

Unresolved Anger: Prosodic analysis and classification of speech from a therapeutic setting

Noam Amir¹, Hansjörg Mixdorff³, Ofer Amir¹, Daniel Rochman², Gary M. Diamond²,
Hartmut R. Pfitzinger⁴, Tami Levi-Isserlish¹, Shira Abramson¹

¹ Department of Communication Disorders, Tel Aviv University, Tel Aviv, Israel

² Psychology Department, Ben Gurion University, Beer Sheva, Israel

³ Department of Computer Sciences and Media, Beuth University of Applied Sciences, Berlin, Germany

⁴ Inst. of Phonetics and Digital Speech Processing, Christian-Albrechts-University Kiel, Germany

noama@post.tau.ac.il; mixdorff@bht-berlin.de

Abstract

This paper describes analyses of a corpus of speech recorded during psychotherapy. The therapy sessions were focused on addressing unresolved anger towards an attachment figure. Speech from the therapy sessions of 22 young adult females was initially recorded, from which 283 stimuli were extracted and submitted for evaluation of emotional content by 14 judges. The emotional content was rated on three scales: Activation, Valence and Dominance. A set of acoustic features was then extracted: statistic features, F0 features based on the Fujisaki model and perceptual speech rate features. The relationship between acoustics and emotional content was examined through correlation analysis and automatic classification. Results of the model-based analysis shows significant correlations between the strength and frequency of accents and Activation, as well between base F0 and dominance. Automatic classification showed that the acoustic features were better at predicting Activation rather than Valence and Dominance, and that the dominant features were those based on F0.

Index Terms: emotional speech, Fujisaki model, emotion classification.

1. Introduction

Emotion in speech has been studied in many contexts, and indeed one of the main conclusions from these studies is that results often depend to a great extent on the specific context. From acted emotional speech, through elicited emotional speech, to truly natural expression of emotions, the variety of emotional expression and styles of emotional expression vary greatly [1, 2].

Many efforts have been made at obtaining the most natural expression of emotion possible, though the associated problems of emotional sparsity, and loss of control over the textual content, among other shortcomings, make this data more difficult to analyze. Nevertheless this is arguably the most interesting and challenging type of data to study.

In this paper we discuss the analysis of subjective judgments and acoustic features of a corpus of emotional speech recorded during sessions of psychotherapy [3]. The structure of these therapy sessions was designed to specifically address unresolved anger experienced by the participants towards an attachment figure, such as a parent or spouse. The sessions were composed of three successive interventions, described below, lasting up to ten minutes each. In comparison to Wizard of Oz (WoZ) type emotion elicitation techniques

[e.g. 4] or call center type data [5] this technique results in a large amount of emotionally charged speech, biased mainly towards negative emotions.

There are many different ways in which an analysis of this type of corpus can be performed. For example, the initial analysis performed by Rochman et al. [3] involved a single coder who examined segments of approximately 45 seconds in length. The authors examined the relationships between several basic acoustic parameters calculated over such segments.

The present study presents an initial attempt at a more fine-grained analysis of the corpus: in segmenting it, judging it, and in analyzing the acoustic variables. The analysis reported here includes several stages:

- Chunking the corpus into small utterances that can be judged to represent one “emotion”,
- Judging the resultant collection of utterances subjectively,
- Extracting acoustic features on several levels and comparing their relationship with the judgments through statistical analysis and automatic classification.

In the following sections we detail the methods and the main results.

2. Methods

Analysis of an emotional speech corpus requires many different decisions at the different stages outlined in the previous section. It is beyond the scope of this paper to debate all the issues that can arise, therefore the reader should keep in mind that this work describes one specific approach, but does not argue for the exclusion of other possibilities. This study is preliminary in nature, being the first in-depth acoustical study of this corpus, and further research is being undertaken.

2.1. Speech data

The complete and detailed description of the speech recording procedure appears in the paper by Rochman [3]. It is presented here in summarized form.

Twenty two women aged 21 to 25 volunteered to participate in the above study, all of them university students at the time. The participants reported experiencing unresolved anger towards an attachment figure and therefore volunteered to participate in a session of psychotherapy. The session was composed of three distinct interventions: (1) *baseline*, in which they discussed ordinary day to day topics; (2) *relation reframe*, in which they discussed the relationship between themselves and the attachment figure; (3) *empty chair*, in

which they expressed their emotions as if the attachment figure was seated in the room with them. Each intervention lasted between 4 and 10 minutes. The first intervention was expected to produce a baseline of vocal parameters, in a relatively neutral mood. The following two interventions produced a variety of mainly negative emotions: anger, sadness, loneliness and longing.

