Emotions in dubbed speech: An intercultural approach with respect to F0

Angelika Braun / Matthias Katerbow

Department of German Linguistics, Phonetics Section
University of Marburg, Germany
Braun3@staff.uni-marburg.de / matthias@katerbow.de

Abstract

A comparative intercultural study on the so-called basic emotions anger, joy, fear, and sadness as well as neutral utterances was carried out based on samples of dubbed speech. The languages studied were the American English original of a popular TV series (*Ally McBeal*) as well as its German and Japanese dubbings. The production by the main male and female characters in all three languages as well as the perception by American, German and Japanese listener groups were examined. The present contribution focuses on results on the production and perception side of F0 and related parameters. The principal findings indicate that there are major cultural differences and also gender differences in encoding and decoding the emotional content of the utterances studied. Differences were found to be larger between linguistically and culturally less related languages than between the more closely related ones.

1. Introduction

Even though there is an abundance of research on the phonetic manifestations of emotionally loaded speech [cf. 1-3 for summaries], certain shortcomings are also to be observed. First of all, the emotionally loaded utterances studied were generally elicited from actors expressly for research purposes. "Natural" externalizations of emotions are difficult to get hold of and therefore rarely analyzed. A famous example is [4]. Furthermore, it is often either the encoding stage (i.e. phonetic features signaling a particular emotional state) OR the decoding stage (i.e. the ability of listeners to recognize a certain emotion) which is considered. Studies involving the complete chain of communication are desirable and yet rare. [5] Finally, there are only very few studies on intercultural differences in the production and perception of emotion [e.g.6;7]. The present research aims to address these shortcomings. First of all, the utterances analyzed are produced by actors but with the intention of sounding natural and not with the intention of conveying an emotion for the purpose of phonetic analysis. Both the encoding and the decoding stage are studied. Intercultural differences on both the production and the perception side are included in the analysis. The following research questions were asked:

- Is there a difference in the encoding of emotions between speakers of different linguistic/cultural backgrounds with respect to F0 and related parameters?
- Is there an additional gender difference in the encoding of emotions between speakers of different linguistic/cultural backgrounds with respect to F0 and related parameters?
- Is there an intercultural difference in the perception of emotions in dubbed speech?

2. Materials and methods

2.1. Materials

The materials examined were isolated from the first series of *Ally McBeal*. It was chosen because it is available in DVD quality in many languages and thus lends itself to intercultural comparison. In addition, a leading male (Billy Thomas) as well as female (Ally McBeal) character were available for analysis. Neutral speech samples as well as four so-called basic emotions which are said occur universally [8] (joy, anger, sadness, and fear) were studied. Within the emotion of anger, a difference was made between hot and cold, since it may be assumed that these two variants of anger are represented by very different acoustic phonetic cues. [5][1] The selection of the stimuli proved more difficult than it would seem at first glance for a number of reasons: for the sake of the perception experiment no stimulus containing a verbalization of the respective emotion could be used; only full utterances (no single words) were considered; since acoustic phonetic analyses were to be carried out no background noise or music could be present; since facial and gestural expressions were to be analyzed, the person speaking had to be the only (or principal) character shown in the picture.

A preliminary selection of potential scenes with respect to emotional content based on these criteria was established independently by three speech scientists. Thereupon, a pre-test was conducted involving 10 German native speakers who were asked to rate the scenes. Only those scenes which were consensually judged by a minimum of 7 out of the 10 raters were used in the experiment. The listeners who participated in the pre-test did not take part in the perception experiment proper.

According to this procedure, 45 stimuli per language were selected: 19 for each of the three male speakers and 26 for each of the three female speakers. The distribution among the different emotions was as follows:

- Male speaker: 5 x anger, 5 x joy, 4 x sadness, 5 x neutral
- Female speaker: 8 x anger, 4 x joy, 5 x sadness, 4 x fear, 5 x neutral.[2]

[1] Cold anger and fear were represented in the female character only, though. Despite an extensive search which reached beyond the first series, no clear representations of male fear or cold anger could be identified.

