Acoustic correlates of frustration in spontaneous speech

Mimmi Forsell[1], Kjell Elenius[1] and Petri Laukka[2]
[1]Speech, Music and Hearing, CSC, KTH, Stockholm
[2]Department of Psychology, Uppsala University, Uppsala

Abstract

The focus of this master's thesis by the first author was to investigate the acoustic attributes of frustration in spontaneous speech. The speech material was recorded from real life Swedish telephone services by the company Voice Provider. The utterances were selected speaker by speaker in order to have at least one of them judged as emotionally neutral by a listener group, while the other utterances of the same speaker were judged as displaying emotional speech. Due to the nature of the speech material most of it was spoken in a neutral way. However, some percent of the utterances displayed various degrees of frustration, mostly anger but also some despondency, and these were the emotions studied in this report. We also studied the emotional intensity of the utterances. Acoustic cues of the emotional speech were compared to those of neutral speech for the same speaker. We found some significant differences between the acoustic cues for neutral and emotional speech. Anger was characterized by a rise of fundamental frequency and an increase in speech amplitude, whereas despondency reduced the syllable rate significantly. The emotional intensity raised the pitch, increased the amplitude and decreased the syllable rate. Correlations were also found between perceived emotions and acoustic speech parameters.

1. Introduction

Recognition of emotions in speech has gathered an increasing interest in the research community over the last years (Juslin & Laukka, 2003; Scherer, 2003). It is a difficult task further complicated by the fact that there is no unambiguous answer to what the "correct" emotion is for a given speech sample.

The vocal emotions explored may have been induced or acted or they may have been elicited from real life contexts. Spontaneous speech from actual telephone services could be counted as "real". The motivation of the latter is often to try to enhance the performance of human-machine interaction systems, such as voice controlled telephone services (e.g., Petrushin, 1999).

Real life speech can have a varying quality and the emotions expressed can be complex (Cowie & Cornelius, 2003; Vidrascu & Devillers, 2005; Wilting et al., 2006). It is important to bear in mind that when using real spontaneous speech uncertainties and imperfections regarding the recordings may affect the outcome of the results. Another difficulty with real emotions is in their labelling, since the actual emotion of the speaker is almost impossible to know with certainty. Furthermore, emotions occurring in spontaneous speech seem to be more difficult to recognize compared to acted speech (Batliner et al., 2003).

2. Method

Bellow follows an explanation of the corpus analysed, a description of the acoustic parameters used and how they were extracted. Finally the listening tests and the analysis of the material are described in detail.

2.1 Speech material

In this paper we investigate the acoustic cues of the emotions of real life speech as spoken to an automatic speech recognition system using various automated services, such as getting information regarding airline traffic and ordering merchandise. This limits the emotions displayed to mostly anger and despondency due to frustration with the service. Other emotions are rare and have therefore not been included.

The utterances were recorded from the automated voice controlled telephone services of the company Voice Provider. A total of 61,078

utterances were labelled by an experienced, senior voice researcher into neutral, emphasized or negative (frustrated) speech. Only 2.21 % were labelled as negative and 1.73 % as emphatic. Out of these the senior researcher and the authors in collaboration selected speakers that had at least one utterance labelled into each of the three categories above. The underlying idea was that utterances with different emotions from the same speaker would facilitate the search for acoustic cues. Due to the emotions chosen to be used in this study, see below, the negative utterances were divided into angry and despondent and utterances from speakers that had one despondent and one neutral utterance were added. Recordings with bad acoustic quality were removed resulting in a total of 268 stimuli. This number was later reduced to 200, since this was considered to be a maximum for the listening tests. The utterances finally chosen came from 64 different speakers.

Parts of our speech material have earlier been used in experiments regarding automatic detection of emotions from the speech signal (see Neiberg et al., 2006).

2.2 Choosing and extracting acoustic cues

The acoustic information chosen for our experiments were: minimum, maximum, standard deviation and median of the amplitude and fundamental frequency, the syllable rate, and the frequency and bandwidth of the first three formants. Praat scripts were used to extract information of the fundamental frequency, amplitude and formants (Boersma et al., 2007). The syllable rate was manually measured using Praat.

The median is considered to be more reliable than the mean when analyzing speech. The median was also used when calculating a discrete value of a perceived emotion from the listening scores.

2.3 Listening experiments

The listening tests were conducted individually using custom software. We decided to ask for the following four categories: anger/irritation (Swedish: ilska/irritation), grief/despondence (sorg/uppgivenhet), neutral (neutral) and level of emotional intensity (känslans intensitet). The subjects were asked to rate the perceived level of each different category on a scale from 0 to 7. During two weeks 20 subjects performed the test. The 200 stimuli were presented in random order.

