A Scale From Frowning To Smiling: Perception And Acoustics Of Short Stimuli

Klaus Härtl

Institute of Phonetics and Speech Communication, University of Munich
haertl@phonetik.uni-muenchen.de

Abstract

The work presented in this paper deals with the listener's ability to distinguish certain nuances in emotional speech. In the first part of this work a perception experiment is described, where stimuli with a duration of only 500 ms were presented to subjects. These stimuli were rated on a scale reaching from "frowning" to "smiling" in order to get a reference value for each presented stimulus.

Since the results of the perception experiment were very conclusive, acoustic analysis was performed to explore the acoustic correlates. In the second part the extraction of acoustic features is presented, which not only includes F_0, formants and related measurements but also voice source parameters derived from the estimation of a LF-model.

A high intra-speaker variation and also the inaccurate extraction of parameters make it difficult to find appropriate acoustic correlates.

1 INTRODUCTION

Research in the area of emotion and speech is tedious because of diversity in terms and definitions. In this study a hearer-oriented apporach is adopted, where it is secondary what the speaker felt at the time he was speaking, but what kind of impression he evokes in the listener [1]. This differentiates this work from speaker-oriented research, where knowledge about the speaker's emotional state and its psychological interpretation is necessary.

Judgements about emotional speech like amusement integrates visual, content and context information, but the listeners' ability to distinguish emotional stimuli from only the auditory channel has been shown in several investigations (e.g. [2], [3], [4]). Typically whole sentences were used as stimuli, so that the judgements of the subjects were also influenced by prosodic cues like intonation on word and phrase level. To minimise these effects stimuli of a length of only 500 ms were used.

Since classification into distinct categories is often insufficient a one-dimensional scale was used. A more real life example would be a speaker narrating a joke, starting off very seriously and earnest, of course overacted, and after a while he starts smiling and ends up in laughing. This investigation explores if listeners are able to distinguish the fine-grained categories in between.

Research before not always brought up the same results, when correlating acoustic measurements like F_0 and formants. Now voice quality is again proposed, since there is a known influence of affect on it [5]. For that purpose an estimation of LF-model parameters [6] is applied.

2 DATA

The speech material used in this investigation was produced by a female native speaker of German reading some 20 times the German story "Die Buttergeschichte" also known from the Kiel Corpus of Read Speech. The recording was taking place in a sound-proof cabin with a high quality condensator microphone Neumann TLM 103 placed in a distance of about 30 cm from her mouth. She was asked to read the story expressing different moods like frowning, anger, smiling and happiness and also in a neutral style. The main idea of this instruction was, to get a wide acoustic variance from negative to positive expressions, hopefully reflected in the acoustics.

The microphone signal was digitized to 48 kHz by a Yamaha O2R digital mixing console and was directly passed on to a computer's digital audio interface. For further processing the speech signal was downsampled to 24 kHz. From these sentences 108 stimuli with a length of 500 ms were randomly extracted. The only condition was that there should be no speaking pause in the stimuli.

On average a stimuli contained three syllables, but there were also stimuli that extended only over a small part of a two-syllabic word or included as much as four complete short words produced with a high speaking rate. For native speakers it was of course possible to recognise that the language in the stimuli was German but it was virtually impossible to retrieve the context. It is worth to mention that from its plot the story does not convey a particular emotion on its own, since it can be interpreted as a mellow fairy-tale as well as a depressing portrayal of German post-war food shortage.

ISBN 1-876346-48-5

3 PERCEPTION EXPERIMENT

3.1 Method

The aim of the perception test was to have the speech samples placed on a scale by the subjects. The interactive computer program used, `per_stim`, has been developed by H. R. Pfitzinger and is described in [7]. It was originally designed for similar tests concerning local speech rate. With the program stimuli can be placed in a 2-dimensional area.

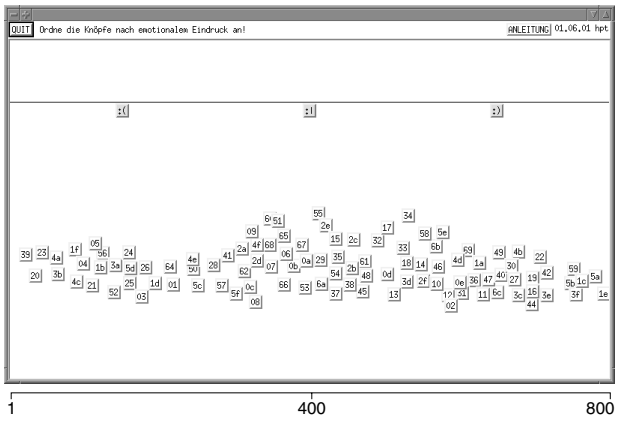

Figure 1: Example screen of the `per_stim` tool with the completed perception experiment.

