Emotional speech recognition: Resources, features, and methods

Dimitrios Ververidis, Constantine Kotropoulos *

Artificial Intelligence and Information Analysis Laboratory, Department of Informatics, Aristotle University of Thessaloniki, University Campus, Box 451, Thessaloniki 54124, Greece

Received 15 July 2004; received in revised form 19 April 2006; accepted 24 April 2006

Abstract

In this paper we overview emotional speech recognition having in mind three goals. The first goal is to provide an up-to-date record of the available emotional speech data collections. The number of emotional states, the language, the number of speakers, and the kind of speech are briefly addressed. The second goal is to present the most frequent acoustic features used for emotional speech recognition and to assess how the emotion affects them. Typical features are the pitch, the formants, the vocal tract cross-section areas, the mel-frequency cepstral coefficients, the Teager energy operator-based features, the intensity of the speech signal, and the speech rate. The third goal is to review appropriate techniques in order to classify speech into emotional states. We examine separately classification techniques that exploit timing information from which that ignore it. Classification techniques based on hidden Markov models, artificial neural networks, linear discriminant analysis, k-nearest neighbors, support vector machines are reviewed.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Emotions; Emotional speech data collections; Emotional speech classification; Stress; Interfaces; Acoustic features

1. Introduction

Emotional speech recognition aims at automatically identifying the emotional or physical state of a human being from his or her voice. The emotional and physical states of a speaker are known as emotional aspects of speech and are included in the so-called paralinguistic aspects. Although the emotional state does not alter the linguistic content, it is an important factor in human communication, because it provides feedback information in many applications as it is outlined next.

Making a machine to recognize emotions from speech is not a new idea. The first investigations were conducted around the mid-1980s using statistical properties of certain acoustic features (Van Bezooijen, 1984; Tolkmitt and Scherer, 1986). Ten years later, the evolution of computer architectures made the implementation of more complicated emotion recognition algorithms feasible. Market requirements for automatic services motivate further research. In environments like aircraft cockpits, speech recognition systems were trained by employing stressed speech instead of neutral (Hansen and Cairns, 1995). The acoustic features were estimated more precisely by iterative algorithms. Advanced classifiers exploiting timing information were proposed (Cairns and Hansen, 1994; Womack and
Hansen, 1996; Polzin and Waibel, 1998). Nowadays, research is focused on finding powerful combinations of classifiers that advance the classification efficiency in real-life applications. The wide use of telecommunication services and multimedia devices paves also the way for new applications. For example, in the projects “Prosody for dialogue systems” and “SmartKom”, ticket reservation systems are developed that employ automatic speech recognition being able to recognize the annoyance or frustration of a user and change their response accordingly (Ang et al., 2002; Schiel et al., 2002). Similar scenarios are also presented for call center applications (Petrushin, 1999; Lee and Narayanan, 2005). Emotional speech recognition can be employed by therapists as a diagnostic tool in medicine (France et al., 2000). In psychology, emotional speech recognition methods can cope with the bulk of enormous speech data in real-time extracting the speech characteristics that convey emotion and attitude in a systematic manner (Mozziconacci and Hermes, 2000).

In the future, the emotional speech research will primarily be benefited by the on-going availability of large-scale emotional speech data collections, and will focus on the improvement of theoretical models for speech production (Flanagan, 1972) or models related to the vocal communication of emotion (Scherer, 2003). Indeed, on the one hand, large data collections which include a variety of speaker utterances under several emotional states are necessary in order to faithfully assess the performance of emotional speech recognition algorithms. The already available data collections consist only of few utterances, and therefore it is difficult to demonstrate reliable emotion recognition results. The data collections listed in Section 2 provide initiatives to set up more relaxed and close to real-life specifications for recording large-scale emotional speech data collections that are complementary to the already existing resources. On the other hand, theoretical models of speech production and vocal communication of emotion will provide the necessary background for a systematic study and will deploy more accurate emotional cues through time. In the following, the contributions of the paper are identified and its outline is given.

1.1. Contributions of the paper

Several reviews on emotional speech analysis have already appeared (Van Bezooijen, 1984; Scherer et al., 1991; Cowie et al., 2001, 2003; Scherer, 2003; Douglas-Cowie et al., 2003). However, as the research towards understanding human emotions increasingly attracts the attention of the research community, the short list of 19 data collections appeared in (Douglas-Cowie et al., 2003) does not adequately cover the topic. In this tutorial, 64 data collections are reviewed. Furthermore, an up-to-date literature survey is provided, complementing the previous studies in (Van Bezooijen, 1984; Scherer et al., 1991; Cowie et al., 2001). Finally, the paper is focused on describing the feature extraction methods and the emotion classification techniques, topics that have not been treated in (Scherer, 2003; Pantic and Rothkrantz, 2003).

1.2. Outline

In Section 2, a corpus of 64 data collections is reviewed putting emphasis on the data collection procedures, the kind of speech (natural, simulated, or elicited), the content, and other physiological signals that may accompany the emotional speech. In Section 3, short-term features (i.e. features that are extracted on speech frame basis) that are related to the emotional content of speech are discussed. In addition to short-term features, their contours are of fundamental importance for emotional speech recognition. The emotions affect the contour characteristics, such as statistics and trends as is summarized in Section 4. Emotion classification techniques that exploit timing information and other techniques that ignore it are surveyed in Section 5. Therefore, Sections 3 and 4 aim at describing the appropriate features to be used with the emotional classification techniques reviewed in Section 5. Finally, Section 6 concludes the tutorial by indicating future research directions.