2.4 Analyzing the results

After making sure the results from the listening test were reliable, using Cronbach's alpha, every utterance was assigned an emotion (anger, despondent or neutral) and a perceived emotional intensity from the listener scores.

The acoustic values extracted with Praat were analyzed speaker dependently. Thus the acoustic values for the different utterances of one speaker were only compared to each other. We compared the following pairs: 29 angry to neutral, 13 despondent to neutral, and 50 high to low emotional intensity. For each pair a number of different acoustic values of the fundamental frequency, amplitude, syllable rate and the first three formants were analyzed.

The *paired samples t-test* was used to test whether there was a significant difference between the acoustic cues for the various emotional pairs or not. The correlations between acoustic cues and emotions were also calculated.

3. Results

Below we first give the significant results of the t-tests, presented according to the pair wise emotion comparisons as described above. After the t-tests we give the results of the correlation tests.

3.1 Angry versus neutral

The fundamental frequency of the voice was higher for the utterances perceived as angry. The t-value was $t(27) = 5.27$, $p < 0.0001$. Thus the difference between neutral and angry utterances was obvious. The increase in fundamental frequency can be seen in Figure 1.

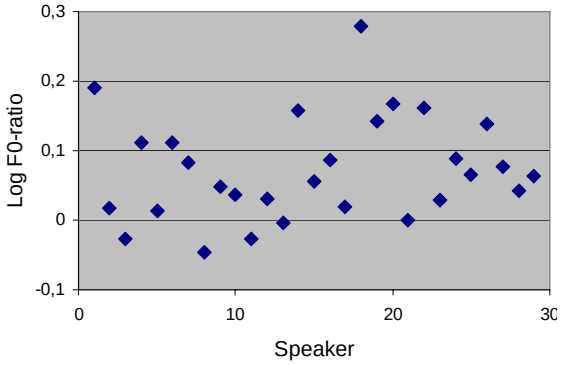

Figure 1. The logarithm of the ratio of the fundamental frequency for angry and neutral utterance pairs. Each pair represents one speaker.

The amplitude also appeared to change as the perception of anger changes. It increased for the angry utterances, $t(27) = 3.66$, $p < 0.001$.

Even though one could expect people to speak faster when being angry, this could not be statistically proven.

For the formant values only the bandwidth seemed to change with the perception of anger. The third and second formant bandwidths decreased as the utterances were perceived as more angry. For the second formant $t(27) = -2.43$, $p < 0.05$, and for the third $t(27) = -2.31$, $p < 0.05$.

3.2 Despondent versus neutral

No measures of the fundamental frequency showed any statistically relevant differences between the despondent and the neutral voice.

However, the standard deviation of the amplitude decreased as the utterances were perceived as more despondent. The t-value for the standard deviation was $t(11) = -2.59$, $p < 0.05$.

The syllable rate also decreased as an utterance got more despondent. The difference can be seen in Figure 2**.** As may be seen, only one of the speakers had a slower syllable rate for the neutral utterance $t(11) = -3.03$, $p < 0.05$.

No significant t-values were found for the formants.

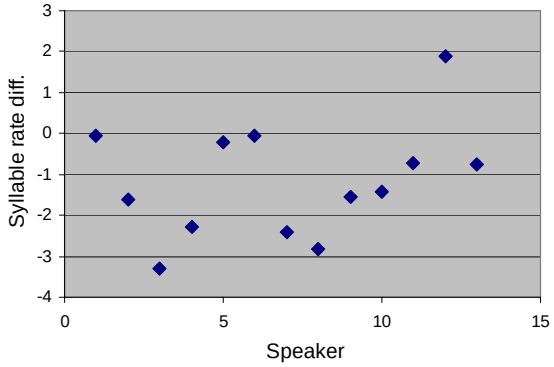

Figure 2. Syllable rate difference between despondent and neutral utterance pairs. Each pair represents one speaker.

3.3 High versus low emotional intensity

The fundamental frequency increased as the intensity rose. The t-value was $t(48) = 2.93$, $p < 0.005$. However, its standard deviation did not show any change.

The amplitude was expected to increase with the perceived intensity and this was also the case, $t(48) = 2.12$, $p < 0.05$.

The syllable rate decreased as the perception of intensity increased, $t(48) = -2.66$, $p < 0.05$.

