Available online at www.sciencedirect.com
ScienceDirect
Journal of Phonetics 37 (2009) 417–435
Journal of Phonetics
www.elsevier.com/locate/phonetics

Factors in the recognition of vocally expressed emotions: A comparison of four languages

Marc D. Pell[a,*], Silke Paulmann[a,b], Chinar Dara[a], Areej Alasseri[a], Sonja A. Kotz[b]

[a] *McGill University, School of Communication Sciences and Disorders, 1266 avenue des Pins Ouest, Montréal, Canada H3G 1A8*
[b] *Max Planck Institute for Human Cognitive and Brain Sciences, Research Group "Neurocognition of Rhythm in Communication" Stephanstraße 1a D-04103 Leipzig, Germany*

Received 29 September 2008; received in revised form 20 April 2009; accepted 21 July 2009

Abstract

To understand how language influences the vocal communication of emotion, we investigated how discrete emotions are recognized and acoustically differentiated in four language contexts—English, German, Hindi, and Arabic. Vocal expressions of six emotions (anger, disgust, fear, sadness, happiness, pleasant surprise) and neutral expressions were elicited from four native speakers of each language. Each speaker produced pseudo-utterances ("nonsense speech") which resembled their native language to express each emotion type, and the recordings were judged for their perceived emotional meaning by a group of native listeners in each language condition. Emotion recognition and acoustic patterns were analyzed within and across languages. Although overall recognition rates varied by language, all emotions could be recognized strictly from vocal cues in each language at levels exceeding chance. Anger, sadness, and fear tended to be recognized most accurately irrespective of language. Acoustic and discriminant function analyses highlighted the importance of speaker fundamental frequency (i.e., relative pitch level and variability) for signalling vocal emotions in all languages. Our data emphasize that while emotional communication is governed by display rules and other social variables, vocal expressions of 'basic' emotion in speech exhibit modal tendencies in their acoustic and perceptual attributes which are largely unaffected by language or linguistic similarity.

1. Introduction

Nonverbal and paralinguistic cues provide a rich source of information about a speaker's emotions and social intentions when engaged in discourse (Wilson & Wharton, 2006). Emotions expressed in the face, and to a much lesser extent the voice, have been studied to elucidate a core set of emotions—most typically joy/happiness, anger, disgust, fear, sadness, and surprise (Ekman, Sorenson, & Friesen, 1969; Izard, 1994). While there are various ways for characterizing the emotional states and affective dimensions which can be expressed in speech, the notion of categorical emotions which are associated with discrete forms of expression is deeply entrenched in the literature (see Cowie & Cornelius, 2003 for a discussion). Many believe that characteristic expressions of basic emotions "erupt" in speech, often involuntarily, as one of the neurophysiological consequences of experiencing the emotion by the 'sender' or *encoder* of the expression. Perhaps owing to the biological significance of these expressions to con-specifics and their importance for adaptive behaviour, these emotional expressions are believed to possess certain invariant properties which allow them to be recognized independent of learning or culture when presented in the face (Ekman & Friesen, 1971) or in the voice (Scherer, Banse, & Wallbott, 2001).

The communicative or expressive aspects of emotional behaviour (e.g., emotional 'display rules', Ekman & Friesen, 1969 are also influenced by socio-cultural dimensions of an interaction. Despite similarities in how emotions are expressed across human cultures, the opportunity to express particular emotions and the form of these displays tend to vary according to cultural norms (Ekman et al., 1987; Elfenbein, Beaupré, Lévesque, & Hess,

*Corresponding author. Tel.: +514 398 4133; fax: +514 398 8123.
E-mail address: marc.pell@mcgill.ca (M.D. Pell).
URL: http://www.mcgill.ca/pell_lab (M.D. Pell).

0095-4470/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2009.07.005

2007; Mesquita & Frijda, 1992; Scherer, 1997). Moreover, cultural rules often dictate how males versus females communicate their emotions in speech or in the face (Atkinson, Tipples, Burt, & Young, 2005; Goos & Silverman, 2002; Hess, Adams, & Kleck, 2005; Hofmann, Suvak, & Litz, 2006; Wallbott, 1988). Finally, it is well recognized that individual encoders within a single language or cultural group further vary in how they regulate their nonverbal behaviour to express emotions in speech (Banse & Scherer, 1996; Wallbott & Scherer, 1986) and to achieve particular social-pragmatic goals, such as to signal dominance, affiliation, or to create other social impressions (Hess et al., 2005). It has been argued that individual differences and personality traits are of central importance for understanding emotional expressive behaviour (Matsumoto, 2006). Collectively, these studies underscore that social and interpersonal variables play an important role in how emotions are encoded in speech and other communication channels, and in how the receiver or *decoder* interprets their emotional meaning in many situations.

1.1. Effects of language on emotional communication

If one concentrates on the vocal expression of emotion in speech (*emotional prosody*), there is a conspicuous lack of research which directly compares how individuals from different linguistic and cultural backgrounds communicate their emotions. In speech, discrete emotion expressions are associated with characteristic variations in the acoustic structure of the speech signal, and the relative perturbation of specific acoustic cues over the course of an utterance, which listeners recognize as an utterance unfolds (Banse & Scherer, 1996; Juslin & Laukka, 2003). The prospect that vocal emotion expressions vary somewhat across languages is suggested by research which has presented vocal emotion expressions for *cross-cultural* recognition. These studies reveal that listeners can accurately detect and categorize vocal emotions when listening to a foreign language (Albas, McCluskey, & Albas, 1976; Pell, Monetta, Paulmann, & Kotz, 2009; Pell & Skorup, 2008; Scherer et al., 2001; Thompson & Balkwill, 2006; Van Bezooijen, Otto, & Heenan, 1983), consistent with the idea of basic human emotions and the existence of shared principles which guide emotional communication (Ekman, 1992). However, these same studies also typically demonstrate an in-group advantage for identifying vocal emotion expressions more accurately when produced by speakers of the same language when compared to speakers of a foreign language (see Elfenbein & Ambady, 2002 for an overview). Thus, based on the cross-cultural data it can be said that vocal emotion expressions seem to exhibit a core set of acoustic-perceptual features which promote accurate recognition across languages, but that there are also language-specific differences which lead to an in-group processing advantage (Pell & Skorup, 2008; Pell et al., 2009).

A separate literature has examined how vocal emotions are encoded and recognized in the context of specific languages (see Juslin & Laukka, 2003 for a comprehensive overview). From this research we know that emotional meanings in the voice are conveyed by concomitant changes in several acoustic parameters of speech, including but not limited to fundamental frequency (pitch), intensity (loudness), duration, rhythm, and different aspects of voice quality (Banse & Scherer, 1996). As demonstrated by Juslin and Laukka's (2003) meta-analysis, most researchers in the acoustic literature have measured changes in vocal pitch, intensity, and speech rate implying that these parameters are critical features of vocal emotion expressions; in particular, a speaker's pitch level (mean), pitch range (or variation), and speech rate appear to differentiate well among discrete emotion categories in both acoustic and perceptual terms (Mozziconacci, 2001; Pell, 2001; Williams & Stevens, 1972). For example, expressions of sadness tend to be produced with a relatively low pitch/fundamental frequency (f0) and slow speaking rate, whereas expressions of anger, fear, and happiness tend to be produced with a moderate or high mean f0 and fast speaking rate. In addition, anger and happiness usually display high f0 variation, whereas fear and sadness often exhibit less f0 variation (Juslin & Laukka, 2003; cf. Banse & Scherer, 1996; Sobin & Alpert, 1999; Williams & Stevens, 1972 for discussion and exceptions to these patterns). Emotions such as disgust and surprise have been studied less in the context of speech and their acoustic-perceptual features are more controversial, although there is evidence that disgust is sometimes produced with a low mean f0 (Banse & Scherer, 1996).

It bears noting that much of the information we have gained is based on analyses of posed or *simulated* exemplars of vocal emotion which were elicited from professional or lay actors who were native speakers of the language of interest. Given the close interplay of emotion and linguistic cues in speech, this investigative approach is often necessary in practical terms to control for variations in the linguistic content of utterances, especially when one of the research goals is to compare acoustic measures of different emotional expressions in speech which can be influenced by the segmental and suprasegmental properties of a language (Pell, 2001). Another characteristic of the present literature is that much of our knowledge of the acoustic properties of vocal emotions derives from major European languages such as English or German. Curiously, there have been few attempts to compare emotional expressions produced under a similar set of conditions by speakers of several different languages, especially languages which vary in their linguistic and/or cultural similarity. Thus, while there appear to be "modal tendencies" in how speakers encode discrete emotions in different languages (e.g, Scherer et al., 2001), this evidence is derived largely from the perceptual literature and/or through indirect comparisons of vocal emotion expressions produced in different languages and with different types of stimulus materials (words, sentences, nonsense speech, or spontaneous dialogue, see Juslin & Laukka, 2003).

Research which has undertaken a systematic, controlled study of the acoustic and perceptual properties of vocal emotion expressions in several languages in tandem is still rare (Burkhardt et al., 2006).

1.2. Research objectives

In the present study, our goal was to directly compare patterns for expressing and recognizing vocal emotion expressions which are assumed to possess certain invariant properties in four distinct language contexts. Given certain evidence that linguistic and/or cultural similarity could play a role in how vocal emotions are recognized (Scherer et al., 2001), we focused on four distinct languages which varied in a systematic manner in their "linguistic proximity" and typology: English, German, Hindi, and Arabic. Whereas English and German are considered closely related in both linguistic and cultural terms (i.e., both from the Germanic branch of Indo-European languages), Hindi is a more distantly related language from the Indo-European family, and Arabic comes from an entirely distinct language group (Semitic). In each language condition, a common procedure was followed: male and female encoders produced utterances in their native language to convey a standard set of different emotions by using their voice; and, recordings of the vocal stimuli were presented to a native listener group (half male, half female) who judged the intended emotion of the speaker for items produced in the *same* language. By following the same methods in each language condition, our data allowed us to examine patterns of vocal emotion recognition in each language context separately and through direct cross-language comparisons. We also extracted basic acoustic measures of the items presented in each language to compare the acoustic data with emotion recognition rates across languages. Although it was *not* the purpose of this study to present vocal expressions of emotion to listeners in their non-native language, complementary studies of this nature are ongoing (e.g., Pell et al., 2009).