2. Data collections

A record of emotional speech data collections is undoubtedly useful for researchers interested in emotional speech analysis. An overview of 64 emotional speech data collections is presented in Table 1. For each data collection additional information is also described such as the speech language, the number and the profession of the subjects, other physiological signals possibly recorded simultaneously with speech, the data collection purpose (emotional speech recognition, expressive synthesis), the emotional states recorded, and the kind of the emotions (natural, simulated, elicited).
<table>
<thead>
<tr>
<th>Reference</th>
<th>Language</th>
<th>Subjects</th>
<th>Other signals</th>
<th>Purpose</th>
<th>Emotions</th>
<th>Kind</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abelin and Allwood (2000)</td>
<td>Swedish</td>
<td>1 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Fr, Jy, Sd, Se, Dt, Dom, Sy</td>
<td>Simulated</td>
</tr>
<tr>
<td>Alpert et al. (2001)</td>
<td>English</td>
<td>22 Patients, 19 healthy</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Hs, Nl</td>
<td>Natural</td>
</tr>
<tr>
<td>Alter et al. (2000)</td>
<td>German</td>
<td>1 Female</td>
<td>EEG</td>
<td>Recognition</td>
<td>Ar, Dt, Fr, Nl, Se</td>
<td>Simulated</td>
</tr>
<tr>
<td>Ambrus (2000), Interface</td>
<td>English, Slovenian</td>
<td>8 Actors</td>
<td>LG</td>
<td>Synthesis</td>
<td>Ar, Dt, Fr, Jy, Sd</td>
<td>Simulated</td>
</tr>
<tr>
<td>Amir et al. (2000)</td>
<td>Hebrew</td>
<td>40 Students</td>
<td>LG, M, G, H</td>
<td>Recognition</td>
<td>Ar, Dt, Fr, Jy, Sd</td>
<td>Natural</td>
</tr>
<tr>
<td>Ang et al. (2002)</td>
<td>English</td>
<td>Many</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, At, Nl, Fd, Td</td>
<td>Natural</td>
</tr>
<tr>
<td>Banse and Scherer (1996)</td>
<td>German</td>
<td>12 Actors</td>
<td>V</td>
<td>Recognition</td>
<td>H/C Ar, Hs, Sd,…</td>
<td>Simulated</td>
</tr>
<tr>
<td>Batliner et al. (2004)</td>
<td>German, English</td>
<td>51 Children</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Bm, Jy, Se</td>
<td>Elicited</td>
</tr>
<tr>
<td>Bulut et al. (2002)</td>
<td>English</td>
<td>1 Actress</td>
<td>–</td>
<td>Synthesis</td>
<td>Ar, Hs, Nl, Sd</td>
<td>Simulated</td>
</tr>
<tr>
<td>Burkhardt and Sendlmeier (2000)</td>
<td>German</td>
<td>10 Actors</td>
<td>V, LG</td>
<td>Synthesis</td>
<td>Ar, Fr, Jy, Nl, Sd, Bm, Dt</td>
<td>Simulated</td>
</tr>
<tr>
<td>Caldognetto et al. (2004)</td>
<td>Italian</td>
<td>1 Native</td>
<td>V, IR</td>
<td>Synthesis</td>
<td>Ar, Fr, Jy, Jy, Sd, Se</td>
<td>Simulated</td>
</tr>
<tr>
<td>Choukri (2003), Groningen</td>
<td>Dutch</td>
<td>238 Native</td>
<td>LG</td>
<td>Recognition</td>
<td>Unknown</td>
<td>Simulated</td>
</tr>
<tr>
<td>Chuang and Wu (2002)</td>
<td>Chinese</td>
<td>2 Actors</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Ay, Hs, Fr, Se, Sd</td>
<td>Simulated</td>
</tr>
<tr>
<td>Clavel et al. (2004)</td>
<td>English</td>
<td>18 From TV</td>
<td>–</td>
<td>Recognition</td>
<td>Nl, levels of Fr</td>
<td>Simulated</td>
</tr>
<tr>
<td>Cole (2005), Kids’ Speech</td>
<td>English</td>
<td>780 Children</td>
<td>V</td>
<td>Recognition, Synthesis</td>
<td>Unknown</td>
<td>Natural</td>
</tr>
<tr>
<td>Cowie and Douglas-Cowie (1996), Belfast Structured</td>
<td>English</td>
<td>40 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Fr, Hs, Nl, Sd</td>
<td>Natural</td>
</tr>
<tr>
<td>Douglas-Cowie et al. (2003), Belfast Natural</td>
<td>English</td>
<td>125 From TV</td>
<td>V</td>
<td>Recognition</td>
<td>Various</td>
<td>Semi-natural</td>
</tr>
<tr>
<td>Edgington (1997)</td>
<td>English</td>
<td>1 Actor</td>
<td>LG</td>
<td>Synthesis</td>
<td>Ar, Bm, Fr, Hs, Nl, Sd</td>
<td>Simulated</td>
</tr>
<tr>
<td>Engberg and Hansen (1996), DES</td>
<td>Danish</td>
<td>4 Actors</td>
<td>–</td>
<td>Synthesis</td>
<td>Ar, Hs, Nl, Sd, Se</td>
<td>Simulated</td>
</tr>
<tr>
<td>Fernandez and Picard (2003)</td>
<td>English</td>
<td>4 Drivers</td>
<td>–</td>
<td>Recognition</td>
<td>Nl, Ss</td>
<td>Natural</td>
</tr>
<tr>
<td>Fischer (1999), Verbmobil</td>
<td>German</td>
<td>58 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Dn, Nl</td>
<td>Natural</td>
</tr>
<tr>
<td>France et al. (2000)</td>
<td>English</td>
<td>70 Patients, 40 healthy</td>
<td>–</td>
<td>Recognition</td>
<td>Dn, Nl</td>
<td>Natural</td>
</tr>
<tr>
<td>Hansen (1996), SUSAS</td>
<td>English</td>
<td>32 Various</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Ld eff., Ss, Ti</td>
<td>Natural, simulated</td>
</tr>
<tr>
<td>Hansen (1996), SUSC-0</td>
<td>English</td>
<td>18 Non-native</td>
<td>H, BP, R</td>
<td>Recognition</td>
<td>Nl, Ss</td>
<td>A-stress</td>
</tr>
<tr>
<td>Hansen (1996), SUSB-1</td>
<td>English</td>
<td>20 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Nl, Ss</td>
<td>P-stress</td>
</tr>
<tr>
<td>Hansen (1996), DLIP</td>
<td>English</td>
<td>15 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Nl, Ss</td>
<td>C-stress</td>
</tr>
<tr>
<td>Hansen (1996), DCEIM</td>
<td>English</td>
<td>Unknown</td>
<td>–</td>
<td>Recognition</td>
<td>Nl, Sleep deprive</td>
<td>Elicited</td>
</tr>
<tr>
<td>Heuft et al. (1996)</td>
<td>German</td>
<td>3 Native</td>
<td>–</td>
<td>Synthesis</td>
<td>Ar, Fr, Jy, Sd,…</td>
<td>Simulated, elicited</td>
</tr>
<tr>
<td>Iida et al. (2000), ESC</td>
<td>Japanese</td>
<td>2 Native</td>
<td>–</td>
<td>Synthesis</td>
<td>Ar, Jy, Sd</td>
<td>Simulated</td>
</tr>
<tr>
<td>Iriondo et al. (2000)</td>
<td>Spanish</td>
<td>8 Actors</td>
<td>–</td>
<td>Synthesis</td>
<td>Fr, Jy, Sd, Se,…</td>
<td>Simulated</td>
</tr>
</tbody>
</table>

(continued on next page)
Table 1 (continued)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Language</th>
<th>Subjects</th>
<th>Other signals</th>
<th>Purpose</th>
<th>Emotions</th>
<th>Kind</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberman (2005), Emotional Prosody</td>
<td>English</td>
<td>Actors</td>
<td>–</td>
<td>Unknown</td>
<td>Anxty, H/C Ar, Hs, Nl, Pe, Sd, Se,…</td>
<td>Simulated</td>
</tr>
<tr>
<td>Linnankoski et al. (2005)</td>
<td>English</td>
<td>13 Native</td>
<td>–</td>
<td>Recognition</td>
<td>An, Ar, Fr, Sd,…</td>
<td>Elicited</td>
</tr>
<tr>
<td>Makarova and Petrushin (2002), RUSSLANA</td>
<td>Russian</td>
<td>61 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Hs, Se, Sd, Fr, Nl</td>
<td>Simulated</td>
</tr>
<tr>
<td>Martins et al. (1998), BDFALA</td>
<td>Portuguese</td>
<td>10 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Dt, Hs, Iy</td>
<td>Simulated</td>
</tr>
<tr>
<td>McMahon et al. (2003), ORESTEIA</td>
<td>English</td>
<td>29 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Ae, Sk, Ss</td>
<td>Elicited</td>
</tr>
<tr>
<td>Montanari et al. (2004)</td>
<td>English</td>
<td>15 Children</td>
<td>V</td>
<td>Recognition</td>
<td>Unknown</td>
<td>Natural</td>
</tr>
<tr>
<td>Montero et al. (1999), SES</td>
<td>Spanish</td>
<td>1 Actor</td>
<td>–</td>
<td>Synthesis</td>
<td>Ar, Dt, Hs, Sd</td>
<td>Simulated</td>
</tr>
<tr>
<td>Mozziconacci and Hermes (1997)</td>
<td>Dutch</td>
<td>3 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Bm, Fr, Jy, Nl, Sd</td>
<td>Simulated</td>
</tr>
<tr>
<td>Niimi et al. (2001)</td>
<td>Japanese</td>
<td>1 Male</td>
<td>–</td>
<td>Synthesis</td>
<td>Ar, Jy, Sd</td>
<td>Simulated</td>
</tr>
<tr>
<td>Nordstrand et al. (2004)</td>
<td>Swedish</td>
<td>1 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Fr, Jy,…</td>
<td>Simulated</td>
</tr>
<tr>
<td>Nwe et al. (2003)</td>
<td>Chinese</td>
<td>12 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Fr, Dt, Jy,…</td>
<td>Simulated</td>
</tr>
<tr>
<td>Petrushin (1999)</td>
<td>English</td>
<td>30 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Fr, Hs, Nl, Sd</td>
<td>Simulated, Natural</td>
</tr>
<tr>
<td>Polzin and Waibel (1998)</td>
<td>English</td>
<td>5 Drama students</td>
<td>LG</td>
<td>Recognition</td>
<td>H, R, BP, BL</td>
<td>Natural</td>
</tr>
<tr>
<td>Rahurkar and Hansen (2002), SOQ</td>
<td>English</td>
<td>6 Soldiers</td>
<td>H, R, BP, BL</td>
<td>Recognition</td>
<td>5 Stress levels</td>
<td>Natural</td>
</tr>
<tr>
<td>Scherer (2000b), Lost Luggage</td>
<td>Various</td>
<td>109 Passengers</td>
<td>V</td>
<td>Recognition</td>
<td>Ar, Hr, Ie, Sd, Ss</td>
<td>Natural</td>
</tr>
<tr>
<td>Scherer (2000a)</td>
<td>German</td>
<td>4 Actors</td>
<td>–</td>
<td>Ecological</td>
<td>Ar, Dt, Fr, Jy, Sd</td>
<td>Simulated</td>
</tr>
<tr>
<td>Scherer et al. (2002)</td>
<td>English, German</td>
<td>100 Native</td>
<td>–</td>
<td>Recognition</td>
<td>2 Tl, 2 Ss</td>
<td>Natural</td>
</tr>
<tr>
<td>Schiel et al. (2002), SmartKom</td>
<td>German</td>
<td>45 Native</td>
<td>V</td>
<td>Recognition</td>
<td>Ar, Dfn, Nl</td>
<td>Natural</td>
</tr>
<tr>
<td>Schröder and Grice (2003)</td>
<td>German</td>
<td>1 Male</td>
<td>–</td>
<td>Synthesis</td>
<td>Soft, modal, loud</td>
<td>Simulated</td>
</tr>
<tr>
<td>Schröder (2000)</td>
<td>German</td>
<td>6 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Bm, Df, Wy,…</td>
<td>Simulated</td>
</tr>
<tr>
<td>Slaney and McARoberts (2003), Babyears</td>
<td>English</td>
<td>12 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Al, An, Pn</td>
<td>Natural</td>
</tr>
<tr>
<td>Stibbard (2000), Leeds</td>
<td>English</td>
<td>Unknown</td>
<td>–</td>
<td>Synthesis</td>
<td>Wide range</td>
<td>Natural, elicited</td>
</tr>
<tr>
<td>Tato (2002), AIBO</td>
<td>German</td>
<td>14 Native</td>
<td>–</td>
<td>Synthesis</td>
<td>Ar, Bm, Hs, Nl, Sd</td>
<td>Elicited</td>
</tr>
<tr>
<td>Tokkmitt and Scherer (1986)</td>
<td>German</td>
<td>60 Native</td>
<td>–</td>
<td>Recognition</td>
<td>Cognitive Ss</td>
<td>Elicited</td>
</tr>
<tr>
<td>Wendt and Scheich (2002), Magdeburger</td>
<td>German</td>
<td>2 Actors</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Dt, Fr, Hs, Sd</td>
<td>Simulated</td>
</tr>
<tr>
<td>Yildirim et al. (2004)</td>
<td>English</td>
<td>1 Actress</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Hs, Nl, Sd</td>
<td>Simulated</td>
</tr>
<tr>
<td>Yu et al. (2001)</td>
<td>Chinese</td>
<td>Native</td>
<td>–</td>
<td>Recognition</td>
<td>Ar, Hs, Nl, Sd</td>
<td>Simulated</td>
</tr>
<tr>
<td>Yuan (2002)</td>
<td>Chinese</td>
<td>9 Native</td>
<td>from TV</td>
<td>Recognition</td>
<td>Ar, Fr, Jy, Nl, Sd</td>
<td>Elicited</td>
</tr>
</tbody>
</table>


