CROSS-LANGUAGE PERCEPTION OF HEBREW AND GERMAN AUTHENTIC EMOTIONAL SPEECH

Hartmut R. Pfitzinger[a], *Noam Amir*[b], *Hansjörg Mixdorff*[c] *& Jessica Bösel*[a]

[a]Inst. of Phonetics and Digital Speech Processing, Christian-Albrechts-University Kiel, Germany;
[b]Dept. of Communication Disorders, Tel Aviv University, Tel Aviv, Israel;
[c]Dept. of Computer Sciences and Media, Beuth University of Applied Sciences, Berlin, Germany
hpt@ipds.uni-kiel.de; noama@post.tau.ac.il; mixdorff@beuth-hochschule.de;
jb@ipds.uni-kiel.de

ABSTRACT

This cross-language study investigates differences in the assessment of emotional spontaneous speech from two real-life situations. These comprise a Hebrew corpus of utterances recorded during psychotherapy and a German corpus from a setting of online gaming. The emotional content was judged by 83 listeners of the two languages and scored on three scales: Activation, Valence, and Dominance. While cross-language correlation of Activation perception of German stimuli achieved an exceptionally high Pearson r of 0.95, Valence in Hebrew stimuli showed the lowest correlation (r=0.60). It turned out that the judgments on Valence differed considerably due to a remarkable amount of German as well as Hebrew stimuli judged positively by Germans, but negatively by Hebrew listeners. Thus, cross-language perception differences of emotional speech are not symmetrical.

Keywords: authentic emotions, spontaneous emotional speech, cross-culture, cross-language, perception

1. INTRODUCTION

Research on cross-language differences of emotional speech is very rare. One of the first investigations was by Albas, et al. [1] who presented acted emotions to Canadian and Cree Indian subjects. Both groups performed significantly better than chance but emotions with a similar Activation level led to high confusion values. Further studies by van Bezooijen, et al. [4], Scherer, et al. [13], Maekawa [9], and Thompson & Balkwill [14] were also based on acted emotions.

We found only three cross-language studies based on authentic emotional speech. Yang & Campbell [15] presented 21 Mandarin emotional speech stimuli to 13 Chinese and 5 American listeners. Both groups recognized 71% of the stimuli correctly. The authors concluded that prosodic features alone are sufficient to identify emotional categories. In a perception experiment of Ibrakhim [8] 30 emotional utterances of a Japanese female were judged by Japanese and Russian listeners, yielding considerable perceived attitude differences. Finally, Erickson [6] presented 40 single word or isolated vowel stimuli produced by a Korean female to 13 Japanese, 12 Korean and American listeners. Erickson assumed that native speakers approach the listening task linguistically while the other listener groups only use acoustic information.

With respect to acoustic features, Frick [7] suggested that Activation is signaled by increased F0 height, F0 range, loudness, and rate. Also, Bänziger & Scherer [3] showed that mean level and range of F0 contours vary strongly as a function of the degree of Activation.

Amir, et al. [2] investigated the relationship between prosodic features (e.g. F0, intensity, voice quality, duration, speech rate) and emotional content judged by 14 Hebrew listeners. They found a correlation of r=0.74 between a combination of features and Activation. It remains to be seen whether there exists an upper threshold limiting the performance of automatic recognition methods which do not consider the linguistic content, but operate entirely on the acoustic properties of the speech signal. This would imply that a considerable amount of emotional information is conveyed by verbal content. As a consequence, however, not only word recognition would be required, but also a correct interpretation of recognized words. This is a capability which, especially in the realm of emotions, may overtax even a human being.

In order to investigate this problem we adopt two approaches: On the one hand Hebrew emotional speech stimuli are judged by Hebrew listeners as well as Germans who do not know any Hebrew, on the other hand, since we are dealing with the cross-language as well as cross-cultural aspects of emotional judgments, we create German stimuli and have them rated by Germans, as well as Hebrew listeners who do not know German. This strategy enables us to examine whether a completely symmetrical behaviour of cultures can be observed, that is, for instance, whether Hebrew listeners as compared with Germans, judge German stimuli as reflecting less Activation and vice versa.

2. METHOD

The following four perception experiments are based on two sets of emotional speech: 283 stimuli from 22 speakers of Hebrew and 144 stimuli from 6 speakers of German. There exist very few accessible databases of authentic emotional speech. For this reason the two corpora employed in this study are also very different with respect to their domain, as well as age, number, and gender of subjects. Despite these differences, the validity of the present investigation is not affected as it is fully symmetrical viz. Hebrew and German listeners judged both Hebrew and German stimuli, and we are only interested in the differences of cross-language and cross-culture emotion judgment.