The resultant speech corpus is large and prohibitively difficult to annotate and judge for emotional content in its entirety. Several studies have undertaken emotional corpora of this size [4,5], though at this initial stage we decided to extract a subset from this corpus for detailed analysis. Two experienced research assistants therefore extracted 283 utterances that were judged to span a relatively wide range of emotional expressions. Interestingly, as shown below, the baseline interventions supplied a balanced set of positive emotions. Durations ranged between 0.32 and 8.7 seconds, with an average of 2.13 seconds ($SD=1.21s$). These utterances were then submitted to judges for subjective evaluation.

2.2. Subjective judgment

Emotional labeling is a complex issue in itself. Though attempts have been made to use categorical labeling [4] on non-acted speech, in this study we adopted a dimensional approach found in numerous studies. This form of judgment lends itself more easily to the complicated mixtures and nuances of emotion found in non-acted speech. All utterances were therefore labeled on three 5-point scales: Activation, Valence and Dominance (sometimes referred to as PAD – Pleasure, Activation, Dominance) [1].

Fourteen judges participated in the judgment task. Each judge was presented with all 283 utterances in randomized order. In order to assess the judges' performance, the ratings of each judge were correlated with the averages of all the other judges' rating, for each scale separately. Four of the 14 judges who had very low correlation values on one or more scales were thus removed.

2.3. Acoustic Analysis

The data was segmented on the word, syllable as well as the phone level. Acoustic analysis was performed at two degrees of abstraction: Feature-based and model-based.

Initially, a large set of fairly common acoustic features was extracted from each utterance. The features were designed to represent the various components of prosody: F0, intensity, duration, and voice quality/spectrum. A fully detailed description is beyond the scope of this paper, therefore we present only a general outline of the types of features by group:

F0: F0 contours were calculated using Praat software and corrected manually. The contours were then normalized per speaker by the mean of all utterances taken from the baseline interventions, and converted to semitones. A set of statistical features was then extracted – mean, STD, interquartile ranges, regression coefficients and so forth.

Intensity: Intensity was calculated and normalized similarly to the F0 contours. An identical set of statistical features was then calculated.

Duration: utterances were segmented manually into syllables. Maximum, mean and StD durations were calculated.

Voice quality: voice quality features were calculated using the Praat "voice report". This contains standardized features such as jitter, shimmer, NHR, autocorrelation. In addition

some original measures of microprosody were calculated, based on the FFT of the pitch and intensity contours.

Spectrum: several features were extracted from the Long Term Average Spectra (LTAS) of each utterance, as calculated by Praat. In addition, features were extracted from averages taken from frame based spectral analysis, using both a linear frequency scale and a Bark frequency scale.

Fujisaki Modeling: F0 contours were further parameterized using the Fujisaki model [6]. This model decomposes a given log F0 contour into a base frequency Fb , a phrase component, capturing slower changes in the F0 contour as associated with intonation phrases, and an accent component that reflects faster changes of F0 associated with accents and boundary tones. The phrase and accent components can be interpreted as smooth responses of the model to impulse-wise phrase commands and box-shaped accent commands. Whereas phrase command magnitudes A_p are related to the amount of F0 reset occurring at the onset of a new phrase, accent command amplitudes A_a are related to the strength of accents. In earlier work by the first two authors it was shown that the model is applicable to Modern Hebrew [7]. To the best of our knowledge, these features have not been used before in analysis of emotional speech. Extraction was performed using the automatic approach by Mixdorff and parameters were then corrected manually in the *FujiParaEditor* if necessary [8]. Unlike the standard practice which treats Fb as a speaker constant in a given context we were interested to see whether the emotional condition might influence this parameter and let it vary utterance-by-utterance. Time constants α and β , however, were set to constant values of 2 and 20 Hz, respectively.

