Voice Source and Vocal Tract Variations as Cues to Emotional States Perceived from Expressive Conversational Speech

Hiroki Mori¹, Hideki Kasuya²

¹Faculty of Engineering, Utsunomiya University, Japan
²International University of Health and Welfare, Japan

hiroki@speech-lab.org

Abstract

Speech parameters originating from voice source and vocal tract were analyzed to find acoustic correlates of dimensional descriptions of emotional states. To achieve this goal best, we adopted the Utsunomiya University Spoken Dialogue Database, which was designed for studies on paralinguistic information in expressive conversational speech. Analyses for four female and two male speakers showed: (i) Prosodic parameters were highly correlated especially with the activation dimension, (ii) The aperiodicity-related voice source parameter showed that breathy phonation was mainly used in unpleasant utterances for three females, (iii) Due to smiling facial expression, formant frequencies were higher in pleasant utterances for a female. Index Terms: paralinguistic information, emotional state, voice source characteristics, formant

1. Introduction

Understanding the nature of speech under various emotional states is a fascinating topic. Traditionally, many studies on emotional speech have assumed some set of emotion categories. A typical example is the “big six” emotions, including anger, happiness, and sadness. Such approaches simplify corpus design and are well suited for controlled studies where actors are expressing the emotions, but one of the obvious drawbacks of category labels is that often there is no appropriate choice for an utterance when evaluating emotional states for spontaneous expressive speech.

The idea of emotion dimensions is an alternative to categorical descriptions. Dimensional descriptions have a long history and are well established in psychology. A number of studies stated that two or three dimensions are sufficient to account for a good portion of emotional variation. Among all, activation and evaluation dimensions have been regarded as fundamental [1, 2]. Dimensional description allows gradual representation of emotional states for any utterances, not only for ones under prototypical emotion. Techniques based on the framework of emotion dimensions are therefore promising candidates for speech applications that can handle daily emotional expressions in a flexible way.

Although there have been a number of studies on emotional speech, a large part of the vocal correlates of emotion dimensions is still unclear. Schröder[3] surveyed the literature and concluded that the correlation of activation with mean $f_0$, mean intensity and speech rate can be assured, but there is less evidence regarding the acoustic correlates of evaluation. So he conducted his own analysis of the Belfast Naturalistic Emotion Database and found a strong correlation between the activation dimension and $f_0$ parameters as in the literature, but the correlation with the evaluation dimension was reportedly much less stable.

The aim of this paper is to find acoustic correlates of emotion dimensions for expressive, spontaneous dialogue speech. As described above, there have been few reports on the acoustic correlates of the evaluation dimension. To overcome this, acoustic parameters reflecting voice quality variations originating both from voice source and vocal tract are introduced in addition to traditional prosodic parameters. The manipulation of these parameters can easily be incorporated in the ARX-based speech synthesizer[4].

To achieve this goal best, we adopt the Utsunomiya University (UU) Spoken Dialogue Database for Paralinguistic Information Studies as speech material. The UU database is especially intended for use in understanding the usage, structure and effect of paralinguistic information in expressive conversational speech. It is a collection of carefully-designed, task-oriented spoken dialogue with a wide variety of affect expression. We introduce the UU database in the subsequent section, describing its design concepts, specifications, and properties of emotional state ratings.

2. Corpus

The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies[5] is a collection of natural, spontaneous dialogues of college students consisting of seven pairs (12 females, 2 males). The participants and pairing were selected carefully to ensure that both people in each pair were of the same grade and able to get along well with each other.

The task of the dialogues, namely “four-frame cartoon sorting,” was carefully designed to stimulate expressively-rich and vivid conversation. In this task, four cards each containing one frame of a four-frame cartoon were shuffled, and each participant had two cards out of the four. Then they were asked to estimate the original order by communicating by voice, without looking at the remaining cards. The task proved to motivate the participants quite well because most Japanese students like cartoons and would be eager to know the true story. Each pair participated in three to seven independent sessions, using different cartoon materials for different sessions. Among the various cartoon combinations used for students, two cartoons were shown to all student pairs (except one pair where one of the participants knew the original cartoon). The recording was carried out in a soundproof room. In total, dialogue speech was recorded for 27 sessions, lasting about 130 minutes. The recordings were then transcribed with some markups (e.g. backchannel responses, fillers, discourse markers, etc.) The whole speech signal was segmented into 4737 utterances, where an utterance is defined...
3. Speech parameters

3.1. Prosody

Fundamental frequency (\(f_0\)) is a speech parameter that has been widely reported to be linked to emotional expressions, especially to the activation dimension[9]. \(f_0\) analyses were performed using Praat, and erroneous values were eliminated by visual inspection. Then the mean \(f_0\) and \(f_0\) range were calculated for each utterance in the logarithmic domain.