One of the unique methodological challenges of studying vocal communication of emotion is how to isolate processes related to the encoding/decoding of emotions in the voice from those of processing linguistic-contextual cues of the utterance which accompany vocal emotion expressions; this potential confound affects any investigation of how emotions are recognized from vocal cues in speech because listeners may attend to corresponding linguistic features which bias or conflict with the meaning of the vocal cues. One way to circumvent the "isolation problem" is to require speakers to express emotions in "pseudo-utterances" which mimic the phonotactic and morpho-syntactic properties of the language of interest, in the absence of meaningful lexical-semantic information (Pell & Baum, 1997; Scherer, Banse, Wallbott, & Goldbeck, 1991). It has been shown that such stimuli can be produced in a relatively natural manner by encoders to portray a range of vocal emotions. The recorded utterances can then be judged by listeners, allowing inferences about the processing of vocal emotions in different languages in a controlled context where listeners must base their judgements strictly on vocal parameters of the utterances. We adopted this approach here to determine precisely how vocal cues operate during emotional encoding and recognition in each of our four language contexts.

Since there is little precedence for this research in the vocal literature, firm predictions about the influence of language on emotion recognition patterns or on the major acoustic cues involved could not be made with certainty. Based on our literature review, we expected that individual speakers/encoders in our experiment would display somewhat different abilities and patterns for encoding emotions in the voice, and that this would be true for each of our language conditions under study. We also predicted that listeners would be capable of identifying each of the target emotions from pseudo-utterances in their native language at levels exceeding chance, although one may expect variations in recognition accuracy and error confusion patterns when the language conditions are compared. Overall, it was expected that expressions of anger and sadness would yield the highest recognition rates in each language, and that disgust might lead to relatively poor recognition rates, although more precise patterns for identifying emotions and their relationship among the four language conditions was unclear. It was assumed that the major acoustic parameters of emotion expressions—mean f0, f0 range, and speech rate—would contribute significantly to differences among the emotion categories in each of the four languages. In light of evidence that vocal emotions can be recognized across languages, we expected to find qualitatively similar tendencies in how the acoustic parameters were associated with specific emotional expressions in the four languages, although the relationship between acoustic and perceptual measures of vocal emotion recognition has not previously been described in this way.

2. Methods

For each of the four languages under study (English, German, Hindi, Arabic), a common set of procedures was adopted to construct and validate the stimuli in each condition. For each language separately, an emotion elicitation study was first carried out to produce recordings of vocal emotion expressions from native speakers. At a second stage, an emotion perception study was undertaken to measure how the vocal stimuli are perceived by a group of *native* listeners and acoustic analyses were performed on a subset of the stimuli. (Listeners were only presented stimuli produced in their native language and never in one of the three foreign languages.) All procedures relating to the English, Hindi, and Arabic stimuli were executed in Montréal, Canada, whereas the German stimuli were prepared in Leipzig, Germany. Since the stimuli representing each of the four languages were constructed at different

Table 1
Example of pseudo-utterances produced by speakers of each language in the corresponding emotion elicitation study.

Language condition	Lexicalized utterance (e.g., *fear*)	Pseudo-utterance (e.g., *fear*)
English	The convict is holding a knife	The dirms are in the cindabal
German	Sie hat die Messer geschliffen und gezogen (She has sharpened and pulled the knife)	Mon set die Sonität verfüget ind geschweugen
Hindi	उस अपराधी के हाथ मे छुरा है। (That convict is holding a knife)	उस कोदी को मीगा।
Arabic	قنيكس كسام دلولا (The boy is holding a knife)	أغلاضْ الأخْوام صَبيرَة

Corresponding utterances with lexical content which biased the emotion were used to facilitate accurate vocal portrayals of the pseudo-utterances and are shown for comparative purposes only.

points in time and for slightly different experimental purposes, there were some differences in the number and duration of tokens analyzed when the four languages are compared (see below).

2.1. Emotion elicitation study

Participants—A total of 16 'encoders', two female and two male speakers of each language, were recruited to produce emotional expressions in their native language (4 encoders × 4 languages = 16 total). All encoders were native speakers who learned the target language from birth and continue to use that language in their home environment (many of the encoders knew additional languages as well). All English speakers were native to Eastern Canada (Québec/Ontario dialect) and all German speakers spoke standard high German. The Hindi speakers spoke standard Hindi as spoken in the central zone of India; all had grown up in India, moved to Montreal as adults, and continued to use Hindi as their dominant language in their social and work arenas. Arabic speakers were native to the Middle-East (Syrian/Jordanian dialects) but were studying at McGill University and had been in Montréal for less than three years. All encoders were young adults (mean age in years: English = 22.3; German = 29.5; Hindi = 28.8; Arabic = 24.8) and were selected for having lay experience in acting (e.g., in community theatre) or in public speaking (e.g., radio, member of a public speaking group).

Materials—In each language, a separate list of pseudo-sentences (English = 30, German = 40, Hindi = 35, Arabic = 20) was constructed by one of the authors who was a native speaker of that language. Pseudo-utterances were designed to be produced only by native speakers of the target language, in a non-emotional (neutral) manner and in six distinct emotional tones: anger, disgust, fear, sadness, happiness, and pleasant surprise. Sentences averaged seven syllables in length (range: 6–14 syllables) for all languages except for German, where these were longer (averaging 12 syllables).[1] For each language, pseudo-utterances were constructed by replacing all content words with sound strings that were phonologically licensed by the language but semantically meaningless to listeners. Because pseudo-utterances contained appropriate phonological and some grammatical properties of the target language and were therefore quite "language-like" to native speakers and listeners, they could be produced by native speakers to effectively communicate emotions following minimal practice.

To facilitate production of pseudo-utterances which were emotionally inflected in a way that was as natural as possible, a list of "lexicalized" utterances was also constructed for each language to convey each of the target emotions through both prosody and the verbal-semantic content of the sentence. Lexicalized stimuli, although not the subject of this report, were useful in the elicitation study as a means for helping the encoders produce pseudo-utterances with naturalistic inflections to specific emotions that resembled normal speaking conditions. The semantic content of lexicalized stimuli varied somewhat from language to language to ensure that these stimuli contained appropriate cultural references for each group under examination. Examples of pseudo-utterances and lexicalized utterances constructed for each language are shown in Table 1.

Elicitation and recording procedure—Each encoder was recorded separately in a sound-attenuated room. Pseudo-utterances conveying neutral affect and each of the six emotions were recorded in a separate block during the elicitation study. The order for recording specific emotion categories was varied across encoders. For each emotion, the encoder first practiced by producing the lexicalized utterances to express the target emotion. Following this, each pseudo-utterance from the list was presented one at a time in written format; encoders were instructed to first read and learn each target sentence, often by repeating it aloud, and then to produce the pseudo-sentence to express the target emotion (as if talking to the examiner). Encoders were encouraged to speak in a way that was natural for them, avoiding exaggeration. During recording, the examiner provided clues to help facilitate production of the target emotion, particularly at the onset of each emotion block; for example, the examiner described culturally appropriate situations which are likely to elicit the target

[1]The German lexicalized and pseudo-utterances were longer because they were simultaneously designed for presentation in ERP experiments, unlike the stimuli for English, Hindi, and Arabic.

emotion (Borod et al., 1998). At no time did the examiner model possible vocal features of the target emotion to participants. After each emotion block was completed, a break was imposed during the recording session to facilitate the transition between different modes of emotion expression.

All utterances were recorded onto digital media (audio or videotape) using a high-quality fixed (head-mounted or lapel) microphone. Encoders were paid for their participation. Recordings captured during the elicitation study were digitally transferred to a computer and saved as individual sound files, edited to mark the onset and offset of the sentence. The average duration of the vocal expressions was 1.73 s for English, 2.99 s for German, 1.33 s for Hindi, and 1.50 s for Arabic.

2.2. Perceptual-acoustic study

For each language separately, the edited pseudo-utterances were entered into a perceptual rating study to determine how each item was perceived by a group of native listeners. Then, based on the perceptual data, a subset of the utterances in each language condition were subjected to acoustic analysis as described in detail below.

Participants—A total of 87 listeners or 'decoders' took part in the study, divided into four separate groups of young adults who were native speakers of each language (English: $n = 24$; German: $n = 24$; Hindi: $n = 20$; Arabic: $n = 19$). The native listener group tested for each language condition was equal in the number of female and male decoders (Arabic: 9 females, 10 males). Participants in the four groups were also roughly equivalent in mean age (English = 24.9 ± 7.9; German = 24.2 ± 2.8; Hindi = 21.55 ± 3.0; Arabic = 23.9 ± 5.1) and in years of formal education (English = 16.6 ± 1.9; German = 16.6 ± 4.1; Hindi = 16.8 ± 2.5; Arabic = 15.8 ± 3.3). Most decoders in each listener group spoke more than one language (e.g., all German, Hindi, and Arabic decoders also knew English; many English decoders knew French). However, as each decoder only judged emotional pseudo-utterances produced by encoders of the *same* language, and all communication with the participant during testing always occurred in the native language, the fact that some of the decoders may have known one of the other languages included in the study (and possibly other languages) should not play a role in our experimental design.

Materials and procedure—All of the emotional pseudo-utterances produced by the four encoders of a given language were randomly combined and entered into a perception study, separately by language condition. Due to differences in the number of pseudo-utterances designed for each language condition, and after removing a small number of stimuli for which there were recording artifacts, a total of 840 utterances were entered into the perception study for English (30 sentences × 7 emotions × 4 speakers), 1120 utterances for German (40 sentences × 7 emotions × 4 speakers), 980 utterances for Hindi (35 sentences × 7 emotions × 4 speakers), and 555 utterances for Arabic (20 sentences × 7 emotions × 4 speakers). Each decoder judged a randomized sequence of the items produced in the same language over a series of blocks presented in two testing sessions. Decoders were tested individually in an experimental laboratory during two sessions, separated by a 1 week interval. During the experiment, each utterance was played a single time over headphones controlled by the computer; after each item, the decoder was instructed to make two judgements in sequence. First, the participant categorized which emotion was being expressed by the speaker from seven alternatives (the six emotions plus neutral) by selecting the corresponding emotion term from a printed list on the computer screen or on the button-press panel.[2] Once the emotion of the voice was categorized, a five-point rating scale appeared on the computer screen and the participant was required to rate the intensity of the emotional meaning selected for that item. Decoders used a button press to indicate each of their decisions and the data were saved automatically by the computer. Only the data on how listeners judged the emotion category (and not intensity) of each stimulus were analyzed for this report.

Perceptual data analysis—Analysis of the perceptual data began by inspecting the individual performance of the four encoders in the elicitation study for each language condition. This allowed us to highlight the extent to which there were individual differences in the ability to encode vocal emotions, as this is commonly observed in both natural settings and artificial paradigms (Scherer et al., 1991). Even more critically, inspecting individual performance features allowed us to estimate how many of the recorded tokens were likely representative of the intended target emotion which was critical to our present goals; we expected that many of the vocal expressions recorded would not adequately portray the intended emotional target due to difficulties adhering to the elicitation/simulation procedure, yielding a number of utterances which were unnatural sounding or emotionally ambiguous. We did not want these experimental performance factors to influence our description of the acoustic-perceptual features of emotion recognition in the four language conditions.

