



The Relation Between Pitch Perception Preference and Emotion Identification

Marie Nilsenová, Martijn Goudbeek, Luuk Kempen

Department of Communication and Information Sciences, Tilburg University, The Netherlands

{m.nilsenova; m.b.goudbeek; l.j.i.kempen}@uvt.nl

Abstract

In our study, we explore the effect of synthetic vs analytic listening mode on the identification of emotions. Numerous psychoacoustic studies have shown that listeners differ in how they process complex sounds; some listeners focus on the fundamental frequency while others attend to the higher harmonics. The difference appears to have a neurological basis, expressed in a leftward (for F_0 listeners) or rightward (for spectral listeners) asymmetry of gray matter volume in the lateral Heschl's gyrus. In our experiment we found that spectral listeners performed better in an emotion judgment task, which is what we expected based on the fact that the processing of emotional prosody is relatively right-hemisphere lateralized.

Index Terms: emotion identification, pitch perception

1. Introduction

Why are some people better at detecting emotions than others? Emotional intelligence is considered to be a complex set of an individual's abilities to perceive emotions [1], fundamentally based on the capacity to process prosodic information present in the speaker's voice, face and gesture/posture [2]. Arguably, vocal expression is the most important predictor of emotions in everyday life [3]. A number of features – none of them sufficient or necessary [4] – add up in a strongly speaker-dependent manner to signal a speaker's emotional state [5], [4], [6]. For example, happiness, but also anger, can be characterized by fast production rate, high intensity level, high intensity variability, high frequency energy, high F_0 level and variability, rising F_0 contours and fast voice onsets [4]. Listeners processing the emotional speech have to make use of fine distinctions in temporal, pitch and spectral values in order to evaluate correctly the complex set of features.

Interestingly, there are individual differences in how people perceive complex sounds such as speech. As shown in a number of psychoacoustic and neurological studies, for different listeners different aspects of the signal come across as more or less dominant ([7], [8], [9], [10], [11]). On the one hand, some listeners are sensitive to the information encoded in the fundamental frequency and its change in time (so called F_0 listeners, or synthetic listeners). Other listeners are primarily affected by the overall spectral information in the signal (spectral, or analytical listeners). The listener type apparently has a neural basis expressed primarily in a rightward/leftward asymmetry of the lateral Heschl's gyrus ([10], [11], [12]). In particular, MRI and magnetoencephalography studies have shown that spectral listeners, compared to F_0 listeners, appear to have a rightward (rather than a leftward) asymmetry of gray matter volume in what has been called the 'pitch processing center' ([14], [10]). Until fairly recently, the dominant listening mode has only been examined in the context of musical training; the research of Wong and his colleagues, however, indicates that it

may also affect linguistic performance ([13], [12], [15]).

The effect of dominant listening mode has so far not been explored in the context of emotion identification. In our study, we explore the relation between listening mode and the judgments of eight emotions and compare the identification of basic emotions (anger, panic fear, joy, and sadness) to non-basic ones (relief, pride, anxiety, pleasure). Given that the processing of emotional prosody is relatively right-hemisphere lateralized ([16], [17]), we hypothesized that analytic (spectral) listening would be related to a better performance in an emotion judgment task.