**Abbreviations for other signals:** BP: Blood pressure, BL: Blood examination, EEG: Electroencephalogram, G: Galvanic skin response, H: Heart beat rate, IR: Infrared Camera, LG: Laryngograph, M: Myogram of the face, R: Respiration, V: Video.

**Other abbreviations:** H/C: Hot/cold, Ld eff.: Lombard effect, A-stress, P-stress, C-stress: Actual, Physical, and Cognitive stress, respectively, Sim.: Simulated, Elic.: Elicited, N/A: Not available.
From Table 1, it is evident that the research on emotional speech recognition is limited to certain emotions. The majority of emotional speech data collections encompasses five or six emotions, although the emotion categories are much more in real life. For example, many words “with emotional connotation”, originally found in the semantic Atlas of Emotional Concepts, are enlisted in (Cowie and Cornelius, 2003). In the early 1970s, the pallet theory was proposed by Anscombe and Geach in an attempt to describe all emotions as a mixture of some primary emotions like what exactly happens with colors (Anscombe and Geach, 1970). This idea has been rejected and the term “basic emotions” is now widely used without implying that such emotions can be mixed to produce others (Eckman, 1992). It is commonly agreed that the basic emotions are more primitive and universal than the others. Eckman proposed the following basic emotions: anger, fear, sadness, sensory pleasure, amusement, satisfaction, contentment, excitement, disgust, contempt, pride, shame, guilt, embarrassment, and relief. Non-basic emotions are called “higher-level” emotions (Buck, 1999) and they are rarely represented in emotional speech data collections.

Three kinds of speech are observed. Natural speech is simply spontaneous speech where all emotions are real. Simulated or acted speech is speech expressed in a professionally deliberated manner. Finally, elicited speech is speech in which the emotions are induced. The elicited speech is neither neutral nor simulated. For example, portrayals of non-professionals while imitating a professional produce elicited speech, which can also be an acceptable solution when an adequate number of professionals is not available (Nakatsu et al., 1999). Acted speech from professionals is the most reliable for emotional speech recognition because professionals can deliver speech colored by emotions that possess a high arousal, i.e. emotions with a great amplitude or strength.

Additional synchronous physiological signals such as sweat indication, heart beat rate, blood pressure, and respiration could be recorded during the experiments. They provide a ground truth for the degree of subjects’ arousal or stress (Rahurkar and Hansen, 2002; Picard et al., 2001). There is a direct evidence that the aforementioned signals are related more to the arousal information of speech than to the valence of the emotion, i.e. the positive or negative character of the emotions (Wagner et al., 2005). As regards other physiological signals, such as EEG or signals derived from blood analysis, no sufficient and reliable results have been reported yet.

The recording scenarios employed in data collections are presumably useful for repeating or augmenting the experiments. Material from radio or television is always available (Douglas-Cowie et al., 2003). However, such material raises copyright issues and impedes the data collection distribution. An alternative is speech from interviews with specialists, such as psychologists and scientists specialized in phonetics (Douglas-Cowie et al., 2003). Furthermore, speech from real-life situations such as oral interviews of employees when they are examined for promotion can be also used (Rahurkar and Hansen, 2002). Parents talking to infants, when they try to keep them away from dangerous objects can be another real-life example (Slaney and McRoberts, 2003). Interviews between a doctor and a patient before and after medication was used in (France et al., 2000). Speech can be recorded while the subject faces a machine, e.g. during telephone calls to automatic speech recognition (ASR) call centers (Lee and Narayanan, 2005), or when the subjects are talking to fake-ASR machines, which are operated by a human (wizard-of-OZ method, WOZ) (Fischer, 1999). Giving commands to a robot is another idea explored (Batliner et al., 2004). Speech can be also recorded during imposed stressed situations. For example when the subject adds numbers while driving a car at various speeds (Fernandez and Picard, 2003), or when the subject reads distant car plates on a big computer screen (Steeneken and Hansen, 1999). Finally, subjects’ readings of emotionally neutral sentences located between emotionally biased ones can be another manner of recording emotional speech.

3. Estimation of short-term acoustic features

Methods for estimating short-term acoustic features that are frequently used in emotion recognition are described hereafter. Short-term features are estimated on a frame basis

\[ f_s(n; m) = s(n)w(m - n), \]

where \( s(n) \) is the speech signal and \( w(m - n) \) is a window of length \( N_w \) ending at sample \( m \) (Deller et al., 2000). Most of the methods stem from the front-end signal processing employed in speech recognition and coding. However, the discussion is focused on acoustic features that are useful for
emotion recognition. The outline of this section is as follows. Methods for estimating the fundamental phonation or pitch are discussed in Section 3.1. In Section 3.2 features based on a non-linear model of speech production are addressed. Vocal tract features related to emotional speech are described in Section 3.3. Finally, a method to estimate speech energy is presented in Section 3.4.

3.1. Pitch

The pitch signal, also known as the glottal waveform, has information about emotion, because it depends on the tension of the vocal folds and the subglottal air pressure. The pitch signal is produced from the vibration of the vocal folds. Two features related to the pitch signal are widely used, namely the pitch frequency and the glottal air velocity at the vocal fold opening time instant. The time elapsed between two successive vocal fold openings is called pitch period $T$, while the vibration rate of the vocal folds is the fundamental frequency of the phonation $F_0$ or pitch frequency. The glottal volume velocity denotes the air velocity through glottis during the vocal fold vibration. High velocity indicates a music-like speech like joy or surprise. Low velocity is in harsher styles such as anger or disgust (Nogueiras et al., 2001). Many algorithms for estimating the pitch signal exist (Hess, 1992). Two algorithms will be discussed here. The first pitch estimation procedure that prevents the first formant interfering with the pitch is invented. In the second pass, an autocorrelation of center-clipped frame is used with low error in emotion classification (Tolkmitt and Scherer, 1986; Iida et al., 2003). However, it is argued that this method of extracting pitch is affected by the interference of the first formant in the pitch frequency, no matter which the parameters of the center clipping are (Tolkmitt and Scherer, 1986). The clipping of small signal values may not remove the effect of the non-linear propagation of the air through the vocal tract which is an indication of the abnormal spectral characteristics during emotion.

