Acoustic-prosodic Analysis of Attitudinal Expressions in German

Hansjörg Mixdorff [1], *Angelika Hönemann*[2], *Albert Rilliard*[3]

[1] Beuth University Berlin, Germany
[2] University of Bielefeld, CITEC, Germany
[3] LIMSI-CNRS, Orsay, France

mixdorff@bht-berlin.de, ahoenemann@techfak.uni-bielefeld.de, Albert.Rilliard@limsi.fr

Abstract

This paper presents results from the prosodic analysis of short utterances of German produced with varying attitudinal expressions. It is based on the framework developed by Rilliard et al. eliciting 16 different kinds of social and/or propositional attitudes which place the subjects in various social interactions with a partner of inferior, equal or superior status, respectively as well as positive, neutral or negative, valence. Prosodic variations are analyzed in the framework of the Fujisaki model with respect to *F0*, as well as other prosodic features, such as duration, intensity and measures related to changes of voice quality. An analysis regarding the features that set apart two attitudes is presented. Expressive changes are discussed in light of previous results on US-English, and relative to universal codes proposed in the literature.

Index Terms: social attitudes, prosodic analysis, Fujisaki model

1. Introduction

Early work analyzing prosodic expressions mainly concerned their linguistic functions such as sentence modality or focus (see, for instance, [1]). As we move into the realm of para-linguistics, however, categories become more difficult to define. In the study of expressions of social affect early works mostly examined impromptu productions of “happy”, “sad” or “bored” versions of the same sentence, often by actors, to find acoustic differences [2].

However, almost always human communication has a social goal. Above and beyond pure linguistics, information about e.g. the mental state, emotions, mood or attitudes of the collocutors is exchanged during the dialog. The affective state is influenced, for instance, by the situation or roles of the dialog partners. Mutual understanding of the social intention between communication partners should not be difficult as long as they belong to the same language or culture. In contrast, interaction between partners from different cultures sometimes leads to wrong interpretations of social expressions. It has been shown that the verbal and non-verbal expressions depend, to some extent, on the culture in which we grow up. A study by Shochi et al. investigated twelve social attitudes e.g. surprise, irritation, command-authority for prosodic effects in the languages British English, French and Japanese [3]. They found similarities across these languages, but also some culture-specific uses of prosodic parameters. The similarities may be explained within the framework of a theory such as the frequency code [4] – a code phylogenetically derived that (roughly) proposes the use of pitch level as a marker inverse to dominance. Other codes have been proposed [5] that may refine the predicted use of fundamental frequency for communicative purposes. Conversely, culture-specific uses have been documented [6]. Intercultural comparison of linguistic and paralinguistic effects has enjoyed growing attention as the knowledge about how verbal and non-verbal social affects are expressed in different languages is paramount for mutual understanding between different cultures.

A main obstacle to the ecological study of social affect lies in the need to record such data with reasonably high quality while keeping the setting for the subjects as natural as possible, maintaining a certain level of spontaneity. To this effect and to facilitate the speaker’s task, [7] proposes to place target sentences in affectively loaded texts; similarly, [8] recorded attitudinally-neutral sentences embedded into dialogues that prepare the speaker to perform an adequate expression for the target sentence. An important issue here is the adequate labeling of attitudes elicited as the associated terminology will vary between languages.

The current work is based on the framework developed by Rilliard et al. [9] in which attitudes are characterized by a situational description of between whom and where they occur. An important difference from [8] is that recordings also concern the visual channel, as facial gestures are known to be a vital part of attitudinal expressions [10].

In the following section this approach will be discussed in more detail. Based on an underlying protocol two instances of 16 different attitudes were elicited from a total of 20 native speakers of German. In a recent paper we had native German subjects rate the credibility of the expressions portrayed by the first 10 of the subjects [11].

The focus of the current paper is the acoustic analysis of the target utterances in search for features that distinguish attitudes from one another. We will examine macro-prosodic features such as fundamental frequency, speech rate and intensity, as well as voice quality features such as harmonics-to-noise ratio, jitter, and shimmer. For the first group of ten speakers we selected the one instance of two that was rated better in [11], that is, more convincing. For the second group of ten speakers we selected that instance ourselves.

