Acoustic-prosodic Analysis of Attitudinal Expressions in German

Hansjörg Mixdorff¹, Angelika Hönemann², Albert Rilliard³

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 tests; 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).

<table>
<thead>
<tr>
<th>ADMI</th>
<th>admiration</th>
<th>OBVI</th>
<th>obviousness</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARRO</td>
<td>arrogance</td>
<td>POLI</td>
<td>politeness</td>
</tr>
<tr>
<td>AUTH</td>
<td>authority</td>
<td>QUES</td>
<td>neutral</td>
</tr>
<tr>
<td>CONT</td>
<td>contempt</td>
<td>SEDU</td>
<td>seductiveness</td>
</tr>
<tr>
<td>DECL</td>
<td>neutral</td>
<td>SINC</td>
<td>sincerity</td>
</tr>
<tr>
<td>DOUB</td>
<td>doubt</td>
<td>SURP</td>
<td>surprise</td>
</tr>
<tr>
<td>IRRI</td>
<td>irritation</td>
<td>WOEG</td>
<td>walking-on-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>eggs</td>
</tr>
</tbody>
</table>

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-

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 “Marie”, “tanze”, “eine” and “Binane” or the phrase-final syllables when they perform the function of a question-final rise.

Table 3 shows means and standard deviations of 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.

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.

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

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

Table 7: Means and standard deviations of local jitter in % measured in vowels for all sixteen attitudes (N=180).
Whitney-U tests. Figure 2 shows a matrix of significant differences between attitudes, that is, test results with \( p < 0.01 \). The prosodic features discussed above were tested with differing significantly between two attitudes. The brightness of colors indicates the number of features that are separated by our Mann-Whitney-U tests with \( p < 0.01 \). The clusters of attitudes which are not separated by our prosodic features point in that direction. Future work 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.

<table>
<thead>
<tr>
<th>Attitude</th>
<th>mean</th>
<th>s.d.</th>
<th>Attitude</th>
<th>mean</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARRO</td>
<td>0.12</td>
<td>0.06</td>
<td>DOUB</td>
<td>0.14</td>
<td>0.07</td>
</tr>
<tr>
<td>IRRI</td>
<td>0.11</td>
<td>0.06</td>
<td>SINC</td>
<td>0.12</td>
<td>0.06</td>
</tr>
<tr>
<td>POLI</td>
<td>0.12</td>
<td>0.06</td>
<td>QUES</td>
<td>0.13</td>
<td>0.06</td>
</tr>
<tr>
<td>WOEG</td>
<td>0.12</td>
<td>0.06</td>
<td>SEDU</td>
<td>0.13</td>
<td>0.07</td>
</tr>
<tr>
<td>DECL</td>
<td>0.12</td>
<td>0.06</td>
<td>IRON</td>
<td>0.13</td>
<td>0.06</td>
</tr>
<tr>
<td>CONT</td>
<td>0.12</td>
<td>0.06</td>
<td>OBVI</td>
<td>0.13</td>
<td>0.07</td>
</tr>
<tr>
<td>UNCE</td>
<td>0.12</td>
<td>0.06</td>
<td>SURP</td>
<td>0.13</td>
<td>0.06</td>
</tr>
<tr>
<td>AUTH</td>
<td>0.12</td>
<td>0.06</td>
<td>ADMI</td>
<td>0.13</td>
<td>0.07</td>
</tr>
<tr>
<td>ARRO</td>
<td>0.12</td>
<td>0.06</td>
<td>DOUB</td>
<td>0.14</td>
<td>0.07</td>
</tr>
</tbody>
</table>

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.

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 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 Mixdorf for a research stay at LIMSI. Recordings of attitudes were funded through German DLR research grant no. 01DN13007 ARG.
7. References