The second method for estimating the pitch uses the wavelet transform (Cairns and Hansen, 1994). It is a derivation of the method described in (Kadambi and Boudreaux-Bartels, 1992). The pitch period extraction is based on a two pass dyadic wavelet transform over the signal. Let $b$ denote a time index, $2^l$ be a scaling parameter, $s(n)$ be the sampled speech signal, and $\phi(n)$ be a cubic spline wavelet generated with the method in (Mallat and Zhong, 1989). The dyadic wavelet transform is defined by

$D_{y}WT(b, 2^l) = \frac{1}{2^l} \sum_{n=-\infty}^{n=\infty} s(n) \phi \left( \frac{n - b}{2^l} \right).$  \hspace{1cm} (5)

It represents a convolution of the time-reversed wavelet with the speech signal. This procedure is repeated for three wavelet scales. In the first pass, the result of the transform is windowed by a 16 ms rectangular window shifted with a rate of 8 ms. The pitch frequency is found by estimating the maxima of $D_{y}WT(b, 2^l)$ across the three scales. Although the method tracks the pitch epochs for neutral speech, it skips epochs for stressed speech. For marking the speech epochs in stressed speech, a second pass of wavelets is invented. In the second pass, the same wavelet transform is applied only in the
intervals between the first pass pitch periods found to have a pitch epoch greater than 150% of the median value of the pitch epochs measured during the first pass. The result of the second wavelet transform is windowed by a 8 ms window with a 4 ms skip rate to capture the sudden pitch epochs that occur often in stressed speech.

The pitch period and the glottal volume velocity at the time instant of vocal fold opening are not the only characteristics of the glottal waveform. The shape of the glottal waveform during a pitch period is also informative about the speech signal and probably has to do with the emotional coloring of the speech, a topic that has not been studied adequately yet.

3.2. Teager energy operator

Another useful feature for emotion recognition is the number of harmonics due to the non-linear air flow in the vocal tract that produces the speech signal. In the emotional state of anger or for stressed speech, the fast air flow causes vortices located near the false vocal folds providing additional excitation signals other than the pitch (Teager and Teager, 1990; Zhou et al., 2001). The additional excitation signals are apparent in the spectrum as harmonics and cross-harmonics. In the following, a procedure to calculate the number of harmonics in the speech signal is described.

Let us assume that a speech frame \( f_s(n; m) \) has a single harmonic which can be considered as an AM–FM sinewave. In discrete time, the AM–FM sinewave \( f_s(n; m) \) can be represented as (Quatieri, 2002)

\[
\begin{align*}
  f_s(n; m) &= \alpha(n; m) \cos(\Phi(n; m)) \times \cos \left( \omega_c n + \omega_h \int_0^n q(k) dk + \theta \right) \\
  &= \alpha(n; m) \cos(\Phi(n; m)) \times \cos \left( \omega_c n + \omega_h q(n) \right)
\end{align*}
\]

with instantaneous amplitude \( \alpha(n; m) \) and instantaneous frequency

\[
\omega_i(n; m) = \frac{d\Phi(n; m)}{dn} = \omega_c + \omega_h q(n), \quad |q(n)| \leq 1
\]

where \( \omega_c \) is the carrier frequency, \( \omega_h \in [0, \omega_c] \) is the maximum frequency deviation, and \( \theta \) is a constant phase offset.

The Teager energy operator (TEO) (Teager and Teager, 1990)

\[
\Psi[f_s(n; m)] = (f_s(n; m))^2 - f_s(n+1; m)f_s(n-1; m)
\]

when applied to an AM–FM sinewave yields the squared product of the AM–FM components

\[
\Psi[f_s(n; m)] = x^2(n; m) \sin(\omega_c^2(n; m))
\]

The unknown parameters \( \alpha(n; m) \) and \( \omega_i(n; m) \) can be estimated approximately with

\[
\omega_i(n; m) \approx \arcsin \left( \frac{\Psi[A_2]}{4\Psi[f_s(n; m)]} \right)
\]

and

\[
\alpha(n; m) \approx \frac{2\Psi[f_s(n; m)]}{\sqrt{\Psi[A_2]}}
\]

where \( A_2 = f_s(n+1; m) - f_s(n-1; m) \). Let us assume that within a speech frame each harmonic has an almost constant instantaneous amplitude and constant instantaneous frequency. If the signal has a single harmonic, then from (9) it is deduced that the TEO profile is a constant number. Otherwise, if the signal has more than one harmonic then the TEO profile is a function of \( n \).

Since it is certain that more than one harmonic exist in the spectrum, it is more convenient to break the bandwidth into 16 small bands, and study each band independently. The polynomial coefficients, which describe the TEO autocorrelation envelope area, can be used as features for classifying the speech into emotional states (Zhou et al., 2001). This method achieves a correct classification rate of 89% in classifying neutral versus stressed speech whereas MFCCs yield 67% in the same task.

Pitch frequency also affects the number of harmonics in the spectrum. Less harmonics are produced when the pitch frequency is high. More harmonics are expected when the pitch frequency is low. It seems that the harmonics from the additional excitation signals due to vortices are more intense than those caused by the pitch signal. The interaction of the two factors is a topic for further research. A method which can be used to alleviate the presence of harmonics due to the pitch frequency factor is to normalize the speech so that it has a constant pitch frequency (Cairns and Hansen, 1994).

3.3. Vocal tract features

The shape of the vocal tract is modified by the emotional states. Many features have been used to describe the shape of the vocal tract during emotional speech production. Such features include

- the formants which are a representation of the vocal tract resonances,
• the cross-section areas when the vocal tract is modeled as a series of concatenated lossless tubes (Flanagan, 1972),
• the coefficients derived from frequency transformations.

The formants are one of the quantitative characteristics of the vocal tract. In the frequency domain, the location of vocal tract resonances depends upon the shape and the physical dimensions of the vocal tract. Since the resonances tend to “form” the overall spectrum, speech scientists refer to them as formants. Each formant is characterized by its center frequency and its bandwidth. It has been found that subjects during stress or under depression do not articulate voiced sounds with the same precision as subjects during slackened articulated speech. Next, we describe methods to estimate formant center frequency via

\[ f^*_c(m) = \frac{1}{2\pi N_w} \sum_{n=m-N_w+1}^{m} \omega(n; m), \quad (13) \]

where \( f^*_c(m) \) is the formant center frequency during iteration \( l + 1 \). If the distance between \( f^*_c(m) \) and \( f^*_c(m) \) is smaller than 10 Hz, then the method stops and \( f^*_c(m) \) is the formant frequency estimate. In detail, \( f^*_c(m) \) is estimated by the formant frequency estimation procedure that employs LPCs. The signal is filtered with a bandpass filter in order to isolate the band which includes the formant. Let \( G(n) \) be the impulse response of a Gabor bandpass filter

\[ G_n = \exp[-(\beta n T)^2] \cos(2\pi f_n T n), \quad (14) \]

where \( f_n \) is the center frequency, \( \beta \) the bandwidth of the filter, and \( T \) is the sampling period. If \( f_n < 1000 \) Hz, then \( \beta \) equals to 800 Hz, otherwise \( \beta = 1100 \) Hz. The value of \( \beta \) is chosen small enough so as to not have more than one formant inside the bandwidth and large enough to capture the change of the instantaneous frequency. Then, \( f^*_c(m) \) is estimated by (13). If the criterion \( |f^*_c(m) - f^*_c(m)| < 10 \) is satisfied, then the method stops, otherwise the frame is re-filtered with the Gabor filter centered at \( f^*_c(m) \). The latter is re-estimated with (13) and the criterion is checked again. The method stops after a few iterations. However, it is reported that there are a few exceptions where the method does not converge. This could be a topic for further study.

The second feature is the cross-section areas of the vocal tract modeled by the multi-tube lossless model (Flanagan, 1972). Each tube is described by its cross-section area and its length. To a first approximation, one may assume that there is no loss of energy due to soft wall vibrations, heat conduction, and thermal viscosity. For a large number of tubes, the model becomes a realistic representation of the vocal tract, but it is not possible to be computed in real time. A model that can easily be computed is that of 10 cross-section areas of fixed length (Mrayati et al., 1988). The cross-section area near the glottis is indexed by \( A_1 \) and the others are following sequentially until the lips. The back vocal tract area \( A_2 \) can be used to discriminate the neutral
speech from that by anger colored, as \( A_2 \) is greater in the former emotion than in the latter one (Womack and Hansen, 1996).

