Visualizing acoustic similarities between emotions in speech: an acoustic map of emotions

Khiet P. Truong and David A. van Leeuwen

TNO Human Factors, Dept. of Human Interfaces
P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
{khiets.truong, david.vanleeuwen}@tno.nl

Abstract

In this paper, we introduce a visual analysis method to assess the discriminability and confusability between emotions according to automatic emotion classifiers. The degree of acoustic similarities between emotions can be defined in terms of distances that are based on pair-wise emotion discrimination experiments. By employing Multidimensional Scaling, the discriminability between emotions can then be visualized in a two-dimensional plot that is relatively easy to interpret. This ‘map of emotions’ is compared to the well-known ‘Feeltrace’ two-dimensional mapping of emotions. While there is correlation with the ‘arousal’ dimension of Feeltrace, it appears that the ‘valence’ dimension is difficult to relate to the acoustic map.

Index Terms: emotion, emotion detection, acoustic map.

1. Introduction

From previous studies that have performed emotion classification experiments with the aim to automatically detect emotions in speech, we have learned that some emotions are often confused with each other. For example, acoustic classifiers often confuse sadness with boredom or neutralness, happiness is often confused with anger, while sadness is almost never confused with anger [1, 2, 3]. Usually, we can observe confusions of emotions in a confusion matrix which is obtained as output of an emotion classification experiment. However, it may take some effort for the reader to interpret the numbers and to determine the type of confusions. Instead of interpreting the numbers in a confusion matrix, we would like to have an alternative way to displaying confusions and present an acoustic map of emotions in which we can relatively easily interpret the confusions between emotions in terms of distances. Emotions are placed in this map such that the distance can be interpreted as a similarity measure: the closer the emotions are to each other, the more similar they are, and thus, the more difficult it is to automatically discriminate between them.

The goal of this paper is to present a visual analysis method to assess discriminability between emotions. How can we visualize the degree of acoustic similarity between emotions and how does this visualization relate to the well-known arousal-valence space? In Section 2, we describe the data we used in this study. The speech features and learning technique used to train emotion models are described in Section 3. In Section 4 we compare classification and detection as analysis methods. Section 5 illustrates how we measure the degree of confusion between emotions in terms of distances. The results are shown in Section 6 where we present a visualization of the degree of similarity between emotions. Finally, in Section 7, we discuss our results and conclusions.

2. Data

We used the Berlin Emotional Speech database [4] which contains emotional speech from five female and five male actors. Each actor uttered ten different sentences (plus some extra versions) in seven different emotions: neutral (Ne), anger (An), fear (Fe), joy (Jo), sadness (Sa), disgust (Di) and boredom (Bo). This resulted in a total of 800 emotional utterances that were normal sentences which could be used in everyday life. In a validation experiment, 20 subjects were asked to recognize the emotion and to rate the naturalness of the emotional utterance.

In our analyses, we only used those utterances that had a human recognition accuracy of more than 80% and a naturalness of more than 60% which resulted in a total of 535 utterances (see Table 1).

<table>
<thead>
<tr>
<th>Emotion</th>
<th>N</th>
<th>Emotion</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger (An)</td>
<td>127</td>
<td>Boredom (Bo)</td>
<td>81</td>
</tr>
<tr>
<td>Disgust (Di)</td>
<td>46</td>
<td>Fear (Fe)</td>
<td>69</td>
</tr>
<tr>
<td>Joy (Jo)</td>
<td>71</td>
<td>Sadness (Sa)</td>
<td>62</td>
</tr>
</tbody>
</table>

We are well aware of the fact that the database contains acted data and so-called full blown emotions and that these are rare in everyday life. Therefore, we should realize that these are results obtained in an optimal situation with extreme emotions and interpret the results of the analyses as such.

3. Method

3.1. Features

We used RASTA-PLP [5, 6] as speech features. For each 32 ms-frame, 12 RASTA-PLP coefficients plus 1 log energy component were computed with a forwardshift of 16 ms. In addition, deltas were calculated by taking the first order derivatives of the 13 features over five consecutive frames. The features were normalized such that the mean and standard deviation of all features of each utterance are zero and one respectively. Other spectral features such as Mel-Frequency Cepstrum Coefficients (MFCCs) would also have been good candidates, but mainly for practical reasons and their good performance in our (related) speaker recognition system, we chose RASTA-PLP.

