Atten Percept Psychophys (2013) 75:1799–1810
DOI 10.3758/s13414-013-0518-x

Psychoacoustic abilities as predictors of vocal emotion recognition

Eitan Globerson · Noam Amir · Ofer Golan · Liat Kishon-Rabin · Michal Lavidor

Published online: 27 July 2013

Abstract Prosodic attributes of speech, such as intonation, influence our ability to recognize, comprehend, and produce affect, as well as semantic and pragmatic meaning, in vocal utterances. The present study examines associations between auditory perceptual abilities and the perception of prosody, both pragmatic and affective. This association has not been previously examined. Ninety-seven participants (49 female and 48 male participants) with normal hearing thresholds took part in two experiments, involving both prosody recognition and psychoacoustic tasks. The prosody recognition tasks included a vocal emotion recognition task and a focus perception task requiring recognition of an accented word in a spoken sentence. The psychoacoustic tasks included a task requiring pitch discrimination and three tasks also requiring pitch direction (i.e., high/low, rising/falling, changing/steady pitch). Results demonstrate that psychoacoustic thresholds can predict 31% and 38% of affective and pragmatic prosody recognition scores, respectively. Psychoacoustic tasks requiring pitch direction recognition were the only significant predictors of prosody recognition scores. These findings contribute to a better understanding of the mechanisms underlying prosody recognition and may have an impact on the assessment and rehabilitation of individuals suffering from deficient prosodic perception.

Keywords Psychoacoustics · Music cognition · Sound recognition · Audition

E. Globerson (✉) · M. Lavidor
Gonda Multidisciplinary Brain Research Center,
Bar-Ilan University, Ramat-Gan, 52900 Jerusalem, Israel
e-mail: gleitan@zahav.net.il

M. Lavidor
e-mail: michal.lavidor@gmail.com

E. Globerson
Academy of Music and Dance, Givaat Ram Campus, Jerusalem, Israel

N. Amir
Department of Communication Disorders, the Sackler Faculty of Medicine, Tel-Aviv University, PO 39040, Ramat Aviv, Tel Aviv 69978-39040, Israel
e-mail: noama@post.tau.ac.il

O. Golan
Department of Psychology, Bar-Ilan University, Ramat-Gan 52900, Israel
e-mail: golano1@mail.biu.ac.il

L. Kishon-Rabin
Department of Communication Disorders, the Sackler Faculty of Medicine, Tel-Aviv University, Ramat Aviv, Tel Aviv 69978-39040, Israel
e-mail: lrabin@post.tau.ac.il

Introduction

When we speak, we can express pragmatic and emotional meaning, not only through words, but also by changing certain attributes of our voice, such as fundamental frequency (*f*0), perceived as pitch, intensity (perceived as loudness), and duration. An all-encompassing term for such acoustic attributes of speech is *prosody*. Indeed, identical phrases may convey completely different pragmatic or emotional information, depending on their prosody. For example, an utterance conveying a joyful feeling is likely to be loud and display a relatively large *f*0 range, whereas a sad utterance would tend to be softer with a smaller *f*0 range (Sobin & Alpert, 1999).

Prosody recognition has been described as a multistep process, involving both sensory and cognitive mechanisms. Auditory perceptual mechanisms perform an initial acoustic analysis of the speech signal, followed by higher cognitive mechanisms, which derive pragmatic and emotional meaning from the acoustic components, employing preexisting socio-emotional scripts (Schirmer & Kotz, 2006). Studies

investigating these cognitive mechanisms have associated vocal emotion recognition with the ability to comprehend others' mental and emotional states, commonly referred to as *theory of mind* (Kleinman, Marciano, & Ault, 2001; Rutherford, Baron Cohen, & Wheelwright, 2002) or *emotional intelligence* (Trimmer & Cuddy, 2008). These studies have viewed vocal emotion recognition as part of a more general emotion recognition mechanism, guiding the attention and perception of emotional cues through the different sensory channels (Adolphs, 2003).

A relatively limited number of studies have highlighted the importance of *auditory* perceptual abilities to the perception of prosody. These studies have focused mainly on populations representing the extreme ends of auditory capacities. For example, research conducted with hearing-impaired individuals showed that intonation, stress, and emphasis in voice are difficult to perceive by many individuals with severe hearing loss (Gold, 1987), including individuals with cochlear implants (Most & Peled, 2007). Other studies demonstrated that musicians, who represent better than average auditory capabilities, exhibit enhanced performance in identifying emotion in vocal utterances (Thompson, Schellenberg, & Husain, 2004).

The results of the above studies, focusing on unique populations, support an association between auditory abilities and the ability to perceive prosodic cues in speech. This association, however, has not been previously examined in the general population. The reason for this absence of data may be the common belief that auditory perceptual abilities (such as pitch discrimination) of individuals with intact hearing are sufficient to extract the acoustic information needed to infer meaning from the prosodic attributes of speech. A closer look, however, at prior research investigating vocal emotion recognition in the general population shows rather low scores (approximately 60%) for recognition of emotions conveyed in voice (e.g., Fenster, Blake, & Goldstein, 1977; Hammerschmidt & Jurgens, 2007; Scherer, 2003; Sobin & Alpert, 1999). There are two possible explanations for this phenomenon. The immediate explanation would be that the subjective character of emotion recognition dictates a great variability in the way individuals interpret emotional messages. Another possible explanation is that individual differences in lower level perceptual mechanisms underlie the low consensus in vocal emotion recognition. This second conjecture is supported by evidence showing large variability in pitch discrimination thresholds in the general population (Johnson, Watson, & Jensen, 1987; Kidd, Watson, & Gygi, 2007; Surprenant & Watson, 2001). Further support of this approach is provided by findings demonstrating better abilities of musicians in pitch discrimination (Dankovicova, House, Crooks, & Jones, 2007; Kishon-Rabin, Amir, Vexler, & Zaltz, 2001; Magne, Schon, & Besson, 2006; Schon, Magne, & Besson, 2004), as well as in recognition of emotions in voice (Thompson et al., 2004).