[2] Originally, 4-5 instances of each emotion were aimed at. Among the female anger samples, 5 were hot anger and 3 were cold anger

2.2. Methods

The acoustic measurements were carried out using a MEDAV Spectro 3000 computer with its built-in SIFT F0 algorithm. This stand-alone device was originally developed for forensic phonetic purposes and is thus particularly robust to any kind of disturbance. Frame length was set to 16 ms for female voices and 20 ms for male voices. The lower limit of analysis was 60 Hz; the upper limit was set between 300 and 500 Hz depending on the individual stimulus. All measurements were checked by hand. They were: Average F0, F0 standard deviation, and F0 range. (Due to limited space, results on F0 standard deviation will not be discussed in this paper.) Measurements were originally carried out in terms of Hz but later converted to semitones (ST).

The perception experiment involved three different listener groups who had American English, German and Japanese as a native language. Subjects were undergraduate students who received credit for their participation but were not paid. Originally, 148 subjects participated in the listening experiment. 65 had German as their native language, 40 were native speakers of AE, and 43 had Japanese as their first language. They were presented with the audio segments in question and asked to rate them in terms of emotion. Ten per cent of the stimuli were presented twice in a test-retest paradigm. Any subject that did not reach a test-retest agreement of 80% was excluded from further analysis.

The listeners remaining after applying this criterion were 60 Germans, 35 Americans, and 34 Japanese. Preliminary analyses demonstrated no differences in judgment according to listener sex; thus the perception results reported in this paper are averaged over listener sex. This observation is consistent with the findings reported in [7].

3. Results

3.1. Production

The production data are presented in two different ways: Absolute measurements of F0, its standard deviation and range are represented in Hertz (Hz), whereas differences between emotionally loaded utterances and neutral ones are represented in semitones (ST) in order to compensate for absolute gender or language differences.

3.1.1. F0 mean

Figures 1 and 2 summarize the results for the male speakers.

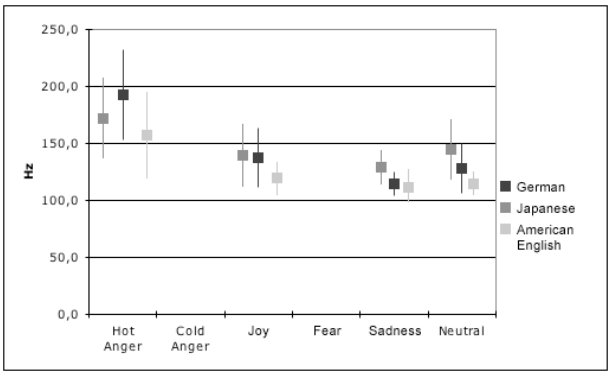

Figure 1: F0 means and standard deviations

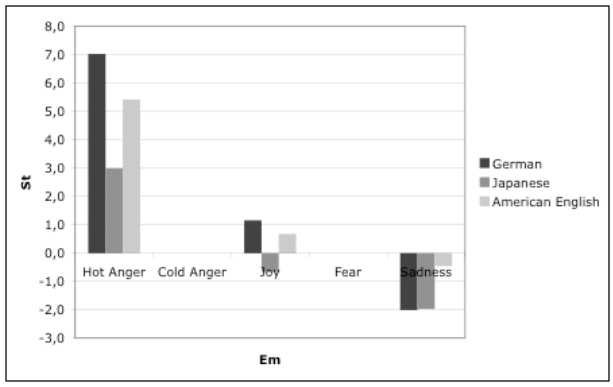

Figure 2: Change in F0 means for different emotions

As far as average F0s are concerned, the American English original speaker always exhibits the lowest values. The Japanese speaker exhibits the highest values with the exception of hot anger, which is most clearly marked by the German speaker.