To control for these variables inasmuch as possible, we adopted a criterion to limit all perceptual and acoustic data analyses to items which obtained a reasonable consensus about the emotion conveyed by the native listener group. Previous work suggests that vocal emotions (excluding surprise and neutral) are recognized at rates approximating four times chance (Scherer et al., 1991). Accordingly, to include as many "valid" exemplars as possible, our criterion here was set at a minimum native-group

[2]The precise verbal labels used in each language were: for English, anger, disgust, fear, sadness, happiness, pleasant surprise, neutral; for German, Ärger, Ekel, Angst, Trauer, Freude, freudige Überraschung, Neutral; for Arabic, قدي احم, قدي عس ةئ اجافم, قدا عس, نز ح, فـو خ, فـرق, بـض غ and for Hindi, गुस्सा, घिन, भय, उदास, खुश, सुखःद आश्चर्य, बिना भाव के.

Table 2
Frequency of valid tokens (i.e., stimuli with >42.85% emotional target recognition) included in the main experiment, distributed over actors and emotions in each language condition. The proportion of valid tokens retained in each condition is shown in parentheses.

Emotion target	Language/speaker (sex)																		Sum/ emotion
	English (/30)				German (/40)				Hindi (/35)				Arabic (/20)						
	NA (F)	SL (F)	DF (M)	MG (M)	K (F)	V (F)	C (M)	S (M)	JC (F)	RM (F)	AO (M)	KS (M)	FA (F)	YN (F)	IF (M)	MH (M)			
Anger	23 (0.77)	30 (1.0)	30 (1.0)	30 (1.0)	40 (1.00)	40 (1.00)	39 (0.98)	39 (0.98)	31 (0.89)	0 (0.00)	31 (0.89)	35 (1.00)	18 (0.90)	0 (0.00)	14 (0.70)	6 (0.30)	406 (0.81)		
Disgust	25 (0.83)	16 (0.53)	30 (1.0)	30 (1.0)	19 (0.48)	23 (0.58)	39 (0.98)	40 (1.00)	0 (0.00)	1 (0.03)	32 (0.91)	30 (0.86)	4 (0.20)	7 (0.35)	7 (0.35)	8 (0.40)	311 (0.62)		
Fear	28 (0.93)	30 (1.0)	29 (0.97)	30 (1.0)	36 (0.90)	36 (0.90)	20 (0.50)	38 (0.95)	4 (0.11)	2 (0.06)	32 (0.91)	34 (0.97)	9 (0.45)	13 (0.65)	12 (0.60)	14 (0.70)	367 (0.73)		
Sad	30 (1.0)	30 (1.0)	30 (1.0)	29 (0.97)	40 (1.00)	40 (1.00)	20 (0.50)	30 (0.75)	35 (1.00)	35 (1.00)	34 (0.97)	31 (0.89)	12 (0.60)	20 (1.00)	15 (0.75)	20 (1.00)	451 (0.90)		
Happy	29 (0.97)	30 (1.0)	18 (0.60)	25 (0.83)	29 (0.73)	25 (0.63)	31 (0.78)	5 (0.13)	26 (0.74)	29 (0.83)	15 (0.43)	8 (0.23)	9 (0.45)	15 (0.75)	6 (0.30)	7 (0.35)	307 (0.61)		
Surprise	17 (0.57)	16 (0.53)	30 (1.00)	30 (1.0)	29 (0.73)	1 (0.03)	19 (0.48)	34 (0.80)	28 (0.80)	31 (0.89)	12 (0.34)	17 (0.49)	0 (0.0)	1 (0.05)	2 (0.10)	4 (0.20)	271 (0.54)		
Neutral	30 (1.0)	29 (0.97)	30 (1.00)	27 (0.90)	40 (1.00)	40 (1.00)	38 (0.95)	40 (1.00)	0 (0.00)	11 (0.31)	35 (1.00)	34 (0.97)	16 (0.80)	13 (0.65)	18 (0.90)	5 (0.25)	406 (0.81)		
Sum/speaker	182 (0.87)	181 (0.86)	197 (0.94)	201 (0.96)	233 (0.83)	205 (0.73)	206 (0.74)	226 (0.81)	124 (0.51)	109 (0.45)	191 (0.78)	189 (0.77)	68 (0.49)	69 (0.49)	74 (0.53)	64 (0.46)	2519 (0.72)		
Sum/ Language	761 (0.91)							870 (0.78)				613 (0.63)				275 (0.49)			

consensus for each item of three times chance performance in the seven-choice emotion recognition task, or 42.86% accuracy per item. The frequency and ratio of tokens considered perceptually valid for each of the 16 encoders for each emotion type are summarized in Table 2.

For English, applying this criterion led to the inclusion of 91% (761/840) of the original tokens representing the seven emotions (participant NA = 87%, SL = 86%, DF = 94%, MG = 96%). The majority of English items which failed to reach the validity criterion (>42.86%) fell in the category of "pleasant surprise" (approximately 1/3 of all omitted items). For German, a total of 78% (870/1120) of tokens were retained ($K = 83\%$, $V = 73\%$, $C = 74\%$, $S = 81\%$); for this stimulus set, discarded items fell predominantly in the "pleasant surprise" and "happiness" categories (32% and 27% of omitted items). For Hindi, a total of 63% (613/980) of the original tokens were retained (JC = 51%, RM = 45%; AO = 79%; KS = 77%) which mostly affected the categories of "disgust", "fear" and "happiness". For Arabic, a total of 49% (275/555) of the original items were retained (FA = 49%, YN = 49%, IF = 53%, MH = 46%). For Arabic, items which did not reach the perceptual criterion fell in several categories: "pleasant surprise" (25% of omitted items), "disgust" (19%), "anger" (15%), and "happy" (15%). Across languages, approximately 1/3 of items deemed perceptually invalid according to our criterion were meant to communicate pleasant surprise, and there were several cases where a single encoder did not produce a valid exemplar for at least one emotion type (although this did not systematically occur for the same emotion). In total, our analyses were based on 2519 perceptually valid emotional exemplars across languages. Within each language condition, there was a relatively equal contribution of tokens from each of the four encoders, although tokens were unevenly distributed as a combination of encoder and emotion type.

Acoustic data analysis—Acoustic analyses were performed on all valid exemplars of vocal emotion in each of the four languages (English = 761 tokens, German = 870 tokens, Hindi = 613 tokens, Arabic = 275 tokens). Given the broad scope of this investigation which elicited six discrete emotions and neutral exemplars in four language contexts (2519 tokens in total), acoustic analyses were limited to three critical parameters that frequently differentiate among vocal emotion categories: mean fundamental frequency (f0Mean, in Hertz); fundamental frequency range (f0Range, in Hertz), and speaking rate (SpRate, in syllables per second). Acoustic analyses were performed using Praat speech analysis software and were always derived from the whole utterance. At the first step, several acoustic measures (mean, minimum, and maximum f0; utterance duration) were computed automatically by Praat for each utterance; then, all frequency measures were manually inspected and corrected by one of the investigators when the algorithm led to obvious 'doubling' or 'halving' errors in pitch tracking (approximately .05% of all tokens); finally, the three acoustic measures of interest

were calculated and normalized before conducting any statistical analyses to allow for valid insights about emotion expression in different languages which average *across* speakers.

Notably, mean vocal frequency varies naturally among individual speakers (especially as a function of sex) and absolute differences in f0 range vary as an index of the speaker's meanf0. To standardize our two f0 variables (f0Mean, f0Range), we chose a speaker's "resting frequency" as the anchor point for referencing all observed f0 values for that speaker. There is evidence that when speaking in a non-emotional/affirmative manner, each speaker returns to a highly stable "resting frequency" or end-point f0 at the end of their utterances which is characteristic for that individual (Menn & Boyce, 1982). Accordingly, we normalized the f0 data for each speaker in reference to the average *minimum* f0 observed for all utterances produced by that speaker in the neutral condition. A single Resting Frequency (in Hz) was identified for each of the 16 encoders and all normalized measures for a given speaker were then expressed as the *proportional distance* of the observed value in reference to the speaker's natural resting frequency.[3] To normalize f0Mean, the computed mean f0 of each utterance in Hertz was standardized as follows: $\text{f0Mean}_{\text{Norm}} = (\text{f0Mean}_{\text{observed}} - \text{Resting frequency})/\text{Resting frequency}$. To normalize f0Range, the observed maximum and minimum f0 of an utterance were each standardized in reference to the speaker's resting frequency using the same formula, and then the f0Range was computed by subtracting the normalized f0Min from the normalized f0Max, where $\text{f0Range}_{\text{Norm}} = ((\text{f0Max}_{\text{observed}} - \text{Resting frequency})/\text{Resting frequency}) - ((\text{f0Min}_{\text{observed}} - \text{Resting frequency})/\text{Resting frequency})$. For both normalized frequency measures (f0Mean, f0Range), this meant that a value of 1 for a given utterance represents a doubling or 100% increase in a speakers' resting frequency in that instance which was a standardized distance across speakers and languages. Similarly, a normalized f0Range value of 1 would always mean that a speaker used an expressive range that was twice the speaker's resting frequency in that particular condition (thus normalizing for the effects of f0Mean on f0Range). Finally, speaking rate was calculated by taking the number of syllables in an utterance and dividing by the corresponding utterance duration (in seconds); this yielded a SpRate measure expressed in syllables per second which could be reliably compared across speakers, emotions, and language conditions.

Statistical analyses—All statistical analyses referred to data obtained for the valid exemplars of emotion for each language. Within each language condition, we first sought to exemplify how the six emotions and neutral utterances (i.e., seven emotion categories) could be differentiated both perceptually and acoustically in that language; univariate and/or multivariate analysis of variance (ANOVA/MANOVA) was performed on the emotional target hit rates (% accuracy), normalized f0Mean (in Hz), normalized f0Range (in Hz), and Speaking Rate (syllables/s). In addition, a discriminant function analysis was performed on the data in each language condition to determine how well the seven emotion categories could be classified in each language based on the three acoustic measures. These analyses, which were not cross-validated, focused on differences among the items (rather than participants) since our main purpose was to generalize about the perceptual and acoustic features of valid emotional expressions within each language. (For comparative purposes, emotion target recognition patterns were simultaneously examined by subjects with Emotion as a repeated measure in the analysis.) Significant effects emerging from the ANOVAs, whenever relevant, were always elaborated through Tukey's (HSD) post hoc comparisons on the marginal means ($p < .01$). After describing factors in the recognition of vocal emotion in each language, a final analysis compared emotion recognition patterns in the four languages directly.