Based on the results of a pilot test with four subjects three stimuli were selected which were used as anchor stimuli in the main test. There was one anchor each for "neutral", "frown" and "smile". The neutral anchor was placed in the middle of the scale, the others each in a distance of about 1/5 of the whole plane from the left and right border. This allows subjects to place stimuli beyond the anchors. The anchor stimuli should assist especially at the beginning of the test by providing orientation references. The anchors were labelled with the well known emoticons `:(`, `:|` and `:)` that are also commonly used in emails. Since the `per_stim` tool restricts the labels for each button to two ASCII characters, such simple indicators had to be used.

At the beginning of an experiment run, the remaining 105 stimuli appeared in the upper part of the window. The subjects should now place the stimuli in the lower part by a drag-and-drop of buttons that have only been labelled with a number. Each button represented one stimulus. The subjects could listen to the stimuli by clicking the button as often as they wanted, and were advised to change the position of the stimuli until they were satisfied with the relative location to other stimuli and to the anchor stimuli. There should then emerge an order based on perceived emotional colouring and multiple comparisons. Rearranging was allowed during the whole run of the experiment and as often as the subjects wanted. They were instructed to place those samples, where they could *hear* a smile more on the right and those where they could *hear* a frown more on the left. The use of the y-axis was left to the free choice of the subjects, e.g. for making an order according to the assumed reliability of the judgement.

At the very beginning of the test, the experiment is more like an identification test, because a single stimulus has to be placed in the appropriate region, by identifying similarities between anchors and the actual stimulus. Later the experiment becomes more and more a discrimination test, because subjects often listen to neighbours of already placed stimuli and only rearrange the order or the distance between them.

For most participants it took between 20 and 40 minutes to accomplish the test. A typical sample from the result is shown in fig. 1. Each stimulus was then assigned to a value between 1 and 800, according to its pixel position on the x-axis in the window of the computer screen. All subjects used a high quality headphone of the same brand and type (Beyerdynamic DT770).

3.2 Results

So far 19 subjects took part in the perception test. The subjects' age ranged between 23 and 54 years and most of the subjects are staff members or students at the phonetics department. Since the work described here is an ongoing study more subjects will be included in the analysis as soon as there are new volunteers.

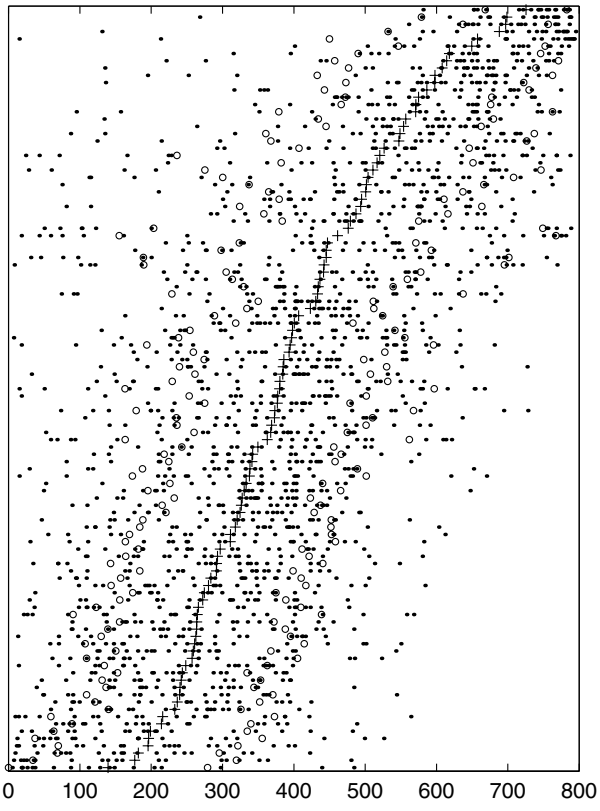

Figure 2: All results of the experiment, each dot is one rating, collected in one line per stimulus.
+: mean, ○: standard deviation

ISBN 1-876346-48-5

In fig. 2 all ratings for one stimulus are represented by dots in the same line. The cross marks the mean value of all ratings for this stimulus and the circles show the standard deviations. The stimuli are sorted by mean ratings, i.e. the stimulus, that in average was placed rightmost on the screen can be found at the top, and that one placed leftmost can be found at the bottom. It can be seen that the standard deviation is almost anywhere greater than 100 units, but there is also a significant trend without abrupt changes.