Perceptual local speech rate (PLSR) is a psychophysical measure which was developed [9] because earlier measures such as the local syllable rate and the local phone rate are not well-correlated, meaning that they represent different aspects of speech rate. Perception experiments with short stretches of speech being judged on a rate scale revealed that neither syllable rate nor phone rate is sufficient to predict the perception results. Subsequently it was shown that a linear combination of the two measures yielded a correlation of $r=0.91$ and a mean deviation of 10% which is accurate enough to successfully extract PLSR from large spoken language corpora. The result is a smooth contour of local speech rate values aligned with the speech signal where a value of 100% represents a typical average speech rate while 50% being approx. half of it and 200% being roughly twice the average. It was also shown [10] that the language background affects the perception of local speech rate. German and Japanese listeners overshoot the speech rate of the respective unknown language by 7.5% or 9.1%, respectively. This additional deviation seems to be small enough to apply the PLSR extraction method also to Hebrew speech.

2.4. Statistical Analysis

Several types of statistical analysis were applied: correlation analysis, linear regression, and automatic classification. The first two methods were used mainly to compare the different types of F0 features – the purely statistical features to those based on the Fujisaki model. However, it is difficult to apply this type of statistical analysis to very large feature sets. An alternative and useful method, carried out in many studies of emotional speech, is to evaluate the discriminative power of the feature set by feeding it into an automatic classification algorithm.

Automatic classification was carried out using the publicly available *Weka* software [11].

Out of the large number of classifiers available therein, we chose the Random Forest classifier throughout, as the purpose here was not to evaluate different classification algorithms. Classification results were evaluated using 10-fold cross validation. The emotional classes were obtained by quantizing the mean judgments on each scale separately. Each scale was quantized once into two classes, so that the number of occurrences in each such class was balanced. This scheme can be elaborated further into joint classes, such as quadrants derived jointly from two scales, though in this preliminary study we did not yet attempt this.

Aside from the classification results obtained using the full set of features, we also carried out classification experiments using each group of acoustic parameters separately. Finally, with the aid of the correlation based subset feature selection algorithm, using best forward selection, as implemented in *Weka*, we looked for a subset of the most discriminative features in each case.

3. Results

3.1. Automatic classification

Table 1 summarizes the results of the automatic classification experiments. Classification was carried out twice for each set of features: with and without feature selection. The best overall results were obtained after feature selection on the entire feature set. Several interesting observations can be made, based on this table.

First, the best recognition rates were always obtained for Activation, for nearly all feature groupings. The highest percentage achieved was 80%, which is far above chance. Recognition rates for Valence and Dominance were appreciably lower, though well above chance at their highest levels, and mostly similar to each other.

Feature selection usually improved the results. In addition, feature selection usually reduced the number of features relevant to Dominance quite drastically, less so for valence, and even less so for Activation. Thus it seems that acoustic information is more useful for predicting Activation and spread out over more features.

As discussed before, there is a large amount of redundancy in the features, though the combined set did give highest recognition percentages. The most important features were those based on F0, followed by Bark Spectrum feature, STFT features, and Intensity features.

3.2. Statistical Analysis

Figure 1 shows a result of analysis from the database. The figure displays from the top to the bottom: The speech waveform, extracted (+) and modeled F0 contour (-), the perceptual local speech rate (PLSR) contour, as well as the underlying phrase and accent commands. Since every stimulus is assigned a single value for Activation, Valence and Dominance, respectively, but there are often several phrase and accent commands in an utterance, we calculated the following average parameters which produced the best correlations with the emotion judgments: *Fb*, *mean Ap*, *mean Aa*, mean accent command distance *acc_dist* (calculated as the duration of speech segments in an utterance divided by the number of accent commands). Table 2 shows the correlations between the aforementioned parameters and the emotion

judgments, as well as those for some of the raw F0 features such as mean F0, SD of F0, and F0 range in semitones.

Table 1: Percent of correct classifications for the different scales, using various feature types (in bold font). In parenthesis: the number of participating features