3.2. Modeling technique

Gaussian Mixture Models (GMM) were used to model the emotional speech. Considering the little amount of speech...
data we had for training and testing, the GMMs were trained with four Gaussian components. Further, five iterations of the Expectation-Maximization algorithm (EM) were used to estimate the parameters of the GMMs. In testing, a maximum likelihood criterion was used. Depending on the type of experiment (classification or detection), we used the log-likelihood or the log-likelihood ratio as a ‘soft detector’ score (see Section 4.2).

4. Emotion classification vs. detection

4.1. Emotion classification in seven emotions

In order to have a feeling about the possible confusions between emotions that acoustic classifiers can make, we first performed emotion classification experiments in seven emotions. Training and testing was performed through a cross-validation procedure based on the speaker identity: speakers that are present in the training set are never included in the test set. GMMs were trained for each emotion class. During testing, the emotion class is determined by the corresponding GMM that produces the highest log-likelihood for a given trial. As a result, we obtained the following confusion matrix in Table 2.

### Table 2: Confusion matrix

<table>
<thead>
<tr>
<th>Classified as</th>
<th>An</th>
<th>Bo</th>
<th>Di</th>
<th>Fe</th>
<th>Jo</th>
<th>Ne</th>
<th>Sa</th>
</tr>
</thead>
<tbody>
<tr>
<td>An</td>
<td>81</td>
<td>4</td>
<td>10</td>
<td>0</td>
<td>17</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Bo</td>
<td>1</td>
<td>37</td>
<td>2</td>
<td>7</td>
<td>16</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Di</td>
<td>9</td>
<td>1</td>
<td>18</td>
<td>4</td>
<td>9</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Fe</td>
<td>3</td>
<td>12</td>
<td>7</td>
<td>18</td>
<td>14</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Jo</td>
<td>16</td>
<td>3</td>
<td>8</td>
<td>7</td>
<td>33</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Ne</td>
<td>1</td>
<td>22</td>
<td>10</td>
<td>8</td>
<td>38</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Sa</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>46</td>
</tr>
</tbody>
</table>

Usually, an averaged accuracy is given as a performance measure which can be defined as the number of correctly classified trials (the diagonal in Table 2) divided by the total number of trials. In our case, we obtained an averaged accuracy of 50.7%. However, disadvantages to this accuracy measure are that it depends on the number of emotion classes and the proportions of the test trials of each emotion class. Further, it is very likely that in real life applications, the classifier can encounter other emotions than the assumed number of emotion classes from which the classifier can choose. Finally, this measure makes it difficult to compare different studies to each other because other databases may contain different types and numbers of emotions.

4.2. Emotion detection: emotion $X$ vs. emotion $\neg X$

Given the disadvantages of classification as discussed above, we prefer to perform emotion detection within a detection framework. In detection, the task is defined as a two-class problem: is this emotion $X$ or not? Typically, one class represents the emotion $X$ that one would like to detect; we call this the target emotion. And the other class represents the class that one would not want to detect, e.g., $\neg X$; this is called the non-target. In our case, we performed speaker-independent detection experiments for each emotion class according to a cross-validation procedure based on the speaker’s identity. For each detection experiment, a target model was trained on emotion $X$ and a non-target model was trained on $\neg X$, i.e., all other emotions excluding the target emotion $X$. A ‘soft detector’ score is obtained by determining the log-likelihood ratio of the data given the target and non-target GMMs respectively. The detector can make two types of errors: misses and false alarms. By examining the tradeoff between these two errors, we can identify the Equal Error Rate (EER), the point at which the miss rate is equal to the false alarm rate. EER is often used as a single performance measure where lower EER means better discrimination.

In Fig. 1 we show in a DET plot [7] the detection results for the seven emotions. We can observe that Sa has a relatively low EER while Fe has a high EER; Sa is much easier to detect than Fe. In terms of our “acoustic map of emotions”, this would mean that Sa lies “further away” from the other emotions, while Fe is more “in the middle”. We will also be able to observe the confusability between individual emotions in our “map of emotions” which is not observable in Fig. 1.