The third feature is the energy of certain frequency bands. There are many contradictions in identifying the best frequency band of the power spectrum in order to classify emotions. Many investigators put high significance on the low frequency bands, such as the 0–1.5 kHz band (Tolkmitt and Scherer, 1986; Banse and Scherer, 1996; France et al., 2000) whereas other suggest the opposite (Nwe et al., 2003). An explanation for both opinions is that stressed or colored by anger speech may be expressed with a low articulation effort, a fact which causes formant peak smoothing and spectral flatness as well as energy shifting from low to high frequencies in the power spectrum. The Mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980) provide a better representation of the signal than the frequency bands since they additionally exploit the human auditory frequency response. Nevertheless, the experimental results have demonstrated that the MFCCs achieve poor emotion classification results (Zhou et al., 2001; Nwe et al., 2003), which might be due to the textual dependency and the embedded pitch filtering during cepstral analysis (Davis and Mermelstein, 1980). Better features than MFCCs for emotion classification in practice are the log-frequency power coefficients (LFPCs) which include the pitch information (Nwe et al., 2003). The LFPCs are simply derived by filtering each short-time spectrum with 12 band-pass filters having bandwidths and center frequencies corresponding to the critical bands of the human ear (Rabiner and Juang, 1993).

### 3.4. Speech energy

The short-term speech energy can be exploited for emotion recognition, because it is related to the arousal level of emotions. The short-term energy of the speech frame ending at \( m \) is

\[
E_s(m) = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} |f_t(n; m)|^2. \tag{15}
\]

### 4. Cues to emotion

In this section, we review how the contour of selected short-term acoustic features is affected by the emotional states of anger, disgust, fear, joy, and sadness. A short-term feature contour is formed by assigning the feature value computed on a frame basis to all samples belonging to the frame. For example, the energy contour is given by

\[
e(n) = E_s(m), \quad n = m - N_w + 1, \ldots, m. \tag{16}
\]

The contour trends (i.e. its plateaux, its rising or falling slopes) is a valuable feature for emotion recognition, because they describe the temporal characteristics of an emotion. The survey is limited to those acoustic features for which at least two references are found in the literature (Van Bezooijen, 1984; Cowie and Douglas-Cowie, 1996; Pantic and Rothkrantz, 2003; Gonzalez, 1999; Heuft et al., 1996; Iida et al., 2000; Iriondo et al., 2000; Montero et al., 1999; Mozziconacci and Hermes, 2000; Murray and Arnott, 1996; Pollerman and Archinard, 2002; Scherer, 2003; Ververidis and Kotropoulos, 2004; Yuan, 2002). The following statistics are measured for the extracted features:

- mean, range, variance, and the pitch contour trends;
- mean and range of the intensity contour;
- rate of speech and transmission duration between utterances.

The speech rate is calculated as the inverse duration of the voiced part of speech determined by the presence of pitch pulses (Dellaert et al., 1996; Banse and Scherer, 1996) or it can be found by the rate of syllable units. The speech signal can be segmented into syllable units using the maxima and the minima of energy contour (Mermelstein, 1975).

In Table 2, the behavior of the most studied acoustic features for the five emotional states under consideration is outlined. Anger is the emotion of the highest energy and pitch level. Angry males show higher levels of energy than angry females. It is found that males express anger with a slow speech rate as opposed to females who employ a fast speech rate under similar circumstances (Heuft et al., 1996; Iida et al., 2000). Disgust is expressed with a low mean pitch level, a low intensity level, and a slower speech rate than the neutral state does. The emotional state of fear is correlated with a high pitch level and a raised intensity level. The majority of research outcomes reports a wide pitch range. The pitch contour has falling slopes and sometimes plateaux appear. The lapse of time between speech segments is shorter than that in the neutral state. Low levels of the mean intensity and mean pitch
are measured when the subjects express sadness. The speech rate under similar circumstances is generally slower than that in the neutral state. The pitch contour trend is a valuable parameter, because it separates fear from joy. Fear resembles sadness having an almost downwards slope in the pitch contour, whereas joy exhibits a rising slope. The speech rate varies within each emotion. An interesting observation is that males speak faster when they are sad than when they are angry or disgusted.

The trends of prosody contours include discriminatory information about emotions. However, very few the efforts to describe the shape of feature contours in a systematic manner can be found in the literature. In (Leinonen et al., 1997; Linnankoski et al., 2005), several statistics are estimated on the syllables of the word ‘Sarah’. However, there is no consensus if the results obtained from a word are universal due to textual dependency. Another option is to estimate feature statistics on the rising or falling slopes of contours as well as at their plateaux at minimal/maxima (McGilloway et al., 2000; Ververidis and Kotropoulos, 2004; Bänziger and Scherer, 2005). Statistics such as the mean and the variance are rather rudimentary. An alternative is to transcribe the contour into discrete elements, i.e. a sequence of symbols that provide information about the tendency of a contour on a short-time basis. Such elements can be provided by the ToBI (Tones and Breaks Indices) system (Silverman et al., 1992). For example, the pitch contour is transcribed into a sequence of binary elements L, H, where L stands for low and H stands for high values, respectively. There is evidence that some sequences of L and H elements provide information about emotions (Stibbard, 2000). A similar investigation for 10 elements that describe the duration and the inclination of rising and falling slopes of pitch contour also exists (Mozziconacci and Hermes, 1997). Classifiers based on discrete elements have not been studied yet. In the following section, several techniques for emotion classification are described.

5. Emotion classification techniques

The output of emotion classification techniques is a prediction value (label) about the emotional state of an utterance. An utterance \( u_\xi \) is a speech segment corresponding to a word or a phrase. Let \( u_\xi, \xi \in \{1,2,\ldots,\Xi\} \) be an utterance of the data collection. In order to evaluate the performance of a classification technique, the cross-validation method is used. According to this method, the utterances of the whole data collection are divided into the design set \( D_s \) containing \( N_{D_s} \) utterances and the test set \( T_s \) comprised of \( N_{T_s} \) utterances. The classifiers are trained using the design set and the classification error is estimated on the test set. The design and the test set are chosen randomly. This procedure is repeated for several times defined by the user and the estimated classification error is the average classification error over all repetitions (Efron and Tibshirani, 1993).

The classification techniques can be divided into two categories, namely those employing

- prosody contours, i.e. sequences of short-time prosody features;
- statistics of prosody contours, like the mean, the variance, etc. or the contour trends.

The aforementioned categories will be reviewed independently in this section.

5.1. Classification techniques that employ prosody contours

The emotion classification techniques that employ prosody contours exploit the temporal

<table>
<thead>
<tr>
<th>Pitch</th>
<th>Intensity</th>
<th>Timing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Range</td>
<td>Variance</td>
</tr>
<tr>
<td>Anger</td>
<td>&gt;&gt;</td>
<td>&gt;&gt;</td>
</tr>
<tr>
<td>Disgust</td>
<td>&lt;</td>
<td>&gt;&gt;_M&lt;_F</td>
</tr>
<tr>
<td>Fear</td>
<td>&gt;&gt;</td>
<td>&gt;&gt;</td>
</tr>
<tr>
<td>Joy</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td>Sadness</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
</tbody>
</table>

Explanation of symbols: >>: increases, <<: decreases, =: no change from neutral, /\: inclines, \/: declines. Double symbols indicate a change of increased predicted strength. The subscripts refer to gender information: M stands for males and F stands for females.
information of speech, and therefore could be useful for speech recognition. Three emotion classification techniques were found in the literature, namely a technique based on artificial neural networks (ANNs) (Womack and Hansen, 1996), the multi-channel hidden Markov Model (Womack and Hansen, 1999), and the mixture of hidden Markov models (Fernandez and Picard, 2003).

In the first classification technique, the short-time features are used as an input to an ANN in order to classify utterances into emotional states (Womack and Hansen, 1996). The algorithm is depicted in Fig. 1. The utterance \( u \) is partitioned into \( Q \) bins containing \( K \) frames each. \( Q \) varies according to the utterance length, whereas \( K \) is a constant number. Let \( x_{nq} \) denote a bin of \( u \), where \( q \in \{1,2,\ldots, Q\} \). \( x_{nq} \) is classified automatically to a phoneme group, such as fricatives (FR), vowels (VL), semi-vowels (SV), etc. by means of hidden Markov Models (HMMs) (Pellom and Hansen, 1996). Let \( H_k \) denote the \( k \)th phoneme group, where \( k=1,2,\ldots, K \). From each frame \( t=1,2,\ldots, K \) of the bin \( x_{nq} \), \( D \) features related to the emotional state of speech are extracted. Let \( y_{ndt} \) be the \( d \)th feature of the \( t \)th frame for the bin \( x_{nq} \), where \( d \in \{1,2,\ldots, D\} \). The \( K \times D \) matrix of feature values is rearranged to a vector of length \( KD \), by lexicographic ordering of the rows of the \( K \times D \) matrix. This feature vector of \( KD \) feature values extracted from the bin \( x_{nq} \) is input to the ANN described in Section 5.2. Let \( \Omega \) be an emotional state, where \( c \in \{1,2,\ldots, C\} \). An ANN is trained on the \( c \)th emotional state of the \( \lambda \)th phoneme group. The output node of the ANN denotes the likelihood of \( x_{nq} \) given the emotional state \( \Omega \) and the phoneme group \( \Theta \). The likelihood of an utterance \( u \) given the emotional state \( \Omega \) is the sum of the likelihoods for all \( x_{nq} \in u \) given \( \Omega \) and \( \Theta \).

\[
P(u|\Omega) = \sum_{q=1}^{Q} \sum_{\lambda=1}^{A} P(x_{nq}|\Omega, \Theta) P(\Theta).
\]

The aforementioned technique achieves a correct classification rate of 91% for 10 stress categories using vocal tract cross-section areas (Womack and Hansen, 1996). An issue for further study is the evolution of the emotional cues through time. Such a study can be accomplished through a new classifier which employs as input the output of each ANN.