Intensity is another parameter that is believed to be a strong vocal correlate to emotional expressions. Again, Praat was used to obtain intensity curves, then peak intensity was obtained for each utterance. The intensity values are comparable only within each speaker.

as a speech continuum bounded by either silence (> 400 ms) or slash unit boundaries[6, 7].

In the UU database, each utterance is assigned a six-dimension vector that represents the perceived emotional state of the speaker. The dimensions are pleasant-unpleasant, aroused-sleepy, dominant-submissive, credible-doubtful, interested-indifferent, and positive-negative. The first three dimensions are compatible with the representation of emotional states in [3], i.e., evaluation, activation and power, respectively. This paper focuses on the first two dimensions: pleasant-unpleasant and aroused-sleepy.

Multiple annotators rated the perceived emotional state of the speaker for each dimension with a value from 1 to 7, where 4 corresponds to neutral. They listened to the stimuli using an identical environment (PC, headphones, playback level). In our previous works[8], the consistency and inter-annotator agreement of the emotional state rating was extensively examined with 22 annotators for the “core” subset of the database, and the results were used to conduct a screening test for newly hired annotators. Consequently, three out of six annotators were selected according to our criteria (consistency, correlation with the average, distinction of the dimensions), who then rated the emotional states for the rest of the corpus. Complete (3 of 3) and partial (2 of 3) agreement were 22.09% (chance: 2.04%) and 83.92% (chance: 38.78%), respectively.

Figure 1 shows the distribution of averaged ratings for pleasant-unpleasant and aroused-sleepy dimensions. The ratings are distributed over a broad range, which means the database covers a wide variety of expressive speech.

3.2. Voice quality variations derived from voice source

Voice quality is another important aspect of expressive speech, but it always suffers from the technical difficulty of obtaining reliable measurements[10]. Inverse filtering of the speech signal is necessary for evaluating voice quality variations that originate from voice source and vocal tract separately. However, such analyses are inherently difficult, especially for spontaneous dialogues where surprisingly high-pitched voice or non-modal voice qualities are used as paralinguistic cues. Thus we adopted a robust speech analysis algorithm based on the ARX speech production model[4], which simultaneously estimates the voice source and vocal tract filter as shown in Fig. 2.

Among the various kinds of voice source parameters that can be automatically obtained from the speech signal, we focused on the one related to the breathiness of voice source, because glottal noise is associated with suspicious and disappointed utterances from the viewpoints of both speech production and perception [11, 12], which are likely to be related to unpleasantness. As the parameter reflecting glottal noise, we used \(\hat{f}_{aperiodic}\)[13]; this is defined as the boundary frequency that partitions the whole band of the estimated source signal into a lower band in which harmonic components dominate and a higher band in which aperiodic components dominate, as illustrated in Fig. 3. Lower \(\hat{f}_{aperiodic}\) indicates a relative increase of the aperiodic component to the periodic one. For each utterance, mean \(\hat{f}_{aperiodic}\) is defined as the average of \(\hat{f}_{aperiodic}\) for voiced frames.
Table 1: Correlation coefficients for prosodic parameters (the aroused-sleepy dimension).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>FKC</th>
<th>FUE</th>
<th>FTS</th>
<th>FMS</th>
<th>MKK</th>
<th>MKO</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean $f_0$</td>
<td>0.47</td>
<td>0.58</td>
<td>0.79</td>
<td>0.67</td>
<td>0.83</td>
<td>0.65</td>
</tr>
<tr>
<td>$f_0$ range</td>
<td>0.68</td>
<td>0.70</td>
<td>0.74</td>
<td>0.72</td>
<td>0.64</td>
<td>0.56</td>
</tr>
<tr>
<td>peak intensity</td>
<td>0.80</td>
<td>0.78</td>
<td>0.88</td>
<td>0.81</td>
<td>0.91</td>
<td>0.74</td>
</tr>
</tbody>
</table>

3.3. Voice quality variations derived from vocal tract

Last but not least, there should be a systematic change in the filter characteristics that reflect paralinguistic information. For example, Maekawa[14] showed that the vowel /a/ in acted suspicious utterances had higher $F_2$ values than admiring ones. However, this kind of analysis is extremely difficult for spontaneous speech because the linguistic contents of the utterances cannot be controlled. In this study, therefore, we limit the target to interjections “a” and “aa”, which are used in the same way as “oh” in English. The meaning of “a” and “aa” varies (e.g. noticing something, acknowledgment, understanding, etc.) depending on prosody and voice quality.

For parameters reflecting vocal tract variations, the first and second formant frequencies ($F_1$, $F_2$) were obtained using the ARX speech analysis algorithm described above, then averaged $F_1$ and $F_2$ were calculated for each “a” and “aa.” Nasalized vowels were eliminated because the current implementation of the ARX analysis forcedly approximates the vocal tract filter with an all-pole system.

4. Results and Discussions

In the current study, speech parameters were analyzed for 1439 utterances of four females and two males. Most of them were talking very vividly and expressively, so it seems appropriate to start with this subset.