Twenty-two Hebrew women aged 21 to 25, all of them university students at the time, volunteered to participate in a session of psychotherapy since they reported experiencing unresolved anger towards an attachment figure. The resultant speech corpus is large and prohibitively difficult to annotate and judge for emotional content in its entirety, thus, two experienced research assistants extracted 283 utterances that were judged to span a relatively wide range of emotional expressions. The data are described in detail by Rochman, et al. [12] and were already used for prosodic analysis and automatic emotion recognition (Amir, et al. [2]).

As described above, the pre-existing Hebrew corpus was compiled in a real-life setting of psychotherapy which could not be easily replicated for German. As a consequence, a completely different scenario, namely a setting from an ego shooter video game was adopted (see Fig. 1). In contrast to the Hebrew corpus, there is a certain degree of control of the subjects' emotional states because they can be related to the stages in the game and the associated situations of winning and losing, as well as the element of surprise. Six male subjects aged 24 to 31, were recorded producing 28 hours of German speech data. An experienced research assistant selected 144 stimuli to cover a wide range of emotional colorings.

Figure 1: Recording authentic emotional speech during a LAN party with six German male subjects engaging in multi-player computer games.

Eighteen Hebrew listeners took part in the German speech perception task. Their mean experimental duration was 28 minutes and 46 seconds. 26 German listeners judged the Hebrew data, 25 the German data, and 14 Hebrew listeners participated in the Hebrew perception test. Their task was to judge Activation, Valence, and Dominance of each stimulus on three different 5-point scales represented by Self-Assessment Manikin (SAM, Bradley & Lang [5]). The subjects were free to listen to each stimulus as often as they liked before they made their choice.

3. RESULTS

3.1. Duration, response time and repetition of stimuli

The mean duration of the Hebrew stimuli (2.3s, s.d.=1.2s) is approx. twice as long as of the German stimuli (1.1s, s.d.=0.4s). Because this might be a possible explanation for our cross-language perception differences, and because we cannot investigate the effect of this duration difference on emotion perception directly, we analyzed the correlation of German stimulus duration and judgment difficulty.

Fig. 2 shows the distribution of the Hebrew listeners' response times to German stimuli. We regard response times greater than ca. 30s as short recovery phases. In order to capture the difficulty of judgments that the listeners experienced, we recorded response times as well as the number of repetitions per stimulus during the listening experiment. Fig. 3 shows that these measurements are moderately correlated (r=0.57), meaning that both represent slightly overlapping aspects of judgment difficulties.

Figure 2: Histogram of the Hebrew listeners' response times to German stimuli.

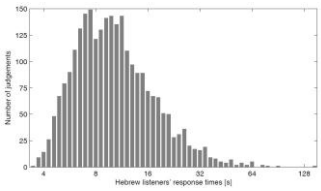
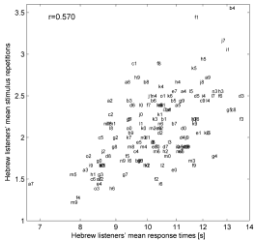


Figure 3: Scatter plot of the Hebrew listeners' mean number of stimulus repetitions vs. mean response times.



However, Fig. 4 shows that the variation in stimulus duration accounts for only 5.7% of the

variation in mean response time and only 6.0% in mean number of stimulus repetitions. Thus, it turned out that stimulus duration is not related to the degree of judgment difficulty.

Figure 4: Scatter plots of German stimulus durations versus the Hebrew listeners' mean response times (left) and mean stimulus repetitions, respectively (right).

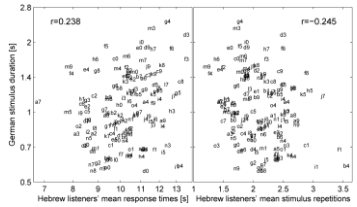
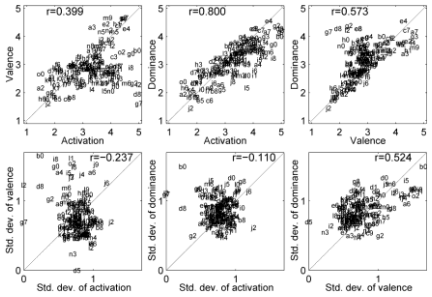


Figure 5: Scatter plots between mean judgments of Activation, Valence, and Dominance (top row) and between their standard deviations (bottom row) of 144 German stimuli judged by 18 Hebrew listeners.