2. Speech Data Elicitation

Attitudes such as arrogance, politeness, doubt or irritation - see Table 1 for abbreviations henceforth used in this paper - were elicited through short dialogs which ended in the target sentences 'Eine Banane' (engl. *a banana*) or 'Marie tanzte' (engl. *Marie was dancing*). Preceding the target dialog a test dialog was performed in order to prepare the speakers and help them immerse themselves in the context of the attitude. These dialogs were designed according to different social situations differing in social and linguistic aspects such as the type of speech act (propositional/social), hierarchical distance, social distance or valence of speech act (positive/negative).

ADMI	admiration	OBVI	obviousness
ARRO	arrogance	POLI	politeness
AUTH	authority	QUES	neutral question
CONT	contempt	SEDU	seductiveness
DECL	neutral statement	SINC	sincerity
DOUB	doubt	SURP	surprise
IRON	irony	UNCE	wncertainty
IRRI	irritation	WOEG	walking-on-eggs

Table 1: *List of sixteen attitudes.*

All 20 native German subjects (11 female, 9 male) participating had academic background, were asked to produce the sixteen attitudes twice and paid for their time. Ages ranged from 20 to 60 with a median of 31.5 years.

3. Method of Analysis

Following the work presented in [9] we first examined the prosodic features fundamental frequency, speech rate and intensity. All target utterances were force-aligned on the phone level using the LINGWAVES alignment tool [12] and then manually checked inside the *PRAAT* TextGrid[14].
F0 contours were extracted at a step of 10 ms using the *PRAAT* default pitch extraction settings and subjected to manual inspection and correction.

We performed Fujisaki model parameter extraction [15]. Figure 1 displays examples of the utterance "eine Banane" uttered by male speaker 04, uttered with attitudes (from the top to the bottom) SURP, ADMI, DECL and QUES. Each panel displays from the top to the bottom: The speech wave form, the *F0* contour (extracted +++, modelled ---) the underlying impulse-wise phrase commands and box-shaped accents commands of the Fujisaki model [16]. Phone boundaries are indicated by dotted vertical lines and the transcription is given in German SAMPA.

The Fujisaki model approximates natural *F0* contours by superimposing three components: A constant base frequency *Fb* (indicated by the dotted horizontal line), exponentially decaying phrase components which are the responses to the phrase commands and accent components which are the smoothed responses to the accent commands. *Fb* therefore indicates the *F0* floor of the whole utterance. *Ap* expresses the global slope of the *F0* pattern, the accent command amplitudes *Aa* reflect the magnitude of local *F0* gestures and onset and offset times of accent commands their alignment with the underlying segments. Since its components are superimposed in the log *F0* domain, the model performs a normalization of the raw *F0* contour.

When we apply the Fujisaki model to reading style speech we generally assume that lexically stressed syllables of content words are candidates for the assignment of accent commands. Syllables preceding a prosodic boundary are also potential locations, as they may exhibit a high boundary tone [17]. In the case of affective speech, however, any syllable can be associated with an *F0* gesture [18]. Think, for instance, of 'ba-na-na!' being produced in an extremely irritated manner.