In order to classify listener preference, we made use of the classical pitch discrimination task with missing fundamental ([8], [9]). In this task, subjects are presented with two successive complex tones, A and B, both of which consist of the same number of upper harmonic tones with the same highest harmonic but different levels of virtual pitch (derived from the harmonics as the best fit, [18]) and spectral pitch (based on the lowest harmonic). For example, the sequence A = (800 Hz + 1000 Hz) and B = (667 Hz + 1000 Hz) would be perceived as rising by synthetic listeners who derive the missing F_0 to be 200 Hz and 333 Hz, respectively; it would be perceived as falling by analytic listeners who focus on the 800 Hz and 667 Hz harmonics. The listener's performance on the discrimination task is evaluated with the formula for an 'index of pitch perception preference' $\delta_p = \frac{(N_{F_0} - N_{F_{sp}})}{(N_{F_0} + N_{F_{sp}})}$, where N_{F_0} is the number of trials the subject processed in the synthetic mode and $N_{F_{sp}}$ the number of trials processed in the analytic (spectral) mode [10]. The outcome of the discrimination task is somewhat dependent on the experimental conditions [20]; for instance, the number of harmonics in the experimental stimuli favors either the synthetic (for 4 harmonics) or the analytic interpretation (for 2 harmonics), [17]. In our study, we focused on the effect of the duration of the complex tones, since tone length has been shown to have an effect on the perception of pitch at least for some listeners [21], with longer stimuli favoring the synthetic (F_0) interpretation. In the past experiments using the pitch discrimination task, the duration varied between 160 ms ([8]), 400 ms ([17], [9]) to 500 ms ([11]). Our aim was to explore possible within-subject shifts in the analytic/synthetic listening mode caused by duration change. Given that the length of the tone appears to be positively correlated with pitch salience [21], we also expected that subjects would perform better on trials with longer durations relative to short ones.

2. Method

2.1. Participants

The participants were 110 BA students of Communication and Information Sciences drawn from the participant pool of Tilburg University, who participated in the experiment for a partial

course credit. Their age varied from 18 to 44 ($M=23$, $S.D.=4.2$), 80 were female, 30 male, 15 were left-handed and 73 reported at least a moderate experience with music. None of the participants had any history of hearing problems.

2.2. Design

The study consisted of two parts; in the first part, the preferred listening mode of our participants was investigated with a pitch discrimination task. The dependent variable was the participants' δ_p , which reflects their listening mode on a scale from -1 (exclusively spectral, i.e., analytic) to +1 (exclusively $F0$, i.e., synthetic). As an independent factor, the tones were presented at two stimulus durations: 160 ms and 600 ms.

The second part of the study consisted of an emotion judgment task. In this task, participants had to indicate by forced choice what emotion they perceived in sound clips of eight emotions that were expressed with the vowel /a/. The dependent variable in this study was the percentage correct identification for each expressed emotion.

2.3. Materials

For the pitch discrimination task, we constructed 72 pairs of complex harmonic tones that consisted of two, three, or four harmonics. The harmonic composition of the stimuli was derived from the procedure employed by [9]. A half of the tone pairs were control stimuli in that they had an unambiguous tonal progression (e.g., the second tone was unambiguously higher or unambiguously lower in pitch than the first). The other 36 pairs contained a missing fundamental and were constructed in a way that the perceived sequence of the first tone to the second tone depended on the preferred listening mode of the participants. For 18 pairs a spectral interpretation would result in a higher- lower judgment and a fundamental frequency interpretation would result in a lower – higher judgment. For the other 18 pairs, a spectral interpretation would result in a lower – higher judgment and a fundamental frequency interpretation would result in a higher – lower judgment.

For the emotion judgment test, emotional expressions using the sustained vowel /a/ were drawn from the Geneva Multimodal Emotion Portrayals [22]. The expressions in this corpus have been carefully selected from a larger set recorded in an interactive setting and subjected to several large rating studies to obtain the best selection of emotion expressions. The corpus as a whole contains eighteen emotions expressed by ten actors. We selected eight emotions for our study: joy, anger, pride, panic fear, pleasure, irritation, anxiety and relief. These emotions include four of the so called basic emotions (anger, panic fear, joy, and sadness) and four of the so called non-basic emotions (relief, pleasure, pride, and anxiety) and represent a balanced set in terms of their arousal and valence properties. Each actor's best portrayal of each emotion (as determined by the rating studies conducted at the University of Geneva) was entered into the emotion judgment task. This resulted in 80 stimuli (10 actors x 8 emotions).

