. National Training Systems Association, Arlington (2002)
3. Fuller, C., Biros, D.P., Adkins, M., Burgoon, J.K., Nunamaker, J.F., Coulon, S.: Detecting Deception in Person-of-Interest Statements. In: Proceedings of the IEEE International Conference on Intelligence and Security Informatics, May 23-24, 2006, pp. 504–509, San Diego (2006)
4. Susca, M.: Connecting Stuttering Measurement and Management: II. Measures of Cognition and Affect. International Journal of Language & Communication Disorders 41(4), 365–377 (2006)
5. Yacoub, S., Simske, S., Lin, X., Burns, J.: Recognition of Emotions in Interactive Voice Response Systems. In: Proceedings of the European Conference on Speech Communication and Technology, September 1-4, 2003, pp. 729–732, Geneva, Switzerland (2003)
6. Maroney, T.A.: Law and Emotion: A Proposed Taxonomy of an Emerging Field. Law and Human Behavior 30, 119–142 (2006)
7. Cowie, R., Douglas-Cowie, E., Cox, C.: Beyond Emotion Archetypes: Databases for Emotion Modelling Using Neural Networks. Neural Networks 18, 371–388 (2005)
8. Hozjan, V., Kačič, Z.: A Rule-Based Emotion-Dependent Feature Extraction Method for Emotion Analysis from Speech. Journal of Acoustic Society of America 119(5), 3109–3120 (2006)
9. Burns, H.L., Capps, C.G.: Foundations of Intelligent Tutoring Systems: An Introduction. In: Poison, M.C., Richardson, J.J. (eds.) Foundations of Intelligent Tutoring Systems, pp. 1–19. Lawrence Erlbaum, London (1988)
10. Picard, R.W.: Affective Computing. MIT Press, Cambridge (1997)
11. Scherer, K.R.: Vocal Communication of Emotion: A Review of Research Paradigms. Speech Communication 40, 227–256 (2003)

12. Solan, L.M., Tiersma, P.M.: Falling on Deaf Ears. Legal Affairs (November/December 2003) Available from as of (August 30, 2004) http://www.legalaffairs.org/issues/November-December-2003/story_solan_novdec03.html
13. Forgas, J.: Handbook of Affect and Social Cognition. Lawrence Erlbaum Publishers, New York (2001)
14. Christianson, S.: Emotional Stress and Eyewitness Memory: A Critical Review. Psychological Bulletin 112, 284–309 (1992)
15. Taylor, J.G., Fragopanagos, N.: The Interaction of Attention and Emotion. Neural Networks 18(4), 353–369 (2005)
16. Jackson, B.S.: Making Sense in Law. Deborah Charles Publications, Liverpool (1995)
17. Tulving, E., Thomson, D.M.: Encoding Specificity and Retrieval Processes in Episodic Memory. Psychological Review 80, 352–373 (1973)
18. Olsson, N,: Realism of Confidence in Witness Identification of Faces and Voices. Unpublished doctoral dissertation, Uppsala University, Uppsala, Sweden (2000)
19. Loftus, E., Ketcham, K.: Witness for the Defense. St. Martin's Press, New York (1991)
20. Loftus, E.: Our Changeable Memories: Legal and Practical Implications. Nature Reviews: Neuroscience 4, 231–234 (2003)
21. Deffenbacher, K.A., Bornstein, B.H., Penrod, S.D., McGorty, K.: A Meta-Analytic Review of the Effects of High Stress on Eyewitness Memory. Law and Human Behavior 28(6), 687–706 (2004)
22. Berk-Seligson, S.: The Miranda Warnings and Linguistic Coercion: The Role of Footing in the Interrogation of a Limited-English-Speaking Murder Suspect. In: Cotterill, J. (ed.) Language in the Legal Process, pp. 127–143. Palgrave Macmillan Ltd, New York (2002)
23. Russell, S.: 'Three's a Crowd': Shifting Dynamics in the Interpreted Interview. In: Cotterill, J. (ed.) Language in the Legal Process, pp. 111–126. Palgrave Macmillan Ltd, New York (2002)
24. Wierzbicka, A.: Emotions across Languages and Cultures. Cambridge University Press, Cambridge (1999)
25. Semin, G.R., Görts, C.A., Nandram, S., Semin-Goossens, A.: Cultural Perspectives on the Linguistic Representation of Emotion and Emotion Events. Cognition & Emotion 16(1), 11–28 (2002)
26. Tsai, J.L., Chentsova-Dutton, Y.: Variation among European Americans in Emotional Facial Expression. Journal of Cross-Cultural Psychology 34(6), 650–657 (2003)
27. Ekman, P., Friesen, W.V., O'Sullivan, M., Chan, A., Diacoyanni- Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E., Scherer, K.R., Tomita, M., Tzavaras, A.: Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion. Journal of Personality and Social Psychology 53(4), 712–717 (1987)
28. Ekman, P., Keltner, D.: Universal Facial Expressions of Emotion: An Old Controversy and New Findings. In: Segerstrale, U., Molnár, P. (eds.) Nonverbal Communication: Where Nature Meets Culture, pp. 27–46. Lawrence Erlbaum Associates, Mahwah (1997)
29. Markus, H., Kitayama, S.: Culture and the Self: Implications for Cognition, Emotion, and Motivation. Psychological Review 98, 224–253 (1991)