The second emotion classification technique is called multi-channel hidden Markov model (Womack and Hansen, 1999). Let \( s_i, i=1,2,\ldots, V \) be a sequence of states of a single-channel HMM. By using a single-channel HMM, a classification system can be described at any time as being in one of \( V \) distinct states that correspond to phonemes, as is presented in Fig. 2(a) (Rabiner and Juang, 1993). The multi-channel HMM combines the benefits of emotional speech classification with a traditional single-channel HMM for speech recognition. For example a \( C \)-channel HMM could be formulated to model speech from \( C \) emotional states with one dimension allocated for each emotional state, as is depicted in Fig. 2(b). In detail, the multi-channel

![Fig. 1. An emotion classification technique that employs HMMs for phoneme classification and ANNs for emotion classification.](image1)

![Fig. 2. Two structures of HMMs that can be used for emotion recognition: (a) a single-channel HMM and (b) a multi-channel HMM.](image2)
HMM consists of states $s_{cv}$, $v = 1, 2, \ldots, V$, $c = 1, 2, \ldots, C$. The states $s_{cv}$, $c = 1, 2, \ldots, C$ form a disc. Transitions are allowed from left to right as in a single-channel HMM, across emotional states within the same disc, and across emotional states in the next disc. It offers the additional benefit of a subphoneme speech model at the emotional state level instead of the phoneme level. The overall flexibility of the multi-channel HMM is improved by allowing a combined model where the integrity of each dimension is preserved (Womack and Hansen, 1999). In addition to a C mixture single-channel HMM it offers separate state transition probabilities.

The training phase of the multi-channel HMM consists of two steps. The first step requires training of each single-channel HMM to an emotional state, and the second step combines the emotion-dependent single-channel HMMs into a multi-channel HMM. In order to classify an utterance, a probability measurement is constructed. The likelihood of an utterance given an emotional state $\Omega_c$ is the ratio of the number of passes through states $s_{cv}$, $v = 1, 2, \ldots, V$ versus the total number of state transitions. The multi-channel HMM was used firstly for stress classification, and secondly for speech recognition on a data collection consisting of 35 words spoken in four stress styles. The correct stress classification rate achieved was 57.6% using MFCCs, which was almost equal to the stress classification rate of 58.6% achieved by the single-channel HMM using the same features. A reason for the aforementioned performance deterioration might be the small size of the data collection (Womack and Hansen, 1999). However, the multi-channel HMM achieved a correct speech classification rate of 94.4%, whereas the single-channel HMM achieved a rate of 78.7% in the same task. The great performance of the multi-channel HMM in speech recognition experiments might be an indication that the proposed model can be useful for stress classification in large data collections. A topic for further investigation would be to model the transitions across the disks with an additional HMM or an ANN (Bou-Ghazale and Hansen, 1998).

The third technique used for emotion classification is the so-called mixture of HMMs (Fernandez and Picard, 2003). The technique consists of two training stages. In the first stage, an unsupervised iterative clustering algorithm is used to discover $M$ clusters in the feature space of the training data, where it is assumed that the data of each cluster are governed by a single underlying HMM. In the second stage, a number of HMMs are trained on the clusters. Each HMM is trained on the $c$th emotional state of the $m$th cluster, where $c = 1, 2, \ldots, C$ and $m = 1, 2, \ldots, M$. Both training stages and the classification of an utterance which belongs to the test set are described next.

In the first training stage, the utterances of the training set are divided into $M$ clusters. Let $\gamma^{(l)} = \{\gamma_1^{(l)}, \ldots, \gamma_{m}^{(l)}, \ldots, \gamma_{M}^{(l)}\}$ be the clusters at the $l$th iteration of the clustering algorithm, $\delta^{(l)} = \{\delta_1^{(l)}, \ldots, \delta_{m}^{(l)}, \ldots, \delta_{M}^{(l)}\}$ be the HMM parameters for the cluster set $\Gamma^{(l)}$, $P(u_c|\delta_m^{(l)})$ be the likelihood of $u_c$ given the cluster with HMM parameters $\delta_m^{(l)}$, and

$$P(l) = \sum_{m=1}^{M} \sum_{u_c \in \gamma_m^{(l)}} \log P(u_c|\delta_m^{(l)})$$

be the log-likelihood of all utterances during the $l$th iteration. The iterative clustering procedure is described in Fig. 3. In the second training stage, the utterances which have already been classified into a cluster $\gamma_m$ are used to train $C$ HMMs, where each HMM corresponds to an emotional state. Let $P(\delta_m|\Omega_c)$ be the ratio of the utterances that were assigned to cluster $\gamma_m$ and belong to $\Omega_c$ over the number of the training utterances. In order to classify a test utterance $u_c$ into an emotional state the Bayes classifier is used. The probability of an emotional state $\Omega_c$ given an utterance is

$$P(\Omega_c|u_c) = \sum_{m=1}^{M} P(\Omega_c|\delta_m)P(\delta_m|u_c|\Omega_c) = \sum_{m=1}^{M} P(u_c|\Omega_c, \delta_m)P(\delta_m|\Omega_c)P(\Omega_c),$$

Step 1. Assign randomly the utterances to obtain the initial clusters $\Gamma^{(0)}$. Calculate $\Delta^{(0)}$ given $\Gamma^{(0)}$ using the Viterbi algorithm (Rabiner and Juang, 1993). Estimate $P^{(0)}$ using (18).

Step 2. Re-assign the utterances using $\Delta^{(0)}$ to the cluster with the highest likelihood in order to obtain $\Gamma^{(1)}$, i.e. assign $u_c$ to $\gamma_k^{(1)}$, where $k = \arg \max_m P(u_c|\delta_m^{(0)})$. Calculate $\Delta^{(1)}$ from $\Gamma^{(1)}$. Estimate $P^{(1)}$.

Step 3. Re-assign the utterances using $\Delta^{(l-1)}$ to obtain $\Gamma^{(l)}$. Calculate $\Delta^{(l)}$ from $\Gamma^{(l)}$. Estimate $P^{(l)}$.

Step 4. If $|P^{(l)} - P^{(l-1)}| < \epsilon$ stop, where $\epsilon$ is a user-defined threshold. The procedure stops when (18) reaches a maximum. Otherwise, $l = l + 1$ and go to Step 3.

Fig. 3. A clustering procedure that it is based on HMMs.
where \(P(u|\Omega, \delta_m)\) is the output of the HMM which was trained on the emotional state \(\Omega_c\) of the cluster \(\gamma_m\), and \(P(\Omega, \delta)\) is the likelihood of each emotional state in the data collection. The correct classification rate achieved for four emotional states by the mixture of HMMs was 62\% using energy contours in several frequency bands, whereas a single-channel HMM yields a smaller classification rate by 10\% using the same features. A topic of future investigation might be the clustering algorithm described in Fig. 3. It is not clear what each cluster of utterances represents. Also, the convergence of the clustering procedure has not been investigated yet.

5.2. Classification techniques that employ statistics of prosody contours

Statistics of prosody contours have also been used as features for emotion classification techniques. The major drawback of such classification techniques is the loss of the timing information. In this section, the emotion classification techniques are separated into two classes, namely those that estimate the probability density function (pdf) of the features and those that discriminate emotional states without any estimation of the feature distributions for each emotional state. In Table 3, the literature related to discriminant classifiers applied to emotion recognition is summarized. First, the Bayes classifier when the class pdfs are modeled either as Gaussians, or mixtures of Gaussians, or estimated via Parzen windows is described. Next, we briefly discuss classifiers that do not employ any pdf modeling such as the k-nearest neighbors, the support vector machines, and the artificial neural networks.

The features used for emotion classification are statistics of the prosody contours such as the mean, the variance, etc. A full list of such features can be found in (Verberidis and Kotropoulos, 2004). Let \(y = (y_1, y_2, \ldots, y_p)\) be the measurement vector containing \(y_{\delta d}\) statistics extracted from \(u\), where \(d = 1, 2, \ldots, D\) denotes the feature index.

