Perception of synthetic emotion expressions in speech: Categorical and dimensional annotations

Judith M. Kessens\textsuperscript{a}, Mark A. Neerincx\textsuperscript{a,b}, Rosemarijn Looije\textsuperscript{a}, Melanie Kroes\textsuperscript{c}, Gerrit Bloothooft\textsuperscript{c}

\textsuperscript{a}TNO Defense, Safety and Security
P.O. Box 23
3769 ZG Soesterberg
The Netherlands
judith.kessens@tno.nl

\textsuperscript{b}Delft University of Technology
P.O. Box 5031
2600 GA Delft
The Netherlands
http://mmi.tudelft.nl

\textsuperscript{c}University of Utrecht
Trans 10
3512 JK Utrecht
The Netherlands
http://www-uilots.let.uu.nl

Abstract

In this paper, both categorical and dimensional annotations have been made of neutral and emotional speech synthesis (anger, fear, sad, happy and relaxed). With various prosodic emotion manipulation techniques we found emotion classification rates of 40%, which is significantly above chance level (17%). The classification rates are higher for sentences that have a semantics matching the synthetic emotion. By manipulating the pitch and duration, differences in arousal were perceived whereas differences in valence were hardly perceived. Of the investigated emotion manipulation methods, EmoFilt and EmoSpeak performed very similar, except for the emotion fear. Copy synthesis did not perform well, probably caused by suboptimal alignments and the use of multiple speakers.

1. Introduction

For natural human-computer interaction, affective computing is indispensable [1]. An example is a robot that serves as a personal assistant for children with a chronic illness to adhere to a healthy lifestyle. In order to make the interaction more natural and effective it is important that the computer assistant reacts in a proper way to the emotions of the children. Also, the more human-like the robot will be, the more persuasive, engaging and fun the interaction will be. An important part of the communication between robot and child is the vocal interaction. Besides using emotional word expressions, also non-verbal aspects of emotion expression play a role. If a synthetic robot-voice is used, the content of the message can be adjusted freely, but the speaking style cannot be adjusted. Therefore, there is a need for flexible manipulation of emotional aspects of the speech. In the last decades, various methods have been proposed for making the synthetic speech sound more expressive. For instance, Schröder ([2], [3]) and Murray and Arnott [4] present an overview of emotional speech synthesis methods. As noted by Schröder [2], there exists a trade-off between the flexibility of manipulating speech parameters and the perceived naturalness. For instance, emotion manipulation is relatively simple for formant synthesis as the speech voice source and vocal tract parameters can be manipulated freely, but formant synthesis sounds unnatural and ‘robot-like’. On the other hand, state-of-the-art concatenative speech synthesis systems sound very natural, but manipulation of the speech parameters is less straightforward.

One approach which is used for concatenative synthesizers is to record emotional speech databases, from which emotional speech units can be selected [4]. In line with this approach, ‘Hidden-Markov-Model’-based speech synthesis can be trained on emotional speech data [6]. Another approach exists in deriving rules for modifying speech parameters that are correlated to emotion expression (e.g. duration, pitch and voice quality parameters) [7]. Finally, stochastic methods (voice conversion technology) have been used for transforming emotion in speech [8].

As we will use the expressive voice in robots, e.g. a rescuing robot that operate in disaster areas, or computer assistant that assists diabetic children, the voice may sound unnatural. Moreover, in order to make clear to the user that he/she is communicating with a robot, we prefer to use a robot-like voice. As a lot of research on emotional speech synthesis has been done in the past, we decided to start of with comparing different available emotional speech synthesis methods. For diphone synthesis, we found two emotion manipulation editors (EmoFilt [7] and EmoSpeak [2]) that we could directly use for our purposes. A disadvantage of using diphone synthesis is that it is difficult to manipulate ‘voice-quality’, which is considered as an important aspect for emotion expression. However, contradictory results has been found on how well emotions can be recognized without controlling voice quality [3]. The importance of voice quality for emotion expressions appears to be
emotion [9] and speaker dependent [10]. Therefore, the research in this paper will address the question whether synthetic emotion express is feasible without manipulation of voice quality parameters. To this end, a perception experiment is performed in which various emotion manipulation techniques are compared. Both categorical and dimensional annotations are made of semantically neutral and emotional sentences. Sentences are selected from a natural speech database containing emotion annotation, which enables us to compare human perception of the synthetic and natural emotion expressions.