3. Results

Table 3 presents the mean recognition rates and Table 4 presents the acoustic data for valid exemplars of each emotion averaged across the four encoders, per language condition. As expected, despite eliminating tokens which were poorly recognized owing to presumed difficulties in the ability to consistently pose emotion expressions, there was marked variability in how accurately decoders recognized specific emotions from the voice in each of the four languages of interest. Based on qualitative inspection of data in Table 3, it was noted that "pleasant surprise" (hereafter, simply "surprise") tended to result in poor recognition overall. With a few exceptions, anger, sadness, and fear tended to result in relatively good recognition across languages. A tendency for accuracy to be lower in the Arabic language condition overall when compared to the English, German, and Hindi conditions was observed.

3.1. Characterizing vocal emotion expressions by language

3.1.1. English

Perceptual data—Emotion recognition patterns for English are presented in the top panel of Table 3. Visual inspection of the error patterns suggests that for English, surprise was frequently confused for happiness (77% of all error responses to surprise expressions). To evaluate whether emotional expressions in English differed in how well they were recognized, a one-way ANOVA was run on the mean target recognition rates (% correct) as a function of Emotion type (anger, disgust, fear, sadness, happiness, surprise, neutral). The effect of Emotion on recognition

[3]Speakers had the following Resting Frequencies (in Hz): for English, NA = 154, SL = 150, DF = 103, MG = 91; for German, $V = 180$, $K = 144$, $S = 83$, $C = 91$; for Hindi, JC = 137, RM = 167, AO = 91, KS = 97; for Arabic, FA = 161, YN = 200, IF = 106, MH = 102.

Table 3
Percentage of English ($n = 24$), German ($n = 24$), Hindi ($n = 20$), and Arabic ($n = 19$) decoders who identified pseudo-utterances according to each emotion target (averaged for the four decoders).

Language	Emotion target	Percentage of responses						
		Anger	Disgust	Fear	Sadness	Happiness	Surprise	Neutral
English	Anger	**88.4**	4.9	0.5	0.4	2.2	1.2	2.4
	Disgust	2.4	**76.4**	1.1	13.3	2.5	1.3	3.0
	Fear	0.4	1.3	**87.4**	4.2	1.1	4.6	1.0
	Sadness	0.2	4.2	2.1	**90.5**	0.2	0.0	2.8
	Happiness	2.3	4.7	0.6	1.4	**79.6**	3.6	7.8
	Surprise	2.5	2.7	0.5	0.4	21.9	**71.5**	0.5
	Neutral	1.9	3.3	0.1	9.8	8.1	0.4	**76.4**
German	Anger	**88.0**	0.6	0.8	0.1	2.4	4.3	3.8
	Disgust	3.3	**76.6**	2.2	1.5	3.7	4.4	8.3
	Fear	0.2	3.4	**70.8**	20.5	0.1	3.0	2.0
	Sadness	0.2	4.8	17.4	**72.6**	0.2	0.5	4.3
	Happiness	3.3	0.4	0.5	0.0	**59.6**	24.8	11.4
	Surprise	3.1	0.9	3.4	0.2	21.9	**68.8**	1.7
	Neutral	0.8	1.3	0.4	1.5	0.9	0.8	**94.3**
Hindi	Anger	**74.4**	12.1	3.4	0.6	4.5	2.6	2.4
	Disgust	20.7	**63.9**	2.1	1.2	6.1	4.0	2.0
	Fear	6.6	1.8	**75.6**	4.9	2.6	4.8	3.7
	Sadness	1.0	2.2	8.6	**75.7**	0.5	0.9	11.1
	Happiness	7.4	2.5	2.0	2.8	**67.1**	12.6	5.6
	Surprise	11.1	5.1	3.9	1.0	17.6	**57.9**	3.4
	Neutral	2.0	1.6	2.0	26.1	1.4	0.8	**66.1**
Arabic	Anger	**63.1**	13.8	4.0	4.2	4.6	1.9	8.4
	Disgust	12.4	**55.0**	5.0	8.8	4.8	2.2	11.8
	Fear	6.4	3.3	**62.3**	10.0	4.7	7.4	5.9
	Sadness	2.5	5.3	4.3	**74.7**	2.2	0.6	10.4
	Happiness	2.1	4.8	4.5	5.9	**59.9**	8.5	14.3
	Surprise	7.6	3.4	7.6	0.8	26.0	**50.4**	4.2
	Neutral	6.7	6.9	4.4	14.1	3.2	1.2	**63.5**

accuracy was significant when analyzed by subjects, $F(6, 138) = 17.15$, $p < 0.0001$, and by items, $F(6, 754) = 31.44$, $p < 0.0001$. Post hoc (Tukey's) comparisons performed on the subject data revealed that expressions of sadness (91%), anger (88%) and fear (87%) were recognized significantly more accurately from English pseudo-utterances than expressions of happiness (80%), neutral (76%), disgust (76%), and surprise (72%). Happy expressions were also recognized significantly better than those conveying surprise.

Acoustic data—To characterize the relationship among the perceptually valid exemplars of each emotion and the three acoustic measures of interest, a one-factor, between-subjects multivariate analysis of variance (MANOVA) was carried out on the 761 English items. The three normalized acoustic measures (f0Mean, f0Range, SpRate) served as the dependent variables in the analysis, and the seven emotion types served as the independent variable. Results of the MANOVA were statistically significant according to Wilks' $\Lambda(0.08)$, $F(18, 2130) = 165.91$, $p < 0.001$. Univariate analyses indicated that the influence of Emotion was significant for f0Mean, $F(6, 754) = 435.20$, $p < 0.0001$, f0Range, $F(6, 754) = 271.89$, $p < 0.0001$, and SpRate, $F(6, 754) = 142.81$, $p < 0.0001$. Post hoc Tukey's tests performed separately for each acoustic parameter can be summarized as follows: for f0Mean, surprise was expressed with a very high f0Mean, which surpassed fear, then anger, and then happiness (all contrasts were significantly different). Sadness, disgust, and neutral expressions were produced with a significantly lower f0Mean than all the other emotions, and neutral expressions also exhibited a lower f0Mean than sad expressions for English. For f0Range, surprise was produced with the widest f0Range, followed by anger, followed by fear (all three contrasts were significant). These three emotions had a significantly wider f0Range than happiness, disgust, sadness, and neutral expressions. Happy expressions also surpassed sad and neutral expressions in their f0Range, and disgust was greater than neutral. For SpRate, it was noted that fear was produced significantly faster than all other emotion expressions, and disgust was expressed with a significantly slower rate than all other emotions. After fear, neutral expressions were spoken with the quickest rate, significantly greater than happy and surprise expressions, which in turn were significantly faster than anger and sad expressions (which all exceeded disgust). The manner in

Table 4
Normalized acoustic measures of valid emotional expressions produced in English ($n = 761$), German ($n = 870$), Hindi ($n = 613$), and Arabic ($n = 275$), by emotion.

Language	Emotion	Acoustic measure		
		f0Mean (Hz)	f0Range (Hz, Max-Min)	Speaking rate (syllables/s)
English	Anger	0.72	1.39	3.91
	Disgust	0.33	0.86	3.20
	Fear	1.16	1.21	5.58
	Sadness	0.36	0.72	3.91
	Happiness	0.49	1.01	4.41
	Surprise	1.80	2.86	4.44
	Neutral	0.24	0.59	4.75
German	Anger	1.08	1.67	4.43
	Disgust	0.68	1.28	3.58
	Fear	0.79	0.92	4.13
	Sadness	0.51	0.79	4.00
	Happiness	1.04	2.01	4.65
	Surprise	1.63	2.38	4.26
	Neutral	0.44	1.28	3.91
Hindi	Anger	1.25	1.86	6.39
	Disgust	0.75	1.56	4.42
	Fear	1.55	1.26	6.19
	Sadness	0.41	0.75	4.08
	Happiness	0.95	1.65	4.92
	Surprise	1.44	1.84	5.43
	Neutral	0.24	0.67	5.15
Arabic	Anger	0.35	0.81	5.01
	Disgust	0.43	1.06	4.33
	Fear	0.85	0.88	6.12
	Sadness	0.25	0.49	4.94
	Happiness	0.57	1.11	4.38
	Surprise	1.14	1.72	4.97
	Neutral	0.33	0.81	5.57

which speakers varied their fundamental frequency (i.e., f0mean, f0range) and speaking rate to communicate emotions in English is furnished in Figs. 1a and 2a respectively, expressed in reference to neutral expressions which were always plotted at 0.

Discriminant analysis—A discriminant function analysis was then run to estimate how well the three acoustic parameters of interest could account for differences among the seven emotion categories (where all valid exemplars of each emotion were designated by a categorical grouping variable). The discriminant analysis produced three significant canonical functions (Function 1, $F(18, 2130) = 164.81$, $p < 0.0001$; Function 2, $F(10, 1508) = 85.43$, $p < 0.0001$; Function 3, $F(4, 755) = 22.93$, $p < 0.0001$). The first canonical function explained 75% of the variance and correlated positively with both f0Mean ($r = 0.98$) and f0Range ($r = 0.84$). The second function accounted for 23% of the remaining variance and correlated negatively with SpRate ($r = -0.84$). The third canonical function accounted for 2% of the variance and was positively associated with both SpRate ($r = 0.34$) and f0Range ($r = 0.29$). Overall, contributions of the three acoustic parameters in this model led to accurate emotional classification of 58% (443/761) of the perceptually valid exemplars in English. The success of the classification function was unevenly distributed across emotion categories: only 20% (20/102) of happy expressions, 35% (40/113) of anger expressions, and 40% (48/119) of sad expressions were correctly classified according to changes in the three acoustic parameters. In contrast, fear (87%, or 102/117), surprise (86%, 80/93), disgust (73%, 74/101), and neutral (68%, 79/116) expressions were predicted by the acoustic data at relatively high levels for English.

3.1.2. German

Perceptual data—Patterns for identifying vocal emotions in German are reported in the middle of Table 3. Cursory examination of the error patterns implies that fear and sadness were often confusable for German listeners in both directions, and happiness and surprise were also confusable in both directions. The ANOVA performed on the emotion target recognition scores yielded a significant Emotion effect, $F_{\text{Subjects}}(6, 126) = 50.73$, $p < 0.0001$, $F_{\text{Items}}(6, 863) = 95.78$, $p < 0.001$. Post hoc comparison of the subject means showed that recognition of neutral expressions in German pseudo-utterances (94% correct) was significantly more accurate than anger (88%) and disgust (77%). These three emotions were recognized significantly better than sadness (73%), fear (71%), and surprise (69%), which were associated with comparable recognition rates. Happiness in German was recognized significantly more poorly than all other emotions (60% correct).