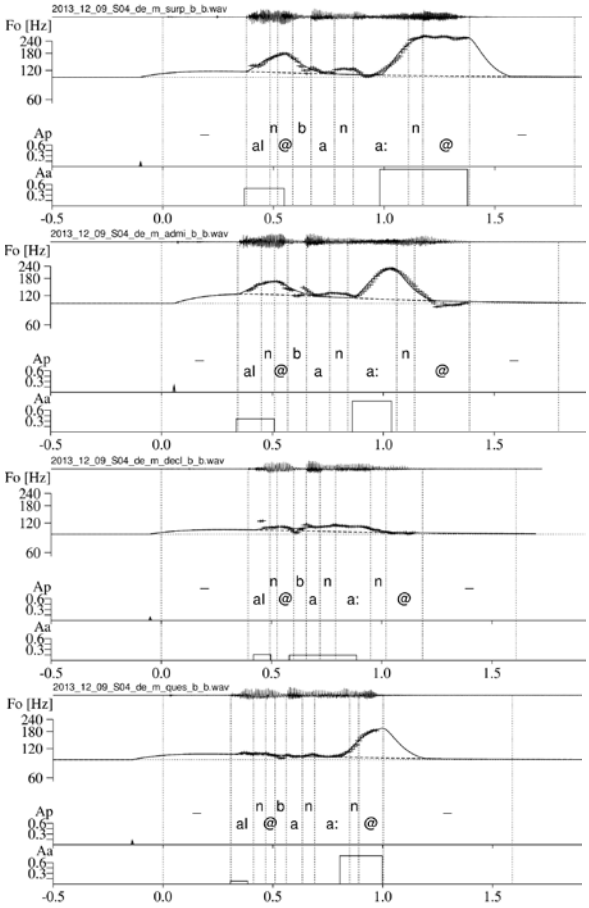

Figure 1: *Results of Fujisaki model-based analysis of F0 contours. Attitudes (from the top to the bottom) SURP, ADMI, DECL, QUES.*

Intensity contours were extracted in *PRAAT* with default settings, and mean intensities in dB, as well as maxima employing parabolic interpolation were determined for each phone. In addition to these features we used *PRAAT* to extract mean harmonics-to-noise ratio, jitter and shimmer for all vowels in the target utterances, applying default settings.

4. Results of Analysis

As shown in Figure 1, some attitudes exhibit rather distinct *F0* patterns. Attitudes such as SURP and DOUB are often associated with large accent command amplitudes whereas CONT and ARRO use narrow *F0* range. We begin the discussion of results with tables of means and standard deviations for all features examined and found to be significantly different for pairs of attitudes.

In all our data accent commands were either aligned with the lexically stressed syllables (in bold script) of "Ma**rie**", "**tanz**te", "**ei**ne" and "Ba**na**ne" or the phrase-final syllables when they perform the function of a question-final rise.

attitude	mean	s.d.	attitude	mean	s.d.
CONT	.145	.184	POLI	.232	.252
WOEG	.163	.178	SEDU	.250	.248
DECL	.166	.186	OBVI	.253	.307
AUTH	.166	.193	ADMI	.277	.305
ARRO	.185	.237	UNCE	.298	.306
SINC	.187	.204	QUES	.561	.487
IRON	.223	.245	DOUB	.589	.468
IRRI	.223	.217	SURP	.665	.475

Table 2: *Means and s.d. of accent command amplitude Aa for all sixteen attitudes (N=120).*

Table 2 displays the means and s.d. for accent command amplitude *Aa* averaged over all speakers and the two target phrases. As can be seen, QUES, DOUB and SURP are clearly set apart from all other attitudes by their large *F0* range. At the bottom end we find CONT and WOEG with very flat *F0* contours next to the neutral statement DECL. As can be seen, the standard deviation of *Aa* is large, hinting at considerable speaker-individual differences in *F0* range.
In contrast to our expectations, base frequency *Fb* and *F0* contour slope in terms of phrase command amplitude *Ap* are not significantly affected by the underlying attitude.

attitude	mean	s.d.	attitude	mean	s.d.
DECL	86	48	IRON	100	57
QUES	86	48	UNCE	100	64
POLI	86	48	WOEG	102	58
AUTH	88	49	SEDU	104	65
SINC	89	50	SURP	105	66
ARRO	92	52	ADMI	107	69
OBVI	95	58	DOUB	108	71
CONT	98	59	IRRI	110	69

Table 3: *Means and standard deviations of phone duration in ms for all sixteen attitudes (N=380).*

Table 3 shows means and standard deviations of phone durations depending on the attitude. As can be seen, neutral statements DECL and questions QUES are uttered at the highest phone rate whereas DOUB and IRRI are at the opposite end of the range and produced most slowly.