According to the Bayes classifier, an utterance \(u\) is assigned to emotional state \(\Omega_c\), if

\[
\hat{c} = \arg \max_{c=1}^{C} \{P(y|\Omega_c)P(\Omega_c)\},
\]

where \(P(y|\Omega_c)\) is the pdf of \(y\) given the emotional state \(\Omega_c\), and \(P(\Omega_c)\) is the prior probability of having the emotional state \(\Omega_c\). \(P(\Omega_c)\) represents the knowledge we have about the emotional state of an utterance before the measurement vector of that utterance is available. Three methods for estimating \(P(y|\Omega_c)\) will be summarized, namely the single Gaussian model, the mixture of Gaussian densities model or Gaussian Mixture Model (GMM), and the estimation via Parzen windows.

Table 3

<table>
<thead>
<tr>
<th>Classifier</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>With pdf modeling</td>
<td></td>
</tr>
<tr>
<td>Bayes classifier using one Gaussian pdf</td>
<td>Dellaert et al. (1996), Schüller et al. (2004)</td>
</tr>
<tr>
<td>Bayes classifier using one Gaussian pdf with linear discriminant analysis</td>
<td>France et al. (2000), Lee and Narayanan (2005)</td>
</tr>
<tr>
<td>Bayes classifier using pdfs estimated by Parzen windows</td>
<td>Dellaert et al. (1996), Verberidis et al. (2004)</td>
</tr>
<tr>
<td>Without pdf modeling</td>
<td></td>
</tr>
<tr>
<td>K-nearest neighbors</td>
<td>Dellaert et al. (1996), Petrushin (1999), Picard et al. (2001)</td>
</tr>
</tbody>
</table>

\[
P(y|\Omega_c) = g(y; \mu_c, \Sigma_c) = \frac{\exp\left[-\frac{1}{2}(y - \mu_c)^T\Sigma_c^{-1}(y - \mu_c)\right]}{(2\pi)^{D/2}|\det(\Sigma_c)|^{1/2}},
\]

where \(\mu_c, \Sigma_c\) are the mean vector and the covariance matrix, and \(\det\) is the determinant of a matrix. The Bayes classifier, when the class conditional pdfs of the energy and pitch contour statistics are modeled by (21), achieves a correct classification rate of a...
56% for four emotional states (Dellaert et al., 1996). The benefit of the Gaussian model is that it is estimated fast. Its drawback is that the assumption of Gaussian distributed features may not be true for real data. Linear discriminant analysis is a method to improve the classification rates achieved by the Bayes classifier, when each $P(y|\Omega_c)$ is modeled as in (21).

In linear discriminant analysis the measurement space is transformed so that the separability between the emotional states is maximized. We will focus on the problem of two emotional states $\Omega_1$ and $\Omega_2$ to maintain simplicity. Let $N_1$ and $N_2$ be the number of utterances that belong to $\Omega_1$ and $\Omega_2$, respectively. The separability between the emotional states can be expressed by several criteria. One such criterion is the

$$J = \text{tr}(S_w^{-1} S_b),$$

(22)

where $S_w$ is the within emotional states scatter matrix defined by

$$S_w = \frac{N_1}{N_s} \Sigma_1 + \frac{N_2}{N_s} \Sigma_2,$$

(23)

and $S_b$ is the between emotional states scatter matrix given by

$$S_b = \frac{N_1}{N_s} (\mu_1 - \mu_0)(\mu_1 - \mu_0)^T$$
$$+ \frac{N_2}{N_s} (\mu_2 - \mu_0)(\mu_2 - \mu_0)^T,$$

(24)

where $\mu_0$ is the gross mean vector. A linear transformation $z = A^T y$ of measurements from space $Y$ to space $Z$ which maximizes $J$ is sought. The scatter matrices $S_{bZ}$ and $S_{wZ}$ in the $Z$-space are calculated from $S_b$ and $S_w$ in the $Y$-space by

$$S_{bZ} = A^T S_b A,$$

$$S_{wZ} = A^T S_w A.$$

(25)

Thus, the problem of transformation is to find $A$ which optimizes $J$ in the $Z$-space. It can be shown that the optimum $A$ is the matrix formed by the eigenvectors that correspond to the maximal eigenvalues of $S_w^{-1} S_b$. A linear discriminant classifier achieves a correct classification of 93% for two emotional classes using statistics of pitch and energy contours (Lee and Narayanan, 2005). Linear discrimination analysis has a disadvantage. The criterion in (22) may not be a good measure of emotional state separability when the pdf of each emotional state in the measurement space $Y$ is not a Gaussian (21) (Fukunaga, 1990).

In the GMM, it is assumed that the measurement vectors $y_t$ of an emotional state $\Omega_c$ are divided into clusters, and the measurement vectors in each cluster follow a Gaussian pdf. Let $K_c$ be the number of clusters in the emotional state $\Omega_c$. The complete pdf estimate is

$$P(y|\Omega_c) = \sum_{k=1}^{K_c} g(y; \mu_{ck}, \Sigma_{ck})$$

(26)

which depends on the mean vector $\mu_{ck}$, the covariance matrix $\Sigma_{ck}$, and the mixing parameter $\pi_{ck}$ ($\sum_{k=1}^{K_c} \pi_{ck} = 1$, $\pi_{ck} \geq 0$) of the $k$th cluster in the $c$th emotional state. The parameters $\mu_{ck}, \Sigma_{ck}, \pi_{ck}$ are calculated with the expectation maximization algorithm (EM) (Dempster et al., 1977), and $K_c$ can be derived by the Akaike information criterion (Akaike, 1974). A correct classification rate of 75% for three emotional states is achieved by the Bayes classifier, when each $P(y|\Omega_c)$ of pitch and energy contour statistics is modeled as a mixture of Gaussian densities (Slaney and McRoberts, 2003). The advantage of the Gaussian mixture modeling is that it might discover relationships between the clusters and the speakers. A disadvantage is that the EM converges to a local optimum.

By using Parzen windows an estimate of the $P(y|\Omega_c)$ could also be obtained. It is certain that at $y_c$ corresponding to $u_c \in \Omega_c$, $p(y_c|\Omega_c) \neq 0$. Since an emotional state pdf is continuous over the measurement space, it is expected that $P(y|\Omega_c)$ in the neighborhood of $y_c$ should also be non-zero. The further we move away from $y_c$, the less we can say about the $P(y|\Omega_c)$. When using Parzen windows for class pdf estimation, the knowledge gained by the measurement vector $y_t$ is represented by a function positioned at $y_t$ and with an influence restricted to the neighborhood of $y_t$. Such a function is called the kernel of the estimator. The kernel function $h(\cdot)$ can be any function from $\mathbb{R}^+ \rightarrow \mathbb{R}^+$ that admits a maximum at $y_t$ and it is monotonically increasing as $y \rightarrow y_t$. Let $d(y, y_t)$ be the Euclidean, Mahalanobis or any other appropriate distance measure. The pdf of an emotional state $\Omega_c$ is estimated by (van der Heijden et al., 2004)

$$P(y|\Omega_c) = \frac{1}{N_c} \sum_{y_t \in \Omega_c} h(d(y, y_t)).$$

(27)

A Bayes classifier achieves a correct classification rate of 53% for five emotional states, when each
$P(y|\Omega_c)$ of pitch and energy contour statistics is estimated via Parzen windows (Ververidis et al., 2004). An advantage by estimating $P(y|\Omega_c)$ via Parzen windows is that a prior knowledge about the conditional pdf of the measurement vectors is not required. The pdfs of the measurement vector for small data collections are hard to find. The execution time for modeling a conditional pdf by Parzen windows is relatively shorter than by a GMM estimated with the EM algorithm. A disadvantage is that the estimate of $P(y|\Omega_c)$ has a great number of peaks that are not present in the real pdf.

A support vector classifier separates the emotional states with a maximal margin. The margin $\gamma$ is defined by the width of the largest 'tube' not containing utterances that can be drawn around a decision boundary. The measurement vectors that define the boundaries of the margin are called support vectors. We shall confine ourselves to a two-class problem without any loss of generality. A support vector classifier was originally designed for a two-class problem, but it can be expanded to more classes.