These results show that increased emotional intensity is accompanied by higher pitch, increased speech amplitude and slower speaking rate.

The formant measures did not change significantly between high and low emotional intensity speech.

3.4 Correlations between acoustic cues and listener ratings

We also checked for correlations between the perceptual scores and the acoustic measures. All of them at the 0.05 significance level or lower are presented in the following.

The pitch compared to the degree of despondency for both the despondent/neutral, $r = -0.66$, $p < 0.05$, see Figure 3, and the high/low emotional pairs, $r = -0.51$, $p < 0.001$ gave a negative correlation.

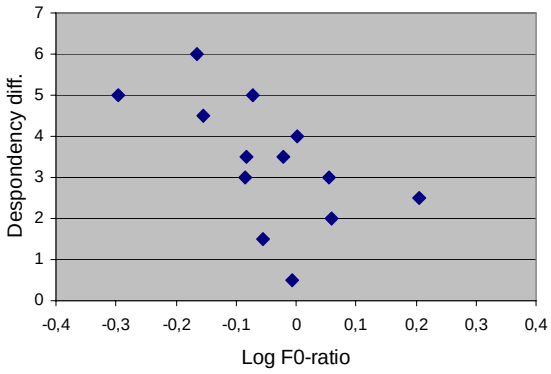

Figure 3. Perceived despondency difference between despondent and neutral utterance pairs as a function of the logarithm of the F0-ratio between the respective utterances. Each pair represents one speaker.

The pitch compared to the intensity showed a positive correlation for the angry/neutral pairs, $r = 0.38$, $p < 0.05$.

For the emotional high/low pairs the standard deviation of the amplitude had a negative correlation with the degree of despondency, $r = -0.49$, $p < 0.001$, and the amplitude had a positive correlation to the degree of emotional intensity, $r = 0.30$, $p < 0.05$.

The pitch correlated to the amplitude had a positive correlation for the high/low emotional pairs of utterances, $r = 0.39$, $p < 0.01$.

No correlation was found between the median level of amplitude and anger.

The strongest correlation of all, at the 0.001 significance level, was anger with the level of

emotional intensity for the angry/neutral, $r = 0.62$, $p < 0.001$ and the high/low emotional pairs, $r = 0{,}60$, $p < 0.001$. This positive correlation was also significant for the despondent/ neutral pairs, $r = 0.56$, $p < 0.05$.

4. Conclusion

To summarize, the results show that anger is visible in the rise of the fundamental frequency and the speech amplitude. Despondency results in decreased syllable rate and amplitude standard deviation. Emotional intensity increases the pitch and the speech amplitude, and it also decreases the syllable rate

Our results show that the expression of anger, despondency and level of emotional intensity is observable in the acoustic cues of real life spontaneous speech. These results are generally in line with the ones found in previous studies using acted speech (e.g., Juslin & Laukka, 2003), as well as spontaneous speech (e.g., Amir et al., 2003). The present study also introduces some novel aspects, e.g., investigating emotional intensity in spontaneous speech.

Acknowledgements

We want to thank Inger Karlsson for the emotion labelling of all the 61,078 utterances from Voice Provider. We also thank Jonas Lindh for his help with the Praat scripts used.

References

Amir N, Ziv S & Cohen R (2003). Characteristics of authentic anger in Hebrew speech. In: *Eurospeech-2003,* 713-716.

Batliner A, Fischer K, Hubera R, Spilker J & Nöth E (2003). How to find trouble in communication, *Speech Communication: 40,* 117–143.

Boersma P & Weenink D (2007). *Praat: doing phonetics by computer* (Version 4.5.14) [Computer program]. Retrieved from http://www.praat.org/

Cowie R & Cornelius RR (2003). Describing the emotional states that are expressed in speech. *Speech Communication: 40,* 5-32.

Juslin PN & Laukka P (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin: 129,* 770-814.

Neiberg D, Elenius K & Laskowski K (2006). Emotion recognition in spontaneous speech using GMMs. In: *Interspeech 2006,* 809-812.

Petrushin VA (1999). Emotion in speech: Recognition and application to call centers. In: *Proc. ANNIE ´99,* 7-10.

Scherer KR (2003). Vocal communication of emotion: A review of research paradigms, *Speech Communication: 40,* 227-256.

Vidrascu L & Devillers L (2005). Detection of real-life emotions in call centers. In: *Interspeech-2005,* 1841-1844.

Wilting J, Krahmer E & Swerts M (2006). Real vs. acted emotional speech. In: *Interspeech 2006.*