attitude	mean	s.d.	attitude	mean	s.d.
SEDU	73.7	6.9	ARRO	75.6	6.7
DECL	74.3	6.5	SURP	75.8	6.7
WOEG	74.7	6.4	OBVI	75.8	6.5
UNCE	74.7	6.4	SINC	75.9	6.1
QUES	74.8	6.6	CONT	75.9	6.4
POLI	75.0	6.8	AUTH	76.1	6.0
DOUB	75.3	6.7	IRON	76.1	6.5
ADMI	75.3	7.2	IRRI	78.1	5.5

Table 4: *Means and standard deviations of maximum phone intensity in dB for all sixteen attitudes (N=380).*

Table 4 displays means and standard deviations of maximum phone intensity. Seductiveness SEDU, neutral statement DECL and WOEG are located at the bottom end whereas IRRI, IRON and AUTH exhibit the highest intensity. However, we need to take into account that usually +3dB are required to perceive an increase in loudness. Most of the results stated so far are in line with outcomes reported in [9].
Now we turn to the measurements which are related to voice quality, namely harmonics-to-noise ratio, mean local jitter and shimmer. We report these features for all vowels in our utterances.

attitude	mean	s.d.	attitude	mean	s.d.
SURP	8.95	4.77	AUTH	9.84	4.25
ADMI	9.21	4.68	SEDU	9.89	4.85
DOUB	9.32	4.86	QUES	10.23	4.12
IRON	9.36	4.45	POLI	10.65	4.14
CONT	9.61	4.45	DECL	10.80	4.09
ARRO	9.67	5.05	IRRI	10.82	4.56
SINC	9.67	4.34	UNCE	10.97	4.47
OBVI	9.70	4.79	WOEG	11.09	4.53

Table 5: *Means and standard deviations of mean vowel harmonics-to-noise ratio in dB for all sixteen attitudes (N=180).*

attitude	mean	s.d.	attitude	mean	s.d.
IRRI	2.18	1.67	QUES	2.84	2.18
AUTH	2.36	1.72	DOUB	2.90	2.06
WOEG	2.51	1.85	OBVI	2.91	2.03
POLI	2.57	1.90	ADMI	2.92	2.37
CONT	2.60	2.09	ARRO	2.94	2.95
SINC	2.68	1.82	IRON	2.96	2.26
DECL	2.71	2.15	SURP	2.99	2.01
UNCE	2.72	2.33	SEDU	3.19	2.75

Table 6: *Means and standard deviations of local jitter in % measured in vowels for all sixteen attitudes (N=180).*

Table 5 shows mean harmonics-to-noise ratios and their standard deviations. At the bottom end we find SURP, ADMI as well as DOUB and IRON whereas neutral questions QUES and statements DECL are located in the upper half. It is somewhat puzzling that WOEG and UNCE register next to IRRI at the top of the scale. However, if we consider that breathiness or even devoicing causes the harmonics-to-noise ratio to drop it might well be the case that irritation brought forward in a clear tone of voice – as does DECL - has a similar effect as a soft, hesitating manner of speaking associated with doubt and embarrassment.
Table 6 presents the results for local jitter in %, that is, local fluctuations of *F0*. Whereas IRRI and AUTH exhibit the lowest values – in close company with WOEG – SEDU and SURP are at the top of the scale, a rise of more than 50%. Although authority can be thought of as being expressed with a stern, unfaltering voice and seductiveness and surprise are accompanied with jittery excitement, these results should be interpreted with extreme caution. Still the tendencies have certain plausibility to them since neutral statements and questions are located in the center part of the lineup.

The last feature examined is local shimmer. This parameter measures amplitude fluctuations between consecutive fundamental periods. Table 7 shows that differences between attitudes are relatively small. The extreme ends are somewhat the inverse of what we find for harmonics-to-noise ratio (Table 5), that is, breathier voice quality implies a higher local shimmer.

Attitude	mean	s.d.	attitude	mean	s.d.
IRRI	0.11	0.06	SINC	0.12	0.06
POLI	0.12	0.06	QUES	0.13	0.06
WOEG	0.12	0.06	SEDU	0.13	0.07
DECL	0.12	0.06	IRON	0.13	0.06
CONT	0.12	0.06	OBVI	0.13	0.07
UNCE	0.12	0.06	SURP	0.13	0.06
AUTH	0.12	0.06	ADMI	0.13	0.07
ARRO	0.12	0.06	DOUB	0.14	0.07

Table 7: Means and standard deviations of local shimmer in dB for all sixteen attitudes (N=180).