2.4. Procedure

All participants first completed the pitch discrimination task and then completed the emotion judgment task. To control for possible order effects, we constructed four different stimulus lists of each task and randomly assigned participants to a stimulus list. To lower the memory demands for the participants, each stimulus pair was presented twice, separated by a short burst

of noise that signaled a reset to the auditory short term memory. The stimuli were presented over high quality headphones (Sennheiser EH 250 and Beyer Dynamic DT 250). Each participant completed the test with paper and pencil, individually in a quiet room with the experimenter present. For the pitch discrimination task, the participants were asked to circle the picture best representing the sequence of their choice (e.g., higher or lower); for the emotion judgment task, they could select the emotion of their choice or choose a neutral option.

3. Results

3.1. Listening mode and emotion judgment

Based on the performance on the pitch discrimination task, we first examined the preferred listening mode of our experimental participants (calculated as δ_p). As figure 1 shows, our sample was predominantly composed of spectral listeners.

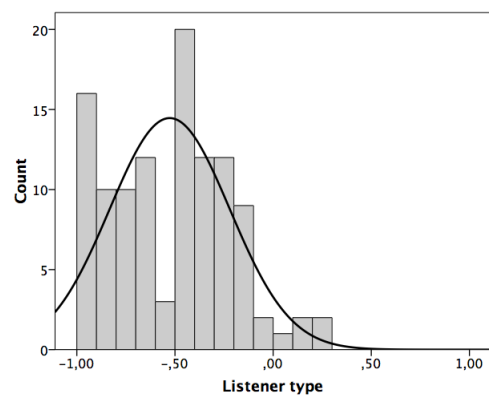


Figure 1: Histogram of δ_p of all 110 participants. The leftward skew indicates that most of our listeners had a spectral interpretation of the stimuli.

For subsequent analyses, we included only those participants who exhibited good listening skills (a percentage correct on the control stimuli higher than 80%); this resulted in a sample of seventy participants ($M_{\delta_p} = -0.69$, $SD = .22$, $Min_{\delta_p} = -1.00$, $Max_{\delta_p} = -.22$). We then investigated the relationship between listening mode and the ability to correctly identify vocal emotional expression. Figure 2 depicts the relation between listener type (for both stimulus durations) and emotion identification. From Figure 2, a moderate negative relationship between δ_p and proportion correct can be determined: participants with a spectral listening mode tend to be better at emotion identification. To quantify this observation, we calculated the correlation between each participant's δ_p -expressing their preference for either spectral interpretations or fundamental frequency interpretations- and their mean proportion correct emotion identification. We also calculated this correlation for each participant's δ_p for the 160 ms and 600 ms stimuli separately. Table 1 shows a significant moderate correlation between listener preference and proportion correct emotion identification. This correlation is stronger and more significant for the 600 ms stimuli. There was no significant correlation between general listening skills (measured with control stimuli in the pitch discrimination task) and percentage correct on the emotion identification task ($r = .06$, $p = .625$). Finally, we also tested for possible gender effects. A one-way analysis of covariance with δ_p as the covariate showed that female participants performed

slightly better in emotion identification than male participants, $F(1,67) = 4.061, p < .05$.

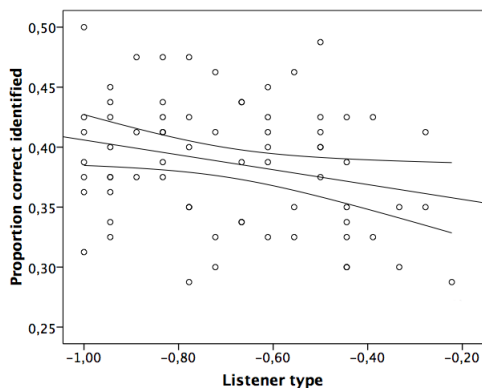


Figure 2: Scatterplot depicting the relationship between listener type and emotion identification for all stimulus durations. The lines represent the mean and the 95 % confidence interval.