Acoustic data—The independent effect of Emotion type on acoustic measures derived from the 870 perceptually valid German items was examined in a one-factor MANOVA using the three acoustic measures as dependent factors in the analysis. The MANOVA was significant according to Wilks' $\Lambda(0.28)$, $F(18, 2436) = 77.61$, $p < 0.01$. Follow-up, univariate analyses demonstrated a significant effect of Emotion on each acoustic measure: f0Mean, $F(6, 863) = 170.36$, $p < 0.0001$; f0Range, $F(6, 863) = 153.20$, $p < 0.0001$; and SpRate, $F(6, 863) = 27.78$, $p < 0.0001$. Tukey's HSD post hoc comparisons showed that for f0Mean, surprise was significantly higher than anger and happiness, which were in turn significantly higher than disgust and fear. Neutral and sad expressions were expressed with the lowest f0Mean (significantly less than all other emotions). For f0Range, surprise exhibited the greatest variability, which was significantly greater than happiness, which exceeded anger. These three emotions exhibited a significantly greater f0Range than disgust and neutral expressions (which did not differ), which in turn was greater than fear and sadness (which also did not differ). For SpRate, happy and angry expressions in German were spoken most quickly, at a rate which significantly exceeded that of fear, surprise, sadness, and neutral utterances (which did not differ). Expressions of disgust were significantly slower when compared to all

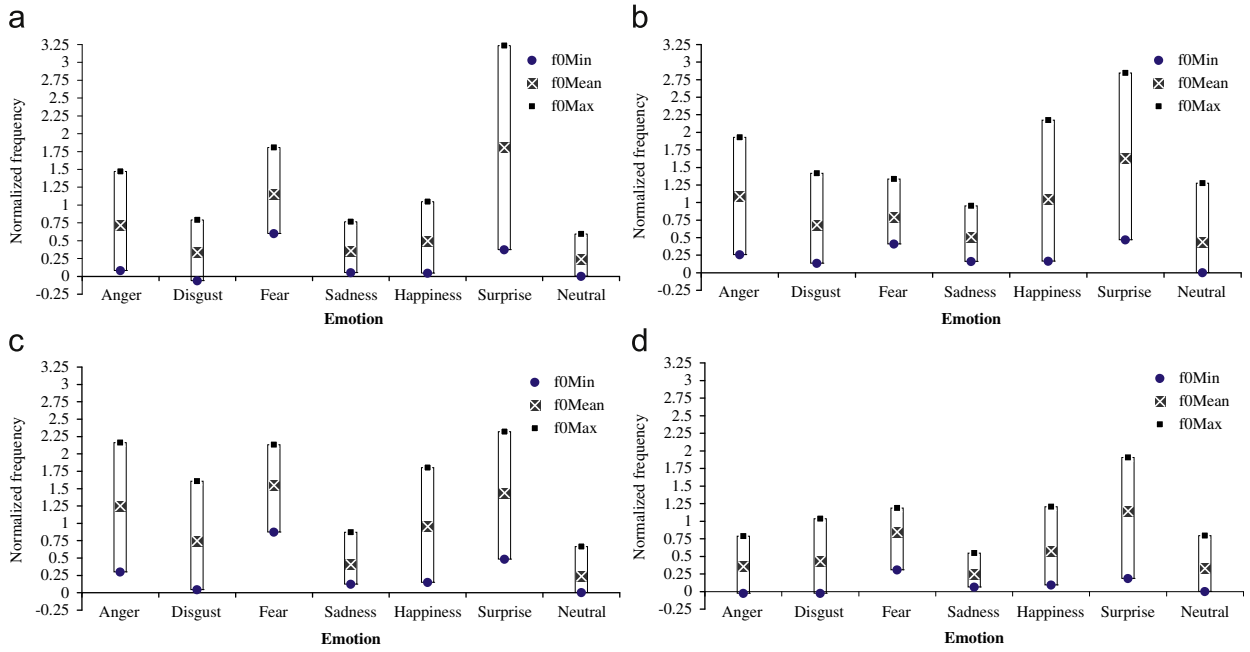

Fig. 1. Mean fundamental frequency (f0) and f0 range (maximum–minimum) associated with each of the seven emotions for (a) English ($n = 761$ tokens), (b) German ($n = 870$ tokens), (c) Hindi ($n = 613$ tokens), and (d) Arabic ($n = 275$ tokens). Measures were normalized for each speaker in reference to the neutral category (where minimum f0 for neutral = 0) and then averaged across speakers within each category.

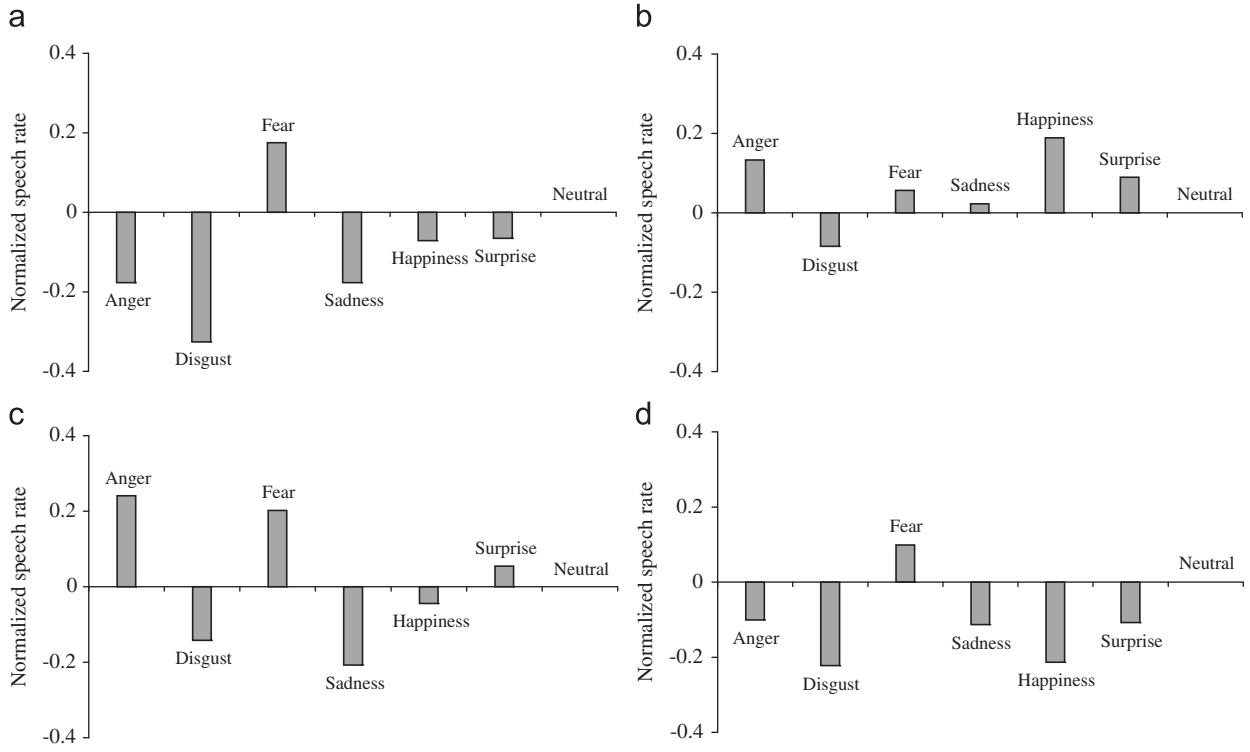

Fig. 2. Mean speaking rate associated with each of the seven emotions for (a) English ($n = 761$ tokens), (b) German ($n = 870$ tokens), (c) Hindi ($n = 613$ tokens), and (d) Arabic ($n = 275$ tokens). Speaking rate was normalized in reference to the neutral category (where speaking rate = 0) within each language condition.

other emotions. The acoustic findings for German are illustrated in Fig. 1b (for f0Mean and f0Range) and Fig. 2b (for SpRate).

Discriminant analysis—Analysis of the 870 perceptually valid German items, with the three acoustic measures as predictor variables, produced three significant canonical

functions (Function 1, $F(18, 2435) = 77.62$, $p < 0.0001$; Function 2, $F(10, 1724) = 34.66$, $p < 0.0001$; Function 3, $F(4, 863) = 21.51$, $p < 0.0001$). The first canonical function accounted for 79% of the variance and correlated positively with both f0mean ($r = 0.94$) and f0Range ($r = 0.69$). The second function accounted for 16% of the variance and was associated with f0Range ($r = 0.67$), whereas the third function explained 5% of the variance and was positively correlated with SpRate ($r = 0.93$). Overall, the discriminant function for German correctly predicted the emotional category of 49% (424/870) of the perceptually valid items recognized by decoders. Again, there were marked differences in how the acoustic data predicted specific emotion categories: the most poorly classified emotions in German were anger (28%, 44/158), disgust (32%, 39/121), and fear (41%, 53/130). The model was relatively more successful at classifying neutral (53%, 84/158), sadness (59%, 77/130), happiness (60%, 54/90), and surprise (88%, 73/83) in German.

3.1.3. Hindi

Perceptual data—Patterns of emotion recognition for Hindi are furnished in the middle of Table 3. Visual inspection of these data reveals that disgust was often confused with anger (57% of all errors to disgust stimuli), and neutral expressions were frequently confused for sadness (78% of all errors to neutral stimuli). The ANOVA performed on the recognition data yielded a significance influence of Emotion when analyzed by subjects, $F_{\text{Subjects}}(6, 114) = 2.64$, $p < 0.05$, and by items, $F_{\text{Items}}(6,606) = 24.37$, $p < 0.001$. For the subject data, post hoc tests showed that sadness (76%), fear (76%), and anger (74%) were recognized most accurately by Hindi decoders, significantly better than happiness (67%), neutral (66%), and disgust (64%). Surprise was recognized significantly worse than all other emotions (58%).

Acoustic data—The three acoustic measures obtained from the 613 emotional expressions for Hindi were entered as dependent variables in a MANOVA with Emotion as the independent factor. Results of the MANOVA were statistically significant, Wilks' $\Lambda(0.17)$, $F(18, 1709) = 83.81$, $p < 0.001$. One-way ANOVAs on each acoustic measure yielded a significant Emotion effect on all three parameters: f0Mean, $F(6, 606) = 159.66$, $p < 0.0001$, f0Range, $F(6, 606) = 73.55$, $p < 0.0001$, and SpRate, $F(6, 606) = 125.06$, $p < 0.0001$. Following post hoc (Tukey's) tests, it was found that Hindi speakers expressed fear and surprise by elevating their f0Mean; these two emotions displayed a significantly higher f0Mean than anger, which in turn surpassed happiness and disgust. Sadness and neutral expressions were produced with a significantly lower f0Mean than all other emotions. For f0Range, surprise, anger, and happiness exhibited the widest f0Range, which was significantly greater than disgust and fear. Neutral and sadness exhibited a significantly reduced f0Range when compared to all other emotions. For SpRate, expressions of anger and fear displayed the highest speaking rate for Hindi, which was significantly greater than surprise, neutral, and happiness (the speaking rate of surprise was also greater than happiness). Disgust and sadness demonstrated a significantly slower speaking rate than all other emotions. These patterns are presented graphically in Figs. 1c and 2c.