5. Defining acoustic distances between emotions

5.1. Pair-wise emotion discrimination experiments

In order to be able to visualize the emotions in an acoustic map, we performed speaker-independent pair-wise emotion detection experiments and subsequently, computed acoustic distances between emotions based on the obtained EERs. GMMs were trained for each emotion class. In each discrimination experiment, the task was to discriminate between two specific emotion classes (e.g., An vs Bo, An vs Fe). So in testing, we tested only with trials coming from the pair of emotion classes under discrimination. The log-likelihood ratios obtained with the two GMMs are used as ‘soft detector’ scores to determine EER. The EERs obtained with these pair-wise discrimination experiments represent discrimination performance between two emotions and can be interpreted as a similarity measure: the higher the EER, the more similar the two emotions are. However, EERs are not actual distance metrics: EERs have a scale from 0 to 1 (with increasing performance) while distances have a range of 0 and up.
5.2. Defining a distance metric

A proper distance representation of the EER is a quantity known from signal detection theory as \( d' \) (‘d-prime’). Assuming equal variance of the target and non-target score distributions (i.e., the obtained log-likelihood ratios of the target and non-target test trials), \( d' \) is defined as the difference in mean between the distributions expressed in terms of the standard deviation, 

\[
d' = \frac{\mu_{tar} - \mu_{non}}{\sigma}
\]

Under the assumption of Gaussian score distributions DET curves are straight lines perpendicular to the equal-probability diagonal. The EER \( P_{EER} \) and \( d' \) are related through the inverse cumulative normal distribution, or probit function \[8\]:

\[
d' = -2 \text{probit}(P_{EER}) = -2\sqrt{2}\text{erf}^{-1}(2P_{EER} - 1) \quad (1)
\]

This probit function, expressed here in terms of the inverse error function, is just the warping function of the DET axes, so that \( d' \) varies linearly in the DET plot from 0 in the upper-right corner to about 6 lower-left corner.

6. Results

6.1. EERs and distances

Table 3 shows the distances \( d' \) computed on the basis of \( P_{EER} \) obtained in the pair-wise emotion discrimination experiments. After closer inspection of Table 3, we can observe that Ne and Fe have some close neighbors, while Sa does not have any (relatively) close neighbors. This is in agreement with the observations made from Fig. 1. In addition, we observe in Table 3 small distances between An and Jo, and Ne and Bo which indicate easy confusability between these pairs of emotions. We can visualize the \( \binom{n}{2} = 21 \) numbers of distances, which are rather difficult to grasp in Table 3, in a two-dimensional plot by performing Multidimensional Scaling.

Table 3: Distances based on \( d' \) at EER

<table>
<thead>
<tr>
<th></th>
<th>An</th>
<th>Bo</th>
<th>Di</th>
<th>Fe</th>
<th>Jo</th>
<th>Ne</th>
<th>Sa</th>
</tr>
</thead>
<tbody>
<tr>
<td>An</td>
<td></td>
<td>3.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bo</td>
<td>3.07</td>
<td></td>
<td>2.07</td>
<td>1.88</td>
<td>0.77</td>
<td>2.29</td>
<td>4.29</td>
</tr>
<tr>
<td>Di</td>
<td>2.07</td>
<td>2.22</td>
<td></td>
<td>1.92</td>
<td>2.43</td>
<td>0.74</td>
<td>1.87</td>
</tr>
<tr>
<td>Fe</td>
<td>1.88</td>
<td>1.29</td>
<td>1.92</td>
<td></td>
<td>1.58</td>
<td>1.74</td>
<td>2.44</td>
</tr>
<tr>
<td>Jo</td>
<td>0.77</td>
<td>2.43</td>
<td>1.58</td>
<td>1.58</td>
<td></td>
<td>0.48</td>
<td>2.12</td>
</tr>
<tr>
<td>Ne</td>
<td>2.29</td>
<td>0.74</td>
<td>1.74</td>
<td>0.48</td>
<td>0.48</td>
<td></td>
<td>4.33</td>
</tr>
<tr>
<td>Sa</td>
<td>4.29</td>
<td>1.87</td>
<td>2.44</td>
<td>2.12</td>
<td>2.12</td>
<td>4.33</td>
<td></td>
</tr>
</tbody>
</table>

