Emotional thin-slicing: a proposal for a short- and long-term division of emotional speech

Daniel Oliveira Peres¹, Dominic Watt², Waldemar Ferreira Netto¹

¹University of São Paulo, Brazil
²University of York, United Kingdom

doliveiraperes@gmail.com, dominic.watt@york.ac.uk, waldemar.ferreiranetto@gmail.com

Abstract

Human listeners are adept at successfully recovering linguistically- and socially-relevant information from very brief utterances. Studies using the ‘thin-slicing’ approach show that accurate judgments of the speaker’s emotional state can be made from minimal quantities of speech. The present experiment tested the performance of listeners exposed to thin-sliced samples of spoken Brazilian Portuguese selected to exemplify four emotions (anger, fear, sadness, happiness). Rather than attaching verbal labels to the audio samples, participants were asked to pair the excerpts with averaged facial images illustrating the four emotion categories. Half of the listeners were native speakers of Brazilian Portuguese, while the others were native English speakers who knew no Portuguese. Both groups of participants were found to be accurate and consistent in assigning the audio samples to the expected emotion category, but some emotions were more reliably identified than others. Fear was misidentified most frequently. We conclude that the phonetic cues to speakers’ emotional states are sufficiently salient and differentiated that listeners need only a few syllables upon which to base judgments, and that as a species we owe our perceptual sensitivity in this area to the survival value of being able to make rapid decisions concerning the psychological states of others.

Index Terms: speech perception; acoustic phonetics; intonation; emotion

1. Introduction

This study investigates the perception of emotional speech by native and non-native speakers by using the thin-slicing method [1, 2, 3, 4]. Previous studies of normal and delexicalized speech using a variety of ‘basic emotions’ and ‘dimensional’ approaches have yielded diverse results. Differences between the types of experimental stimuli used, in terms of the amount of information conveyed by the speech signal (lexis, intonation, voice quality, etc.), have been a source of variability in the findings [5, 6, 7, 8]. Understandably, the different methodological and theoretical approaches taken have likewise affected the findings. In the light of the broad spectrum of results from earlier studies, we propose an experimental design that reduces the problems associated with labeling, e.g. the labels for basic emotion names and dimensions, and with the kinds of stimuli, e.g. low-pass filtered speech and other types of content-masked speech. On the basis of the results of the present experiment, we argue in favor of a short- and long-term division of emotional speech. This division can be a useful tool in understanding how emotion is conveyed by speech.

1.2 Different approaches, different results

Studies of emotional speech are highly diverse in nature. For this reason, significant variation among theoretical viewpoints, methodologies, and types of stimuli is to be expected. The ‘basic emotions’ and ‘dimensional’ approaches have been the most frequently followed and have been undeniably fruitful, but as ways to analyze emotional speech they are in some sense antagonistic towards one another. Moreover, they are each characterized by considerable internal variability. The number of basic emotions proposed varies significantly not just in quantity, but also in the way in which they have been chosen by the researchers. [9], for example, postulated six basic emotions based on facial expressions (anger, fear, sadness, happiness, disgust and surprise); [10] proposed eleven emotions based on readiness for action (anger, aversion, courage, dejection, desire, despair, fear, hatred, hope, love, and sadness); [11] argued for four basic emotions based on body aspects (fear, pain, love, and anger). Inconsistency in how emotions are labeled inevitably influences the responses given by the participants in perception experiments such as those using a multi-alternative forced choice (mAFC) design. The dimensional perspectives also present two problems: the number of dimensions required, and the labels for each dimension. The number of dimensions varies markedly across prior studies. [12] proposed three dimensions (pleasantness-unpleasantness, rest-activation and relaxation-attention), while [13, 14] postulated two, one with two dimensions (pleasantness-unpleasantness and attention-rejection) and a second with three dimensions (the two previous ones, plus activation). The required labels for the extreme points of each dimension may also vary. Activation, for example, can be a continuum between calm, sleep or rest (low activation) and arousal, tension or excitement (high activation).

Another source of variation is the type of stimuli used in perception experiments: real versus acted emotional speech, for instance, or normal versus delexicalized speech. Among these categories there is variability in terms of how the emotions are elicited (e.g. scenarios for acted speech, as in [6]; use of MIP – the mood induction procedure – as in [15]; and spontaneous contexts for authentic emotional speech). The same applies to the numerous delexicalization methods devised to date, e.g. randomized splicing, reversed speech, or PURR (Prosody Unveiling through Restricted Representation) [16]. By way of illustration, [8] analyzed a set of stimuli (normal versus PURR-delexicalized speech) by using basic emotions and dimensional approaches. The results showed that basic and dimensional
approaches as well as normal and delexicalized speech affected the performance of native and non-native participants in perception experiments. The problems discussed above can be summarized as follows: (1) the amount of information available in the stimuli and its impact on participants’ responses; (2) the influence of the terminology used in the experiment (specifically, the labels given to emotions and dimensions).