	Activation	Valence	Dominance
<i>All features</i>	78.4 (226)	65.4 (226)	63.2 (226)
<i>Selected, All</i>	80.2 (41)	68.9 (18)	66.2 (11)
<i>Pitch</i>	71.4 (46)	62.9 (46)	66.8 (46)
<i>Selected, Pitch</i>	72.8 (20)	67.5 (10)	62.2 (6)
<i>VQ</i>	58.7 (16)	54.4 (16)	57.2 (16)
<i>Selected, VQ</i>	48.1 (2)	51.2 (3)	49.5 (2)
<i>Intensity</i>	68.6 (37)	54.1 (37)	56.2 (37)
<i>Selected, Int.</i>	67.1 (10)	54.8 (4)	56.5 (2)
<i>Bark</i>	69.6 (56)	55.8 (56)	59.0 (56)
<i>Selected, Bark</i>	74.6 (15)	57.6 (2)	58.3 (8)
<i>STFT</i>	72.1 (56)	58.3 (56)	52.7 (56)
<i>Selected STFT</i>	71.7 (14)	59.4 (3)	52.7 (5)
<i>LTAS</i>	68.2 (4)	49.5 (4)	62.5 (4)
<i>Selected LTAS</i>	62.5 (3)	55.1 (4)	56.5 (2)
<i>Syl. Duration</i>	53.0 (4)	48.8 (4)	55.9 (4)
<i>Selected Dur.</i>	53.0 (4)	52.7 (2)	47.1 (4)
<i>Fujisaki</i>	63.3 (8)	60.4 (8)	55.9 (8)
<i>Selected Fuji.</i>	60.8 (4)	56.9 (8)	47.1 (2)

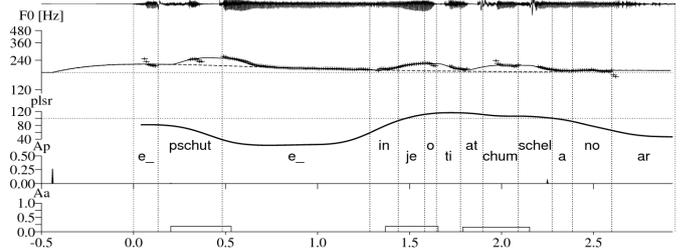


Figure 1: Example of Fujisaki model-based F0 contour decomposition: From the top to the bottom: The speech waveform, the F0 contour (+extracted, -modeled), perceptual local speech rate (PLSR) contour and the underlying phrase and accent commands.

Table 2: Correlations between Fujisaki model-based parameters, some raw F0 features and the emotion judgments.

Parameter (N=283)	Activation	Valence	Dominance
<i>mean Aa</i>	,378**	-,233**	-,038
<i>mean Ap</i>	,236**	-,056	,149**
<i>acc_dist</i>	-,363**	,112*	-,048
<i>Fb</i>	,081	-,188**	-,345**
<i>mean F0</i>	,453**	-,286**	,023
<i>SD of F0</i>	,491**	-,312**	-,050
<i>F0 range</i>	,472**	-,322**	-,124*

As can be seen, Activation is mostly associated with stronger F0 resets (*mean Ap*), stronger (*mean Aa*) and more frequent accents (*acc_dist*), whereas Valence is negatively correlated with the latter two factors, but to a lesser degree. Interestingly we find a negative correlation between Dominance and *Fb*. It is striking that each of the raw F0 features shows much higher correlations with Activation and Valence, respectively, than the Fujisaki model-based parameters. We have to bear in mind, however, that the Fujisaki model parameters are based

on a parsimonious decomposition of the same F0 contours that the raw features were extracted from. Furthermore, the raw features are strongly correlated with each other (mean F0 vs. SD of F0, $r = .551$; mean F0 vs. F0 range, $r = .438$). If we introduce the four Fujisaki model-based parameter to a multiple linear regression (MLR) model of the judgments of Activation, all parameters add significant contributions and explain 31.1% of the variance ($r = .558$), whereas a similar model based on mean F0 and F0 range (which give the best prediction provided that all factors be significant) explains 29.8% ($r = .546$). Hence, the F0 information is still captured by the model parameters, although on a higher level of abstraction and decomposed into the contributing factors at the utterance (*Fb*), phrase (*Ap*) and word levels (*Aa*).

Perceptual local speech rate was extracted for all 283 stimuli from 6501 manually segmented phone boundaries and 2628 syllable centers. The resulting contours were reduced to mean PLSR, its standard deviation, overall slope, slope averaged over the phrase slopes, median PLSR, minimum, maximum, as well as the mean PLSR and standard deviation averaged over all utterances of each speaker. These nine parameters were submitted to MLR of the perception data for activation, valence, and dominance.

Table 3: MLR correlation for PLSR and Fujisaki model-based features.

MLR vars.	Activation	Valence	Dominance
PLSR	0.485	0.266	0.379
Fujisaki model*	0.682	0.436	0.405
combined	0.738	0.490	0.525
*contains a total of nine features, the first four parameters from Table 1, augmented by, among others, SD and maximum of <i>Aa</i> .			