2. Method

2.1 Emotions

Besides neutral synthesis, synthesis of the following five emotions have been investigated: anger, fear, sad, happy and relaxed. The first four emotions have been chosen as these four emotions occur most in research on emotional speech synthesis and as they are denoted as ‘basic emotions’ in psychological research (e.g. [11]). The emotion ‘relaxed’ has been added in order to cover the full (valence, arousal) space that we used for dimensional annotation.

2.2 Emotion annotation

Emotions have been annotated both categorically and dimensionally. For dimensionally annotation, the dimensions ‘valence’ (positive/negative) and ‘arousal’ (active/passive) were annotated. The third dimension ‘power’ (amount of control) is not used as other research [12] has shown that the speech is less affected by this dimension. The dimensions ‘arousal’ and ‘valence’ were independently annotated on a scale ranging from (-50,+50). For categorically annotation, listeners had to make a forced choice between six alternatives: the five emotions (see 2.1) or neutral emotion.

2.3 Generation of emotional and neutral speech

In this experiment, we used diphone synthesis, which consists of concatenation of diphone-units (sound transitions). The speech synthesis system MARY [13] is used to automatically generate from text: the speech sounds, their duration and intonation pattern (pho-files). Besides Neutral Synthesis (NS), three methods are used to manipulate the pho-files:

- **EmoFilt (EF) of Burkhardt [7]**. The following parameters are manipulated: Pitch (level, range, variability, stress contour) duration (tempo and durations of voiced fricatives, vowels, focus syllables and pauses).
- **EmoSpeak (ES) of Schröder [2]**. The following parameters are manipulated: Pitch (level, range, variability, stress contour) duration (tempo and durations of voiced fricatives, vowels, focus syllables and pauses).
- **Copy Synthesis (CS)**. Both phoneme durations and intonation is estimated from natural speech. The phoneme durations are obtained by alignment with an English version of the TNO speech recognizer (based on Sonic [14]). The intonation is estimated with Praat [15].

The neutral and emotional pho-files are subsequently used with MBROLA diphones (male voice, US2) [16] in order to obtain the neutrally and emotionally synthesized speech.

2.4 Test conditions

The sentences have been selected from the Belfast Naturalistic Database [17]. This database contains recordings from TV-shows and has been annotated both categorically and dimensionally. For each annotated emotion in the database, sentences were selected with neutral semantics and with semantics that matches the annotated emotion. Examples of sentences are: “He’s sitting outside of my daughter’s house” (neutral), “I haven’t got a life with you” (angry). Only if two independent raters fully agreed on the semantics, the sentence has been selected.

The 34 conditions that have been tested are shown in Table 1. The emotion ‘relaxed’ is not tested with EmoFilt, as no standard settings for ‘relaxed’ are available. For each test condition, two sentences were synthesized, resulting in 68 audio fragments. In addition, we randomly selected 10% of the fragments that were presented twice in order to estimate intra-rater reliability.

2.5 Experimental setup

The experiment consisted of two parts of 30 minutes each. In the first part, the audio fragments were annotated dimensionally, whereas in the second part they were annotated categorically. Listeners were instructed to listen to the audio fragments and to make the
2.6 Participants

Twenty participants (7 male, 13 female) performed the annotation task. Their age ranged from 19 to 29 (mean = 22.3), and none of them reported any hearing problems. Two speakers were native speaker of English, the other participants have a Dutch mother tongue, but all had a high language proficiency in English.

3. Results

3.1 Participants

3.1.1 Inter-rater agreement
Inter-rater agreement is calculated for all data. For the categorical annotations a slight inter-rater agreement was found (Fleiss’ kappa = 0.198). For the dimensional annotations, inter-rater agreement was significant (p<0.001). For arousal and valence, Kendall’s W was .07 and 0.12 respectively.

3.1.2 Subjective ratings
On a 5-point Likert scale, participants answered 6 questions. The first question was asked before the experiment, questions 2 and 3 after the categorical and dimensional part of the experiment, and questions 4 to 6 after the experiment. The results presented in Table 2 show that the participants do not have much experience with synthetic speech. The dimensional annotation task was rated as significantly more difficult than the categorical annotation task (p<0.05), but participants liked both tasks. The intelligibility and quality of the speech is rated as low, and the naturalness as moderate.

