Perception of levels of emotion in speech prosody

Kostis Dimos[a], Leopold Dick[b] , Volker Dellwo[c]

[a,c] Phonetic Lab, Department of Comparative Linguistics, University of Zurich
[b] Centre for R+D, Bern University of Arts
kostis.dimos@uzh.ch, leodick@hotmail.com, volker.dellwo@uzh.ch

ABSTRACT

Prosody conveys information about the emotional state of the speaker. In this study we test whether listeners are able to detect different levels in the emotional state of the speaker based on prosodic features such as intonation, speech rate and intensity.

We ran a perception experiment in which we ask Swiss German and Chinese listeners to recognize the intended emotions that the professional speaker produced. The results indicate that both Chinese and Swiss German listeners could identify the intended emotions. However, Swiss German listeners could detect different levels of happiness and sadness better than the Chinese listeners. This finding might show that emotional prosody does not function categorically, distinguishing only different emotions, but also indicates different degrees of the expressed emotion.

Keywords: Emotional prosody, speech perception, theatrical speech, Swiss German, Chinese

1. INTRODUCTION

1.1. Emotions in speech

Emotion in speech is being transmitted via verbal and vocal components. Early studies have focused on the relation between emotions and various suprasegmental speech features such as loudness, pitch, speech rate etc. [13, 15]. Previous studies have associated prosodic characteristics such as high F0 mean and range and higher speech rate with specific emotions such as fear rather than sadness [27]. Whiteside has also noticed differences in jitter (%) and shimmer (dB) between different types of anger ("hot" and "cold").

However, different listeners may focus on different acoustic parameters [12] and different speakers may also use different acoustic features to express emotions, especially in cases that might be considered more subconscious, thus less controlled, for example when the speaker is under stress [24].

Biological [6] and cultural factors contribute to the production and perception of emotional prosody. Emotions can be identified in content-free speech by speakers from different countries and language backgrounds [21, 23]. Some of these parameters, such as the pitch range may carry emotional information in non-verbal communication such as laughter [25].

Moreover, studies on emotion recognition in music have also shown a relation between the level of loudness and tempo and the perceived emotion [8, 10, 11, 22, 28]. High tempo and increased loudness are associated with happiness while low sound level and slow tempo has been associated with sadness [9]. Additionally, listeners show preference to less urgent speech, therefore less "annoying" speech [4]. Of course, researchers have also described inter-speaker and inter-listener variability in the production and interpretation of vocally expressed emotions [17].

Global prosodic features, such as the F0 level and range can lead to automatic discrimination of emotions at levels comparable to humans' recognition for the four, considered as, basic or "primary" emotions, that is, happiness, anger, sadness and neutral [18, 26].

An interesting issue is the definition of the emotions themselves. It is common to define some "basic" emotions such as happiness, sadness, anger, fear [15] and consider them as the basis for other, complex emotions. It is however difficult to understand how emotions work and even more to understand how emotional prosody is being processed neurologically.

In an emotion identification study Pell detected a contribution of both hemispheres in the identification and recognition of different emotions in real and nonsense utterances, revealing an increased role of the right hemisphere on the recognition of emotional prosody features [19, 20]. However, emotional recognition appears to be a rather bilateral process, rather than focused in one area [1, 2, 3, 5, 14].

2. RESEARCH APPROACH

In this study we want to focus on the ability of a professional speaker, specifically an actress, to use emotional prosody to intentionally make the listeners perceive the selected emotion. More specifically we will try to answer the following questions:

1. Can a professional speaker produce different levels of emotion based only on prosodic cues?

2. Can listeners recognize the emotion and emotion level the speaker intentionally produces based only on the perception of her prosody?

3. Can non-German speakers perceive the emotional prosody of the speaker and the intended emotion? Will their perception differ from the native listeners?

3. METHODOLOGY

2.1. Production experiment

2.1.1. Material

We used a set of 20 short sentences (8-10 syllables). The first ten were real German sentences, elicited from the OLSA database [16] and the other ten were pseudo-sentences created by a native Swiss German speaker and professional composer. The sentences were semantically neutral, and syntactically simple.

e.g.1 a. *Peter bekommt neun kleine Blumen*. (real sent.)
b. *Pita sarft dei ulta Finge*. (pseudo-sent.)