The fundamental frequency (*f*0) of the speech signal, perceived as pitch, is one of the most informative acoustic parameters, carrying affective and pragmatic information in speech. Various features extracted from the *f*0 contour (such as slope, standard deviation, mean, range, and others) serve as important cues enabling us to deduce the emotional state of our dialogue partner (Hammerschmidt & Jurgens, 2007; Monnot, Orbelo, Riccardo, Sikka, & Rossa, 2003; Pell & Baum, 1997). Certain features of pitch can discriminate between emotions involving high levels of arousal (such as happiness, anger, and fear) and emotions involving relatively lower levels of arousal (such as sadness and tenderness). For example, rising *f*0 contours and an increase in mean *f*0 are associated with high arousal, whereas falling *f*0 contours and a decrease in mean *f*0 may be associated with lower arousal (Juslin & Laukka, 2003). In order to successfully differentiate between different emotion types, it is therefore necessary to distinguish between high and low pitch, as well as rising versus falling pitch contours. These perceptual abilities involve recognition of large, as well as small, pitch fluctuations. The importance of small pitch changes as informative cues for emotion recognition has long been acknowledged. Lieberman and Michaels (1962) demonstrated that smoothing the *f*0 contours with a 40-ms time constant resulted in a drop of emotion recognition scores from 47% to 38%. Further smoothing with a larger 100-ms time constant reduced the recognition rate to 25%. They concluded that microperturbations are an essential cue to some expressions of emotions in speech (Lieberman & Michaels, 1962).

Pitch features also play an important role in pragmatic prosody recognition. For example, pitch serves as a major cue in determining the position of lexical stress and word accent in a sentence (Amir, Almogi, & Mixdorff, 2008; Eady, Cooper, Klouda, Mueller, & Lotts, 1986; Hasegawa & Hata, 1992). Typically, an accented word is characterized by heightened average *f*0 in the stressed syllable, as well as higher intensity and longer duration. Details of the *f*0 contour in this syllable may differ, depending on the location of the accented word in the sentence: It could be rising, falling, or peaked (Hasegawa & Hata, 1992). In some languages, pitch cues carry over to the rest of the sentence in the form of *postfocal compression*, which is the tendency to have a lower pitch range in the rest of the utterance after the accented syllable (Xu, 2011; Xu & Xu, 2005). Therefore, in order to perceive the location of an accented word in a sentence, the ability to distinguish between high and low pitch seems highly relevant.

The above cited studies imply that prosody recognition, pragmatic and affective alike, relies on a range of pitch-processing abilities. The purpose of the present study was to examine whether this dependence would be supported by

empirical evidence. Two experiments were performed. In the first experiment, four psychoacoustic tests, as well as two prosody recognition tests, were employed. The psychoacoustic tests consisted of both steady tone and gliding tone tasks. Single thresholds were obtained in a pitch discrimination task, using an odd-ball paradigm and in pitch identification tasks (high/low pitch, rising/falling pitch, changing/steady pitch). The prosody recognition tests included a vocal emotion recognition task, as well as a pragmatic prosody task, in which participants had to choose the meaning of the spoken sentence according to the accented word.

The goal of the second experiment was to determine whether the associations between certain psychoacoustic abilities and prosodic perception, which were revealed in the first experiment, would be retained in a multithreshold psychoacoustic assessment protocol. This was performed in order to exclude the possibility that the results of the first experiment represent task-specific procedural factors, which may influence the first threshold assessments (Karni & Bertini, 1997). The psychoacoustic tests employed in this experiment were the tasks that exhibited the most pronounced association with prosodic perception in the first experiment.

Experiment 1

Method

Participants

A group of 60 participants (30 males and 30 females), 20–36 years of age (M = 25.3, SD = 4.25) took part in the experiment. All participants were native Hebrew speakers with no known neurological or psychiatric conditions. Participants were recruited from the general population. No professional musicians were included. Participants were questioned about the exact duration of instrumental and theoretical musical studies. Of the participants, 41.7% had no prior musical training. The median of the duration of musical training was 1.75 years (interquartile range: 0–3 years). Participants underwent hearing screening at 0.5–4 kHz and demonstrated bilateral pure-tone air-conduction thresholds within normal limits (<15 dBHL) (ANSI, 1996). Participants had no previous experience in psychoacoustic testing and signed a consent form prior to the beginning of the experiment.

Stimuli and tasks

A total of six tasks were employed, including four psychoacoustic tasks, a four-block vocal emotion recognition (VER) task, and a focus perception (FP) task. No feedback was given in any of the tasks, except in the training sessions. All assignments were implemented as a graphic user interface (GUI) programmed in MATLAB and involved using a mouse to press the appropriate button in the GUI. The number of buttons in the GUI varied according to the task.

Psychoacoustic tasks The thresholds in all psychoacoustic tasks were obtained using a two-down, one-up adaptive staircase procedure, converging at a performance level of 70.7% (Levitt, 1971). The initial step size was 40 Hz. The step size was divided by 2 after each reversal, until a final step size of 1 Hz was reached. Assessment terminated after 10 reversals with the final step size. Thresholds were calculated using the arithmetic mean of the last 8 reversals. All stimuli were 300 ms in duration, (including 25-ms rise and fall ramps) with a 500-ms interstimulus gap. The sound level was adjusted for each participant, according to his or her individual comfort. All tests were preceded by a training session, in which participants had to achieve five subsequent successes in order to begin the experiment (binomial calculation yields a probability of p = .03 for obtaining five consecutive correct answers by chance). This was performed in order to ensure that participants fully understood the procedure, thus reducing the possible influence of procedural factors, related to the task. The stimuli in the training session employed a single easy-to-detect frequency interval of 200 Hz between the reference tone and the test tone or a 200-Hz frequency range of the gliding tones. The employment of a suprathreshold frequency difference in the training session was meant to ensure that the training process only clarified the operation of the task's interface and that the threshold obtained reflected primarily untrained auditory perceptual abilities. Each psychoacoustic task consisted of one threshold assessment.

The following psychoacoustic tasks were employed (see also Fig. 1):

Steady tone tasks

1. *Two-tone discrimination task* (2TDT): Two constant-frequency pure tones (PTs) were presented on each trial. One of the tones was a 1-kHz PT, and the other had a larger frequency value (i.e., higher in pitch). Participants had to indicate the tone that was higher. The initial frequency difference between the tones was 200 Hz.
2. *Oddball paradigm task* (OPT): A three-interval two-alternative forced choice paradigm was employed. A fixed reference PT of 1 kHz was followed by two other tones, one of which was a repetition of the reference and the other was higher in pitch. Participants had to indicate which tone (second or third) was different from the reference tone. The initial frequency difference between the deviant stimulus and the other two stimuli was 200 Hz.