	δ_p	δ_{p160}	δ_{p600}	PC_{emo}	
Total	δ_p	1.00	0.95***	0.87***	-0.26*
	δ_{p160}	0.95***	1.00	0.68***	-0.21
	δ_{p600}	0.87***	0.68***	1.00	-0.29**
	PC_{emo}	-0.26*	-0.21	-0.29**	1.00

Table 1: Correlations between the preferred listening modes and the mean proportion of correctly identified emotions, $N = 70$ (* indicates $p < .05$, ** $p < .01$, *** $p < .001$).

Figure 3 shows the results of the emotion judgment task for all eight emotions in the study. The high scores for anger and relief are indicative of the fact that the acoustic profiles of these emotions are best preserved when the emotion is expressed with just the vowel /a/. For anger, this is likely caused by the high loudness and pitch of the acoustic realization of this emotion and for relief, the breathy aspiration caused by the sighs that often accompany the expression of relief can also easily be preserved in a sustained vowel.

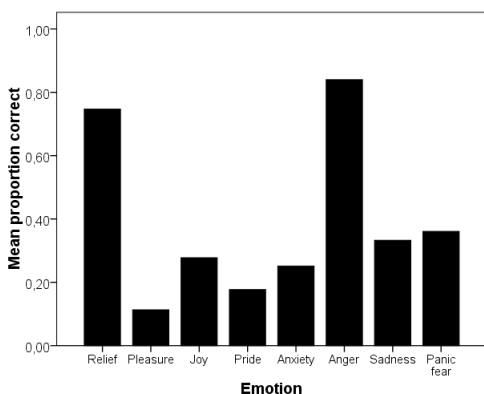


Figure 3: The mean proportion correct for all eight emotions.

Emotions that are typically expressed by changing the prosodic parameters of an utterance were less easily recognized. Nevertheless, with the exception of pleasure, a one sample t-test revealed that all emotions were recognized above the chance level of 0.11 ($t_{min} [69] = 8.71, p < 0.001$).

We tested if basic emotions would be better recognized than non-basic emotions by comparing each participant’s mean proportion correct for the basic emotions (anger, panic fear, joy, sadness) with the mean proportion correct for the non-basic emotion (relief, pride, anxiety, pleasure). A paired samples t-test revealed a significant mean difference of 0.14 between the respective means ($t [69] = 10.31, p < 0.01$), indicating that basic emotions expressed in sustained vowels are indeed, on average, better recognized than non-basic emotions.

3.2. Effect of length on pitch discrimination

To test our hypothesis that longer stimulus durations would affect the participants’ judgments, we investigated whether overall performance was better with the 600 ms stimuli. Figure 4 shows the proportion correct of the control stimuli judgments separately for the two stimulus durations. Judging by the data in figure 4, choosing the correct sequence is easier with longer stimulus durations. An analysis of variance with proportion correct as dependent variable and stimulus duration as independent variable indicated that the advantage for longer stimulus durations was significant ($F(1,69) = 19.01, p < 0.001, \eta^2 = 0.22$).

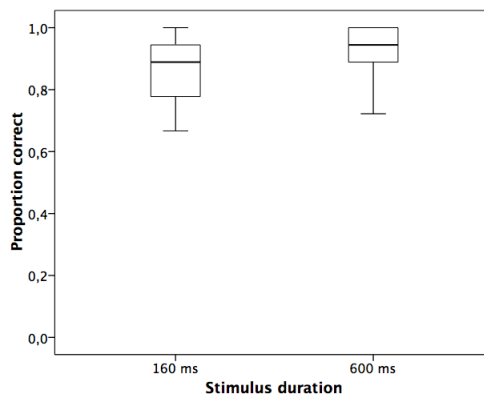


Figure 4: The mean proportion correct for the control stimuli for the two stimulus durations.

We also analyzed the relation between stimulus duration and listening mode by conducting a within-subjects analysis of variance with listening type (expressed in δ_p) as dependent variable and stimulus duration as within-subjects factor. Figure 5 shows the mean, standard deviation and 95% confidence interval for this analysis. As can be seen from the figure, stimulus durations of 600 ms result in a mean δ_p of -0.80, while stimulus durations of 160 ms result in a mean δ_p of -0.59. Both means indicate a preference for a spectral interpretation, but this preference is significantly stronger for the longer stimuli ($F [1,69] = 71.28, p < 0.001, \eta^2 = 0.51$).

