Context and priming effects in the recognition of emotion of old and young listeners

Martijn Goudbeek, Marie Nilsenová

Tilburg Center for Communication and Cognition, Tilburg University, Tilburg, The Netherlands

m.b.goudbeek@uvt.nl, m.nilsenova@uvt.nl

Abstract

The development of our ability to recognize (vocal) emotional expression has been relatively understudied. Even less studied is the effect of linguistic (spoken) context on emotion perception. In this study we investigate the performance of young (18-25) and old (60-85) listeners on two tasks: an emotion recognition task where emotions expressed in a sustained vowel (/a/) had to be recognized and an emotion attribution task where listeners had to judge a neutral fragment that was preceded by a phrase that varied in speech rate and/or loudness. The results of the recognition task showed that old and young participants do not differ in their recognition accuracy. The emotion attribution task showed that young listeners are more likely to interpret neutral stimuli as emotional when the preceding speech is emotionally colored. The results are interpreted as evidence for diminished plasticity later in life.

Index Terms: vocal emotion decoding, context effects

1. Introduction

The accuracy with which emotions are recognized in the voice develops over the lifespan. In a study comparing young (mean age 23.44, $\sigma = 2.06$) and middle aged (mean age 42.63, $\sigma = 3.02$) listeners, Paulmann, Pell and Kotz [1] showed that young listeners are better than middle aged listeners at judging vocal emotional expressions. The authors point to a number of possible explanations for the performance difference between old and young listeners: they mention general cognitive factors such as a decline in working memory performance or a decline in the ability to sustain attention [1]. However, Paulmann et al. favor the idea that emotion specific factors are the reason for the decline in performance. One such factor is the degradation of specific neuroanatomical networks in the right hemisphere that are responsible for emotion recognition, another possible factor is the use of different prosodic cues and processes by both age groups when decoding an emotional utterance [1].

Here, we would like to argue for a possibility not explicitly mentioned by Paulmann et al., namely decreasing perceptual plasticity over the lifespan [2]. If older listeners are not as good at using information from the context, such as the characteristics of the speaker and the linguistic context, they should be outperformed by young listeners, especially when there is a lot of contextual information, such as in emotional sentences in dialog. On the other hand, when there is little or no additional information about the speaker or the linguistic context, such as in an isolated sustained vowel, old and young listeners should perform equally well.

Emotion expression recognition is usually studied by presenting listeners with isolated emotional expressions that subsequently have to be judged on their emotional content [3]. However, emotional expressions do seldom occur in isolation. They are usually part of an ongoing interaction between speakers. A growing body of evidence suggests that in these interactions, speakers adapt to the phonetic and paralinguistic peculiarities of each other's speech in order to promote lexical access [4, 5]. In the domain of facial emotional expression several studies have shown the role of context in perceiving emotion from facial expressions [6, 7]. It stands to reason that listeners should also adopt the way they interpret the nonverbal cues to the emotional state of their dialog partners. After all, speakers differ individually in the way they express emotions and listeners would be well advised to be sensitive to these differences.

Whether or not speakers take the surrounding context into account when decoding emotional utterances is a topic that has been relatively understudied. The exception is the work of Cauldwell [8] and Scherer, Ladd and Silverman [9]. Both studies show that the decoding of emotional information in the vocal modality is affected by the preceding context. In Cauldwell's study, the interpretation of the phrase "What do you mean" in a dialog between a father and his son depends greatly on whether the listeners are presented with the preceding context or not. Specifically, the phrase is judged as angry far more often without context than within context [8]. Similarly, Scherer, Ladd and Silverman showed that the linguistic context of an expression affects the interpretation of the emotion expressed [9]. We argue that this flexibility in the perception of emotion results from the plasticity of the perception apparatus, comparable perhaps to the adjustments listeners make when confronted with speakers with an accent or a speech impediment [5]. Here, we will investigate the role of different preceding speech contexts on the emotional interpretation of a neutral phrase.

For the speech manipulation in our second task, we choose to vary the speech rate and the loudness of the preceding speech fragment. Both speech rate and loudness have been extensively shown to be important cues for vocal emotional decoding and encoding. Loud and fast speech has been linked to emotions such as anger and joy, whereas soft and slow speech has been linked to emotions such as sadness and tenderness [10, 11, **?**]. If a preceding speech fragment influences the perception of a subsequent neutral fragment, than this influence should be revealed by our parametric manipulation of loudness and speech rate.