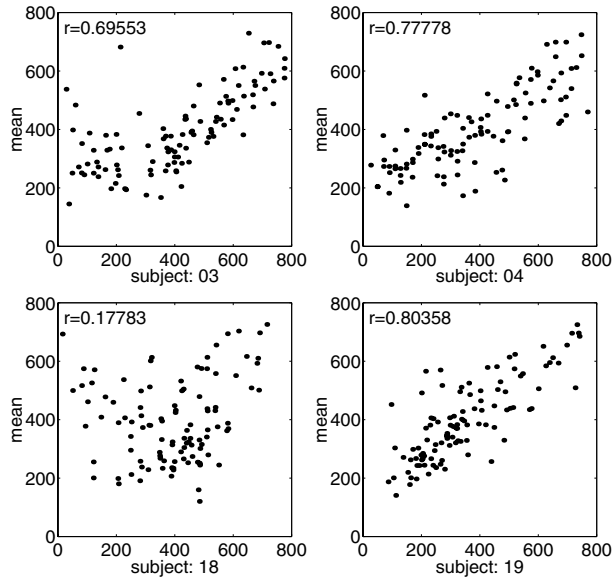

Figure 3: Typical subjects' scatter plots with correlation coefficients r of single subjects vs. mean of all other subjects.

The scatter plots in fig. 3 show examples for correlations between single subjects ratings (x-axis) and the mean rating of all other subjects (y-axis). A test for evaluating the correlation coefficients was performed by carrying out a t-test. Table 1 shows the correlation coefficient for each subject to the mean of all other subjects (with the rating of the particular subject excluded), the t value and its significance level: significant (*: $t_{17;0.05} > 2.11$), high significant (**: $t_{17;0.01} > 2.90$) or very high significant (***: $t_{17;0.001} > 3.96$).

Most subjects show very high significant correlations with the mean of all other subjects. It can be concluded that most listeners have a common concept about the used scale reaching from "frown" to "smile".

Three subjects have produced ratings which do not significantly correlate with the other subjects. It has to be noted that one of these is not phonetically trained and one is very shortly trained. The third subject was done in only ten minutes and mentioned that he had been halfhearted.

From now on the mean rating for each stimulus is taken as the reference value p of the stimulus on the frown-smile scale. The mean ratings were not transformed or normalized, e.g. to a -1 to +1 scale because it cannot be assumed that the presented stimuli cover really the ends of this scale. In addition, the units on this scale are arbitrary and do not bias the analysis.

4 ACOUSTIC ANALYSIS

4.1 Formants, F_0, Duration, Intensity

For further acoustic analysis the stimuli were manually segmented on the phone level. Duration and intensity values were computed for all segments. For voiced sounds, like vowels and nasals, values for mean F_0, its range and slope, and for jitter and the harmonics-to-noise ratio (HNR) were obtained. Additionally the values for the first five formants were calculated. All the acoustic features mentioned above were calculated with `PRAAT`, since the program offers standard techniques like autocorrelation for extracting F_0 and linear prediction coding for formant estimation.

In addition, parameters of the Liljencrants-Fant-Model (LF-Model) were estimated for voiced segments. The procedure is described in the following section, the reliability of its results will be discussed later.

4.2 LF-Model of Glottal Airflow

The Liljencrants-Fant-Model [6] has three parameters for describing the shape of one glottal pulse. The three parameters are R_a, the relative duration of the return phase, R_g, the relative opening branch time, and R_k, the symmetry between opening and closing phase. All three parameters are relative to F_0, which is an extra parameter. Together with the amplitude these parameters can be used for estimating the glottal flow of one pulse. The often used open quotient can be determined from R_k and R_g: $OQ = (1+R_k)/2R_g$. Estimating the parameters is an iterative process:

1. Calculation of the inverse filtered signal.
2. Inital estimation of the LF-model parameters from previous model or with reasonable values.
3. Calculation of the LF-model.
4. Determination of the distance between a single glottal pulse from the inverse filtered signal and the calculated model.
5. If result is not good enough, change of parameters and back to step 3.