30. Mesquita, B., Markus, H.R.: Culture and Emotion: Models of Agency as Sources of Cultural Variation in Emotion. In: Frijda, N.H., Manstead, A.S.R., Fisher, A. (eds.) Feelings and Emotions: The Amsterdam Symposium, pp. 341–358. Cambridge University Press, Cambridge (2004)
31. Kopelman, S., Rosette, A.S., Thompson, L.: The Three Faces of Eve: Strategic Displays of Positive, Negative, and Neutral Emotions in Negotiations. Organizational Behaviour and Human Decision Processes 99, 81–101 (2006)
32. Karstedt, S.N.E.: Emotions and Criminal Justice. Theoretical Criminology 6(3), 299–317 (2002)
33. Vidmar, N.: Case Studies of Pre- and Midtrial Prejudice in Criminal and Civil Litigation. Law and Human Behavior 26(1), 73–105 (2002)
34. Yang, L., Campbell, N.: Linking Form to Meaning: The Expression and Recognition of Emotions through Prosody. In: Proceedings of the 4th ISCA Workshop on Speech Synthesis. August 29 - September 1, 2001, Perthshire, Scotland (2001)
35. Breitenstein, C., Van Lancker, D., Daum, I.: The Contribution of Speech Rate and Pitch Variation to the Perception of Vocal Emotions in a German and an American Sample. Cognition and Emotion 15, 57–79 (2001)
36. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: FEELTRACE: An Instrument for Recording Perceived Emotions in Real Time. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 19–24. ISCA, Belfast, Ireland (2000)
37. Douglas-Cowie, E., Cowie, R., Schröder, M.: A New Emotion Database: Considerations, Sources and Scope. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 39–44. ISCA, Belfast, Ireland (2000)
38. Laukka, P., Juslin, P.N., Breslin, R.: A Dimensional Approach to Vocal Expression of Emotion. Cognition and Emotion 19(5), 633–653 (2005)
39. Mullennix, J.W., Bihon, T., Bricklemyer, J., Gaston, J., Keener, J.M.: Effects of variation in emotional tone of voice on speech perception. Language and Speech 45(3), 255–283 (2002)
40. Scherer, K.R., Banse, R., Wallbott, H.G.: Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. Journal of Cross-Cultural Psychology 32, 76–92 (2001)
41. Thompson, W.F., Balkwill, L.L.: Decoding Speech Prosody in Five Languages. Semiotica 158(1/4), 407–424 (2006)
42. Wurm, L.H., Vakoch, D.A., Strasser, M.R., Calin-Jageman, R., Ross, S.E.: Speech Perception and Vocal Expression of Emotion. Cognition and Emotion 15(6), 831–852 (2001)
43. Polzin, T., Waibel, A.: Emotion-Sensitive Human-Computer Interface. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 201–206. ISCA , Ireland, Belfast (2000)
44. Schröder, M., Cowie, R., Douglas-Cowie, M., Westerdijk, E., Gielen, S.: Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis. In: Proceedings of Eurospeech, pp. 87–90. ISCA, Geneva, Switzerland (2001)
45. Tickle, A.: English and Japanese Speakers' Emotion Vocalisation and Recognition: A Comparison Highlighting Vowel Quality. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 104–109. ISCA, Belfast, Ireland (2000)
46. Iida, A., Campbell, N., Higuchi, F., Yasamura, M.: A Corpus-Based Speech Synthesis System with Emotion. Speech Communication 40, 161–187 (2003)
47. Westermann, R., Stahl, G., Hesse, F.W.: Relative Effectiveness and Validity of Mood Induction Procedures: A Meta-analysis. European Journal of Social Psychology 26, 557–580 (1996)