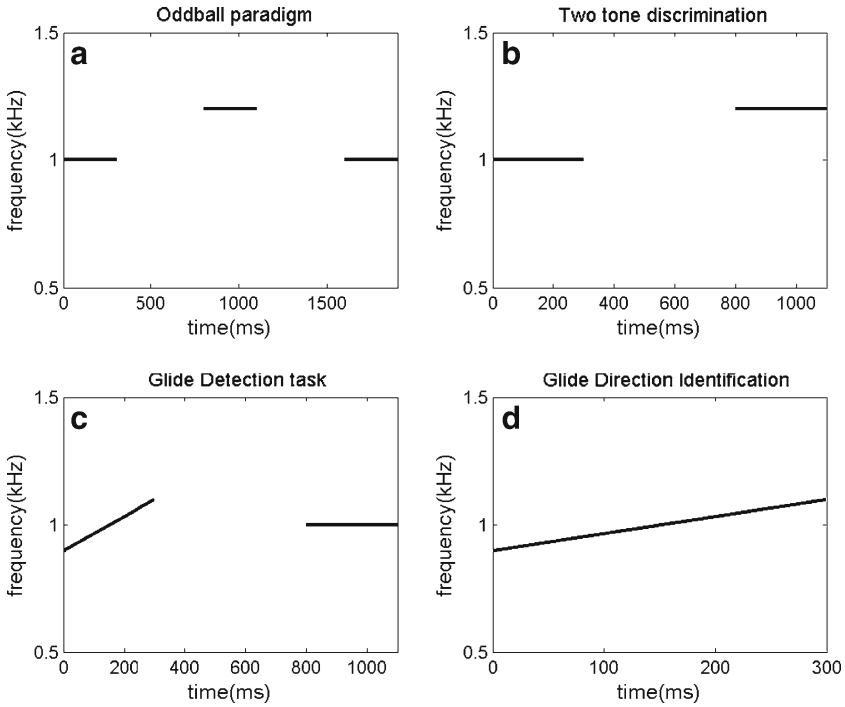

Fig. 1 Illustration of the psychoacoustic tasks. **a** Oddball paradigm task: A fixed reference tone was followed by two other tones. Participants were asked to indicate which tone was different from the reference tone. **b** Two-tone discrimination task: Two constant frequency pure tones were presented. Participants were asked to indicate the tone that was higher in pitch. **c** Glide detection task: Participants were asked to point to one of two auditory stimuli that was changing in pitch (gliding vs. constant). **d** Glide direction identification task: Participants were asked to decide whether a gliding tone was rising or falling in frequency

Gliding tone tasks The stimuli in this part of the experiment were PTs with a linear, unidirectional change of frequency over time. The gliding tones were either ascending or descending PT glides, 300 ms in length, with a center frequency of 1 kHz. The initial glide range was 200 Hz in both assignments.

1. *Glide detection task* (GDT): Participants heard two PT signals. One of them was a gliding tone, and the other one was a 1-kHz steady tone. A two-alternative forced choice task was employed, in which the participant was asked to point to one of two auditory stimuli that was changing in pitch (i.e., gliding). The slope of the gliding tone was gradually reduced until the participant's threshold was reached.
2. *Glide direction identification task* (GDIT): Participants heard one gliding tone on each trial. A one-interval two-alternative forced choice task was employed, in which the participants were asked to decide whether the gliding tone was rising or falling in frequency. As in the previous task, the slope of the gliding tone was gradually reduced until the participant's threshold was reached.

Vocal emotion recognition task All the stimuli for the vocal emotion recognition (VER) task were recorded in a professional recording studio by four professional actors (two female, and two male). A total of 966 stimuli were recorded. Stimuli were equalized using the following procedure: The root mean square (RMS) value of the samples in each sound file was calculated. Subsequently, the sound file was normalized to achieve an equal RMS value for all stimuli. All words and sentences had no linguistic emotional content. The stimuli represented four basic emotions: happiness, sadness, anger, and fear. The stimuli included nonsense monosyllabic utterances, nonsense polysyllabic words, Hebrew words, and Hebrew sentences. The inclusion of different stimulus types in the VER task enabled a comprehensive evaluation of daily prosody recognition abilities and an examination of associations between the perception of these distinct utterance types and psychoacoustic abilities.

These stimuli were validated by a panel of 20 independent judges, who were asked to choose the appropriate label for each stimulus from five different emotional labels (see below). The stimuli which were selected for the task received an overall average recognition rate of 79.0% (SD = 15.9%), demonstrating the reliability of the task. The battery included stimuli that were unanimously recognized by the judges, as well as others that were not recognized unanimously but received at least 8/20 correct answers in this validation process ($p < .05$, binomial test). The motivation for employing stimuli that were not unanimously recognized by the judges was to avoid a ceiling effect, thus enhancing the sensitivity of

the task. However, for each of the stimuli included in the final selection, most votes out of the 20 were for the *intended* emotion. This was monitored in order to avoid misleading stimuli. For example, a stimulus in which 8 votes were for the intended emotion and 9 or more for a single other emotion was not included in the final selection. For the final VER task, 302 stimuli were chosen. These included 88 nonsense monosyllabic utterances, 70 nonsense polysyllabic words, 80 Hebrew words, and 64 Hebrew sentences. The Pearson correlation coefficient between correct answer rates for these stimuli at the validation and the experimental stages was $r = .73$ ($p < .0001$). This demonstrates the reliability of the stimulus elimination process.

Participants were asked to decide which emotion was conveyed by the actors. The possible choices were all the emotions mentioned above, in addition to a "neutral" option (for further discussion, see the Calculated Scores section). The battery was divided into blocks, characterized by the nature of the utterances (monosyllabic utterances, words, etc.). The order of the blocks was randomized between participants, as well as the order of the stimuli within the blocks. After naming the emotion, participants also had to rate the intensity of the emotion (except when "neutral" was selected) on a scale of 1 (*low intensity*) to 3.

Focus perception task The FP task examined the sensitivity of the participants to different locations of accented (focused) words in a spoken phrase. The FP task employed six Hebrew sentences, recorded by four actors in a professional recording studio. The sentences were recorded in several different versions. In each version, a different word in the sentence was accented, thus modifying the pragmatic meaning of the phrase. Prior to the experimental process, the recorded stimuli were validated by a panel of 19 judges, thus ensuring that the location of the intended accentuation was expressed reliably. All stimuli chosen for the task received correct responses from at least 10/19 judges ($p < .01$ binomial test). These stimuli received an overall average recognition rate of 80.7% ($SD = 9.5\%$), demonstrating the reliability of the task. Each of the stimuli was played 5 times. The stimuli were played in a randomized order. Participants had to choose between four options describing the specific meaning given to the sentence when different words were accented or when no word was accented. It is important to note that the listeners were not asked explicitly which word was stressed in the phrase but were instructed to focus on the new acquired meaning of the phrase in each recorded version. The purpose of this task was therefore to measure the participants' ability to perceive an intention in a spoken utterance, rather than define the acoustic attributes of that utterance.