The prosodic features discussed above were tested with respect to their attitude-discriminating power applying Mann-Whitney-U tests. Figure 2 shows a matrix of significant differences between attitudes, that is, test results with $p < 0.01$. The brightness of colors indicates the number of features differing significantly between two attitudes.

	ADMI	IRON	OBVI	POLI	QUES	SURP	SEDU	SINC	UNCE	WOEG
ADMI		DI	I						HS	AHI
ARRO	H	HIS	D	I			H	H		
AUTH		DI		I	A		A	AH	S	AS
CONT	I	DI		I	A		AI	AI	DS	AS
DECL	I	I		I	A		AH	A	S	A
DOUB	HI	HIJS	DH	I		DJ	AHI	AHI		A
IRON	DI		DI	I	ADI	DI	AI	I	HIS	AHIS
IRRI	I	DIJ			A	H	AH	AH		A
OBVI	I	DI		D	A		A	AD	HDS	ADHS
POLI		I	D		I	D			I	I
QUES		ADI	A	I		A	H	H	AD	
SEDU		AI	A		H	ADH			AHS	
SINC		I	AD		H	ADH			AHS	
SURP		DI		D	A		ADH	ADH	D	AD

Figure 2: *Acoustic features differing significantly between two attitudes. Abbreviations: A - accent command amplitude Aa, D - phone duration, H - harmonic-to-noise ratio, I - maximum intensity, J - local jitter, S - local shimmer.*

This information permits us to cluster similar attitudes. Whereas ARRO, AUTH, CONT, IRRI form one group with mostly negative connotation, SINC, ADMI, POLI and SEDU cluster in a more positive group. QUES, DOUB and partly UNCE form a group of attitudes typically marked by interrogative sentence mode.

5. Discussion and Conclusions

First of all a word is in order regarding the small size of our sample. Although we recorded 20 subjects we only had two short utterances of each attitude at our disposal. Despite selecting the more acceptable one of two instances we know from our earlier perception test [11] that neither the subjects nor the attitudes are perceived as equally convincing. Still we opted against discarding certain speakers or attitudes in order not to further diminish our data set. What we intended to report here are trends of modification that prosodic features show under the influence of attitudinal expressions. We aimed to identify prosodic features which distinguish the intended attitudes – or do not.

When we compare the German performances with English produced by American L1 speakers and L2 speakers from Japan and France, some common trends, but also differences can be identified. The necessity to express varying degrees of involvement of the speaker in her/his speech act [19] requires a spreading of the attitudes along a voice strength dimension [20] that is common to all speakers. Whereas irritation and surprise are typically expressed with a loud voice implying raised pitch, neutral statements or reserved politeness can be found on the other side of the scale. Orthogonal to the voice strength dimension, a change of pitch for a given level of voice strength may be interpreted under the frequency code hypothesis: higher pitch is regarded as more submissive and lower as more imposing. Thus doubt, question, or surprise are found on the upper end of this scale whereas irritation, contempt and arrogance are found on the lower end – for all language groups. The performance of WOEG by German is characterized by lower speech rate as is the case for L1 US-English speakers and L2 French ones, but not for the Japanese speakers, as the specific attitudinal expression is conceptualized in their language. The strategies of German speakers set them further apart from US-American and French ones, as they appear to use a strongly harmonic voice for WOEG.

It should also be stated that we observed acoustic cues that escape the metric applied in this paper. These include disapproving lip smacks or smile, for instance. Overall the variability observed between speakers was striking, the realizations ranging from inspired, lively to stoic, almost autistic. Speaker-individuality was also reflected by statements like “I’m not good at irony” or “I never get upset”.

Furthermore we need to bear in mind that the current paper only concerned the acoustic channel. Results from [11] suggest that expressions of attitude which are presented along with a visual display usually yield higher ratings.