48. Liscombe, J., Venditti, J., Hirschberg, J.: Classifying Subject Ratings of Emotional Speech using Acoustic Features. In: Proceedings of Eurospeech, pp. 725–728. ISCA, Geneva, Switzerland (2003)
49. Bachorowski, J.A., Owren, M.J.: Production and Perception of Affect-Rated Vocal Acoustics. In: Ekman, P., Campos, J.J., Davidson, R.J., de Waal, F.B.M. (eds.) Emotions Inside Out, pp. 244–265. Annals of the New York Academy of Sciences, New York (2003)
50. Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E.: How to find Trouble in Communication. Speech Communication 40, 117–143 (2003)
51. Cowie, R., Cornelius, R.R.: Describing the Emotional States that are Expressed in Speech. Speech Communication 40, 5–32 (2003)
52. Picard, R.W., Vyzas, E., Healey, J.: Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 1175–1191 (2001)
53. Banse, R., Scherer, K.: Acoustic Profiles in Vocal Emotion Expression. Journal of Personality and Social Psychology 70(3), 614–636 (1996)
54. Mozziconacci, S.: Prosody and Emotions. In: Proceedings of Speech Prosody, pp. 1–9. ISCA, Aix-en-Provence (2002)
55. Oudeyer, P.Y.: The Production and Recognition of Emotions in Speech: Features and Algorithms. International Journal of Human-Computer Studies 59, 157–183 (2003)
56. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech Emotion Recognition using Hidden Markov Models. Speech Communication 41, 603–623 (2003)
57. Bänziger, T., Scherer, K.R.: The Role of Intonation in Emotional Expressions. Speech Communication 46(3-4), 252–267 (2005)
58. McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S.: Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 207–212. ISCA Belfast, Ireland (2000)
59. Williams, C.E., Stevens, K.N.: Emotions and Speech: Some Acoustical Correlates. Journal of the Acoustical Society of America 52, 1238–1250 (1972)
60. Toivanen, J., Väyrynen, E., Seppänen, T.: Automatic Discrimination of Emotion from Spoken Finnish. Language and Speech 47(4), 383–412 (2004)
61. Paeschke, A., Kienast, M., Sendlmeier, W.F.: F0-Contours in Emotional Speech. In: Proceedings of ICPhS, pp. 929–931. Linguistics Department, San Francisco, USA, University of California, Berkeley (1999)
62. Mozziconacci, S., Hermes, D.J.: Role of Intonation Patterns in Conveying Emotion in Speech. In: Proceedings of ICPhS, pp. 2001–2004. Linguistics Department, San Francisco, USA, University of California, Berkeley (1999)
63. Burkhardt, F., Sendlmeier, W.F.: Verification of Acoustical Correlates of Emotional Speech Using Formant-Synthesis. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 151–156. ISCA, Belfast, Ireland (2000)
64. Dellaert, F., Polzin, T., Waibel, A.: Recognizing Emotion in Speech. In: Proceedings of the ICSLP, pp. 896–900. ICSA, Philadelphia (1996)
65. Lee, C.M., Narayanan, S.S., Pieraccini, R.: Recognition of Negative Emotions from the Speech Signal. In: Proceedings of Automatic Speech Recognition and Understanding, pp. 240–243 (2001)
66. Roy, D., Pentland, A.: Automatic Spoken Affect Analysis and Classification. In: Proceedings of the International Conference on Automatic Face and Gesture Recognition, pp. 363–367. IEEE Computer Society Press, Los Alamitos (1996)