For example, in the sentence "there are birds in the park," if the word "birds" was accented, producing the phrase "there are *birds* in the park," the multiple choice answers were the following:

1. There is no new information conveyed beyond the meaning of the words (i.e., there is no accented word in the sentence).
2. There *are* birds in the park, as opposed to "there are *no* birds in the park."
3. There are *birds* in the park, as opposed to *monkeys*.
4. There are birds in the *park*, as opposed to the *zoo*.

In this case, the right answer would be answer 3.Although all sentences had an accented word, the first option ("there is no accented word in the sentence") was supplied for cases in which participants were insensitive to the existence of an accent in the sentence. This option is similar to the neutral option included in the VER task.

Acoustic properties of the speech stimuli

The *f*0 range and standard deviations (*SD*s) of the stimuli in the VER task were measured, using Praat (Table 1). These values were normalized to semitones. The *SD* was normalized to semitones using the average *f*0 of each actor as reference.

The *f*0 ranges and *SD*s of the stimuli in the FP task were also measured, using Praat. The mean *f*0 range of these stimuli was 12.28 semitones (*SD*=3.33 semitones). The deviation from the mean *f*0 of the sentences was measured for the accented words, yielding an average deviation of 3.27 semitones. Figure 2 illustrates an example of two versions of the same sentence, with different accented words. Both examples demonstrate a local peak in the *f*0 contour located on the accented word.

Procedure

Participants were tested individually in a quiet room, where background noise was assessed prior to the experiment with a sound level meter, to ensure a quiet working environment (below 40 dBA). All tasks were presented to the participants binaurally with Sennheiser HD-201 headphones and a Line6 Ux1 external sound card. The order of the assignments was modified between participants. Half of the participants were first tested with the psychoacoustic tasks, followed by the prosodic tasks. These participants were presented with the following order of psychoacoustic tasks: the GDIT, the GDT,

Table 1 Acoustic properties of the emotional utterances, split by emotions

Emotion	Mean *f*0 Range (semitones)	Mean *SD* (semitones)
Anger	9.00	2.63
Fear	7.86	2.17
Happiness	13.65	4.09
Sadness	11.02	2.92
Overall results	10.39	2.97

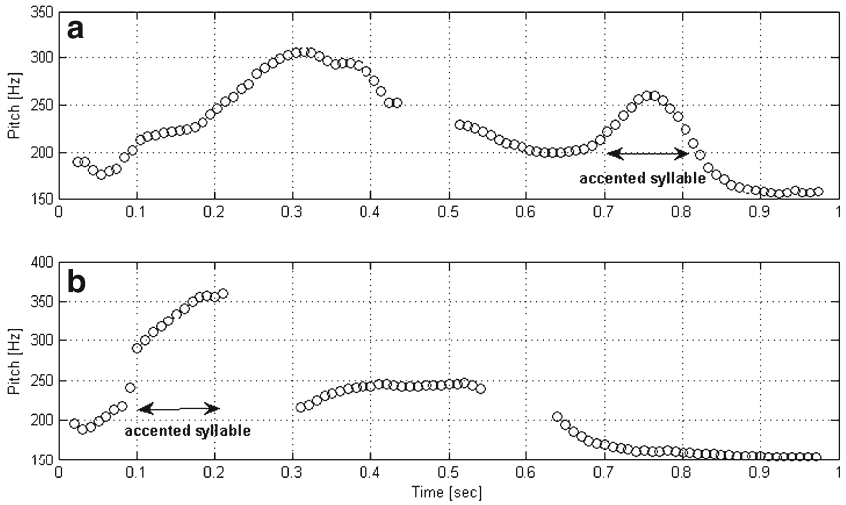

Fig. 2 The pitch contour of the Hebrew sentence "Li ein mezuman" (I don't have cash money). **a** Accent on the last word in the sentence. **b** Accent on the first word in the sentence. Both examples demonstrate a local peak in the pitch contour

and the 2TDT, followed by the OPT. The remaining half of the participants were presented first with the prosodic tests and then with the psychoacoustic tasks in reversed order. Participants were allowed to take as many breaks as needed between the various tasks. On average, participants needed 120 min to complete all tasks (including breaks).

Calculated scores

A total of six scores were obtained for each participant, as follows:

1. The VER task score (in percent correct): the overall percentage of correct responses in the VER task. Any incorrect emotion labeling, including the "neutral" responses, were counted as errors. Since daily speech communication involves a variety of utterance types and emotions, this score is meant to provide an ecologic measure of the participants' overall VER.
2. The FP task score: This score represents the percentage of correct responses in the task.
3. Four psychoacoustic thresholds: the thresholds were $\log_{10}$-transformed.[1]

[1] The employment of a logarithmic transformation was motivated by both theoretical and practical considerations. From a theoretical standpoint, this decision derives from the nature of pitch perception, which is logarithmic in nature (Moore, 2003). The practical reason for employing log transformation considerations is that the logarithmic transformation helps to avoid a violation of the homoscedasticity assumption of the parametric statistical tests, which is a result of extensive variability of frequency discrimination thresholds, which increases with their magnitude. It is important to note that the use of a logarithmic or square-root transformation of frequency discrimination thresholds is common practice in the psychoacoustic literature (Delhommeau, Micheyl, & Jouvent, 2005; Demany & Semal, 2002; Irvine, Martin, Klimkeit, & Smith, 2000; Micheyl, Delhommeau, Perrot, & Oxenham, 2006).

Results

Descriptive statistics

The group mean scores for each of the experimental tasks, as well as ranges and standard deviations, are shown in Tables 2 and 3.

Vocal emotion recognition scores, with breakdown into emotions

Table 4 illustrates the confusion matrix for the total recognition scores of specific emotions in the VER task. Average scores range from 72% (recognition score for anger) to 86.5% (recognition score for fear). Recognition rates for all emotions were well above the chance level of 20%.

Psychoacoustic threshold comparison

Paired-samples *t*-tests were conducted in order to compare the values of the psychoacoustic thresholds. Differences between thresholds were tested separately for the steady tone and gliding tone tasks. A significant difference was found between the two steady tone tasks. Thresholds for the OPT were significantly lower than the thresholds for the 2TDT, $t(59) =$

Table 2 Mean scores, standard deviations, and ranges of the prosody recognition tasks

Task	*M*	*SD*	Range
Vocal emotion recognition (percent correct)	79.07	7.41	60.60–92.05
Focus perception: score (percent correct)	76.5	15.36	30.0–96.66

Table 3 Mean scores, standard deviations, ranges, and log-transformed ranges of the psychoacoustic tasks (1000-Hz reference or center frequency for all psychoacoustic tasks)

Task	*M*	*SD*	Range	Log_{10} Range
*Oddball paradigm (Hz)	9.43	10.29	1–63.50	0–1.8
*Two-tone discrimination (Hz)	27.04	43.84	0.75–200.00	−0.12–2.3
*Glide direction identification (Hz)	23.87	36.14	2.00–199	0.3–2.29
*Glide detection task (Hz)	20.28	20.15	4.50–122.25	0.65–2.08

3.242, $p < .01$. No significant difference was found between the results of the two gliding tone tasks.