Here, we present the results of two tasks that were administered to a group of young and a group of old participants: an emotion *recognition* task, and an emotion *attribution* task. In the emotion recognition task listeners have to judge the emotional tone of sustained vowel. Because such a short fragment does not afford the use of context or perceptual plasticity, we predict that both groups will perform equally well on this task. In the emotion attribution task participants had to judge the answer to a preceding question that was manipulated in terms of speech rate and loudness. Because this is a longer fragment and taking context into account requires plasticity and flexibility, we predict that the manipulated fragments will have a greater influence on the young participants. In other words, because of their higher sensitivity to context, young listeners will be more likely to attribute an emotion to a neutral speech fragment when it is preceded by faster, slower, louder, or softer speech.

2. Method

2.1. Participants

Two groups of participants took part in the emotion recognition and the emotion attribution task; a group of young (18-25, mean age 21,4, σ = 2.31) and a group of old (60-85, mean age 73,6, σ = 8,12) participants. The young participants were students from the University of Tilburg and the old participants were residents of a local nursery home and their acquaintances. All young participants reported normal hearing. Some of the elderly participants were hard of hearing, resulting in a number of unanswered trials in the emotion recognition task, especially with the low intensity emotions sadness and tenderness.

2.2. Materials

Emotion recognition task

The stimuli for the emotion recognition task were 35 expressions of the emotions *anger*, *sadness*, *joy*, and *tenderness* using the vowel /a/ that were drawn from the Geneva Multimodal Emotion Portrayals [13]. Three emotions were expressed by 10 actors (5 males, 5 females), while tenderness was only expressed by 5 actors. Because of their low intensity level and the possibility of hearing difficulties in the elderly population, the intensity of the portrayals of sadness and tenderness was increased to 65 dB.

Emotion identifications and intensity ratings were collected with a paper and pencil version of the Geneva Emotion Wheel (see figure 1) developed by Scherer [3]. Participants were instructed to indicate which emotion they heard by selecting the appropriate spoke in figure 1 and to indicate the intensity of that emotion by selecting the appropriate circle (where larger means more intense).

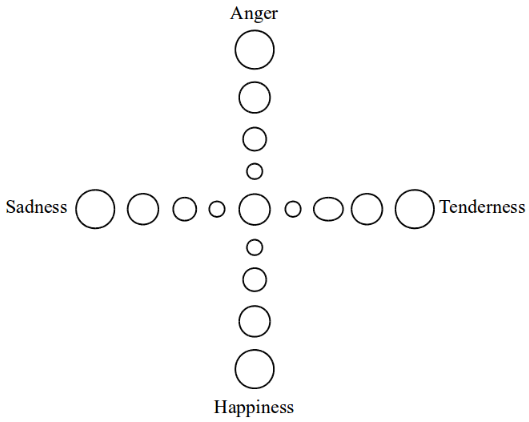

Figure 1: Our paper and pencil version of the Geneva Emotion Wheel.

Emotion attribution task

The stimuli for the emotion attribution task consisted of a question (the prime) and an answer (the target utterance). The interaction was similar to that in Cauldwell [8]: two actors, a male and a female, act out a dialog where the question is always asked by the female. After recording, the utterances were pretested for neutrality and the 11 most neutral expressions were selected. Of these selected expressions, the speech of the woman (the questions) was further processed with PRAAT [14]. In this post processing the speech rate and the loudness of the utterances were parametrically varied at three levels (faster, neutral, slower –using time scaling–, and louder, neutral, softer, respectively), leading to nine conditions (see table 1). The criteria for the manipulations were based on measurements on the expressions of anger, joy, sadness and tenderness in the GEMEP corpus [13]. Based on these measurements, fast speech was defined as an increase in speech rate with a factor of 1.17 while slow speech was defined as a decrease with a factor of 0.76. For loudness, extensive pretesting revealed that adherence to the criteria for soft and loud speech derived from the GEMEP corpus would result in soft speech that would not be audible for the old participants. After further testing, soft speech was defined at 50 dB, neutral speech at 65 dB and loud speech was set at 80 dB.