67. Hubal, R., Frank, G., Guinn, C., Dupont, R.: Integrating a Crisis Stages Model into a Simulation for Training Law Enforcement Officers to Manage Encounters with the Mentally Ill. In: Proceedings of the Workshop on Architectures for Modeling Emotion: Cross-Disciplinary Foundations, American Association for Artificial Intelligence Spring Symposium Series, pp. 68–69. ACM Press, New York (2004)
68. Link, M.W., Armsby, P.P., Hubal, R.C., Guinn, C.I.: Accessibility and Acceptance of Responsive Virtual Human Technology as a Survey Interviewer Training Tool. Computers in Human Behavior 22(3), 412–426 (2006)
69. Gobl, C., Chasaide, A.N.: The Role of Voice Quality in Communicating Emotion, Mood and Attitude. Speech Communication 14, 189–212 (2003)
70. Janniro, M.J., Cestaro, V.L.: Effectiveness of Detection of Deception Examinations using the Computer Voice Stress Analyzer. Report No. DoDPI96-R-0005. Department of Defense Polygraph Institute, Fort McClellan (1996)
71. Oudeyer, P.Y.: Novel Useful Features and Algorithms for the Recognition of Emotions in Human Speech. Sony Computer Science Lab, Paris, France (2002) Available at Internet website: Downloaded on (2004)-03-17, `http://www3.isrl.uiuc.edu/ junwang4/langev/localcopy/pdf/oudeyerprosody2002a.pdf`
72. Petrushin, V.A., Makarova, V.: Parameters and Fricatives and Affricates in Russian Emotional Speech. In: Proceedings of Speech and Communciation (SPECOM), June 25-29, pp. 423–428, St. Petersburg (2006)
73. Schröder, M.: Experimental Study of Affect Bursts. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 132–137. ISCA, Belfast, Ireland (2000)
74. Engel, R.S., Silver, E.: Policing Mentally Disordered Suspects: A Re-examination of the Criminalization Hypothesis. Criminology 39, 225–232 (2002)
75. Anderson, J.R., Reder, L.M., Simon, H.A.: Situated Learning and Education. Educational Researcher 25(4), 5–11 (1996)
76. Paschall, M.J., Fishbein, D.H., Hubal, R.C., Eldreth, D.: Psychometric Properties of Virtual Reality Vignette Performance Measures: A Novel Approach for Assessing Adolescents' Social Competency Skills. Health Education Research: Theory and Practice 20(1), 61–70 (2005)
77. Lee, C.M., Narayanan, S.S., Pieraccini, R.: Classifying Emotions in Human-Machine Spoken Dialogs. In: Proceedings of IEEE ICME, pp. 737–740 (2002)
78. Lee, C.M., Narayanan, S.S.: Toward Detecting Emotions in Spoken Dialogs. IEEE Transactions on Speech and Audio Processing 13(2), 293–303 (2005)
79. Potapova, R., Potapova, V.: Temporal Correlates of Emotions as a Speaker-State Specific Parameters for Forensic Speaker Identification. In: Proceedings of SPECOM 2003, Moscow (2003)
80. Plutchik, R.: The Psychology and Biology of Emotion. Harper-Collins, New York (1994)
81. Jovičic, S.T., Rajkovic, M., Dordecic, M., Kašic, Z.: Perceptual and Statistical Analysis of Emotional Speech in Man-Computer Communication. In: Proceedings of Speech and Communciation (SPECOM), June 25-29, 2006, pp. 409–414, St. Petersburg (2006)
82. Rose, P.: Forensic Speaker Identification. Taylor & Francis, London (2002)
83. Kaiser, J.F.: On a Simple Algorithm to Calculate the 'Energy' of a Signal. In: Proceedings of ICASSP, pp. 381–384 (1990)
84. Slyh, R.E., Nelson, W.T., Hansen, E.G.: Analysis of Mrate, Shimmer, Jitter, and F0 Contour Features Across Stress and Speaking Style in the SUSAS Database. In: Proceedings of ICASSP, pp. 2091–2094 (1999)

# A Appendix

Research in emotion detection from speech acoustics is a four-pronged investigation. The questions surround emotional categories, acoustic parameters, classifiers, and databases. In this Appendix we dig deeper into the emotion detection literature to illustrate these points.