Bivariate correlations

Bivariate Pearson correlations were calculated between all four psychoacoustic thresholds, and the scores obtained in the VER task and the FP task. Bonferroni correction for multiple correlations was applied, and the critical *p* values were readjusted accordingly. Table 5 and Fig. 3 illustrate the correlations between prosodic scores and psychoacoustic thresholds. Psychoacoustic thresholds had a negative relationship with prosodic abilities, such that better psychoacoustic abilities corresponded to better prosodic abilities (the negative correlation is due to the fact that lower thresholds correspond to better abilities). Three psychoacoustic thresholds correlated significantly with prosody recognition scores (both pragmatic and affective). The correlation between the oddball paradigm scores and prosody recognition scores was nonsignificant. The correlations between all psychoacoustic thresholds were significant. A significant correlation was also found between the two prosody recognition tasks.

Stepwise regressions

Psychoacoustic scores as predictors of vocal emotion recognition A forward stepwise regression was performed with all four psychoacoustic scores as predictors of VER score (Table 6). The only two psychoacoustic scores to enter the regression (in this order) were the 2TDT thresholds, $\beta = -.506$, $p < .001$, and the GDIT thresholds, $\beta = -.292$, $p < .05$. Together, these scores explain 31.0% of the variance in the VER scores. It is worth mentioning that the psychoacoustic predictors of VER abilities that entered the regression equation are those tasks in which the participants had to identify the direction of a pitch change. Once these predictors entered the analysis, the contribution of the remaining pitch discrimination tasks became nonsignificant.

Psychoacoustic scores as predictors of focus perception A stepwise regression was performed with all four psychoacoustic scores as predictors of the FP task scores (Table 7).The only two psychoacoustic thresholds to enter the analysis were the 2TDT threshold, $\beta = -.399$, $p < .005$, and the GDT threshold, $\beta = -.313$, $p < .05$. Together, these thresholds explain 38.7% of the variance in the FP task scores.

Table 4 Confusion matrix for emotion recognition (%)

Identified Emotion / Portrayed Emotion	Anger	Happiness	Fear	Sadness	Neutral
Anger	72	4.5	1	3.5	19
Happiness	3.5	81.5	2	0.5	12.5
Fear	0.5	2	86.5	10	1
Sadness	2.5	0.5	8	79.5	9.5

Experiment 2

Rationale

In the first experiment, a series of psychoacoustic tasks was employed, in order to track down specific auditory abilities that correlate with prosody recognition. In this experiment, only one psychoacoustic threshold assessment was performed for each task, due to the large number of psychoacoustic evaluations involved. The results of the first experiment demonstrated a significant correlation between three psychoacoustic thresholds and prosody perception. Linear regression highlighted two of these psychoacoustic measures as the most significant predictors of prosody perception. A significant correlation was found between affective and pragmatic prosody recognition, indicating that these abilities may be supported by the same perceptual mechanism. Therefore, it remained to examine whether the association between psychoacoustic thresholds and affective prosody recognition would be maintained when a multithreshold assessment procedure is employed. This was performed to ensure that the results of the first experiment were not influenced by task-specific factors and reflect a true association between auditory perceptual abilities and prosody recognition.

Table 5 Correlation between prosody recognition scores (in bold) and psychoacoustic thresholds (in italic)

	VER	FP	*OPT*	*2TDT*	*GDT*
FP	.461**				
OPT	−.255	−.304			
2TDT	−.506***	−.561***	.537***		
GDT	−.391*	−.519***	.416*	.516***	
GDIT	−.490**	−.439**	.370	.601***	.581***

Note. VER, vocal emotion recognition; FP, focus perception; OPT, oddball paradigm task; 2TDT, two-tone discrimination task; GDT, glide detection task; GDIT, glide direction identification task

. *Correlation is significant at the .05 level (Bonferroni adjustment for multiple correlations, critical p level = .0033)

**Correlation is significant at the .01 level (Bonferroni adjustment for multiple correlations, critical p level = .00066)

***Correlation is significant at the .001 level (Bonferroni adjustment for multiple correlations, critical p level = .000066)

Method

Participants

A new group of 37 participants (18 males and 19 females), 22–34 years of age (mean, 25.4, *SD*, 2.4), took part in this experiment, with similar exclusion and inclusion criteria as described in the first experiment.

Stimuli and tasks

Psychoacoustic tasks This experiment involved three tasks: a VER score (identical to the task employed in Experiment 1) and two psychoacoustic tasks: the 2TDT and the GDIT. These were the psychoacoustic tasks that demonstrated the strongest correlation with the VER scores in the first experiment. The difference in the experimental protocol was that each threshold assessment was repeated *five* consecutive times. Participants took a break of at least 2 min between assessments.

Results

Table 8 illustrates the group mean score and *SD* for each of the experimental tasks.

Repeated measures ANOVA

A repeated measures ANOVA with a Greenhouse–Geisser correction determined that mean thresholds for the 2TDT differed significantly between assessments, $F(3.16, 113.7) = 6.462$, $p < .001$. Post hoc tests using the Bonferroni correction revealed that the first assessment thresholds were significantly higher (i.e., poorer) than the thresholds obtained in the second, third, and fifth assessments. No significant difference was found between the second, third, fourth, and fifth assessments. The same procedure was employed in order to examine the differences in thresholds for the GDIT. A repeated measures ANOVA with a Greenhouse–Geisser correction determined that the mean threshold for the GDIT differed significantly between assessments, $F(3.16, 123.2) = 11.645$, $p < .00001$. Post hoc tests using the Bonferroni correction revealed that, on average, the first assessment threshold was significantly higher (i.e., poorer) than the thresholds obtained in the second, third, fourth, and fifth assessments. No significant difference was found between the second, third, fourth, and fifth assessments.