Inverse filtering was performed by using LPC coefficents, that were also used for estimating formant frequencies. If the procedure is applied automatically, a substantial amount of ripples remain in the inverse filtered signal, which makes the following estimation inaccurate and sometimes even impossible. In step 2 the initial estimation of LF-model parameters was achievied by time-domain analysis, like detecting zero-crossings and maxima and minima in the inverse filtered signal. As distance measurement mean square error was used. The values for the exit condition in step 5

ISBN 1-876346-48-5

Table 1: List of all r and t values for each subject.

subject	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
r	0.34	0.71	0.70	0.78	0.73	0.52	0.82	0.83	0.66	0.75	0.81	0.16	0.77	0.76	0.73	0.72	0.63	0.18	0.80
t	1.49	4.11	3.99	5.10	4.46	2.48	5.93	6.21	3.60	4.73	5.74	0.67	4.97	4.89	4.36	4.26	3.39	0.75	5.57
level		***	***	***	***	*	***	***	**	***	***		***	***	***	***	**		***

influences mainly runtime, but should be chosen in a way, that there is no significant enhancement in the results anymore. Steps 2 to 5 were performed by an own computer-program written in C++, that produced parameters for over 5900 LF-models.

4.3 Results

All extracted acoustic values where correlated with the reference value p from the perception experiment. The highest r values of around 0.5 were brought up by correlating F_0 and p for the sounds [n], [l], [i:] and [œ]. For some sounds there is a negative correlation between duration and p and a slight positive correlation between intensity and p. Other correlations for the formants, F_0's slope and range, jitter and HNR do not show a common trend. In other research F_2 is often higher in smiled stimuli ([2]), but here this trend could not be observed.

The parameter values derived from the LF-model estimation are highly variable. Calculating the correlation coefficient of the parameter values and the reference value p taken from the perception experiment results in r values not higher than 0.05. It can be assumed that the automatic estimation process produces too many errors. This may be improved by better inverse filtering techniques or by exchanging the LF-model parameter estimation by using a frequency-domain approach for analysing voice source variations.

Overall, there are no clear trends of correlation between the derived acoustic features and quite reliable results from the perception experiment. There are again tendencies to higher F_0 and intensity, but overall no significantly conclusive results.

5 DISCUSSION AND CONCLUSION

In this study the promising results from a perception experiment were correlated with acoustics features describing the vocal tract and the voice source. A comparison of different studies on smiles shows that there are not always the same acoustic effects on formants ([2], [4]). Of course there is an influence on the results by the way the stimuli were produced.

For the LF-model paramter estimation a fully automatic procedure was chooen. This may result in too many ripples remaining in the inverse filtered signal and also in an inaccurate estimation of the model. Other investigation used semi-automatic estimation techniques, both for inverse filtering and for LF-model fitting [5]. Since in the presented study the shape of over 5000 glottal pulses had to be analysed, the manual interaction had to be kept low. So far the retrieved parameters from the LF-model are not treated as precise and methods have to be found to improve the estimation. Another way out is using a frequency-domain analysis approach for the voice source.

Speakers seem to have various strategies to express "frowning" or "smiling". Furthermore, a high intra-speaker variabilitiy can be assumed. High variation is also incorporated in the speech stimuli by mixing several moods. These fine-grained acoustic differences of such "second order emotion categories" [1] are perceptible even in short stimuli by human listeners. However, it seems to be based on a highly complex interaction between several acoustic voice parameters.

REFERENCES

[1] Roddy Cowie. Describing the emotional states expressed in speech. In *Proceedings of ISCA Workshop on Emotion and Speech 2000*, Belfast, 2000.

[2] Vivien C. Tartter and David Braun. Hearing smiles and frowns in normal and whisper registers. *JASA*, 96(4):2101–2107, 1994.

[3] Marc Schröder, Véronique Aubergé, and Marie-Agnès Cathiard. Can we hear smile? In *Proceedings of ICSLP 98*, 1998.

[4] Julie Robson and Janet MackenzieBeck. Hearing smiles — perceptual, acoustic and production aspects of labial spreading. In *Proceedings of ICPhS 99*, 1999.

[5] Christer Gobl and Ailbhe Ní Chasaide. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40:189–212, 2003.

[6] Gunnar Fant, Johan Liljencrants, and Qi-guang Lin. A four-parameter model of glottal flow. *STL-QPSR*, 4:1–13, 1985.

[7] Hartmut R. Pfitzinger. Local speech rate perception in german speech. In *Proceddings of ICPhS 1999*, volume 2, pages 893–896, August 1999.

ISBN 1-876346-48-5