As to the standard deviations shown in the bottom row of Fig. 5 one might expect that they increase with the judgment difficulty represented by stimulus response times and number of repetitions. This would imply that they are well correlated with these two parameters. But none of the standard deviations of the three dimensions shows a higher correlation than r=0.15 with either stimulus duration, response time, or repetition.

3.2. Perception of Activation, Valence, and Dominance

Fig. 6 shows that the correlation of German and Hebrew listeners is the highest when judging the emotional dimension Activation. Especially for German stimuli the listener groups correlated remarkably well (r=0.95) though there is a significant off- set of ca. one-third of a point on the 5-point scale ($t(143)=13.41$, $p<0.001$, two-tailed), indicating that Hebrew listeners, to a small degree but consistently, over-estimated the Activation of the German stimuli. The other five panels show clear clockwise rotations of the data points which means that the Hebrew listeners mostly used smaller ranges of the scales than the Germans.

The Valence scale exhibits the most diverging judgment when comparing the two groups of listeners. This is due to a remarkable amount of German as well as Hebrew stimuli positively judged by Germans but negatively by Hebrew listeners. Particularly, the Hebrew stimuli *I3*, *M3*, *I0* (see Fig. 6, top center panel) which received negative Valence judgments by Hebrew listeners were over-estimated by German listeners by at least two points on the 5-point scale. Conversely, the German stimuli *l2*, *l5*, *a4* (see Fig. 6, bottom center panel) which received positive Valence judgments by German listeners were negatively judged by Hebrew listeners. Obviously, this behaviour is not symmetrical across both languages because of the absence of stimuli judged negatively by German but positively by Hebrew listeners. Generally, the Valence judgments of the Hebrew listeners were more negative than those of the German listeners.

Figure 6: Scatter plots of Hebrew versus German listeners' mean judgments of Activation, Valence, and Dominance for 283 Hebrew (top row) and 144 German emotional speech stimuli (bottom row).

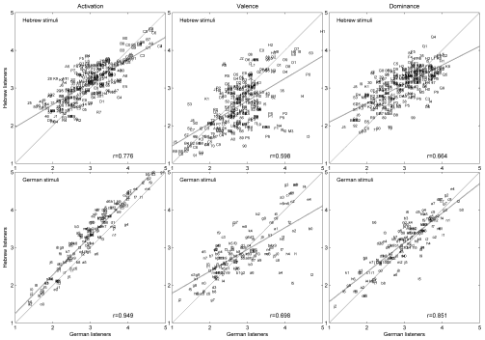
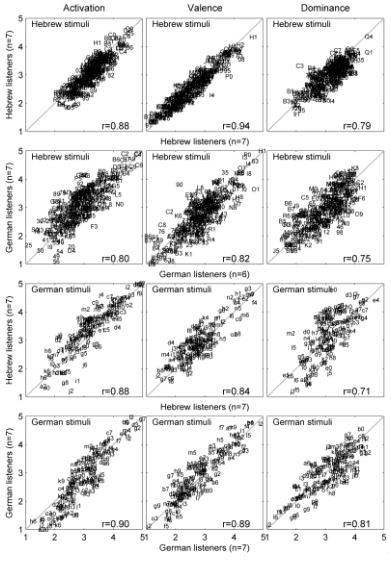


Figure 7: Scatter plots of intra-group correlations. Hebrew stimuli judged by Hebrew listeners (*1.row*) and by German listeners (*2.row*). German stimuli judged by Hebrew listeners (*3.row*) and German listeners (*4.row*).



Regarding Dominance the observed cross-language correlation is quite high for the German stimuli (r=0.85) but only moderately high for the

Hebrew stimuli (r=0.66). This is not surprising when linking the two facts (a) that Fig. 5 reveals a fairly high correlation between Activation and Dominance (r=0.80), and (b) that Activation showed the highest cross-language correlation.

To check consistency within each of the four listener groups we allocated 14 randomly selected listeners into two subgroups and correlated their mean results. Fig. 7 shows that intra-group correlation is between r=0.71 and r=0.94. None of the 12 scatter plots shows extraordinary outliers. It turns out that intra-group correlations are generally higher when listeners judge stimuli from their native language.

4. DISCUSSION

The present, fully symmetric experimental design did not lead to symmetric emotion perception results across languages.

The stimuli taken from two sets of emotional speech differ with respect to their mean durations and were produced by a different number of speakers of opposite gender, and in two completely different scenarios. Therefore, this study is primarily concerned with the difference in judgment by the two groups of listeners, and not with the difference in the datasets.