Table 1: *The two manipulations (speech rate and loudness) resulting in nine primes in the emotion attribution task*

Condition	Speech rate	Loudness
1	-	+
2	-	-
3	-	0
4	+	+
5	+	-
6	+	0
7	0	+
8	0	-
9	0	0

Importantly, the loudness and the speech rate of the *preceding context* (the question) in the dialog was manipulated, the speech rate and loudness of the target utterance (the answer) were left unaltered.

2.3. Procedure

Both groups first performed the emotion recognition task and then did the emotion attribution task. Participants were tested individually in a quiet environment and could do the tasks at their own pace. They were allowed to listen to the stimuli as often as they wanted. In the emotion recognition task, participants indicated on the emotion wheel which emotion (anger, joy, sadness, tenderness or neutral) best fitted each of the 35 emotional /a/'s. In addition, the participants indicated the intensity of the emotion that was expressed. In the emotion attribution task, participants were instructed to listen to 99 dialogs (11 sentences in 9 conditions) and judge the emotion expressed (if any, they could also -and often did- judge a fragment as neutral) in the *answer* to the question. Again, the participants were also asked to indicate the intensity of the emotion on the emotion wheel.

3. Results

3.1. Emotion recognition task

Figure 2 displays the mean recognition accuracy for each emotion for the old and the young listeners. As is often found in vocal emotion recognition studies, basic emotions with a high arousal level such as anger are recognized best and low intensity emotions are recognized less well [9, 10]. Nevertheless, all emotions were recognized significantly better than the chance level of 0.20 (t_{min} [58] = 8.74).

To investigate differences in recognition accuracy between the four emotions and, more importantly, between the two age groups, we performed a mixed effects analysis of variance with emotion as within-subject variable with four levels (anger, joy, sadness, tenderness) and age group as between-subject variable with two levels: old and young, and with recognition accuracy as the dependent variable. This analysis revealed a significant effect of emotion (F [3,165] = 107.38, p $<$ 0.001, η^2 = 0.66) showing that the four emotions present in the study differed in how well they were recognized. Importantly, there was no significant main effect of age group (F [2,17] = 1.15, *ns*). In addition, none of the interactions were significant (all $F < 1.9$), providing clear evidence that old and young listeners recognize emotions expressed in these sustained vowels equally well.

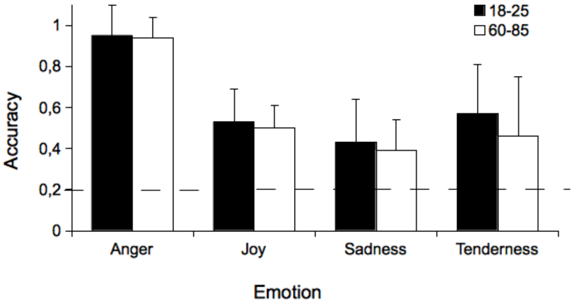

Figure 2: Judged accuracy per emotion for each age group

A similar pattern is found in the intensity ratings depicted in figure 3. Again, there are differences in the intensity levels for the emotions, but this effect of emotion is the same for both age groups. We investigated these effects in a mixed effect analysis of variance like the one above with the intensity rating as dependent variable. As was the case for accuracy, the emotions differed significantly from each other in their accuracy rating (F [3,198] = 269.84, p $<$ 0.001, η^2 = 0.80). In line with the accuracy judgments, there was no difference between the intensity ratings of young and old participants (F [1,64] = 0.026, *ns*). In addition, none of the two- or three-way interactions were significant (all $F < 1.01$), indicating that old and young listeners judge emotional intensity expressed in sustained /a/'s in the same way.

3.2. Emotion attribution task

Table 2 displays the proportion of neutral attributions to the neutral utterance for each preceding context manipulation. In addition, the number of neutral attributions to the utterance without manipulation of the preceding stimulus are shown. Because the intensity ratings in the emotion attribution task were extremely low for every condition, they will not be reported here. Given that the fragment that the listeners had to judge was neutral, any (systematic) deviation from neutral would be evidence for the role of preceding context. As can be inferred from table 2, the preceding context clearly had an effect on the perception of the neutral phrase in that it is judged as neutral in at most 74 % of the time, while the rest of the time, the manipulated preceding context results in the attribution of an emotion to the acoustically neutral speech fragment.