6.2. Visualizing acoustic similarities between emotions using Multidimensional Scaling

We used Multidimensional Scaling \[9\] to visualize the distances as shown in Table 3. Multidimensional Scaling (MDS) is a statistical technique that can visualize distance-like data in a low-dimensional geometric picture. In non-metric MDS, the goal is to minimize the differences between the reproduced distances \( d_{ij} \) in the map and a monotonic transformation of the input distance data \( f(d_{ij}) \) (see Eq. 2). The method uses the relative orderings of the given distances \( d'_{ij} \) (hence the “non-metric”) to construct the metric structure of the input data.

\[
S^2 = \frac{\sum_{i\neq j} f(d_{ij}) - d_{ij}}{\sum_{i\neq j} d_{ij}^2}, \quad (2)
\]

where \( S \) is a stress measure. In Fig. 2, we can observe our non-metric MDS analysis performed with the distance data in Table 3. In this plot, emotions that were difficult to discriminate from each other according to the spectral classifiers are closer to each other, while emotions that were easy to discriminate from each other according to the spectral classifiers are further apart. The advantage of such a visualization is that it in fact summarizes both Fig. 1 and Table 3. We can now actually see what types of confusions are made and how low EERs are related to these confusions. Thus, Fe has a high EER in Fig. 1 because it is often confused with its relatively close neighbors Ne, Bo and Di, while Sa is relatively easy to detect since it has no close neighbors. Note that an artefact of this method is that if all emotion detectors have equal EERs, the acoustic map would be a circle.

As an indication of how good the fit is of the MDS analysis on the input data, we can use the stress measure \( S \) in Eq. 2 to reflect the goodness-of-fit, where a lower stress value means a better fit. Fig. 3 shows that this stress measure rapidly drops with increasing MDS-dimension.

7. Discussion and Conclusions

Note that our visual analysis method to assess discriminability and confusability of emotions is based on a speech database containing acted, extreme emotions which were recorded un-
under ideal conditions. We can expect that when working with more spontaneous emotional speech material, which can consist of less extreme emotions, the discrimination and detection performance will be somewhat worse. In view of our map of emotions, this would lead to smaller distances, resulting in a smaller emotional space.

A striking similarity between two emotions according to our spectral classifiers is the one between Jo and An. We, as humans can make a clear distinction between these two emotions: we tend to place An and Jo at the opposite extremes of the valence scale (negative - positive) in an arousal-valence space, see Fig. 4 [10]. Note that the landmarks in Fig. 4 were established by subjects rating verbal emotion words rather than audiovisual stimuli. Although the dimensions and the rotation of the dimensions in Fig. 2 are arbitrary, we could interpret Dim 2 as an arousal scale (active - passive) since Jo and An are on the opposite side of Sa. The valence scale is less evident, partly because the database contains only one positive emotion Jo and because Jo seems to be difficult to discriminate from its opposite An. Other studies have also reported difficulties in discriminating between An and Jo [2] acoustically. Or more generally, it seems to be difficult to discriminate between negative and positive emotions, while discrimination between active and passive emotions (arousal scale) is relatively easy. Studies have found much stronger acoustic correlates for the arousal dimension than the valence dimension [2, 11].

Future research should investigate how automatic discrimination on the valence dimension can be improved by fusing different levels of information. De Silva et al. [12] suggested that some emotions have a “dominant modality”: some emotions are better recognized by humans in the visual domain than the auditory domain, e.g., anger, happiness and surprise are visually dominant emotions. So there is a chance that valence is better expressed and recognized through the visual channel while arousal is better expressed and recognized through the audio channel. This is one of the reasons why multimodal emotion recognition is increasingly investigated. Further, in the current study, we have used spectral features to train our classifiers. A fusion between spectral and prosodic features, e.g., $F_0$, intensity, speech rate etc., could improve detection performance.

To summarize, we showed how we derived an acoustic map of emotions and how we can interpret detection performances and confusability of emotions in this map. Sa has the lowest EER (see Fig. 1) which causes the MDS to position Sa without any close neighbors (see Fig. 2). On the other hand, Fe has the highest EER which causes the MDS to position it centrally with Ne, Bo and Di as relatively close neighbors. Thus, our acoustic map of emotions can give readily insight in 1) the detection and discrimination performance of the acoustic emotion classifiers and 2) the confusions between emotions that the acoustic emotion classifiers are most likely to make.

8. Acknowledgements

This research is supported by the Dutch BSIK-project MultimediaN http://www.multimedian.nl.

9. References