Correlation between psychoacoustic thresholds and vocal emotion recognition scores

Since there were no significant differences found between the last four assessments for both psychoacoustic

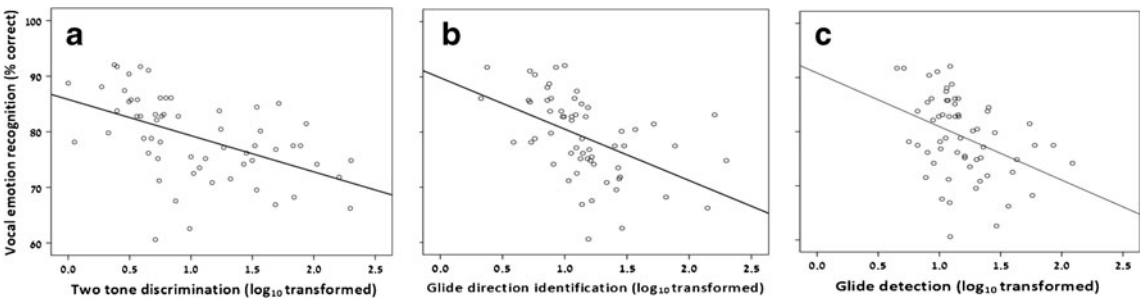

Fig. 3 Vocal emotion recognition task score, plotted against psychoacoustic thresholds (thresholds presented on a log $_{10}$ scale, 1-kHz reference tone for all tasks). **a** Two-tone discrimination task thresholds ($r = -.506$, $p < .001$). **b** Glide direction identification task thresholds ($r = -.490$, $p < .001$). **c** Glide detection task thresholds ($r = -.391$, $p < .01$)

Table 6 Linear regression with psychoacoustic thresholds as predictors of vocal emotion recognition

Model	Predictor	*B*	*SE*	β	*t*	R^2	*F* change
1	(Constant)	85.770	1.716		49.974		
	2TDT	−6.480	1.452	−.506***	−4.462	.256	19.914***
2	(Constant)	89.851	2.546		35.294		
	2TDT	−4.229	1.765	−.330*	−2.396		
	GDIT	−5.542	2.613	−.292*	−2.121	.310	4.497*

Note. 2TDT, two-tone discrimination task; GDIT, glide direction identification task

*$p < .05$

**$p < .01$

***$p < .001$

tasks, the mean of these four assessments was used to study the correlations between psychoacoustic thresholds and VER scores.

A significant correlation was found between the thresholds for both psychoacoustic tasks and the scores for the VER task ($r = -.472$, $p < .01$, for the 2TDT, and $r = -.384$, $p < .05$, for the GDIT). These correlation coefficients are slightly lower than those obtained in the first experiment. However, the Fisher *Z* calculation for correlation coefficient comparison (Fisher, 1915) did not demonstrate a significant difference between correlations in the first and the second experiments, $z = -.206$, $p = .83$, implying a similarity between the results of both experiments.

Stepwise regression: Psychoacoustic scores as predictors of vocal emotion recognition

A stepwise regression was performed with the two psychoacoustic scores (i.e., the mean of the last four measurements) as predictors of VER. The only psychoacoustic score to enter the analysis was the 2TDT threshold, which explained 22.2% of the variance in VER (Table 9).The GDIT thresholds did not enter the regression equation, $\beta = .024$, $p = .859$.

General discussion

The present study was designed to investigate the association between pitch discrimination abilities and the perception of prosody in speech. The results of the first experiment demonstrated a significant positive association between psychoacoustic abilities and the perception of emotional messages in speech, accounting for 31% of the variance in VER. The second experiment, which involved a multiple-threshold design, reaffirmed these results. The associations were similar to those found in the first experiment, albeit having smaller correlation values. This drop in correlation values, although statistically insignificant, could result from an improvement in psychoacoustic thresholds, which is typically observed in multiassessment protocols (e.g., Ari-Even Roth, Avrahami, Sabo, & Kishon-Rabin, 2004; Demany & Semal, 2002; Irvine et al., 2000). The nature of this improvement is debatable. While some investigators have attributed the fast learning observed within the first psychoacoustic assessments to procedural factors (Karni & Bertini, 1997; Wright & Fitzgerald, 2001), others have maintained that this rapid improvement in thresholds is mainly due to perceptual factors (Hawkey, Amitay, & Moore, 2004). Hence, it is not possible to determine, at this

Table 7 Linear regression with psychoacoustic thresholds as predictors of focus perception

Model	Predictor	*B*	*SE*	β	*t*	R^2	*F* change
1	(Constant)	91.905	3.413		26.927		
	2TDT	−14.906	2.888	−.561***	−5.162	.315	26.644***
2	(Constant)	106.953	6.665		16.047		
	2TDT	−10.612	3.217	−.399**	−3.299		
	GDIT	−16.398	6.337	−.313*	−2.588	.387	6.696*

Note. 2TDT, two-tone discrimination task; GDIT, glide direction identification task

*$p < .05$

**$p < .01$

***$p < .001$

Table 8 Mean score, standard deviation, and range for each of the experimental tasks

Task	*M*	*SD*	Range	Log_{10} Range
VER (% correct)	78.5	7.1	62.7–91.4	1.79–1.96
2TDT (Hz), 1	31.9	46.6	1.5–188.6	0.17–2.27
2TDT (Hz), 2	20.9	33.7	1–184	0–2.26
2TDT (Hz), 3	22.02	42.7	1.6–197	0.21–2.29
2TDT (Hz), 4	28.3	52.8	1.4–200	0.13–2.30
2TDT (Hz), 5	25.2	52.9	1–200	0–2.30
GDIT (Hz), 1	28.0	35.1	6.5–177.7	0.81–2.25
GDIT (Hz), 2	21.8	25.5	3.1–108.2	0.49–2.03
GDIT (Hz), 3	19.2	32.7	3.7–200	0.57–2.30
GDIT (Hz), 4	21.6	37.5	1.7–200	0.24–2.30
GDIT (Hz), 5	14.1	10.6	2.1–43.5	0.32–1.63

Note. 2TDT, two-tone discrimination task; GDIT, glide direction identification task

point, whether the lower correlation values between psychoacoustic thresholds and prosody recognition obtained in the second experiment derived from procedural related factors, such as auditory attention, or from the fact that the first thresholds represented raw auditory abilities, which were a better reflection of the participants' daily auditory aptitudes and, therefore, correlated more with prosody recognition.

Our results also demonstrate a positive association between psychoacoustic abilities and the perception of a focused word in a spoken phrase, accounting for 38% of the variance in the focus perception task scores. Additionally, we found a positive association between VER and the recognition of a focused word in a spoken sentence. Furthermore, when the correlation between VER and focus perception scores was recomputed, controlling for psychoacoustic abilities, it became nonsignificant. Taken together, our results may point to the importance of pitch perception as a mechanism supporting both pragmatic and emotional processing of prosody.