1.3 Two solutions

Two reasonable solutions would be: (i) to provide content-masked speech that preserves the acoustic information in the stimulus, i.e. that does not remove phonetic information from the acoustic signal; (ii) to provide a judgment tool that does not involve linguistic information, such as names, labels, etc. The delexicalization methods mentioned earlier remove important features of the speech signal such as the voice quality (PURR) and intonation contour (randomized splicing and reversed speech), and can sound highly unnatural to listeners. For these reasons, we need a method that completely or partially removes linguistic information without any damage to the acoustic signal. Thus, the proposed solution to (i) is to make use of the thin-slicing method; for (ii), it is to deploy the set of averaged facial expressions of emotions, AKDEF [17]. Both are discussed in more detail below.

1.4 Thin-slicing

Studies of perceptual accuracy have shown that subjects are consistently able to recover information from the environment even under limiting circumstances. Research in social psychology has demonstrated that minimizing the amount of information participants are given to work with does not prevent them from making satisfactory judgments about situations based on first impressions. The term thin-slice is first attested in [3], in which it is defined as a brief excerpt of expressive behavior extracted from a behavioral continuum. [1] defines it as any excerpt containing dynamic information of less than five minutes’ duration. According to [1], thin-slices retain much, if not most, of the information contained in the communicative action, even after reducing or sometimes eliminating (i) information encoded within the verbal stream; (ii) the historical context; (iii) the global context in which the behavior is occurring. Thin-slicing is, in other words, a way of isolating stimuli from the general circumstances in which they were produced. This delimitation thus obliges the participant to focus on phonetic cues, without the influence of the speech context and verbal information.

One central question about this method is often raised at this point: is it possible that such restricted information can convey the necessary content to listeners? Can excerpts as short as 400ms, say, be properly perceived? These questions can be answered by looking to the acoustic-perceptual, cognitive, psychological, and neurological evidence. We can, for instance, cite differential thresholds for isolated non-speech sounds with durations of approximately 40 and 250ms [18]. We also find that speech can be stripped of considerable amounts of content and still be intelligible [19], and that different basic emotions can be distinguished in neutrally-worded sentences just 200ms after sentence onset (P200) [20].

1.2 AKDEF

The Averaged Karolinska Directed Emotional Faces (AKDEF) database is a set of 70 pictures of averaged human facial expressions. The material was originally compiled for psychological and medical research, in particular for perception, attention, and emotion experiments. AKDEF is derived from the much larger KDEF (Karolinska Directed Emotional Faces) database [21], which is comprised of 4,900 pictures of human facial expressions of emotion. AKDEF contains 70 images of 35 males and 35 females, each displaying 7 different emotional expressions (angry, afraid, sad, happy, disgusted, surprised, and neutral). The subjects were photographed (twice) from 5 different angles. These images were converted from 32-bit RGB color to 8-bit greyscale. After the conversion, the images were ‘stacked’ (70 different stacks of 7 emotions). For example, two sets of 35 happy male subjects were placed in the same stack. From these stacks, the averages of the images were calculated, keeping the characteristics common to all the (same-sex) actors’ facial images for that emotion (see Figure 1). AKDEF is a creative solution that diminishes the individual influence of each actor while s/he performed the emotions. Since there was a high degree of variation among the actors in terms of how they performed the emotions, AKDEF is the solution best suited to our current purposes. For this study, the emotions chosen were anger, fear, sadness and happiness.

2. Methodology

2.1 Data collection

The audio data for this study were collected from YouTube (www.youtube.com). 48 thin-slices were selected from excerpts of spontaneous emotional speech in Brazilian Portuguese. The excerpts were divided into four basic emotions: anger, fear, sadness and happiness (i.e. 12 excerpts per emotion). They had previously been evaluated in a pilot experiment to verify the agreement between the experimenter (the first author) and lay listeners. The degree of agreement among the Brazilian subjects was from satisfactory to good (K = 0.77, p = 0) while for non-natives, the agreement was satisfactory (K = 0.43, p = 0) [22]. The duration range of the thin-slices was 220-1,369ms.

2.2 Experimental design

76 participants took part in the perception experiment, which was administered online. 38 were natives (i.e. Brazilians) and 38 were non-natives (English, American and Canadian people knowing no Portuguese), giving a total of 3,648 responses. Of the native and non-native participants respectively, 21 and 23 were female; the sex imbalance arose because participants were self-selected volunteers. The effect of participant sex is discussed further below.