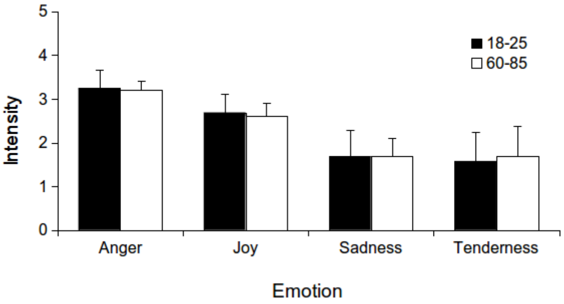

Figure 3: Judged intensity per expressed emotion for each age group

Table 2: *The mean proportion neutral attribution and the standard deviation in the nine primes in the emotion attribution task*

		18-25		**60-85**	
Speed	Loudness	μ	σ	μ	σ
Slow	Loud	0.42	0.24	0.68	0.23
Slow	Soft	0.51	0.27	0.73	0.19
Slow	Neutral	0.51	0.29	0.69	0.26
Fast	Loud	0.51	0.25	0.70	0.23
Fast	Soft	0.47	0.27	0.74	0.26
Fast	Neutral	0.43	0.24	0.66	0.21
Neutral	Loud	0.52	0.29	0.74	0.19
Neutral	Soft	0.44	0.26	0.74	0.26

A mixed effects analysis of variance with percentage neutral attribution as dependent variable and loudness and speech rate as independent within-subject variables (both with three levels; soft/slow, neutral, and loud/fast) and age group as between-subjects variable showed a significant main effect of speech rate (F [2,128] = 6.99, $p < 0.001$, η^2 = 0.10), but not of loudness (F [2,128] = 1.66, $p < 0.19$, η^2 = 0.03), possibly due to our older participants' hearing difficulties. A large significant effect of age group (F [1,64] = 21.45, $p < 0.001$, η^2 = 0.25) showed that young listeners are more influenced by context than old listeners. Crucially, the effect of context is not only larger in young listeners, the way young listeners respond to changes in speech rate and loudness is also different in kind from the response of old listeners, as can be inferred from the significant interactions of age group with loudness (F [2,128] = 3.66, $p < 0.029$, η^2 = 0.05) and age group and speech rate (F [2,128] = 7.46, $p < 0.001$, η^2 = 0.10).

4. Discussion

Two experiments investigated emotion recognition and emotion attribution in old and young listeners. In the emotion recognition task, the participants had to judge the identity and intensity of emotions expressed in the sustained vowel /a/. While the expected differences between the emotions were found, the results revealed that the old and young listeners did not differ with respect to the accuracy of their judgments or in the way they judged the intensity of the emotions. In the emotion attribution task, the participants had to listen to a question that varied in speech rate and loudness and had to judge the possible emotion (if they thought one was present) in a neutrally phrased answer. The results reported here show that the manipulations of the preceding context had a clear effect in that at least 26 % (and often more) of the neutral fragments were perceived to express an emotion. Importantly, this effect of preceding context was much stronger in young listeners, providing evidence for the hypothesis that young listeners are more able to take context into account compared to older listeners.

Our data can be considered to be at odds with the data of Paulmann et al. [1]. Contrary to their finding that younger listeners are better at decoding emotional information from the voice, we find that when little or no context is available, such as in our emotion recognition task, old and young listeners perform equally well. This lack of a difference between old and young listeners could be thought of as a floor effect, where both groups perform equally because they have reached the lower limit of their performance, but the high recognition rates in both groups preclude such an interpretation. The results of the emotion attribution task show that young listeners are more likely than old listeners to attribute an emotion to a neutral stimulus.

However, when the results from the emotion recognition and the emotion attribution task are combined, our findings also provide a way to flesh out the hypothesis of Paulmann et al. [1] by providing a clear-cut cognitive domain where young and old listeners differ in the way they listen to emotional speech; perceptual plasticity. Without contextual information, such as in our emotion recognition task, the diminished perceptual plasticity of older listeners is of no consequence. In contrast, when there is a lot of contextual information present, young listeners take context into account to a much larger extend than old listeners do when they are judging emotional expressions. As a consequence, younger listeners perceive neutral stimuli more often as emotional than old listeners do.