When it comes to differences between emotionally loaded speech and neutral samples, the picture is not entirely clear: The 3 male speakers raise their F0 considerably for the representation of hot anger, all of them lower their F0 in order to indicate sadness. This is in accord with previous findings as summarized in [5]. As opposed to the two others though, the Japanese speaker lowers his average F0 when expressing joy. This can be expected to generate problems for non-Japanese listeners in the decoding process (see below).

The results for the female speakers are contained in Figures 3 and 4.

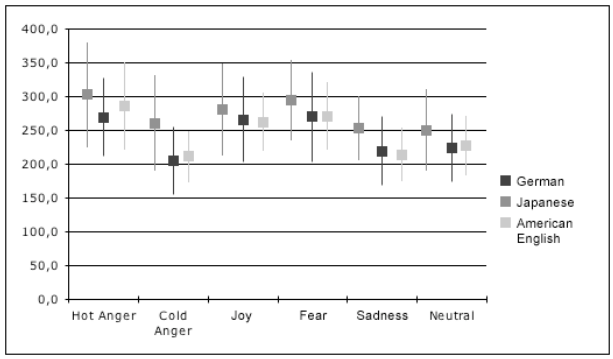

Figure 3: F0 means and standard deviations

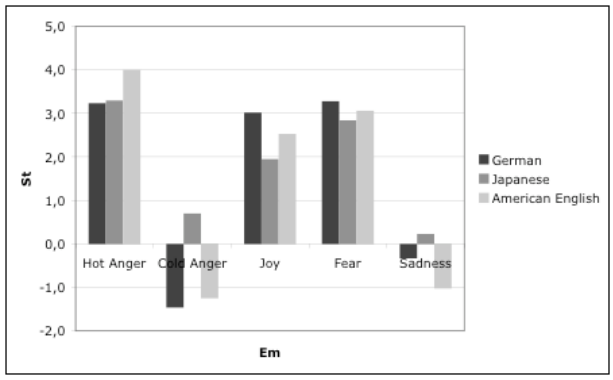

Figure 4: Change in F0 means for different emotions

The Japanese speaker exhibits by far the highest values among the three in all emotions as well as neutral speech. This is in accordance with previous findings [7,9]. The differences between the other two speakers are minor. This finding confirms the hypothesis that linguistically and culturally more similar languages behave in a more similar way than culturally more distant ones. The difference is greatest for cold anger, which will have to be closely observed in the perception study. Furthermore, for the German speaker, there is no clear difference in average F0 between the emotions of hot anger, joy, and fear. Therefore, listeners may encounter problems in correctly identifying these emotions.

The coding of emotionally loaded speech with respect to neutral samples once again shows differences between American English and German on the one hand and Japanese on the other. Whereas all three speakers considerably raise their average F0 in order to signal hot anger, joy, and fear, the American and the German speakers lower their F0 for cold anger and sadness in the magnitude of approximately 1 ST, whereas the Japanese speaker does not. Instead, she slightly raises her F0 for both. Non-Japanese listeners may be expected to encounter a problem with this behavior.

3.1.2. F0 range

Table 1 and Figure 5 depict the F0 range for the male speakers.

Table 1: F0 range in Hz (male speakers)

	Anger (h)	Joy	Sadness	Neutral
G	139	95	40	90
J	128	94	68	92
E	170	61	55	40

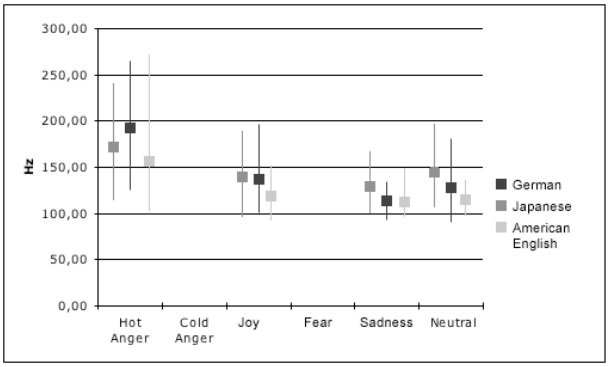

Figure 5: F0 means and ranges (male speakers)

Results are fairly unanimous in that the ranges for hot anger and joy well exceed those for neutral speech and, but they differ with respect to sadness. Here, the American speaker – as opposed to the other two – displays the most monotonous speech. Table 2 and Figure 6 show the results for the female speakers.