## A.1 Emotional Categories

Hundreds of emotion categories have been identified and discussed in the literature. From a practical stance, one needs to choose a small subset of these that suit a particular application. In an instructional context, for example, one might focus on the emotions of *confidence*, *confusion*, and *frustration*. In a judicial context, *anxiety*, *hostility*, or *uncertainty* may be the emotions of interest. In contexts where detecting and distinguishing as few as three emotions may be overly difficult, merely attempting to identify negative emotions from non-negative ones may have to suffice (Lee, et al., 2001, 2002, 2005 [65] [77] [78]).

To give an idea of how wide ranging researchers' views on categories of emotions are, we offer in Table 1 a non-comprehensive, alphabetized list of emotions that have appeared in the literature. One difficulty in compiling this list is that emotions are discussed in terms of both nouns ("happiness") and adjectives ("happy"). Adjectival descriptions have been translated into nominal ones for consistency. We believe the description should be consistent, but feel that the difference between whether one is experiencing the emotion of happiness versus experiencing a happy emotion, in describing a pervading feeling of joy, is of less concern as it relates to juristic implications.

The list in Table 1 is not only incomplete, but also the items are not mutually exclusive, in that the emotions overlap, several emotions may be experienced at the same time, and the emotions are "fuzzy" in the sense that one imagines them to be experienced to a greater or lesser degree, and not to be merely present or absent in all cases. The latter situation is addressed in Gobl and Chasaide (2003) [69] , who present eight sliding scales of emotions (shown in Table 2 in their original, adjectival format). (Potapova & Potapova, 2003 [79], present a similar scheme with their "scaleable subtypes" of emotions, e.g., fear is replaced by consternation - dread - terror.)

Emotions have also been represented on planes of varying dimensionality. For instance, Cowie, et al. (2000) [36] and Douglas-Cowie, et al. (2000) [37] presented their work with two or three dimensions (*activation*, *valence*, and *power*). Listeners were asked to rate stimuli along these scales. (A similar scheme is to plot Plutchik's (1994) [80] circle, as is discussed in Jovicic, et al., 2006 [81]). Laukka, et al. (2005) [38] extended Cowie's set of dimensions with a fourth, *intensity*, to accommodate further separation of emotions.

Correlating these dimensions with acoustic features, however, can be difficult. One approach is described by Jovičic, et al. (2006) [81], who suggest a three-level hierarchy of emotions within their multidimensional framework: primary, secondary, and tertiary. Primary emotions are fundamental and easiest to detect

**Table 1.** List of Emotions

*anger, anxiety, bemusement, bliss, boredom, certainty, complacency, confidence, confusion, contempt, contentedness, delight, depression, despair, disgust, excitement, exhilaration, fear, friendship, frustration, fury, happiness, hostility, impatience, interest, neutrality, outrage, pleasure, politeness, relaxation, sadness, serenity, shame, stressfulness, surprise, terror, timidity, volatility.*

**Table 2.** Sliding Scale of Emotions (from Gobl & Chasaide, 2003 [69])

*relaxed-stressed, content-angry, friendly-hostile, sad-happy, bored-interested, intimate-formal, timid-confident, afraid-unafraid.*

acoustically, for example *fear*. A secondary fear emotion would further subdivide into, say, *anxiety*, *terror*, *phobic*, distinctions that are more difficult to detect reliably from the speech signal. A tertiary fear emotion would presumably identify even finer distinctions, say *mildly anxious* to *severely anxious*, and these would be detected by "micro prosodic features" (p. 413).

### A.2 Acoustic Parameters

There are many acoustic properties of the speech signal discussed in the literature, which reflect the panoply of vocal affects to be linked to the emotional states. We show a non-exhaustive list in Table 3, again noting that the properties are not always independent (orthogonal) amongst themselves, and that some properties may be manifested to a greater or lesser degree. Kaiser, 1990 [83] and Slyh, et al., 1999 [84] are two of a raft of papers on calculating certain acoustic properties.

### A.3 Classifiers

The third consideration in determining emotion from voice has to do with the kinds of classifiers, or statistical tools, used to build models of speakers in which the acoustic parameters of vocal affect are statistically related to the perceived emotions. The underlying assumption is that there is a database of speech that is tagged with the names of the emotion or emotions purportedly evident from the various segments that comprise the speech. Associated with the tags are any acoustic parameters of interest. The classifier is used to build a speaker model from known speech samples, after which the model can be used to determine the emotions portrayed in future speech samples.