The participants’ task was to listen to the stimuli and to relate them to the AKDEF images presented in groups of four on the screen. The listeners were allowed to listen to the excerpts as many times as they thought necessary. Number of lists per excerpt was included as a variable in the analysis. Before starting, participants were given two training examples to familiarize themselves with the experimental format. Each trial screen of the experiment presented a participant with a thin-slice to be related to one of the AKDEF images. It is important

---

1 For this study, we considered only the frontal angle.
to note that, owing to the forced-choice design, participants could not continue undertaking the experiment without clicking on one of the four images (cf. the tick at the top left-hand corner of the happy face in Figure 1). In addition to the figures, the 'not sure/em dúvida' alternative was available in case participants could not decide which image best corresponded to the auditory stimulus.

![Figure 1: Example stimulus screen, showing four AKDEF images (from left to right: angry, sad, happy, fearful).](image)

The auditory stimuli were presented in random order, as were the left-to-right sequences of the facial images. The 'not sure/em dúvida' button always remained in the same position on the screen.

2.3 Data analysis

The data analysis was done in three steps: (i) descriptive analysis; (ii) probabilistic analysis; and (iii) an analysis using d-prime ($d'$) [23]. In order to measure participant sensitivity, Detection Theory analyzes hits (H), i.e. correct judgments, while taking into account false alarms (FA). Via these two measures we obtain the value $d'$, which is computed by subtracting the normal inverse of the FA distribution from the normal inverse of H, i.e. $d' = z(H) - z(FA)$.

3. Results

The number of thin-slice stimuli judged for each emotion was 912 (76 participants x 12 stimuli per emotion, subdivided equally between natives and non-natives (i.e. 456 for each group of 38 participants)). The proportion of correct answers was high for both native and non-native groups. For native and non-native participants, the percentage of correct answers was respectively: anger – 71% and 78%; fear – 41% and 42%; sadness – 66% and 78%; happiness – 77% and 65%.

Table 1 shows that the mean and the median of each emotion have very close values, which suggests a symmetric distribution of data. The asymmetry (skewness) values are close to zero. The coefficient of variation is highest for fear, for both participant groups, demonstrating that the fear data are more heterogeneous than those for the other emotions. This pattern is also evident in Figure 2, which shows the pooled scores for the native and non-native groups.

The off-diagonal (i.e. incorrect response) cells in the fear rows and columns in Tables 2 and 3 reveal that these scores are generally relatively high compared to the corresponding values for the other three emotions, which were overall more homogeneous. For both groups of participants, anger yielded virtually the same number of responses in total across the four emotions (487 and 486 for natives and non-natives, respectively). However, unlike fear, most responses for anger – natives 324 (71%) and non-natives 354 (78%) – were correct. Sadness and happiness showed similar results to anger.

Two one-way ANOVAs were performed to examine the effects of stimulus duration ($F(1,180) = 0.57, p = 0.45$) and listener sex ($F(1, 180) = 0.10, p = 0.99$). In addition, we ran two Kruskal-Wallis tests for participant group (natives versus non-natives) and number of listens per stimulus ($p > 0.05$). None of the tests showed significant results.

The $d'$ indices were calculated for every participant’s judgments of each of the four emotions. The greater the value of $d'$, the higher the listener’s degree of accuracy. Differences in listener sensitivity are reflected in the greyscale tone of each cell in Table 4, whereby larger $d'$ values are assigned darker tones.

![Figure 2: Emotion score distributions for native and non-native participant groups pooled.](image)

Table 1: Descriptive statistics of the data.

<table>
<thead>
<tr>
<th></th>
<th>Non-natives</th>
<th>Natives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Anger</td>
<td>Fear</td>
</tr>
<tr>
<td>Mean</td>
<td>28.3</td>
<td>10.0</td>
</tr>
<tr>
<td>Median</td>
<td>29.3</td>
<td>10.6</td>
</tr>
<tr>
<td>Standard Error</td>
<td>1.6</td>
<td>1.5</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>5.9</td>
<td>6.3</td>
</tr>
<tr>
<td>Coefficient of variation</td>
<td>0.08</td>
<td>0.1</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-0.6</td>
<td>-0.3</td>
</tr>
</tbody>
</table>

Table 2: Confusion matrix: natives.

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Anger</th>
<th>Fear</th>
<th>Sadness</th>
<th>Happiness</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>354</td>
<td>324</td>
<td>56</td>
<td>10</td>
<td>456</td>
</tr>
<tr>
<td>Fear</td>
<td>88</td>
<td>181</td>
<td>62</td>
<td>40</td>
<td>385</td>
</tr>
<tr>
<td>Sadness</td>
<td>11</td>
<td>29</td>
<td>136</td>
<td>67</td>
<td>228</td>
</tr>
<tr>
<td>Happiness</td>
<td>35</td>
<td>25</td>
<td>77</td>
<td>10</td>
<td>137</td>
</tr>
<tr>
<td>Total</td>
<td>486</td>
<td>288</td>
<td>394</td>
<td>118</td>
<td>1204</td>
</tr>
</tbody>
</table>

Table 3: Confusion matrix: non-natives.