Table 2: F0 range in Hz (female speakers)

	Anger (h)	Anger (c)	Joy	Fear	Sadness	Neutral
G	229	190	203	266	201	190
J	299	242	242	228	166	199
E	220	134	154	180	139	158

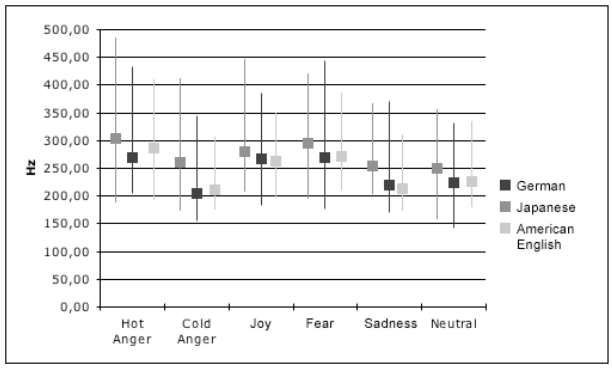

Figure 6: F0 means and ranges (female speakers)

The Japanese speaker exhibits the largest F0 range for all emotions including the neutral samples by far, whereas the German and American speakers are not far apart. This may be a consequence of the generally higher F0 average on the part of the Japanese speaker. Relatively speaking, though, the individual maximum range varies according to language: the Japanese and American speakers show the largest range for hot anger, whereas the German speaker displays the largest range for the emotion of fear. In other words, F0 range in itself cannot be expected to be a reliable cue for listener judgment.

3.2. Perception

There is not nearly enough room in this paper for a full description of the perception results. Two examples, which are suggested by the production data, will be presented here. They concern the encoding of joy by the male Japanese speaker and the encoding of sadness and cold anger by the female Japanese speaker.

A good way of expressing perception performance is by way of confusion matrices. Tables 7-9 show the decoding of the stimuli produced by the male Japanese speaker.

Table 3: Perception of Japanese (male) stimuli by Japanese listeners in percent

Intended↓ Perceived↓	Joy	Neutral	Sadness	Anger
Fear	17	21	24	4
Joy	**40**	0	2	1
Neutral	28	**69**	10	5
Sadness	16	7	**62**	1
Anger	0	23	2	**90**

Table 4: Perception of Japanese (male) stimuli by American listeners in percent

Intended↓ Perceived↓	Joy	Neutral	Sadness	Anger
Fear	19	10	28	6
Joy	**19**	1	3	3
Neutral	27	**76**	9	7
Sadness	33	8	**48**	0
Anger	1	6	11	**84**

The overall recognition rates in the same-language situation are consistent with previous research [5].

Table 5: Perception of Japanese (male) stimuli by German listeners in percent

Intended↓ Perceived↓	Joy	Neutral	Sadness	Anger
Fear	15	9	26	1
Joy	**27**	4	8	3
Neutral	27	**72**	8	15
Sadness	30	12	**53**	0
Anger	1	2	4	**79**

However, neither the German nor the American listeners managed to perceive joy even remotely reliably; instead, sadness was the most frequent answer. This is in perfect agreement with the production data for those two languages in which sadness is the only emotion cued by a lowering of F0. The Japanese listeners chose joy as their most frequent answer, though. This may mean that they did not use F0 as their prime or only cue.