Discriminant analysis—Analysis of the 613 Hindi items in relation to the three acoustic variables of interest produced three significant canonical functions (Function 1, $F(18, 1708) = 83.82$, $p < 0.0001$; Function 2, $F(10, 1210) = 50.74$, $p < 0.0001$; Function 3, $F(4, 606) = 55.57$, $p < 0.0001$). The first canonical function explained 70% of the variance and correlated positively with f0Mean ($r = 0.92$), SpRate ($r = 0.82$) and f0Range ($r = 0.60$). The second function explained 17% of the variance and correlated most strongly with SpRate ($r = 0.55$), whereas the third function (13% of the variance) correlated positively with f0Range ($r = 0.66$). Overall, the three canonical functions for Hindi led to accurate emotional classification of 56% (343/613) of the perceptually valid items. The most poorly classified emotions for Hindi were happiness (18%, or 14/78 items) and surprise (35%, 31/88), whereas the discriminant functions resulted in increasingly accurate classification of disgust (51%, 32/63), fear (60%, 43/72), sadness (68%, 92/135), anger (69%, 67/97) and especially neutral (80%, 64/80).

3.1.4. Arabic

Perceptual data—Patterns for recognizing vocal emotions in Arabic are furnished in the bottom panel of Table 3. Qualitative analysis of confusion patterns among the emotion categories suggests that expressions of surprise often tended to be confused for happiness (53% of all errors to surprise). An ANOVA confirmed that target recognition accuracy in Arabic was influenced significantly by Emotion, $F_{\text{Subjects}}(6, 108) = 4.16$, $p < 0.01$, $F_{\text{Items}}(6, 268) = 15.07$, $p < 0.001$. Post hoc comparisons for the subject data showed that expressions of sadness (75% correct) were recognized best overall, significantly better than all other emotions (which ranged from a high of 63% for anger and neutral to a low of 50% for surprise).

Acoustic data—As before, a one-factor MANOVA was carried out with the seven Emotion types serving as the independent variable, and the three acoustic measures acting as the dependent variables (acoustic measures were obtained from the 275 perceptually valid Arabic utterances). The MANOVA results for Arabic were significant according to Wilks' $\Lambda(0.19)$, $F(18, 753) = 32.98$, $p < 0.01$. Univariate analysis of each dependent variable yielded a significant Emotion effect for f0Mean, $F(6, 268) = 69.18$, $p < 0.0001$, f0Range, $F(6, 268) = 23.94$, $p < 0.0001$, and SpRate, $F(6, 268) = 26.66$, $p < 0.001$. For f0Mean, post hoc tests revealed that Arabic speakers produced surprise with a highly elevated f0Mean which was significantly greater than fear. The f0Mean for all other emotion expressions was significantly lower than that of surprise and fear (happiness was also significantly higher in f0Mean than

anger and neutral). For f0Range, surprise demonstrated a significantly wider f0Range than all other emotions and sadness demonstrated a significantly smaller f0Range than all other emotions. Happiness also displayed a wider f0Range than anger and neutral expressions. For SpRate, fear was expressed with the fastest speaking rate, which was significantly greater than neutral, which in turn was significantly faster than anger, surprise, and sadness. Disgust and happiness were expressed with a significantly slower speaking rate than all other emotions (see Figs. 1 and 2).

Discriminant analysis—This analysis which included the 275 valid tokens in Arabic yielded three significant canonical discriminant functions (Function 1, $F(18, 752) = 32.96$, $p < 0.0001$; Function 2, $F(10, 534) = 19.49$, $p < 0.0001$; Function 3, $F(4, 268) = 11.63$, $p < 0.0001$). The first function accounted for 70% of the total variance and was strongly correlated with f0Mean ($r = 0.94$). The second canonical function explained 23% of the variance and correlated with f0Range ($r = 0.78$) and SpRate ($r = -0.73$) which were inversely related. The third function accounted for 7% of the variance and was positively correlated with f0Range ($r = 0.51$) and SpRate ($r = 0.50$). This model successfully predicted the emotion category of 53% (145/275) of the perceptually valid items for Arabic. Anger expressions were very poorly predicted by these acoustic parameters (5% of all items, or 2/38), followed by disgust (42%, 11/26) and happiness (46%, 17/37). There was comparatively good classification of neutral expressions (58%, 30/52), sadness (69%, 46/67), fear (69%, 33/48), and surprise (86%, 6/7), although the latter emotion was distinct in its very small number of valid tokens in this language condition.

3.2. Characterizing vocal emotion expressions across languages

A final stage of analysis looked at qualitative and relational differences in how the seven emotion categories were communicated when the four languages are compared directly. In addition to showing the emotional target hit rates in each language, Table 3 exemplifies the emotional confusion patterns observed for decoders of English, German, Hindi, and Arabic. Qualitative inspection of these patterns reveals certain cross-language tendencies, and many differences, in the *direction* of errors witnessed during vocal emotion recognition. If one looks simply at the most frequent error response for each emotion category across languages, there was a systematic error observed for only two emotional expression types: surprise was always misjudged as conveying happiness; and neutral utterances were always most frequently mislabeled as conveying sadness. In three of the language conditions, anger was confused with disgust (English, Hindi, Arabic), although sometimes for surprise (German). Fear was confused in relatively equal proportions with surprise and sadness in three languages (English, Hindi, Arabic), although predominantly for sadness in German. Happiness was misjudged as sounding either neutral (English, Arabic) or as surprise (German, Hindi). Results for sadness and disgust were even more variable across languages: sadness was mislabeled as disgust (English), fear (German), or neutral (Hindi, Arabic); whereas the most frequent confusions for disgust were sadness (English), neutral (German), or anger (Hindi, Arabic).

To further understand relational differences among the seven emotion categories across languages, Table 5 summarizes the key perceptual and acoustic measures reported in each separate language condition in the form of significant *ranked differences* between each measure. Inspection of Table 5 permits a number of general observations: first, the recognition of anger, neutral, fear (with the exception of German), and sadness tends to be more accurate in the vocal channel in all languages, whereas surprise and disgust tend to be recognized relatively poorly when compared to the other emotions (happiness tends to assume an intermediary position). As well, one can see that emotions that tend to be recognized

Table 5
Comparative summary of the perceptual and acoustic data for the seven emotions across languages.

Emotion	Target recognition				f0Mean				f0Range				Speech Rate				Classification (%)			
	Eng	Ger	Hin	Ara	Eng	Ger	Hin	Ara	Eng	Ger	Hin	Ara	Eng	Ger	Hin	Ara	Eng	Ger	Hin	Ara
Anger	A	B	A	B	C	B	B	DE	B	C	A	D	D	B	A	C	35	28	69	5
Disgust	C	C	B	CD	E	D	D	D	E	D	B	BC	E	F	D	D	73	32	51	42
Fear	A	DE	A	B	B	C	A	B	C	E	C	CD	A	CD	A	A	87	41	60	69
Sadness	A	D	A	A	E	E	E	E	F	E	D	E	D	DE	E	C	40	59	68	69
Happiness	B	F	B	BC	D	B	C	C	D	B	B	B	C	A	C	D	20	60	18	46
Surprise	D	E	C	D	A	A	A	A	A	A	A	A	C	BC	B	C	86	88	35	86
Neutral	BC	A	B	B	F	E	F	DE	F	D	D	D	B	E	C	B	68	53	80	58
# ranks	4	6	3	4	6	5	6	5	6	5	4	5	5	6	5	4				

Relative differences in the perceptual and acoustic measures are expressed as significant, ranked differences observed in each language condition, where "A" is always the highest value. Means with the same letter were not significantly different, and means with two letters did not significantly differ from the condition specified by each of the two letters (Duncan's Multiple Range Test, $p < 0.05$).

well across languages are not always predicted well by a classification function restricted to variability in f0Mean, f0Range, and speech rate (especially for anger), although this showed similar tendencies across languages. The way that speakers of a particular language used certain acoustic parameters to signal emotions also differed at times (e.g., German speakers tended to express fearful and neutral expressions slowly, whereas speakers of the other languages tended to produce these expressions relatively quickly).

To visualize how valid emotional exemplars separated acoustically and to evaluate whether this was systematic across languages, the 2519 individual tokens were displayed in a scatterplot for each language as shown in Fig. 3. Since f0Mean and f0Range correlated most strongly with the first (and major) canonical function in our discriminant analysis of each language, tokens were plotted according to changes in only these two acoustic variables. Overall, it can be seen that there was considerable overlap among the emotion expressions in each language, with the greatest dispersion of exemplars produced in English (and the least in Arabic). Sad expressions, which exhibited low f0Mean and f0Range, clustered systematically in each language, while fear and surprise expressions were often quite distinct from other exemplars in their high f0Mean and/or f0Range. There were few anger or happy expressions which were acoustically distinct in each language based on combined f0Mean/f0Range.