4. Conclusions & Discussion

In a perceptual experiment consisting of a pitch discrimination task and an emotion judgment task, we found that listeners with a more pronounced analytic (spectral) listening mode were better at identifying both basic (anger, panic fear, joy, sadness) and non-basic emotions (relief, pride, anxiety, pleasure). This finding is in line with earlier research showing an apparent right-hemisphere advantage in processing emotional prosody and the rightward asymmetry in gray matter volume in the ‘pitch processing center’ of analytic listeners. We also found that our listeners were better at identifying the four basic emotions com-

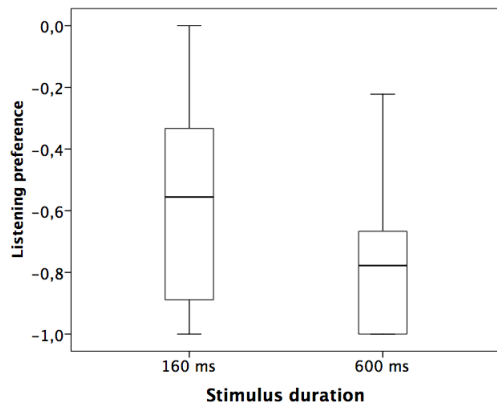


Figure 5: The mean listener type (-1 means exclusively analytic, +1 means exclusively synthetic) for the two stimulus durations.

pared to the four non-basic ones. It remains to be seen if the result regarding the relation between preferred listening mode and emotion identification can be replicated for other emotions that were not included in our experiment.

Contrary to previous experimental results [21], we found that longer stimulus duration (600 ms compared to 160 ms) supports the analytic interpretation of ambiguous tone sequences with a missing fundamental. The contradictory results can probably be accounted for by the fact that the earlier finding was based on an experiment with only four subjects and with a different type of pitch discrimination task. A longer stimulus duration also supported the correct perception of the non-ambiguous trials and correlated more strongly with the number of correctly identified emotions.

In our experiment, we employed the standard pitch discrimination task used in other studies ([8], [9], [10]). In the participant pool, there turned out to be a clear bias in favor of the analytic listening mode. While this was not the case for [10] who present a balanced sample of both synthetic and analytic listeners, some researchers report that the analytic listening mode is, in fact, used by most (90%) of the listeners [18]. It is possible that the instructions the listeners receive in the discrimination task play a role, in that asking subjects to compare two tones in the sense of higher – lower may support the analytic mode, while instructing them to describe the melodic sequence as rising or falling may support the synthetic mode.

5. Acknowledgments

We thank Marcello Mortillaro, Tanja Bänziger and Klaus Scherer for providing the stimulus material from the GEMEP corpus of emotion expressions, to Jan Volín and Radek Skarntzl for a discussion of the experimental methodology and to Marc Swerts and the Interspeech reviewers for comments on an earlier version of the manuscript. Martijn Goudbeek is being funded by Vici grant 277-70-007 “Bridging the gap between computational linguistics and psycholinguistics: the case of referring expressions” from The Dutch Scientific Research Council awarded to Emiel Krahrer.

6. References

[1] Zeidner, M., Matthews, G. and Roberts, R.D., “What We Know about Emotional Intelligence: How It Affects Learning, Work,

Relationships, and Our Mental Health”, MIT Press, Cambridge, MA, 2009.

[2] Mayer, J.D., Salovey, P., and Caruso, D., “Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) User’s Manual”, Multi-Health Systems, Toronto, Canada, 2002.

[3] Planalp, S., “Communicating Emotion in Everyday Life: Cues, Channels, and Processes”, in P.A. Andersen & L.K. Guerrero [Eds.], *Handbook of Communication and Emotion*, 29–48, New York: Academic Press, 1998.