Let us assume that a training set of utterances is denoted by $\{y_i\}_{i=1}^{N_x} = \{(y_i, l_i)\}_{i=1}^{N_x}$, where $l_i \in \{-1, +1\}$ is the emotional state membership of each utterance. The classifier is a hyperplane

$$g(y) = w^T y + b,$$

where $w$ is the gradient vector which is perpendicular to the hyperplane, and $b$ is the offset of the hyperplane from the origin. It can be shown that the margin is inversely proportional to $\|w\|^2/2$. The quantity $l_i g(y_i)$ can be used to indicate to which side of the hyperplane the utterance belongs to. $l_i g(y_i)$ must be greater than 1, if $l_i = +1$ and smaller than −1, if $l_i = −1$. Thus, the choice of the hyperplane can be rephrased to the following optimization problem in the separable case:

$$\begin{align*}
\text{minimize} & \quad \frac{1}{2}w^T w \\
\text{subject to} & \quad l_i (w^T y_i + b) \geq 1, \quad i = 1, 2, \ldots, N_x.
\end{align*}$$

A global optimum for the parameters $w, b$ is found by using Lagrange multipliers (Shawe-Taylor and Cristianini, 2004). Extension to the non-separable case can be made by employing slack variables. The advantage of support vector classifier is that it can be extended to non-linear boundaries by the kernel trick. For four stress styles, the support vector classifier can achieve a correct classification rate of 46% using energy contours in several frequency bands (Fernandez and Picard, 2003).

The $k$-nearest neighbor classifier ($k$-NN) assigns an utterance to an emotional state according to the emotional state of the $k$ utterances that are closest to $u_i$ in the measurement space. In order to measure the distance between $u_i$ and the neighbors, the Euclidean distance is used. The $k$-NN classifier achieves a correct classification rate of 64% for four emotional states using statistics of pitch and energy contours (Dellaert et al., 1996). The disadvantages of $k$-NN is that systematic methods for selecting the optimum number of the closest neighbors and the most suitable distance measure are hard to find. If $k$ equals to 1, then the classifier will classify all the utterances in the design set correctly, but its performance on the test set will be poor. As $k \to \infty$, a less biased classifier is obtained. In the latter case, the optimality is not feasible for a finite number of utterances in the data collection (van der Heijden et al., 2004).

ANN-based classifiers are used for emotion classification due to their ability to find non-linear boundaries separating the emotional states. The most frequently used class of neural networks is that of feedforward ANNs, in which the input feature values propagate through the network in a forward direction on a layer-by-layer basis. Typically, the network consists of a set of sensory units that constitute the input layer, one or more hidden layers of computation nodes, and an output layer of computational nodes. Let us consider an one-hidden layer feedforward neural network that has $Q$ input nodes, $A$ hidden nodes, and $B$ output nodes, as is depicted in Fig. 4.

The neural network provides a mapping of the form $z = f(y)$ defined by

$$v_a = g_1(w_0^T y + w_0),$$

$$z_b = g_2(u_0^T v + u_0),$$

![Fig. 4. An one hidden layer feedforward neural network.](image-url)
where $W = [w_{a\alpha}] = [w_1\cdots|w_{a\alpha}|\cdots|w_{d\alpha}]$ is the weight matrix, $w_a$ is its $a$th column, $w_0$ is the bias, and $g_1(\cdot)$ is the activation function for the input layer. Similarly $U = [u_{b\beta}] = [u_1\cdots|u_{b\beta}|\cdots|u_{B\beta}]$ is the weight matrix for the hidden layer, $u_b$ is its $b$th column, $u_0$ is the bias, and $g_2(\cdot)$ is the activation function for the hidden layer. Usually, $g_1(\cdot)$ is the sigmoid function described by

$$g_1(v) = \frac{1}{1 + \exp(-v)}, \quad (32)$$

and $g_2(\cdot)$ is the softmax function defined by

$$g_2(v) = \frac{\exp(v)}{\sum_{b=1}^{B} \exp(v_b)}. \quad (33)$$

Activation functions for the hidden units are needed to introduce a non-linearity into the network. The softmax function guarantees that the outputs lie between zero and one and sum to one. Thus, the outputs of a network can be interpreted as posterior probabilities for an emotional state. The weights are updated with the back-propagation learning method (Haykin, 1998). The objective of the learning method is to adjust the free parameters of the network so that the mean square error defined by the objective function described by

$$J_{SE} = \frac{1}{2} \sum_{\xi=1}^{N_{\xi}} \sum_{b=1}^{B} (f_b(y_\xi) - l_{\xi,b})^2, \quad (34)$$

where $f_b$ denotes the value of the $b$th output node. The target is usually created by assigning $l_{\xi,b} = 1$, if the label of $y_\xi$ is $Q_b$. Otherwise, $l_{\xi,b}$ is 0. In emotion classification experiments, the ANN-based classifiers are used in two ways:

- an ANN is trained to all emotional states;
- a number of ANNs is used, where each ANN is trained to a specific emotional state.

In the first case, the number of output nodes of the ANN equals the number of emotional states, whereas in the latter case each ANN has one output node. An interesting property of ANNs is that by changing the number of hidden nodes and hidden layers we control the non-linear decision boundaries between the emotional states (Haykin, 1998; van der Heijden et al., 2004). The ANN-based classifiers may achieve a correct classification rate of 50.5% for four emotional states using energy contours in several frequency bands (Fernandez and Picard, 2003) or 75% for seven emotional states using pitch and energy contour statistics of another data collection (Schüller et al., 2004).

6. Concluding remarks

In this paper, several topics have been addressed. First, a list of data collections was provided including all available information about the databases such as the kinds of emotions, the language, etc. Nevertheless, there are still some copyright problems since the material from radio or TV is held under a limited agreement with broadcasters. Furthermore, there is a need for adopting protocols such as those in (Douglas-Cowie et al., 2003; Scherer, 2003; Schröder, 2005) that address issues related to data collection. Links with standardization activities like MPEG-4 and MPEG-7 concerning the emotion states and features should be established. It is recommended the data to be distributed by organizations (like LDC or ELRA), and not by individual research organizations or project initiatives, under a reasonable fee so that the experiments reported using the specific data collections could be repeated. This is not the case with the majority of the databases reviewed in this paper, whose terms of distribution are rather unclear.

Second, our survey has been focused on feature extraction methods that are useful in emotion recognition. The most interesting features are the pitch, the formants, the short-term energy, the MFCCs, the cross-section areas, and the Teager energy operator-based features. Features that are based on voice production models have not fully been investigated (Womack and Hansen, 1996). Non-linear aspects of speech production also contribute to the emotional speech coloring. Revisiting the fundamental models of voice production is expected to boost further the performance of emotional speech classification.

Third, techniques for speech classification into emotional states have been reviewed. The classification rates reported in the related literature are not directly comparable with each other, because they were measured on different data collections by applying different experimental protocols. Therefore, besides the availability of data collections, common experimental protocols should be defined and adopted, as for example in speech/speaker recognition, biometric person authentication, etc. Launching competitions like those regularly hosted by NIST (i.e. TREC, TRECVID, FERET, etc.)
would be worth pursuing. The techniques were separated into two categories, namely the ones that exploit timing information and those ignoring any timing information. In the former category, three techniques based on ANNs and HMMs were described. There are two differences between HMM- and ANN-based classifiers. First, HMM-based classifiers require strong assumptions about the statistical characteristics of the input, such as the parameterization of the input densities as GMMs. In many cases, correlation between the features is not included. This assumption is not required for ANN-based classifiers. An ANN learns something about the correlation between the acoustic features. Second, ANNs offer a good match with discriminative objective functions. For example, it is possible to maximize discrimination between the emotional states rather than to most faithfully approximate the distributions within each class (Morgan and Bourlard, 1995). The advantage of techniques exploiting timing information is that they can be used for speech recognition as well. A topic that has not been investigated is the evolution of emotional cues through time. Such an investigation can be achieved by a classifier that uses timing information for long speech periods. Well-known discrimination classifiers that do not exploit timing information have also been reviewed. Such classifiers include the support vector machines, the Bayes classifier with the class pdfs modeled as mixtures of Gaussians, the k-nearest neighbors, etc. The techniques that model feature pdfs may reveal cues about the modalities of the speech, such as the speaker gender and the speaker identities. One of the major drawbacks of these approaches is the loss of the timing information, because the techniques employ statistics of the prosody features such as the mean, the variance, etc. and neglect the sampling order. A way to overcome the problem is to calculate statistics over rising/falling slopes or during the plateaux at minima/maxima (McGilloway et al., 2000; Ververidis and Kotropoulos, 2005). It appears that most of the contour statistics follow the Gaussian distribution or the $X^2$, or can be modeled by mixture of Gaussians. However, an analytical study of the feature distributions has not been undertaken yet.

Most of the emotion research activity has been focused on advancing the emotion classification performance. In spite of the extensive research in emotion recognition, efficient speech normalization techniques that exploit the emotional state information to improve speech recognition have not been developed yet.

**Acknowledgment**

This work has been supported by the research project 01ED312 “Use of Virtual Reality for training pupils to deal with earthquakes” financed by the Greek Secretariat of Research and Technology.

**References**