4. Discussion

Vocal expression is a primary and phylogenetically significant part of the human repertoire for communicating emotions independent of language (Cosmides, 1983; Wilson & Wharton, 2006). However, most commonly these expressions are realized in the context of speech, according to the prevailing structure of the language in usage, which could influence the physical form of vocal emotion expressions (Pell, 2001) and/or how they are interpreted from one language to another (Juslin & Laukka, 2001). To our knowledge, no previous work has compared the acoustic-perceptual underpinnings of vocal

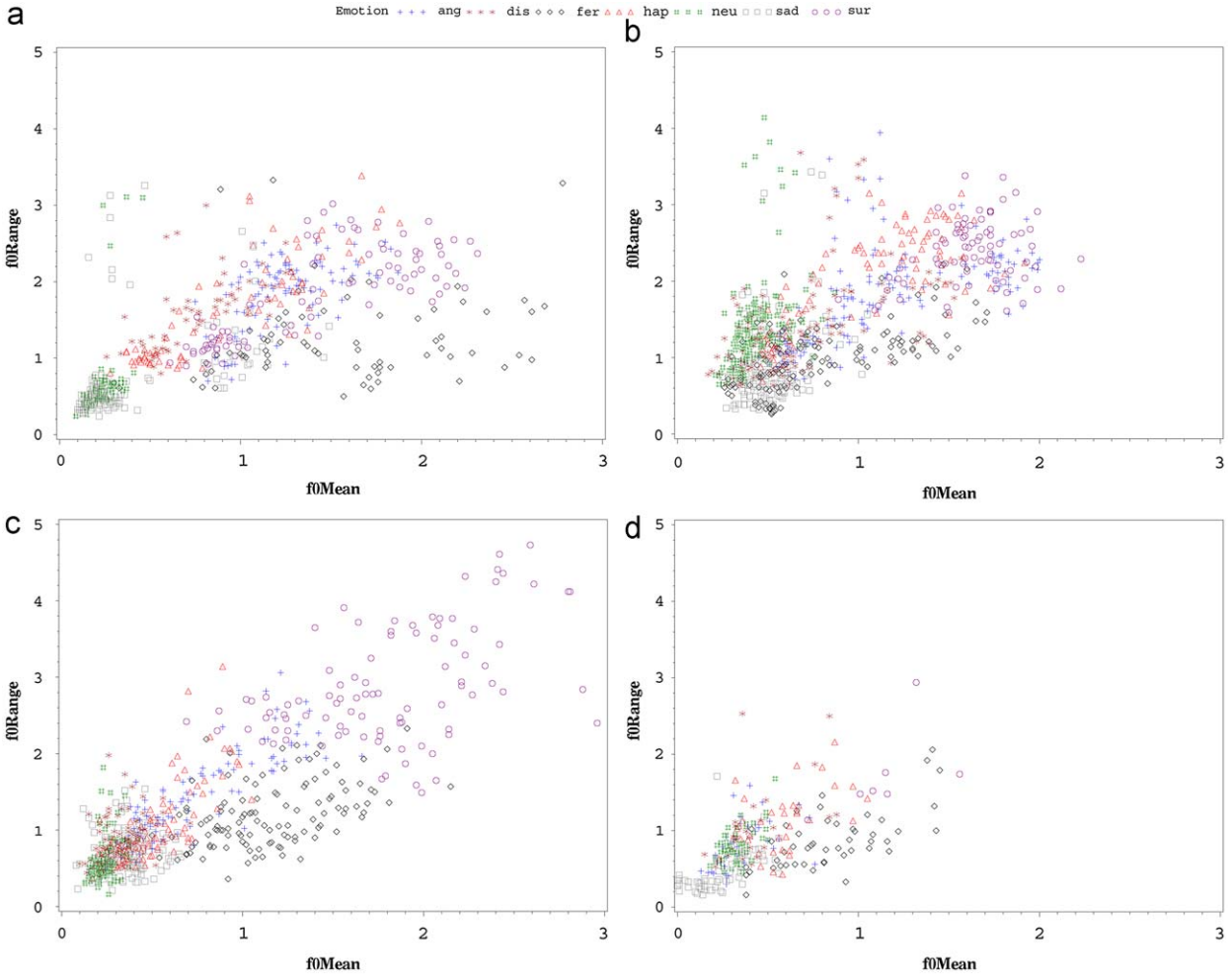

Fig. 3. Scatterplots showing the distribution of all valid emotional exemplars in each language according to mean fundamental frequency (f0Mean) and fundamental frequency variation (f0Range). Measures were normalized for each speaker in reference to the neutral category (where minimum f0 for neutral = 0). (a) English ($n = 761$ tokens), (b) German ($n = 870$ tokens), (c) Hindi ($n = 613$ tokens), and (d) Arabic ($n = 275$ tokens).

emotion expressions in multiple language corpora simultaneously in an empirically rigorous manner (although see Fónagy & Magdics, 1963 for descriptive analysis of vocally expressed emotions in Hungarian, English, German, and French). Our results merit attention because, unlike many previous studies, they are based on a vast number of items (over 2500 sentences) produced by 16 different speakers who each contributed a relatively equal number of tokens within the respective language condition (cf. Banse & Scherer, 1996 who included 16 cases per emotion in German, where 3 of their 12 speakers contributed 88% of the analyzed tokens). Thus, our data act as an important starting point for discussion of how vocal emotions are recognized within and across languages, and how emotion recognition patterns are associated with major acoustic dimensions of spoken language such as f0 and speech rate.

4.1. Recognition of specific emotion types within and across languages

Our initial comparisons focused on four language contexts—English, German, Hindi, and Arabic. In each of the languages studied, our data reinforce the claim that discrete emotions can be identified strictly from vocal inflections of the native language at accuracy levels which are well above chance (Banse & Scherer, 1996; Juslin & Laukka, 2001; Scherer et al., 1991; Sobin & Alpert, 1999). Consistent with many of these reports, the overall emotion recognition rates in each language context ranged between four and six times what would be expected by chance in our task (English = 81%, German = 67%, Hindi = 69%, Arabic = 59%, chance = 14%). This level of accuracy is not as high as one might expect if our stimuli had contained lexical-semantic features which bias the emotional target (e.g., Borod et al., 2000). Here, our stimuli were constructed to restrict emotion recognition processes to the use of vocal cues by employing pseudo-utterances without meaningful semantic language content which sounded "language-like" to each native listener group (see Banse & Scherer, 1996; Laukkanen, Vilkman, Alku, & Okanen, 1996 for similar approaches). Given our methods, there can be little doubt that expressions of presumably 'basic' emotions investigated here have perceptually distinct vocal properties in each of the languages under scrutiny which were familiar to native listeners, irrespective of the anomalous nature of linguistic information present in the stimuli.

Although emotion recognition tended to be reliable overall, the relative ease by which specific emotions could be recognized from vocal attributes of the speech signal varied significantly in the four languages of interest. For English and Hindi listeners, recognition was most accurate for sadness, anger, and fear, whereas Arabic listeners were relatively accurate for these three emotions as well as neutral expressions. For German listeners, neutral and anger expressions were recognized most reliably, followed by disgust, sadness, and fear. Taken together, these patterns imply that certain emotions are systematically recognized with greater accuracy from vocal cues in the four languages: principally, anger and sadness, and to a lesser extent, fear and neutral expressions. The finding that anger and sadness are most accurately identified from emotional prosody has considerable precedence in the literature (Banse & Scherer, 1996; Johnson, Emde, Scherer, & Klinnert, 1986; Kramer, 1964; Murray & Arnott, 1995; Thompson & Balkwill, 2006; Wallbott & Scherer, 1986) and is corroborated clearly here for English, Hindi, and Arabic (and in part for German). Current evidence that fear can be recognized accurately from the voice in different languages is also supported by previous research (Banse & Scherer, 1996), although existing data pertaining to this emotion are mixed (cf. Johnson et al., 1986; Scherer et al., 1991; Schröder, 1999; van Bezooijen et al., 1983). Some researchers argue that fear is the *only* emotion which is recognized preferentially from vocal cues when compared to facial expressions (Levitt, 1964; Zuckerman, Lipets, Koivumaki, & Rosenthal, 1975), implying that expressions of fear possess salient properties in the vocal channel which promote good recognition across languages. Our cross-language comparisons are largely consistent with this idea. As demonstrated by Juslin and Laukka (2001), the ability to recognize fear (as well as anger) is significantly dependent on the *intensity* of these expressions when studied experimentally; this could explain some of the discrepancies among studies in which the intensity of fear expressions may not have been adequately controlled, as well as in our own data for German where fearful expressions were not recognized as well as in the other languages (see below for further comments).

Happiness was recognized by native listeners with moderate accuracy in English, Hindi, and Arabic and very poorly in German, mirroring previous data which show that this emotion is often not reliably detected in the voice (Levitt, 1964; Pell, 2002; Wallbott & Scherer, 1986; Scherer et al., 2001; Zuckerman et al., 1975). It is well known that *facial* expressions of joy/happiness render a processing advantage and/or ceiling effects in many emotion processing tasks (Juth, Lundqvist, Karlsson, & Ohman, 2005; Pell, 2005; Russell, 1994; Wallbott, 1988); this result is probably due to the perceptual salience of the smile which is highly distinctive in the face (Shimamura, Ross, & Bennett, 2006). In speech, our cross-language data underscore that while most studies include only one or two "positive" emotions in their data set, this does not confer an advantage on listeners in the ability to recognize happiness from prosody. Rather, one can speculate that many negative emotions such as anger, fear, and sadness (i.e., grief or despair) are communicated most effectively in the voice because the antecedent events of these expressions—situations of aggression, loss, and physical danger—must be communicated urgently to con-specifics and with high signal value, often over distances when joint visual attention cannot be established (Scherer, 1997). However, since happiness was sometimes confused with

pleasant surprise (especially in German), it is also possible that some decoders failed to differentiate strongly between categories with a positive valence in our task, contributing to a high number of "happy" errors in these conditions.

In relative terms, disgust and surprise were the most difficult emotions to recognize in all four languages (with the exception of disgust in German which yielded moderate levels of recognition). For disgust, this outcome was expected because this emotion is ubiquitously associated with poor recognition when judgments are restricted to vocal attributes of speech (Banse & Scherer, 1996; Juslin & Laukka, 2001; Levitt, 1964; Scherer et al., 1991). Nonetheless, it merits emphasizing that absolute recognition rates for disgust in the four languages were always high (55–77% across languages), and our data show that disgust was not systematically confused with another emotion (with the exception of anger in Hindi), indicating that there is a distinct "voice of disgust" contrary to previous speculation (e.g., Banse & Scherer, 1996).

For surprise, several factors may have contributed to the perceptual data for this emotion. First, it is clear that surprise was the most difficult emotion for speakers to simulate in our encoding study for all four languages, forcing a larger proportion of these stimuli to be discarded prior to our main experiment (particularly for Arabic). This implies that in a laboratory setting, problems at the *encoding* stage are likely to be more pronounced in the case of surprise. In each language condition, native listeners then demonstrated relative difficulties in the ability to recognize surprise when compared to most other emotions (Levitt, 1964; Wallbott & Scherer, 1986), although again absolute recognition rates for this emotion ranged between 50% and 72% across languages. We noted that surprise was frequently identified as "happiness" in relatively equal measures in each of the four languages; this pattern implies that (pleasant) surprise expressions are interpreted as intense expressions of joy or elation by many listeners. Indeed, it is noteworthy that not all emotion theorists consider surprise a 'basic' emotion with strongly invariant properties (Ekman, 1992) which could explain some of the error confusions in our data. Nonetheless, it is undeniable that many decoders of English, German, Hindi, and Arabic could successfully detect vocal cues which differentiate surprise from other emotions in speech, and as shown in Table 2, certain speakers were extremely adept at encoding this emotion in a highly distinctive manner (e.g., DF and MG for English, RM for Hindi).