Conventional evaluations of *f*0 differences in emotional speech measure *f*0 range and *SD*. These measurements, which were also performed in the second experiment, yield values that are far beyond perceptual thresholds (e.g., Banziger & Scherer, 2005; Hammerschmidt & Jurgens, 2007). The present study, which focused on auditory perceptual abilities involving *small* pitch differences, highlights the importance of small pitch fluctuations to prosody recognition. Relatively few studies have focused on the contribution of small pitch changes to emotion recognition in voice in the last decades, since the study of Lieberman and Michaels (1962). The results of the present study suggest that reexamination of this issue could enhance our understanding of emotional production and perception. Three of the four tasks employed in the present study required pitch direction naming (i.e., higher/lower ascending/descending, gliding/nongliding), as well as pitch discrimination. The fourth task (the OPT) required only pitch discrimination. Results demonstrate that psychoacoustic thresholds obtained in the three tasks requiring pitch direction recognition correlated significantly with prosody recognition scores, while no significant correlations were found with thresholds obtained in the fourth task, which required only pitch discrimination abilities. The ability to discriminate between two pitches does not necessarily ensure an equal ability in pitch direction recognition. Previous studies have reported a disparity between pitch discrimination abilities and pitch direction recognition (Mathias, Micheyl, & Bailey, 2010; Semal & Demany, 2006). These findings were replicated in the second experiment, in which thresholds for the steady tone task requiring only pitch discrimination were significantly lower (i.e., better) than those obtained for the steady tone task that also required pitch direction recognition. Moreover, pitch

Table 9 Linear regression with psychoacoustic thresholds as predictors of VER

Model	Predictor	*B*	*SE*	β	*t*	R^2	*F* change
1	(constant)	84.599	2.236		37.830		
	2TDT	−6.373	2.014	−.472**	−3.164	.222	10.014**

Note. VER, vocal emotion recognition; 2TDT, two-tone discrimination task

**$p < .01$

direction recognition has been shown to involve different brain regions than does pitch discrimination. A study performed by Johnsrude, Penhune, and Zatorre (2000) demonstrated that patients with right temporal lobe excision exhibit higher thresholds for pitch direction recognition than do controls, or patients with left temporal lobe removals, while their thresholds for simple pitch discrimination are similar to those obtained by control participants (Johnsrude, Penhune, & Zatorre, 2000). Additional support is provided by the results of studies by Demany et al. (2009), indicating the existence of specialized frequency shift direction detectors, enabling pitch direction detection, even in cases in which the pitch of the reference tone is not consciously perceived (Demany, Pressnitzer, & Semal, 2009). Taken together, these results imply that pitch discrimination and pitch direction naming can be regarded as distinct abilities. Results of the present experiment demonstrate that it is specifically pitch direction recognition that can account for individual differences in the perception of both pragmatic and affective prosody.

In light of the findings obtained in the present study, it is interesting to consider prior studies, which demonstrated better abilities of musicians in identifying emotion in vocal utterances (Thompson et al., 2004). These results may now be partially explained by the fact that musicians demonstrated better pitch discrimination abilities (Kishon-Rabin et al., 2001; Micheyl et al., 2006).

The present study employed exclusively pitch perception aptitudes as predictors of prosodic perception. The decision to use pitch perception as the representative psychoacoustic ability was based on ample evidence demonstrating the importance of pitch as a cue to pragmatic and affective meaning (Amir et al., 2008; Hammerschmidt & Jurgens, 2007; Pell & Baum, 1997; Steinhauer, Alter, & Friederici, 1999). This decision is supported by our results, demonstrating a strong association between pitch perception and the perception of prosodic messages in voice. However, prosodic information in speech is also transmitted through acoustic cues other than pitch, such as voice quality (Scherer, Ladd, & Silverman, 1984), intensity (Hammerschmidt & Jurgens, 2007; Sobin & Alpert, 1999), and the rate of speech (Hammerschmidt & Jurgens, 2007). Therefore, it remains to be explored whether psychoacoustic thresholds related to these acoustic parameters can account for some of the additional, unexplained portion of the variance in prosodic perception.

There could be significant future applications to the results of the present study. The most practical outcome of the present study is the possibility that perceptual auditory learning may improve VER. If an improvement in auditory perceptual abilities results in a corresponding improvement in the perception of emotional prosody, a causal association between the two could be substantiated. A closely related direction of research could involve developmental studies, examining the influence of perceptual auditory training on the development of prosodic skills in children. Finally, our paradigm may be useful in the investigation of the mechanisms underlying the poor VER abilities found in various neuropsychiatric disorders, such as autism (Golan, Baron Cohen, Hill, & Rutherford, 2007) and schizophrenia (Murphy & Cutting, 1990).

In summary, the present study presents evidence for the association between psychoacoustic abilities and prosodic perception, pragmatic and affective alike. Pitch direction recognition abilities can predict 31% of the variance in vocal emotion recognition abilities and 38% of the variance in the scores for the recognition of a focused word in a sentence. Emotional and pragmatic prosody recognition have been found to correlate significantly, supporting a common perceptual mechanism.

Acknowledgments This study was supported by the NARSAD independent investigator award and by grant no. 474/06 from the Israel Science Foundation awarded to M. Lavidor. We are indebted to all participants in the tests for devoting their time and effort to this study. Special thanks to Dr. Michal Ben Shachar for her help and support and to Adi Hoyben and Atal Harush for their assistance in performing the second experiment.

References

Adolphs, R. (2003). Cognitive neuroscience of human social behaviour. *Nature Reviews. Neuroscience, 4*(3), 165–178.

Amir, N., Almogi, B. C., & Mixdorff , H. (2008). A systematic framework for studying the contribution of F0 and duration to the perception of accented words in Hebrew. *Proceedings of the Speech Prosody 4th international conference, Campinas, Brazil, 6-9 May 2008*, 285–288.

ANSI. (1996). *American National Standard Specification for Audiometers.ANSI S3.6-1996*. New York.

Ari-Even Roth, D., Avrahami, T., Sabo, Y., & Kishon-Rabin, L. (2004). Frequency discrimination training: Is there ear symmetry? *Journal of Basic and Clinical Physiology and Pharmacology, 15,* 15–28.

Banziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication, 46,* 252–267.

Dankovicova, J., House, J., Crooks, A., & Jones, K. (2007). The relationship between musical skills, music training, and intonation analysis skills. *Language and Speech, 50*(Pt 2), 177–225.

Delhommeau, K., Micheyl, C., & Jouvent, R. (2005). Generalization of frequency discrimination learning across frequencies and ears: implications for underlying neural mechanisms in humans. *Journal of the Association for Research in Otolaryngology, 6*(2), 171–179.

Demany, L., Pressnitzer, D., & Semal, C. (2009). Tuning properties of the auditory frequency-shift detectors. *Journal of the Acoustical Society of America, 126*(3), 1342–1348.