In the emotion attribution task the fragment to be judged was always neutral, so that every judgment that perceives the fragment as expressing a particular emotion can be thought of as an incorrect attribution. Interpreted in this way, old listeners have a higher accuracy than young listeners (because our older listeners are less likely to attribute emotion to the neutral phrase). However, this is not our favored interpretation, since we consider the sensitivity of listeners to context when judging emotional expressions an important strategy when decoding emotional expressions. Thus, while the plasticity and sensitivity to context of the young participants leads to the incorrect attribution of emotion in this case, it is beneficial to emotion perception in general [6, 7].

In this paper, we have shown that the preceding linguistic context influences the interpretation of a neutral succeeding phrase and that this influence is stronger in younger listeners. The precise influence of the preceding context, which emotions are triggered by which preceding context manipulations, has not been investigated yet. The design of our experiment enables an analysis of the influence of both speech rate and loudness on the attribution of each response category (anger, joy, tenderness, sadness, and neutral). In future work we plan to analyze the pattern of responses related to each stimulus manipulation. This analysis will also elucidate the issue of whether the effect of context should be construed as a form of affective *priming* or a form of *stimulus adaptation*.When the preceding context serves as a prime, the neutral phrase should be interpreted as an emotion fitting to that context. For example, loud and fast speech should result in increased attribution of anger and possibly joy. On the other hand, when listeners perceptually adapt to the preceding context, the neutral phrase will be interpreted in contrast to the preceding utterance. In that case, fast and loud speech should lead to an increase of attributions of tenderness or sadness.

5. Conclusion

Two experiments have provided evidence for the hypothesis that listeners are sensitive to (linguistic) context when interpreting emotional speech and that this sensitivity diminishes with age. When judging isolated vowels that express emotion, old and young listeners perform equally well. However, when they have to judge a neutral utterance that is preceded by a manipulated phrase differing in speech rate and loudness, young listeners are more sensitive to context than old listeners.

6. Acknowledgements

We would first like to thank all participants from nursing home “The Kievitshorst” in Tilburg. In addition, the authors gratefully acknowledge the assistance of Evelien Louvenberg and Amanda Ooijen with the data collection and data analysis. Martijn Goudbeek was supported by funding from the VICI project “Bridging the gap between psycholinguistics and computational linguistics: the case of referring expressions”, funded by the Netherlands Organisation for Scientific Research (NWO grant 277-70-007, awarded to Emiel Krahmer).

7. References

[1] Paulmann, S., Pell, M.D., and Kotz, S.A. “How aging affects the recognition of emotional speech”, Brain Lang, 104:262-269.

[2] Burke, S.N., and Barnes, C.A. “Neural plasticity in the ageing brain”, Nat Rev Neurosci, 70:30-40,2006.

[3] Scherer, K. “What are emotions and how can they be measured?”, Soc Sci Inform, 44: 693-727, 2005.

[4] Eisner, F., and McQueen, J.M. “The specificity of perceptual learning in speech processing”, Perc Psychophys, 67:224-238, 2005.

[5] Kraljic, T, Samuel, A. G., and Brennan, S. E. “First impressions and last resorts: How listeners adjust to speaker variability”, Psych Science, 19:332-338, 2008.

[6] Righart, R., and de Gelder, B. “Recognition of facial expressions is influenced by emotional scene gist”, Cogn Affect Behav Ne, 8:264-272, 2008.

[7] Wallbott, H.G., “In and out of context: Influences of facial expression and context information on emotion attributions”, Brit J Soc Psychol, 27:357-369, 1988.

[8] Cauldwell, R.T. “Where did the anger go? The role of context in interpreting emotion in speech”, SpeechEmotion, 127-131, 2000.

[9] Scherer, K.R., Ladd, D.R., and Silverman, K. “Vocal cues to speaker affect: testing to models”, J Acoust Soc Am, 75:1346-1356, 1984.

[10] Banse, R., and Scherer, K.R. “Acoustic profiles in vocal emotion expression”, J Pers Soc Psychol., 70:614-636, 1996.

[11] Juslin, P., and Laukka, P. “Communication of emotions in vocal expression and music performance: Different channels, same code?”, Psych Bull., 129:770-814, 2003. [12]ES12 Schrder, M. “Expressing degree of activation in synthetic speech”, IEEE T Audio Speech, 14: 1128-1136, 2006.

[13] Bänziger, T., and Scherer, K.R. “Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus”, ACII, 476–487, 2007.

[14] Boersma, P., “Praat, a system for doing phonetics by computer”, Glot International, 5:341-345, 2001.