Table 6: Perception of Japanese (female) stimuli by Japanese listeners in percent

Intended Perceived	Fear	Joy	Neutral	Sadness	Anger (h)	Anger (c)
Fear	**60**	5	2	29	0	7
Joy	0	**55**	1	3	0	1
Neutral	8	30	**79**	10	1	16
Sadness	30	8	1	**51**	1	2
Anger	2	2	17	6	**97**	**74**

Table 7: Perception of Japanese (female) stimuli by American listeners in percent

Intended Perceived	Fear	Joy	Neutral	Sadness	Anger (h)	Anger (c)
Fear	**63**	1	2	13	3	0
Joy	5	**85**	33	14	22	49
Neutral	8	1	**55**	15	2	35
Sadness	21	11	3	**57**	1	2
Anger	4	1	7	1	**73**	**13**

Table 8: Perception of Japanese (female) stimuli by German listeners in percent

Intended Perceived	Fear	Joy	Neutral	Sadness	Anger (h)	Anger (c)
Fear	**58**	1	2	16	2	1
Joy	5	**73**	27	17	22	17
Neutral	9	7	**60**	15	2	57
Sadness	23	18	7	**43**	1	2
Anger	5	1	3	8	**74**	**24**

Quite clearly, American and German listeners are at a loss if cold anger is coded by an increase in F0, as is the case for the Japanese speaker. However, this hardly confuses Japanese listeners. This may indicate that the concept of a division between hot and cold anger does not hold in Japanese.

The perception of sadness is an example for the fact that explanations are not always as easy as those described above. Sadness as vocalized by the Japanese speaker was quite reliably recognized by the American listeners (they outperformed that of the Japanese, in fact). This finding cannot be explained by F0 data alone. It demonstrates that it is most often a combination of features which needs to be taken into account in order to account for the perception results.

4. Discussion

In many respects the present results do confirm previous research. This concerns the principal ways of signaling different basic emotions as well as findings on F0 in Japanese. However, for the first time, intercultural differences could be studied on the encoding and decoding ends of the communication chain using real language samples. Even though the results discussed here only represent a fraction of the results, it can be stated that there is indeed a strong indication that the differences in production and perception between culturally less related languages exceed those between populations, which are culturally and linguistically very close.

5. Conclusions

Even though the present results do reveal cultural differences in the production and perception of emotion in close-to-real settings for the first time it will always have to be kept in mind that the cues described in this paper may well not be the only ones contributing to the perception of a particular emotion. Temporal, voice quality, facial, and gestural features as well as articulatory precision add to the picture, to name only a few and possibly the most important ones. Specifically, the issue of which role the different cues may play in different cultural settings remains as yet unresolved.

6. Acknowledgements

The present research was funded by a grant from the Hessian Ministry of Science and Art, Germany. The authors would like to thank Prof. Satoshi Morimoto, Tenri University, Japan, for his assistance with administering the test to Japanese listeners.

7. References

[1] Scherer, K. R. and Wallbott, H. G., "Ausdruck von Emotionen", *Psychologie der Emotionen* (Scherer, K.R. ed.) Hogrefe, Göttingen etc, 1990.

[2] Scherer, K. R., "Vocal communication of emotion", *Speech Communication, Vol. 40, 2003, p 227-256*.

[3] Kehrein, R., *Prosodie und Emotionen*, Niemeyer, Tübingen, 2002.

[4] Williams, C. E. and Stevens, K. N. (1972): "Emotions and Speech. Some Acoustic Correlates", *J. Acoust. Soc. Amer.,* 52: 1238-1250, 1972.

[5] Banse, R. and Scherer, K. R. "Acoustic Profiles in Vocal Emotion Expression", *Journal of Personality and Social Psychology,* 70: 614-636, 1996

[6] Scherer, K. R., Banse, R. and Wallbott, H. G. "Emotion inferences from vocal expression correlates across languages and cultures", *J. Cross-Cultural Psych,* 32:76-92, 2001.

[7] Van Bezooijen, R., Otto, S. A., Heenan, Th. A., "Recognition of vocal expressions of emotion", *Journal of Cross-Cultural Psychology*, 14: 387-406, 1983.

[8] Ekman, P, "An Argument for Basic Emotions", *Cognition and Emotion*, 6:169-200, 1992.

[9] Van Bezooijen, R., "Sociocultural Aspects of Pitch Differences between Japanese and Dutch Women", *Language and Speech*, 38:253-265, 1995.