<table>
<thead>
<tr>
<th>Question</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Experience with synthetic speech  
(1=very little, 5=very much) | 1.65 | 0.81 |
| 2        |      |    |
| Difficulty of categorical annotation  
(1=very easy, 5=very difficult) | 2.55 | 0.95 |
| Difficulty of dimensional annotation  
(1=very easy, 5=very difficult) | 3.35 | 1.00 |
| 3        |      |    |
| Amusement with categorical annotation  
(1=very annoying, 5=very nice) | 3.65 | 0.67 |
| Amusement with dimensional annotation  
(1=very annoying, 5=very nice) | 3.90 | 0.64 |
| 4        |      |    |
| Intelligibility of synthetic speech  
(1=very good, 5=very bad) | 3.50 | 0.95 |
| 5        |      |    |
| Quality of synthetic speech  
(1=very good, 5=very bad) | 3.15 | 0.81 |
| 6        |      |    |
| Naturalness of synthetic speech  
(1=very unnatural, 5=very natural) | 2.20 | 0.95 |

Table 2: Subjective ratings

3.2 Categorical data

3.2.1 Emotion classification rates
In Figure 1, the emotion classification rates are shown for the various methods, averaged over neutral and semantical emotional sentences. The performance of EmoFilt and EmoSpeak is very similar, except for the emotion ‘fear’ on which EmoSpeak performs poorly. Copy Synthesis has lower classification rates for ‘anger’, ‘fear’ and ‘sad’.

Figure 1: Emotion classification rates (%) for each method

Table 3 shows the confusion matrix of the 6 emotion categories. The mean percentage of correct recognition (diagonals, indicated in bold) is 40%, which is significantly (p<0.001) higher than chance level (17%). Anger is recognized best (49% correct), whereas fear is recognized worst (32%). Furthermore, the confusions can be divided into two groups (indicated with the dashed line); 1) fear, happy and anger, and 2) relaxed, sad and neutral. Confusion of emotions within each group is more frequent than between those two groups.

<table>
<thead>
<tr>
<th></th>
<th>anger</th>
<th>happy</th>
<th>fear</th>
<th>relax</th>
<th>sad</th>
<th>neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>anger</td>
<td>49</td>
<td>17</td>
<td>10</td>
<td>4</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>happy</td>
<td>120</td>
<td>19</td>
<td>37</td>
<td>6</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>fear</td>
<td>120</td>
<td>21</td>
<td>13</td>
<td>32</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>relax</td>
<td>80</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>42</td>
<td>19</td>
</tr>
<tr>
<td>sad</td>
<td>120</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>18</td>
<td>46</td>
</tr>
<tr>
<td>neutral</td>
<td>140</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>23</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 3: Confusion matrix

3.2.2 Neutral vs. emotional text semantics
In Table 4, the percentage of correctly recognized emotions are given for the sentences with emotional and neutral semantics, now averaged over the different methods. It can be seen that the recognition percentages are higher for the sentences with emotional semantics, but none of the differences are significant.

<table>
<thead>
<tr>
<th>emotion</th>
<th>neutral sem.</th>
<th>emotional sem.</th>
</tr>
</thead>
<tbody>
<tr>
<td>anger</td>
<td>40.4</td>
<td>57.5</td>
</tr>
<tr>
<td>fear</td>
<td>22.5</td>
<td>40.8</td>
</tr>
<tr>
<td>happy</td>
<td>37.4</td>
<td>37.5</td>
</tr>
<tr>
<td>relaxed</td>
<td>27.8</td>
<td>56.3</td>
</tr>
<tr>
<td>sad</td>
<td>45.8</td>
<td>45.8</td>
</tr>
</tbody>
</table>

Table 4: Emotion classification rates (%) of sentences with neutral vs. emotional text semantics
3.3 Dimensional data

In Figure 2, the arousal and valence values are given for the different emotion manipulation methods, for neutral synthesis and for the natural speech. Results are presented for all conditions together. The main findings are summarized below.

### Settings

<table>
<thead>
<tr>
<th></th>
<th>Copy synthesis</th>
<th>EmoFilt</th>
<th>EmoSpeak</th>
<th>Neutral synthesis</th>
<th>Natural speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>anger</td>
<td>-35</td>
<td>-30</td>
<td>-25</td>
<td>-20</td>
<td>-15</td>
</tr>
<tr>
<td>fear</td>
<td>-20</td>
<td>-15</td>
<td>-10</td>
<td>-5</td>
<td>0</td>
</tr>
<tr>
<td>happy</td>
<td>-5</td>
<td>0</td>
<td>5</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>relaxed</td>
<td>-15</td>
<td>-10</td>
<td>0</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>sad</td>
<td>-30</td>
<td>-25</td>
<td>-20</td>
<td>-15</td>
<td>-10</td>
</tr>
</tbody>
</table>