Another issue is the question to what extent expressions of attitude are unique and unanimously decoded - irrespective of a communicative context. If we consider results from acoustic prosody studies one and the same form placed in a different context will trigger a different interpretation on the part of the listener. The same may well be true for expressions of attitude. The clusters of attitudes which are not separated by our acoustic features point in that direction. Future work will explore which attitudes benefit from the visual channel. We will also examine how well intended attitudes are identified by native listeners of German and whether clusters of acoustically similar attitudes are confirmed by perception.

6. Acknowledgements

This work was funded through a French *digiteo* grant for Mixdorff for a research stay at LIMSI. Recordings of attitudes were funded through German DLR research grant no. 01DN13007 ARG.

7. References

[1] Altmann, H., Batliner, A., Oppenrieder, W. (Ed.), *Zur Intonation von Modus und Fokus im Deutschen*, Tübingen: Niemeyer. 1989.

[2] Paeschke, A., *Prosodische Analyse emotionaler Sprechweise*, Mündliche Kommunikation. vol. 1 . Logos-Verlag. 2003.

[3] Shochi. T.. Rilliard. A.. Aubergé. V. & Erickson. D.. "Intercultural perception of English. French and Japanese social affective prosody". in S. Hancil (ed.). The Role of Prosody in Affective Speech. Linguistic Insights 97. Bern: Peter Lang. AG. Bern. 31-59. 2009.

[4] Ohala. J. J.. "The frequency codes underlies the sound symbolic use of voice pitch". in Hinton. L.. Nichols. J. & Ohala. J. J. (eds.). Sound symbolism. Cambridge University Press. Cambridge. 325-347. 1994.

[5] Gussenhoven. C., *The Phonology of Tone and Intonation*, Cambridge: Cambridge University Press. 2004.

[6] Léon, P., "*Précis de Phonostylistique. Parole et Expressivité*, Paris: Nathan Université, 1993.

[7] Grichkovtsova. I.. Morel. M.. & Lacheret. A.. "The role of voice quality and prosodic contour in affective speech perception", Speech Communication, 54(3):414-429, 2012.

[8] Gu, W., Zhang, T. & Fujisaki. H., "Prosodic analysis and perception of Mandarin utterances conveying attitudes", Proceedings of Interspeech, Firenze, Italy. 1069-1072, 2011.

[9] Rilliard, A., Erickson, D., Shochi, T., de Moraes, J.A., "Social face to face communication - American English attitudinal prosody", INTERSPEECH 2013. 1648-1652.

[10] Swerts, M. and Krahmer, E., "Audiovisual prosody and feeling of knowing", Journal of Memory and Language 53(1): 81-94, 2005.

[11] Hönemann, A., Mixdorff, H., Rilliard, A. "Social attitudes - recordings and evaluation of an audio-visual corpus in German", Forum Acusticum 2014, Krakow, Poland.

[12] Wierzbicka. A.. "Defining emotion concepts", Cognitive Science 16: 539–581, 1992.

[13] http://www.wevosys.com/products/lingwaves/lingwaves.html

[14] Boersma. P., "Praat, a system for doing phonetics by computer", Glot International 5, 341-345, 2001.

[15] Mixdorff. H., "A new approach to the fully automatic extraction of Fujisaki model parameters", Proc. ICASSP 2000, vol. 3, 1281-1284, Istanbul, Turkey, 2000.

[16] Fujisaki, H., Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", Journal of the Acoustical Society of Japan 5, 233-241, 1984.

[17] Mixdorff, H., Jokisch, O., "Building an integrated prosodic model of German", Proceedings of Eurospeech 2001, vol. 2, pp. 947-950, Aalborg, Denmark (2001).

[18] Amir, N., Mixdorff, H. et al., "Unresolved anger: prosodic analysis and classification of speech from a therapeutical setting", Proceedings of SpeechProsody 2010, Chicago, USA (2010).

[19] Daneš. F, "Involvement with language and in language", Journal of pragmatics, 22(3-4):251-264, 1994.

[20] Liénard, J.-S., Barras, C., "Fine-grain voice strength estimation from vowel spectral cues", INTERSPEECH 2013, 128-132.