2.1.2. Design and Procedure

We asked a female professional speaker to produce the sentences according to two 5-level emotional scales a) from neutral to happy and b) from neutral to sad. The listener produced the same set of 20 sentences two times, once for the happiness and once for the sadness scale. The scales were defined as:

a) Happiness scale:
Neutral, Happy1, Happy2, Happy3, Happy4
b) Sadness scale:
Neutral, Sad1, Sad2, Sad3, Sad4

As we could not know whether the speaker would be able to intentionally produce these multiple levels of happiness and sadness, she was allowed to prepare and practice the utterances as much as she needed and to reproduce them if necessary until she had produced the desired prosody. We also asked her to produce the utterances in a way that it will be possible for the listeners to recognize the different levels of the emotion. The emotional levels were randomly assigned to each of the sentences. For the perception experiment we excluded the 8 Neutral sentences. In total we used 32 utterances.

2.2. Perception experiment

2.2.1. Listeners

28 Swiss German and 23 Chinese listeners participated in the experiment. The participants were students at the University of Zurich and the Chinese participants were students at the Nanjing Institute of Technology (Nanjing) and Renmin University of China (Beijing). None of the Chinese listeners was a German speaker or has ever lived in Switzerland or any other German speaking country.

2.2.2. Design and Procedure

The Swiss German and Chinese native speakers were asked to rate the recorded sentences on a continuous scale from sad to happy. A Praat demo window was used for the experiment. The participants were being introduced to the experiment interface and they had to try a demo of the experiment. The demo was providing 8 stimuli that were part of the experimental sentences and were covering the whole range from "sad 4" to "happy 4" in order to familiarize them with the performer's speech before the experiment. Then the listeners were seated in front of a laptop, in a quiet room.

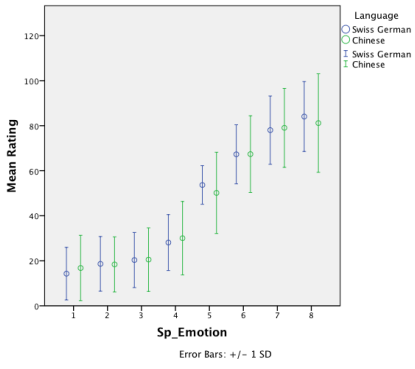
The participants were presented with a screen in which there was the sentence: *Diese person ist…* ("This person is") on the top and a continuous grey area in the middle of the screen. This area was a 0-100 scale consisting of 100 equidistant intervals that were not visible to the listeners [14]. On the left of the scale there was the phrase *eher traurig* ("rather sad") and on the right the phrase *eher fröhlich* ("rather happy"). All the instructions were translated fro the Chinese participants, by a phonetician and native Chinese speaker.

The participants would listen to each of the stimuli and then they would click on the designated gray area from left to right according to the emotion expressed by the speaker. After clicking on the scale the screen was turning blank, then next stimulus was presented and then the screen with the scale would appear again. The order of the stimuli was randomized for each participant. The whole dataset was presented twice, each time in randomized order, for each participant. In total the participants were listening and rating 64 stimuli (32 utterances played twice during the experiment).

4. RESULTS

We used the 0-100 rating scale to compare the listeners' rating to the speaker's intended emotion for every sentence. The scatter plot in Figure 1 shows the correlation between speaker's produced emotion (horizontal axis) and listeners' rating averaged by emotional level (2 real, 2 pseudo-sentences per emotional level).

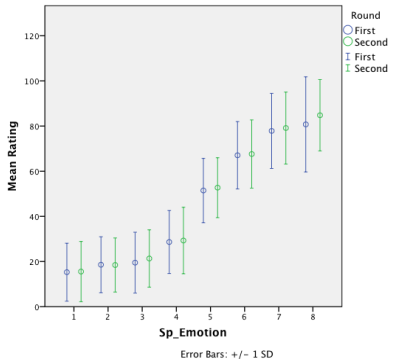
Figure 1: Swiss German and Chinese listeners' perception and speaker's intended emotion.



As we can see, some levels of emotion are better recognized than others. In general we see higher SD in the sadness scale and we can also see that sad2 and sad3 are vey close to each other. On the other side, the different levels of happiness are clearly separated from each other. In the second graph we can also see the differences between the two appearances of the dataset. There is no

significant difference between the first and the second rating of the sentences.