A variety of classifiers has been used by researchers, of which seven are: hidden Markov models, kernel regression, k-nearest neighbors, linear discrimination, maximum likelihood Bayes classifier, neural nets, and vector quantization. The list is hardly complete but it gives a sense of the eclectic tastes of the various

**Table 3.** List of Acoustic Properties

*pitch mean, pitch median, pitch standard deviation, pitch extrema, median duration of falls or rises in pitch, speech rate, mean tune duration*(segments separated by more than 180 milliseconds of silence),*long term average spectrum by frequency, spectral tilt* (a measure of the raising or lowering of the voice), *distribution of energy within various spectral ranges such as below 250Hz, jitter* (variation in pitch period), *shimmer* (variation in amplitude), *per-phoneme first formant mean, per-phoneme second formant mean.*

researchers. Pared down to essentials, given an emotion E and an acoustic parameter A, the model is intended to yield two probabilities: the probability of observing A when E is evident, and the probability of observing A when E is not evident. Mathematically, the former is written as P(A|E) and the latter as P(A|∼E). The ratio of these probabilities – P(A|E)/P(A|∼E) – gives the likelihood, or odds, that when A is detected in the speech, the emotion E is being experienced.

A simple arithmetic example makes this clear. Suppose that in 100 exemplars of speech wherein the speaker is said to experience sadness, the pitch falls 20% or more in 80 of the exemplars. Moreover, suppose that in 1000 exemplars of speech wherein the speaker is said not to experience sadness, the pitch falls 20% or more in 100 of those exemplars. Then the probability of the pitch falling when the speaker is sad is 80/100, i.e., P(A|E)=0.8, while the probability of the pitch falling when the speaker is not said is 100/1000, i.e., P(A|∼E)=0.1. The likelihood that the speaker is sad when the pitch falls 20% is therefore 0.8/0.1 = 8/1. Thus when the speaker's pitch drops 20% the odds are eight to one that the speaker is feeling – and expressing – sadness.

If other acoustic parameters are associated with the sadness emotion, their likelihoods may be computed similarly. If the parameters are known or assumed to be independent, then multiplying the likelihoods gives an overall likelihood ratio for the emotion given the acoustic parameters. (See Rose, 2002, [82] for an excellent, lucid explication of likelihood ratios.)

## A.4 Databases

A fourth thrust of emotion detection research is the development of databases of speech tagged with emotions. The model building and likelihoods discussed above depend on having such a database.

Certainly one way to assemble such a database is to hire actors to exhibit the emotions of interest, and record their speech as they do so. This may be done directly by instructing the actor to speak a given sentence with a feeling of sadness, or it may be done by giving the actor a role to play and lines to read in which sadness is called for. This is a common source of data but one open to much question, perhaps best summed up in Douglas-Cowie, et al. (2000) [37] where the authors write, "At the very least, acted emotion cannot be a sufficient basis for conclusions about the expression of emotion".

Other recourses remain. One is to find “real” people and immerse them in emotion invoking situations while recording their speech. One would presumably not go so far as, say, to threaten to throw people from the top of the Roman Coliseum in order to collect fearful speech, though this suggests that Nero may have had the means to be an effective researcher. Rather, participants might be asked to recall and describe a particularly emotional event in their lives, or perhaps asked to read emotion-invoking passages aloud, either composed for the purpose, or drawn from the literature. Another approach, as taken by Douglas-Cowie, et al. (2000) [37], involves collecting data from media shows, either radio or television, featuring non-actors in verbal interactions that evoke emotions. (In general, negative emotions often come out in chat shows whereas positive emotions derive from religious programs.) For summaries of numerous human-based databases see Cowie, et al. (2005) [7].

A final possibility is to develop a database of emotionally charged synthetic speech (Iida, et al., 2003 [46]). Such a database would be useful for studies in human perception of emotions in the presence of particular acoustic features that could be tightly controlled. But a synthetic database should not be used to identify the acoustic correlates of a particular emotion as perceived by a human as that would clearly be circular, except in the framework of evaluating the efficacy of the database.