Demany, L., & Semal, C. (2002). Learning to perceive pitch differences. *Journal of the Acoustical Society of America, 111*(3), 1377–1388.

Eady, S. J., Cooper, W. E., Klouda, G. V., Mueller, P. R., & Lotts, D. W. (1986). Acoustical characteristics of sentential focus: Narrow vs. broad and single vs. dual focus environments. *Language and Speech, 29*(Pt 3), 233–251.

Fenster, C. A., Blake, L. K., & Goldstein, A. M. (1977). Accuracy of vocal emotional communications among children and adults and the power of negative emotions. *Journal of Communication Disorders, 10*(4), 301–314.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika, 10*(4), 507–521.

Golan, O., Baron Cohen, S., Hill, J. J., & Rutherford, M. D. (2007). The 'Reading the Mind in the Voice' test - Revised: A study of complex emotion recognition in adults with and without Autism Spectrum Conditions. *Journal of Autism and Developmental Disorders, 37*(6), 1096–1106.

Gold, T. G. (1987). Communication skills of mainstreamed hearing-impaired children. In H. Levitt, N. S. McGarr, D. Geffner & (Eds.), *Development of language and communication skills in hearing-impaired children* (Vol. 26, pp. 108–122): ASHA Monograph.

Hammerschmidt, K., & Jurgens, U. (2007). Acoustical correlates of affective prosody. *Journal of Voice, 21*(5), 531–540.

Hasegawa, Y., & Hata, K. (1992). Fundamental frequency as an acoustic cue to accent perception. *Language and Speech, 35*(Pt 1–2), 87–98.

Hawkey, D. J., Amitay, S., & Moore, D. R. (2004). Early and rapid perceptual learning. *Nature Neuroscience, 7*(10), 1055–1056.

Irvine, D. R., Martin, R. L., Klimkeit, E., & Smith, R. (2000). Specificity of perceptual learning in a frequency discrimination task. *Journal of the Acoustical Society of America, 108*(6), 2964–2968.

Johnson, D. M., Watson, C. S., & Jensen, J. K. (1987). Individual differences in auditory capabilities. *Journal of the Acoustical Society of America, 81*(2), 427–438.

Johnsrude, I. S., Penhune, V. B., & Zatorre, R. J. (2000). Functional specificity in the right human auditory cortex for perceiving pitch direction. *Brain, 123*(Pt 1), 155–163.

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.

Karni, A., & Bertini, G. (1997). Learning perceptual skills: Behavioral probes into adult cortical plasticity. *Current Opinion in Neurobiology, 7,* 530–535.

Kidd, G. R., Watson, C. S., & Gygi, B. (2007). Individual differences in auditory abilities. *Journal of the Acoustical Society of America, 122*(1), 418–435.

Kishon-Rabin, L., Amir, O., Vexler, Y., & Zaltz, Y. (2001). Pitch discrimination: Are professional musicians better than nonmusicians? *Journal of Basic and Clinical Physiology and Pharmacology, 12,* 125–143.

Kleinman, J., Marciano, P. L., & Ault, R. L. (2001). Advanced theory of mind in high-functioning adults with autism. *Journal of Autism and Developmental Disorders, 31*(1), 29–36.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49*(2), 467+.

Lieberman, P., & Michaels, S. B. (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *Journal of the Acoustical Society of America, 34,* 922–927.

Magne, C., Schon, D., & Besson, M. (2006). Musician children detect pitch violations in both music and language better than nonmusician children: Behavioral and electrophysiological approaches. *Journal of Cognitive Neuroscience, 18*(2), 199–211.

Mathias, S. R., Micheyl, C., & Bailey, P. J. (2010). Stimulus uncertainty and insensitivity to pitch-change direction. *Journal of the Acoustical Society of America, 127*(5), 3026–3037.

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research, 219*(1–2), 36–47.

Monnot, M., Orbelo, D., Riccardo, L., Sikka, S., & Rossa, E. (2003). Acoustic analyses support subjective judgments of vocal emotion. *Annals of the New York Academy of Sciences, 1000,* 288–292.

Moore, B. C. (2003). *An introduction to the psychology of hearing* (5th ed.). Amsterdam, London: Academic Press.

Most, T., & Peled, M. (2007). Perception of suprasegmental features of speech by children with cochlear implants and children with hearing AIDS. *Journal of Deaf Studies and Deaf Education, 12*(3), 350–361.

Murphy, D., & Cutting, J. (1990). Prosodic comprehension and expression in schizophrenia. *Journal of Neurology, Neurosurgery & Psychiatry, 53*(9), 727–730.

Pell, M. D., & Baum, S. R. (1997). Unilateral brain damage, prosodic comprehension deficits, and the acoustic cues to prosody. *Brain and Language, 57*(2), 195–214.

Rutherford, M. D., Baron Cohen, S., & Wheelwright, S. (2002). Reading the mind in the voice: A study with normal adults and adults with Asperger syndrome and high functioning autism. *Journal of Autism and Developmental Disorders, 32*(3), 189–194.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40,* 227–256.

Scherer, K. R., Ladd, R. D., & Silverman, K. E. A. (1984). Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America, 76,* 1346–1356.

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences, 10*(1), 24–30.

Schon, D., Magne, C., & Besson, M. (2004). The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology, 41*(3), 341–349.

Semal, C., & Demany, L. (2006). Individual differences in the sensitivity to pitch direction. *Journal of the Acoustical Society of America, 120*(6), 3907–3915.

Sobin, C., & Alpert, M. (1999). Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy. *Journal of Psycholinguistic Research, 28*(4), 347–365.

Steinhauer, K., Alter, K., & Friederici, A. D. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience, 2*(2), 191–196.

Surprenant, A. M., & Watson, C. S. (2001). Individual differences in the processing of speech and nonspeech sounds by normal-hearing listeners. *Journal of the Acoustical Society of America, 110*(4), 2085–2095.

Thompson, W. F., Schellenberg, E. G., & Husain, G. (2004). Decoding speech prosody: Do music lessons help? *Emotion, 4*(1), 46–64.

Trimmer, C. G., & Cuddy, L. L. (2008). Emotional intelligence, not music training, predicts recognition of emotional speech prosody. *Emotion, 8*(6), 838–849.

Wright, B. A., & Fitzgerald, M. B. (2001). Different patterns of human discrimination learning for two interaural cues to sound-source location. *Proceedings of the National Academy of Sciences of the United States of America, 98*(21), 12307–12312.

Xu, Y. (2011). *Post-Focus compression: cross linguistic distribution and historical origin.* Paper presented at the International Congress of Phonetic Sciences, Hong Kong.

Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics, 33*(2), 159–197.