Figure 2: Listeners' rating in the first and the second.



A linear regression test was conducted in order to test the relationship between speaker's intended emotion and listeners' rating. A significant regression equation was found ($F(1, 3262)= 8292.483$, $p < .001$ with an $R^2= .718$. Swiss listeners' rating showed higher correlation to the speakers intended emotion. A significant regression equation was found ($F(1, 1790)= 9513.988$, $p < .001$) with an $R^2 = .842$. On the other hand, for Chinese listeners regression equation was also significant ($F (1, 1470)= 2864.338$), with an $R^2= .661$. Bonferroni corrected pairwise comparison indicated non-significant difference in the rating between sad2 and sad3 for both Swiss German ($p > .9$) and Chinese listeners ($p > .9$). Chinese listeners also showed no significant differences in their rating between sad1 and sad2 ($p > .9$) and sad1 and sad3 ($p > .9$). Additionally, Chinese listeners showed no significant differences in their ratings between happy7 and happy8 ($p > .9$).

3. DISCUSSION

This experiment was conducted to test whether speakers can produce multiple levels of emotion using prosodic cues and whether listeners can recognize these levels of emotion. Professional speakers may be able to control their voice in a more sophisticated way in order to demonstrate fine differences between emotions than normal speakers.

The results indicate that both the Swiss-German and the Chinese listeners could identify most of the different degrees of happiness in speech. However, both Swiss German and Chinese listeners performed worse in detecting different levels of sadness. This might be because of different acoustic cues used to express sadness and happiness, making the latter easier to perceive acoustically. Additionally, Chinese listeners also had difficulties distinguishing between the two higher levels of happiness. It is possible that the prosodic and acoustic differences between sad3 and sad4 were not as salient to the Chinese as to the Swiss German listeners.

However, listeners in our study showed the ability to detect differences in short utterances produced by an unknown speaker. These acoustic cues that allowed listeners to distinguish between levels of emotions could be biologically, culturally, or both associated with these emotions. As we have seen, there are arguments for and against biological origins of emotional prosody cues [6]. However, in this study we cannot exclude the effect of cultural differences in the perception of emotions in speech and even if there is a biological base in emotional prosody we do not yet have a clear idea about what part of the paralinguistic or non-linguistic communication is based on biology and evolution and what is culture-specific.

As we included only one speaker in our experiments it is also not easy to understand the acoustic cues that helped the listeners identify the levels of emotion. The first question that arises from these experiments is whether these results would be the same with a different speaker.

Female speakers are usually more expressive in their speech than male speakers, therefore would it be possible for a male speaker to produce the different levels and for the listeners' to perceive them? Additionally, in this experiment we used a professional speaker, a performer that was trained to use her voice in a more accurate and controlled way than an average speaker. An average speaker might not be able to intentionally produce these differences between levels of emotion, however it is possible that we use similar acoustic cues as the professional speaker to express emotions in everyday speech.

Further investigation might reveal some of the acoustic cues the speaker is using to express the levels of emotion. We would expect that loudness, F0 levels and speech rate are the most salient acoustic cues for the listeners since we already know from the literature that they are associated with emotional prosody.

Another aspect of this study is the ability of professional performers to express emotional prosody and to "communicate" with the listeners at a suprasegmental level. Although we know that prosody conveys a lot of information and contributes to everyday communication, we do not know how effective this paralinguistic communication is. These experiments have shown that speakers might be able to elicit much more information from prosodic features than we might expect. Additionally they can elicit this information form a speaker they are not familiar with, as well as from a speaker that speaks an unknown to them language.

Acknowledgments: This research was supported by grant 100016_143874 of the Swiss National Science Foundation.

7. REFERENCES

[1] Adolphs, R., & Tranel, D. 1999. Intact recognition of emotional prosody following amygdala damage. Neuropsychologia, 37, 1285–1292.

[2] Alba-Ferrara, L., Hausmann, M., Mitchell, R. L., & Weis, S. 2011. The Neural Correlates of Emotional Prosody Comprehension: Disentangling Simple from Complex Emotion. PloS One, 6(12), e28701.