While PLSR-based parameters perform significantly worse than an augmented set of nine Fujisaki-model based parameters, the combination of both yielded the highest correlations implying that these two sets of variables represent non-redundant, or partially orthogonal information (see Table 3). This is especially true for the dependent variable Dominance where both speech rate and intonation parameters explain 14-16% of the observed variance while both taken together explain 28%. As these results show the explained variance is still rather low, especially for Valence (24.0%). Since linguistic features were so far ignored in this study, we performed a pilot experiment and asked two native speakers of Modern Hebrew to rate the textual content of all stimuli on a five point scale from negative to positive. Averages of the two ratings are significantly correlated with Valence ($r = .436$). If we combine these ratings with combined features from PLSR and Fujisaki model, MLR analysis yields $r = .613$ for Valence (explained variance 37.6%). Hence the textual content contributes to the perceptual judgment of this parameter, but does not yield an enormous improvement either. This assumption would probably be confirmed even with a larger group of subjects. In our case, the two judges differed 1.15 points with respect to their mean judgments (one showing a stronger tendency towards negative judgments), but their individual correlations with Valence were of the same order.

4. Discussion and Conclusions

This paper presented an acoustic study of authentic emotional speech, establishing the relationship between

perceptual judgment of Activation, Valence and Dominance and acoustic features. It illustrates how complementary methods can provide useful insights into the acoustics of emotional expression.

Sophisticated methods such as the Fujisaki model and PLSR were demonstrated to give compact sets of features which yield interpretable results on higher levels of abstraction than the low level features. On the other hand, low level statistical features do not necessarily miss the relevant acoustical information, and thus are still very useful for automatic classification. They are also easier to compute since extracting the model-based features requires a larger degree of supervision and manual correction than the lower level features.

The results are in line with other work [1,2] in showing that the utilized acoustic features predict judgments of Activation better than those of Valence and Dominance. Results from a pilot study suggest contributions from the text underlying the stimuli. In a future study we therefore intend to employ subjects who do not know Hebrew for judgments of Activation, Valence and Dominance, in order to further assess the contribution of linguistic information.

It will require further research to determine the optimal overall feature set, employing more refined grouping of the judgments into classes, and extraction of more features, both acoustic and linguistic. For example, absolute intensity was not included in the analysis since all stimuli were amplitude-normalized. In a recent study this parameter was found to be relevant [12]. Therefore a more comprehensive study of this corpus is currently being undertaken.

5. References

- [1] Grimm, M., Kroschel, K., Mower, E., Narayanan, S. (2007). Primitive-based evaluation and estimation of emotions in speech. *Speech Communication*. 49, 787-800.
- [2] Cowie, R., Cornelius, R.R. (2003). Describing the emotional states that are expressed in speech. *Speech communication*. 40, 5-32.
- [3] Rochman, D., Diamond, G. M. and Amir, O., "Unresolved anger and sadness: identifying vocal acoustical correlates", *Journal of Counseling Psychology*, 55(4): 505-517, 2008
- [4] Batliner, A., Hacker, C., Steidl, S., Noth, E., D'arcy, S., Russell, M. & Wong, M. (2004). 'You stupid tin box'- Children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. *Proceedings of Fourth International Conference on Language Resources and Evaluation*, 171-174.
- [5] Devillers, L. Vidrascu, L. & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Journal of Neural Networks*, 18, 407-422.
- [6] Fujisaki, H. and Hirose, K. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", *J. of the Acoustical Society of Japan (E)* 5(4), 233-241, 1984.
- [7] Mixdorff, H. and Amir, N. (2002): The Prosody of Modern Hebrew - A Quantitative Study. *Proceedings of Speech Prosody 2002*, pp. 515-518, Aix, France.
- [8] Mixdorff, H. (1/10/2009). FujiParaEditor, <http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html>
- [9] Pfitzinger, H.R. (1999): Local Speech Rate Perception in German Speech. *Proc. ICPhS 1999*, 893-896.
- [10] Pfitzinger, H.R.; Tamashima, M. (2006): Comparing Perceptual Local Speech Rate of German and Japanese Speech. *Proc. of the 3rd Int. Conf. on Speech Prosody*, 105-108.
- [11] <http://www.cs.waikato.ac.nz/ml/weka/>
- [12] Pfitzinger, H.R.; Kaernbach, C. (2008): Amplitude and amplitude variation of emotional speech. *Interspeech 2008*, pp. 1036-1039.