The comparison between Hebrew and German judgments shows better matches for Activitation and Dominance than for Valence, presumably because these parameters can be assessed to a great extent by the prosodic content. The fact that the correlations for the shorter German stimuli with their smaller verbal content are even higher, supports this interpretation. With respect to Valence, it is known that strongly positive and negative emotions, that is, happiness and anger, have been reported to be associated with similar prosodic features, for instance, expanded F0 range [7, 10, 11]. The presence of verbal information might therefore be necessary to fully disambiguate these emotions. We can only speculate whether cross-cultural experiences or expectations influenced the judgments, simply because the listeners presumably were able to identify the language in which the stimuli were presented. More negative Valence judgments by Hebrew listeners with respect to German stimuli and more positive judgments by Germans with respect to Hebrew stimuli might have extra-linguistic, that is, for instance, psychological reasons. These are cross-cultural aspects which will have to be addressed more carefully in later experiments, for instance, by interviewing the subjects about their general impression and their experiences during the experiment.

The Hebrew data were judged less homogeneously. On the average, they exhibit twice as long stimulus durations as the German stimuli. A possible explanation is that longer stimuli require multiple repetitions, leading to longer response times and less consistent judgments across listeners. We tested this hypothesis on a subset of results by examing the relationship between German stimulus duration and Hebrew listeners' response time and number of repetitions. However, the longer stimulus durations do not seem to be the cause for the bias. Rather, the higher verbal content or the less obvious prosodic coding of the emotional content in the Hebrew data is the reason for the lower correlation between German and Hebrew judgments of Hebrew emotional speech.

5. REFERENCES

[1] Albas, D.C., McCluskey, K.W., Albas, C.A. 1976. Perception of the emotional content of speech: A comparison of two Canadian groups. *J. of Cross-Cultural Psychology* 7(4), 481-490.

[2] Amir, N., Mixdorff, H., Amir, O., Rochman, D., Diamond, G.M., Pfitzinger, H.R., Levi-Isserlish, T., Abramson, S. 2010. Unresolved anger: Prosodic analysis and classification of speech from a therapeutic setting. *Proc. of the 5th Int. Conf. on Speech Prosody*, Chicago.

[3] Bänziger, T., Scherer, K.R. 2005. The role of intonation in emotional expressions. *Speech Communication* 46(3-4), 252-267.

[4] van Bezooijen, R., Otto, S.A., Heenan, T.A. 1983. Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics. *J. of Cross-Cultural Psychology* 14(4), 387-406.

[5] Bradley, M.M., Lang, P.J. 1994. Measuring emotion: The self-assessment Manikin and the semantic differential. *J. of Behaviour Therapy and Experimental Psychiatry* 25(1), 49-59.

[6] Erickson, D. 2010. Perception by Japanese, Korean and American listeners to a Korean speaker's recollection of past emotional events: Some acoustic cues. *Proc. of the 5th Int. Conf. on Speech Prosody*, Chicago.

[7] Frick, R.W. 1985. Communicating emotion: The role of prosodic features. *Psychological Bulletin* 97(3), 412-429.

[8] Ibrakhim, I. 2004. Universal and linguistic features of expressing emotional information: Differentiation in the perception level. *Proc. of the 2nd Int. Conf. on Speech Prosody* Nara, Japan, 659-662.

[9] Maekawa, K. 2004. Production and perception of 'Paralinguistic' information. *Proc. of the 2nd Int. Conf. on Speech Prosody* Nara, Japan, 367-374.

[10] Mozziconacci, S.J.L. 2002. Prosody and emotions. *Proc. of the 1st Int. Conf. on Speech Prosody* Aix-en-Provence, 1-9.

[11] Pfitzinger, H.R., Kaernbach, C. 2008. Amplitude and amplitude variation of emotional speech. *Proc. Interspeech* Brisbane, 1036-1039.

[12] Rochman, D., Diamond, G.M., Amir, O. 2008. Unresolved anger and sadness: Identifying vocal acoustical correlates. *J. of Counseling Psychology* 55(4), 505-517.

[13] Scherer, K.R., Banse, R., Wallbott, H.G. 2001. Emotion inferences from vocal expression correlate across languages and cultures. *J. of Cross-Cultural Psychology* 32(1), 76-92.

[14] Thompson, W.F., Balkwill, L.-L. 2006. Decoding speech prosody in five languages. *Semiotica* 158, 407-424.

[15] Yang, L-C., Campbell, N. 2001. Linking form to meaning: The expression and recognition of emotions through prosody. *Proc. of the 4th ISCA Speech Synthesis Workshop (SSW4)* Perthshire, Scotland.