[3] Beaucousin, V., Lacheret, A., Turbelin, M. R., Morel, M., Mazoyer, B., & Tzourio-Mazoyer, N. 2006. FMRI

Study of Emotional Speech Comprehension. Cerebral Cortex, 17(2), 339–352.
[4] Bergman, P., Sköld, A., Västfjäll, D., & Fransson, N. 2009. Perceptual and emotional categorization of sound. The Journal of the Acoustical Society of America, 126(6), 3156.
[5] Buchanan, T. W., Lutz, K., Mirzazade, S., Specht, K., Shah, N. J., Zilles, K., & Jancke, L. 2000. Recognition of emotional prosody and verbal components of spoken language: an fMRI study. Cognitive Brain Research, 9(3), 227–238.
[6] Chuenwattanapranithi, S., Xu, Y., & Thipakorn, B. 2009. Encoding emotions in speech with the size code. Phonetica, 65, 210–230.
[7] Dellwo, V., Leemann, A., & Kolly, M. J. 2015. The recognition of read and spontaneous speech in local vernacular: The case of Zurich German. Journal of Phonetics, 48(C), 13–28.
[8] Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., et al. 2009. Universal Recognition of Three Basic Emotions in Music. Current Biology, 19(7), 573–576.
[9] Juslin, P. N. 2001. Communicating emotion in music performance: A review and a theoretical framework. Oxford University Press.
[10] Juslin, P. N., & Laukka, P. 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin, 129(5), 770–814.
[11] Juslin, P. N., & Laukka, P. 2006. Emotional Expression in Speech and Music. Annals of the New York Academy of Sciences, 1000(1), 279–282.
[12] Lieberman, P., & Michaels, S. B. 1962. Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. The Journal of the Acoustical Society of America, 34(7), 922.
[13] Majid, A. 2012. Current Emotion Research in the Language Sciences. Emotion Review, 4(4), 432–443.
[14] Mitchell, R. L. C. 2013. Further characterization of the functional neuroanatomy associated with prosodic emotion decoding. Cortex, 49(6), 1722– 1732.
[15] Murray, I. R., & Arnott, J. L. 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. The Journal of the Acoustical Society of America, 93(2), 1097–1108.
[16] OLSA Oldenburger Satztest. version 2011
[17] Pakosz, M. 1982. Intonation and attitude. Lingua, 56(2), 153–178.
[18] Park, C. H., & Sim, K. B. 2003. Emotion recognition and acoustic analysis from speech signal. Neural Networks.
[19] Pell, M. 1999. Fundamental Frequency Encoding of Linguistic and Emotional Prosody by Right Hemisphere-Damaged Speakers. Brain and Language, 69(2), 161–192.
[20] Pell, M. D. 2006. Cerebral mechanisms for understanding emotional prosody in speech. Brain and Language, 96(2), 221–234.
[21] Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. 2009. Factors in the recognition of vocally expressed emotions: A comparison of four languages. Journal of Phonetics, 37(4), 417–435.
[22] Peretz, I. (1998). Music and emotion: perceptual determinants, immediacy, and isolation after brain damage. Cognition, 68(2), 111–141.
[23] Scherer, K. R., Banse, R., & Wallbott, H. G. 2001. Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. Journal of Cross-Cultural Psychology, 32(1), 76–92.
[24] Streeter, L. A., Macdonald, N. H., & Apple, W. 1983. Acoustic and perceptual indicators of emotional stress. Journal of the Acoustic Society of America, 73(4), 1354–1360.
[25] Szameitat, D. P., Alter, K., Szameitat, A. J., Wildgruber, D., Sterr, A., & Darwin, C. J. 2009. Acoustic profiles of distinct emotional expressions in laughter. The Journal of the Acoustical Society of America, 126(1), 354.
[26] Toivanen, J., Väyrynen, E., & Seppänen, T. 2004. Automatic Discrimination of Emotion from Spoken Finnish. Language and Speech, 47(4), 383–412.
[27] Whiteside, S. P. 1998. Simulated emotions: an acoustic study of voice and perturbation measures. Icslp.
[28] Zentner, M., Grandjean, D., & Scherer, K. R. 2008. Emotions evoked by the sound of music: Characterization, classification, and measurement. Emotion, 8(4), 494–521.