[4] Juslin, P.N. and Laukka, P., “Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?”, *Psychological Bulletin*, 129(5): 770–814, 2003.

[5] Scherer, K., “Vocal Affect Expression: A Review and a Model for Future Research”, *Psychological Bulletin*, 99:143–165, 1986.

[6] Spackman, M.P., Brown, B.L. and Otto, S., “Do Emotions Have Distinct Vocal Profiles? A Study of Idiographic Patterns of Expression”, *Cognition & Emotion*, 23(8): 1565–1588, 2010.

[7] von Helmholtz, H.L.F., “On the Sensations of Tone”, Longmans, London, 1885.

[8] Smoorenburg, G.F., “Pitch Perception of Two-Frequency Stimuli”, *J. Acoust. Soc. Am.*, 48:924–942, 1970.

[9] Laguitton, V., Demany, L., Semal, C. and Liégeois-Chauvel, C., “Pitch Perception: A Difference Between Right- and Left-Handed Listeners”, *Neuropsychologia*, 36(3): 201–207, 1998.

[10] Schneider, P., Sluming, V., Roberts, N., Scherg, M., Goebel, R., Specht, H.J., Dösch, H.G., Bleeck, S., Stippich, C. and Rupp, A., “Structural and Functional Asymmetry of Lateral Heschl’s Gyrus Reflects Pitch Perception Preference”, *Nature Neuroscience*, 8(9):1241–1247, 2005.

[11] Schneider, P., Sluming, V., Roberts, N., Bleeck, S. and Rupp, A., “Structural, Functional and Perceptual Differences in Heschl’s Gyrus and Musical Instrument Preference”, *Ann. N.Y. Acad. Sci.*, 1060:387–394, 2005.

[12] Wong, P.C.M., Warrier, C.M., Penhune, V.B., Roy, A.K., Sadehh, A., Parrish, T.B. and Zatorre, R.J., “Volume of Left Heschl’s Gyrus and Linguistic Pitch Learning”, *Cerebral Cortex*, 18:828–836, 2008.

[13] Wong, P.C.M. and Perrachione, T.K., “Learning Pitch Patterns in Lexical Identification By Native English-speaking Adults”, *Applied Psycholinguistics*, 28:565–585, 2007.

[14] Griffiths, T.D., “Functional Imaging of Pitch Analysis”, *Ann. NY Acad. Sci.*, 999: 40–49, 2003.

[15] Alexander, J., Wong, P.C.M. and Bradlow, A., “Lexical Tone Perception in Musicians and Nonmusicians”, *Proceedings of Interspeech*, Lisbon, September, 2005.

[16] Mitchell, R.L.C., Elliott, R., Barry, M., Cruttenden, A. and Woodruff, P.W.R., “The Neural Response to Emotional Prosody, As Revealed By Functional Magnetic Resonance Imaging”, *Neuropsychologia* 41:1410–1421, 2003.

[17] Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M.L., Scherer, K. and Vuilleumier, P., “The Voices of Wrath: Brain Responses to Angry Prosody in Meaningless Speech”, *Nature Neuroscience*, 8:145–146, 2005.

[18] Rousseau, L., Peretz, I., Liégeois-Chauvel, C., Demany, L., Semal, C. and Larue S., “Spectral and Virtual Pitch Perception of Complex Tones: An Opposite Hemispheric Lateralization?”, *Brain and Cognition*, 30:303–308, 1996.

[19] Goldstein, J.L., “An Optimum Processor Theory for the Central Formation of the Pitch of Complex Tones”, *J. Acoust. Soc. Am.*, 54:1496–1516, 1973.

[20] Houtsma, A.J.M., “Musical Pitch of Two-Tone Complexes and Predictions by Modern Pitch Theories”, *J. Acoust. Soc. Am.*, 66:87–99, 1979.

[21] Beerends, J.G., “Pitches of Simultaneous Complex Tones”, unpublished Ph.D. diss., University of Eindhoven, 1989.

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