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Anger</th>
<th>Fear</th>
<th>Sadness</th>
<th>Happiness</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>251</td>
<td>354</td>
<td>10</td>
<td>14</td>
<td>455</td>
</tr>
<tr>
<td>Fear</td>
<td>91</td>
<td>176</td>
<td>62</td>
<td>40</td>
<td>375</td>
</tr>
<tr>
<td>Sadness</td>
<td>11</td>
<td>29</td>
<td>136</td>
<td>67</td>
<td>228</td>
</tr>
<tr>
<td>Happiness</td>
<td>35</td>
<td>25</td>
<td>77</td>
<td>10</td>
<td>137</td>
</tr>
<tr>
<td>Total</td>
<td>486</td>
<td>288</td>
<td>394</td>
<td>118</td>
<td>1204</td>
</tr>
</tbody>
</table>

A summary of the results of the $d'$ analysis is provided in Figure 3.
As previously mentioned, speech contains more information than is strictly necessary for successful perception to take place; the image of ‘belt and braces’ is often deployed as an analogy here. Likewise, as speech is plainly one of the principal means by which human emotions are conveyed, only compelling counterevidence would force a revision of our view that emotional speech should be any exception to the redundancy generalization. This hypothesis makes sense if we adopt an evolutionary viewpoint [24], whereby emotions are adaptations which play an important survival role. Redundancy would help interactants to express emotions more effectively. Relatedly, our ability to accurately recognize individual speakers based on informationally-impoverished stimuli (i.e. thin-slices) also supports our contention that the speed with which we make decisions in the context of the perception of emotional speech could be an important factor in survival terms.

To understand the role played by redundancy of information in the perception of emotion, it is also necessary to understand redundancy as a principle that works on two levels: short- and long-term. By long-term, we mean all information that relates to speech dynamics in the temporal domain. As an example, intonation – or more precisely macro-intonation – is a long-term factor that was shown to be an important influence on the results obtained by [5, 6, 7]. Short-term information, by contrast, denotes variations that do not take significant amounts of time, such as voice quality and micro-intonation. Why is this subdivision necessary? Redundancy in the information available to listeners should follow such a division in order to help attain communicative goals. In other words, it must be possible for the emotion to be appropriately decoded by the listener even when important features of the speech signal are absent. Because of this, limitation of information can occur only at one level, either short- or long-term, but should not occur at both simultaneously. The experiments using delexicalized speech, specifically PURR or low-pass filters, limited listeners’ access to short-term features such as voice quality, but left F0 and rhythmic characteristics intact. On the basis of this information, native-speaking participants were able to judge the stimuli with considerable success [6, 7]. On the other hand, thin-slice-based experiments remove long-term features. This does not mean that there is no intonational variation whatever in the stimuli, but that the intonational features of the utterances from which they are extracted is incomplete or underrepresented. In the present case, both groups of participants – native and non-native – were able to judge the stimuli accurately. Thus, intonation as a long-term phenomenon (i.e. macro-intonation) can be understood as a sufficient, but not necessary, characteristic for the perception of emotion. However, in some contexts – e.g. a delexicalized speech experiment – intonation will be a sufficient condition to permit accurate perception. Voice quality could be considered as a sufficient feature because it enables the recognition of emotional thin-slices, but is not necessary, since (with the possible exception of creak and whisper) it is absent from delexicalized stimuli. Testing the predictions of this redundancy model involves manipulation of short- and long-term features to test the limitations they impose upon emotional speech perception. Further research will help to tease these factors apart.

4. Discussion

In general, participants showed a high degree of sensitivity – the tones shown in Table 4 are mostly towards the darker end of the spectrum selected for this graphic – though some low values were found. In the case of fear, the d’ values were not as high as those for the other emotions (Figure 3). The false alarm (FA) value is also an important indicator of listeners’ tendencies to misidentify emotions. For the natives, for example, the errors for anger were distributed as follows: e.g. 67.5% fear, 19.7% sadness, and 12.8% happiness. For fear, they were 46% anger, 30.6% sadness, and 23.4% happiness. For sadness, the figures were 14.3% anger, 50.5% fear, and 35.2% happiness. Lastly, the errors for happiness were 20% anger, 41% fear, and 39% sadness. The results of this experiment confirm that the participants had significant sensitivity to short-duration stimuli, i.e. thin-slices. The findings contribute to the debate about strategies that subjects use when perceiving emotional speech, even when only sparse linguistic information is available.

5. Acknowledgements

This work was supported by CAPES (the Brazilian Federal Agency for Support and Evaluation of Graduate Education). Process: 99999.007276/2014-01.
6. References