4.1. Activation (aroused-sleepy)

Table 1 shows the correlation coefficients between the prosodic parameters and average ratings for the aroused-sleepy dimension. This table reveals that all the prosodic parameters have a strong correlation with the aroused-sleepy dimension. This is exactly what we expected, and is entirely consistent with previous works.

Among all, peak intensity was the most highly correlated, implying that intensity can be the most reliable acoustic correlate to the activation dimension where speech can be recorded in a calibrated environment.

From this result, we conclude that most of the variations of speakers’ perceived activation can be explained by the prosodic parameters such as mean $f_0$, $f_0$ range and peak intensity for our database.

4.2. Evaluation (pleasant-unpleasant)

Table 2 shows the correlation coefficients between the prosodic parameters and average ratings for the pleasant-unpleasant dimension. Although the overall tendency is similar to that of aroused-sleepy, the correlation was much lower for most speakers. Therefore, it seems that prosodic parameters alone are not sufficient to account for the speaker’s perceived evaluation.

Next, we plotted how the aperiodicity-related voice source parameter, mean $f_{aperiodic}$, is correlated with the pleasant-unpleasant dimension in Fig. 4. Roughly speaking, mean $f_{aperiodic}$ below 2 kHz indicates that the utterance is breathy. This figure shows that utterances with low mean $f_{aperiodic}$ tended to be perceived as unpleasant for most female speakers. Specifically, FKC and FMS sometimes used breathy phonation, which was perceived as unpleasant. A moderate correlation was observed between perceived pleasantness and mean $f_{aperiodic}$ for FKC ($r = 0.46$) and FMS ($r = 0.33$). Although they used breathy voice less frequently than FUE, some of their unpleasant utterances sounded very impressive (e.g. FMS’s utterance located at (3.3, 722 Hz) had a harsh voice quality). Another speaker, FUE, used breathy phonation frequently, most of which was perceived as unpleasant. The tendency that utterances accompanied by glottal noise are perceived as unpleasant was the same as the case of FKC and FMS. However, there are three outliers around (5, 1500 Hz). One of these utterances gives the impression that she took a friendly and tender attitude with a soft/whispery voice quality, so it was perceived as pleasant. This implies that FUE used breathy phonation for expressing both pleasant and unpleasant emotional states. In contrast, FTS produced almost no breathy voice. For MKO, there

Figure 4: Distribution of mean $f_{aperiodic}$ in terms of the evaluation dimension.
were few unpleasant utterances, but the mean $f_{aperiodic}$ tended to have lower values because of his inherent breathy voice quality. MKK produced a lot of utterances with low mean $f_{aperiodic}$. This can be explained by the extremely high correlation between peak intensity and $f_{aperiodic}$ for MKK ($r = 0.84$). He tended to produce short utterances with very weak phonation, which consequently had low mean $f_{aperiodic}$ values.

Lastly, the results for the vocal tract variations are shown. Analyses for FKC, FUE, FMS were impossible due to lack of samples ($< 8$), and the correlation was not significant for MKK and MKO. In contrast, “a” and “aa” utterances of FTS have a very characteristic formant distribution as shown in Fig. 5. This figure indicates the tendency that if an utterance is perceived as more pleasant, then $F_1$ and $F_2$ have higher values. Correlation coefficients between average ratings for the pleasant-unpleasant dimension and $F_1 / F_2$ are $0.67 / 0.68$, respectively. As far as the data shown in Fig. 5 are concerned, utterances with higher formant frequencies sound as if the speaker is speaking with a smile. If so, the formant frequency shift described above agrees well with Tartter’s findings that syllables produced while smiling have a significantly higher $F_2$ than neutral syllables, and that listeners can distinguish them[15]. This result implies the existence of a mechanism that systematically changes the vocal tract characteristics even for spontaneous dialogue speech.

5. Conclusion

In this paper, speech parameters originating from voice source and vocal tract were analyzed to find the correlations with the emotion dimensions as a representation of emotional states perceived from expressive conversational speech.

The key finding of the current study is that speakers have their own pattern for changing voice quality according to their emotional states. Acoustic analyses for the UU spoken dialogue database revealed that three out of four female speakers used breathy phonation mainly in unpleasant utterances. On the other hand, one of the female speakers produced pleasant utterances with higher formant frequencies. The mechanisms of these systematic changes in voice quality are consistent with the results of previous controlled studies. The current study has shown the possibility of applying such models of voice quality control to the recognition or synthesis of conversational speech. However, these hypotheses should be supported by further investigations with a larger database.

Acknowledgment

The authors would like to thank Mr. Tomoyuki Satake for his assistance. This work has been supported in part by the Grant-in-Aid for Scientific Research (17700172).

6. References


Figure 5: $F_1$-$F_2$ distribution in terms of the evaluation dimension for FTS.