Arousal

Valence

Figure 2: Mean arousal and valence values for different methods, neutral synthesis and natural speech

3.3.1 Synthetic vs. natural speech

In general, the synthetic emotions are annotated less extreme (i.e. closer to zero) than the natural emotions. Significant differences are found in the arousal of ‘anger’ and ‘happy’ and the valence of all emotions, except ‘sad’ (p<0.05). In general, the valence of the synthetic emotions are close to zero, meaning that participants had difficulty in perceiving valence in the synthetic speech. This might explain why we see mainly confusions between two groups of emotions, namely the groups of emotions that have similar arousal values; group 1 (anger, fear, happy) with positive arousal, and group 2 (neutral, sad relaxed) with zero or negative arousal.

3.3.2 Neutral vs. emotional semantics

We also compared the dimensional annotation results of the semantically neutral and emotional sentences. In general, the dimensional annotations of the sentences with neutral and emotional semantics do not differ significantly.

3.3.3 Differences between methods

Between the emotion manipulation methods, there are not many differences. The following differences are significant (p<0.05) for both arousal and valence:

- EmoSpeak vs. EmoFilt for the emotion ‘fear’
- Copy Synthesis vs. EmoFilt/EmoSpeak for the emotions ‘sad’ and ‘anger’.

For these combination of emotions and methods, we also find the largest differences in classification rates, see Figure 1.

4. Discussion

Although the classification rates of the synthesized emotions are on average 40%, the classification rates of emotion expressions in natural speech are much higher, e.g. Edgington [18] report an average classification rate of 79% for a comparable set of emotions and semantically neutral sentences. A possible explanation can be that not all aspects of vocal expression are modeled, e.g. voice quality is an important parameter which has not been taken into account. The fact that we did not use native listeners possibly influenced our results as well. However, in [19] it has been shown that non-native listeners both positively and negatively influences the emotion classification rates, depending on the emotion and the native language of the listener. Finally, another cause could be that the speech intelligibility and quality of the diphone synthesis is not very high.

Our research shows that without controlling voice quality, classification rates can be obtained which are far above chance level. By manipulating the pitch and duration, differences in arousal were perceived whereas differences in valence were hardly perceived. This is in line with the general finding that more and stronger vocal correlates have been found with the arousal dimension compared to the valence dimension, e.g. [20].

In contrast with results reported by some authors, e.g. [21], copy synthesis did not perform well. A plausible explanation for this finding is that our phone alignments are not optimal, as we used spontaneous speech recordings, containing a lot of assimilations, co-articulations, and reductions/deletions of phones or even complete syllables. These mismatches between the standard and actual pronunciations, resulted in misalignments. Secondly, despite that we used natural and not acted emotions, the emotions are not “full-blown”, only 34% of the clips in the database are rated

---

1 Exceptions: The arousal of relaxed is less extreme for neutral semantics, and the valence of anger is more extreme for neutral semantics
as strong [22]. Thirdly, we used intonation and duration patterns from various speakers, who use different strategies to express emotions, thus creating less consistent emotion manipulation patterns.

Another important finding is that EmoSpeak and EmoFilt perform identical, except for the emotion fear, which could hardly be identified when generated with EmoSpeak. These differences may originate in different settings: for fear, EmoSpeak only increases the F0 by a small amount (+8%) and decreases the speech rate (-3%), whereas EmoFilt increases the F0 much more (+100%) and increases the speech rate (30%).

5. Conclusions

With various prosodic emotion manipulation techniques, we found emotion classification rates of 40%, which is significantly above chance level (17%). The classification rates are higher for sentences that have a semantics matching the synthetic emotion. Our results show that the classification rates depend on the emotion manipulation technique that is used. EmoFilt and EmoSpeak perform very similar, but the performance of copy synthesis was lower, which is probably caused by suboptimal alignment and the use of multiple speakers. However, for all methods we found that arousal differences were perceived with the prosodic manipulation of the speech, whereas differences in valence were hardly perceived. Overall, these prosodic manipulations can have an added value to the human-machine dialogues, for example, helping to motivate or persuade persons to perform their daily activities well (e.g., physical exercises and diets).

6. Acknowledgements

This research has been performed in the framework of the MultimediaN project (www.multimedian.nl).

References