Finally, given evidence that "neutral" utterances were identified reasonably well from the voice in each language, it is remarkable that there has been little concrete debate about the nature of neutral vocal expressions in speech. Vocal stimuli designed to be affectively neutral are routinely employed in forced-choice recognition tasks as they were here, and also as a comparative baseline for understanding behavioural (Pell, 2005), electrophysiological (Paulmann & Kotz, 2008) and hemodynamic (Kotz et al., 2003; Mitchell, Elliott, Barry, Cruttenden, & Woodruff, 2003) responses to specific vocal emotion expressions. The neutral mode is often adopted strategically by encoders to conceal overt cues to emotion or when the speaker's intention is to highlight the propositional message and its information content to the listener (e.g., in certain instructional contexts, news reports, or when attempting to appear "objective"). Our data indicate that neutrality in the voice has unique perceptual properties and acts as a reliable response category for assigning meaning to vocal cues in each of the four languages under study (indeed, this category was associated with the *highest* recognition rates for German, or 93% correct). Recently, we have shown that a subset of these neutral expressions produced by the English, German, and Arabic speakers reported here can be recognized dependably by monolingual Spanish listeners in a cross-cultural setting (Pell et al., 2009). This bolsters the argument that neutral vocal expressions have recognizable properties in the voice much like the basic emotions, such as anger and sadness.

4.2. Acoustic characterization of vocal emotion expressions

The relationship between perceptual recognition patterns and physical properties of vocal emotion expressions was evaluated by focusing on a very small number of acoustic parameters, but those which are widely viewed as critical for differentiating emotions in speech. Given the broad purview of our report on different language types, narrowing our focus to three major acoustic parameters allowed us to constructively infer whether these cues contribute in a similar manner to emotional expressions as a function of language (e.g., through discriminant function analyses). Expectedly, our analyses confirmed that f0Mean (i.e., relative pitch level), f0Range (i.e., long-term pitch variation), and speaking rate each contributed to differences among the seven emotion categories for English, German, Hindi, and Arabic (a main effect of Emotion was highly significant for each parameter in each language condition).

However, as acoustic *patterns* as opposed to individual parameters are thought to govern how discrete emotions are recognized (Sobin & Alpert, 1999), it is even more instructive that f0Mean, f0Range, and SpRate always contributed significantly in a combined manner to classify items according to their perceived emotional meaning when a discriminant function analysis was carried out in each language condition. Interestingly, in each language studied we found that a single canonical function accounted for the large majority (70–80%) of variance in the acoustic data across emotion types; this function was always most strongly correlated with a speaker's f0Mean, in isolation (Arabic) or more typically in combination with f0Range (English, German, Hindi). These results support a well entrenched view: that global settings of f0 level (mean) and deviation around this point (range/variation) are cues of paramount importance for communicating vocal emotions in spoken language (Bachorowski & Owren, 1995;

Mozziconacci, 2001; Williams & Stevens, 1972). At the same time, our findings stress that f0 interacts with other cues such as speech rate to specify emotions even when the number of acoustic cues examined is highly restricted, as was true here.

In fact, given the large number of acoustic parameters which have been attributed to emotional expressions (Banse & Scherer, 1996; Juslin & Laukka, 2003), it is not surprising that changes in f0Mean, f0Range, and SpRate did not account for many of the perceived differences among valid exemplars of the seven emotion categories. The overall success of our discriminant analyses for classifying the 2519 items into their emotion categories was 54%, a rate that was relatively comparable across languages (English = 58%, German = 49%, Hindi = 56%, Arabic = 53%). This outcome compares with data reported by Banse and Scherer (1996) who, in the context of limited items, estimated a classification success rate of around 40% for 14 emotions in German based on 16 different acoustic parameters (including related measures of f0Mean, f0Range, and SpRate). The observation that the three acoustic parameters of interest here explained a relatively similar proportion of the variance among emotion expression types in four distinct languages is novel and potentially important; this finding implies that speakers of English, German, Hindi, and Arabic exploit these parameters in relatively equal measure to differentiate a common set of 'basic' emotions, consistent with the idea that these signaling functions are dictated by modal tendencies independent of language structure (Pell et al., 2009; Scherer et al., 2001).

4.3. Acoustic correlates of specific emotion types

The relationship between the acoustic parameters and specific emotion types demonstrated a number of consistencies with the reported literature. As shown in Table 5/Fig. 3, we observed remarkable consistencies in the expression of sadness across languages: these expressions were uniformly slow in rate, produced with a low f0Mean, and restricted f0 range/variation (Pell, 2001; Sobin & Alpert, 1999). It has been suggested that sadness shares the fewest acoustic properties with other emotion expressions (Sobin & Alpert, 1999) and this level of acoustic distinctiveness in the speech signal could explain why this emotion is consistently recognized accurately in the voice. For the other emotions, we found that fear exhibited a relatively high f0Mean, moderate to narrow f0Range, and with the exception of German, a very fast speech rate (Juslin & Laukka, 2001; Siegman & Boyle, 1993; Williams & Stevens, 1972). Disgust and surprise, although the most difficult emotions to recognize, exhibited rather distinctive acoustic features when our data are examined: disgust always exhibited a very low f0Mean and was produced with the slowest speech rate (Juslin & Laukka, 2001; Scherer, London, & Wolf, 1973), whereas surprise invariably displayed the highest f0mean *and* f0Range of all emotions (Fónagy & Magdics, 1963; Laukkanen et al., 1996; Scherer et al., 1973). Acoustic patterns corresponding to anger and happiness in our data were perhaps the most variable across languages, as is also documented in the wider literature (see Juslin & Laukka, 2003). For example, f0Mean and f0Range of anger expressions were both relatively high in Hindi, moderate in English and German, and low in Arabic.

For anger, acoustic differences may be partially due to the fact that some of our items encoded "hot anger" (i.e., rage or intense frustration) instead of "cold anger" (i.e., threat) despite our attempts to control for this factor in our study (we instructed actors to produce a cold or "controlled" anger in the encoding study). Cold anger tends to exhibit a moderate or low f0Mean and f0Range, whereas hot anger is distinct in its relatively high f0Mean (and increased loudness, Banse & Scherer, 1996; Frick, 1986; Scherer et al., 1973; Whiteside, 1999). Anger expressions produced by our Hindi speakers may have been more reflective of hot anger than in the other language conditions, explaining the use of a high f0Mean and f0Range by these speakers, although we cannot exclude the possibility that these patterns reflect cross-language differences in the use of f0 in this particular context.

Similarly, the observation that German speakers were the only group to produce fear with a slow speaking rate (and a markedly narrow f0Range) suggests that, unlike the English, Hindi, and Arabic speakers, they may have been conveying anguish or sustained fear as opposed to "panic fear"; this is another distinction that is known to produce independent acoustic patterns in speech (Banse & Scherer, 1996; Fónagy & Magdics, 1963). Alternatively, the fact that our German utterances were much longer than those produced by the English, Hindi, and Arabic speakers may have contributed to these differences; quite possibly, it felt unnatural for the German speakers to produce long sentences in a fearful or panicked voice, causing them to express an anguished form of fear with a relatively slow speaking rate. Controlling better for utterance complexity/length and for the "type" and intensity of emotional expressions under study remain an ongoing challenge for researchers, although these factors are likely to influence the acoustic structure of emotional speech in a significant manner.

Whereas sadness and fear tended to be recognized well and exhibited distinct acoustic parameters from other emotions, there is an apparent discrepancy in our perceptual-acoustic comparisons for anger. As shown above, this emotion is detected reliably in the voice although these expressions often do not differentiate well from other emotions in the context of our three acoustic measures (see Sobin & Alpert, 1999 for similar findings when 12 distinct acoustic measures were included). Discriminant analyses further underscored that f0Mean, f0Range, and SpRate do a poor job in specifying *perceived* anger; anger expressions were classified very poorly in

English (35% correct), German (28% correct), and Arabic (5% correct), although not in Hindi (69% correct). Accumulating research highlights the importance of loudness/amplitude variation and especially changes in voice quality and energy distribution for characterizing vocal anger (Banse & Scherer, 1996; Fónagy & Magdics, 1963; Gobl & Chasaide, 2003; Sobin & Alpert, 1999; Williams & Stevens, 1972). Our findings furnish indirect support for the importance of these additional cues which are likely critical for recognizing anger in many languages. The fact that Hindi speakers provided more distinct f0 cues could explain why anger was accurately predicted by the acoustic cues in this one language; more importantly, this observation highlights the likelihood that speakers have certain flexibility in what acoustic strategies they adopt to communicate emotions, both within and across languages.

4.4. Caveats and conclusions

There has been legitimate debate about whether research on simulated emotions, in the face, voice or other channels, should be treated as representative of how humans communicate emotions spontaneously (e.g., Russell, 1994). One reason that we studied vocal emotions simulated by lay actors was to tightly control the nature of the stimuli in each context; by eliciting "pseudo-utterances" stronger claims can be made about how emotional prosody is interpreted independent of language cues in stimuli that are nonetheless "language-like" in many ways. Moreover, we agree with researchers who assume that emotional portrayals are based on natural expressions, and thus, that they contain similar or even identical acoustic properties (Banse & Scherer, 1996; see Juslin & Laukka, 2001 for an analysis). Nonetheless, even professional and lay actors vary in the degree to which they are able to encode vocal emotions (Wallbott & Scherer, 1986) and this point is clearly demonstrated by the results of the encoding study conducted for each language condition (review Table 2). Thus, it is likely that individual preferences and abilities at the stage of *encoding* vocal emotions contributed in an important manner to our data, despite our attempt to focus analyses on only items which were perceptually valid in reference to the intended emotion category.

Another problem we encountered is that the background experience of the 16 "lay actors" was difficult to control and may have been influenced to some extent by cultural factors; whereas most of our encoders in the English, German, and Hindi groups tended to have acting experience (e.g., training or experience in theatrical productions of some nature), this art form and form of training is less culturally prevalent in regions where our Arabic speakers originated. Instead, most Arabic encoders tended to have experience in public speaking. These differences may explain why *absolute* identification rates tended to be lower for most emotions in the Arabic condition relative to the English, German, and Hindi conditions overall. At the same time, it has been suggested that portrayals by trained actors which yield *exaggerated* expressions of emotions tend to interfere with accurate recognition (Wallbott, 1988) and we found no evidence of this trend in our data. Thus, we are confident that the vocal stimuli elicited in this study, while obviously simulated and not the product of spontaneous communication processes, were not exaggerated or unnatural to listeners and that our findings are useful for estimating how vocal emotions are communicated in everyday life.

Another hypothesis which can be examined in light of our data was that linguistic similarity might lead to greater overlap in how vocal emotions are recognized (e.g., Scherer et al., 2001). The four languages examined here can arguably be placed along a continuum of linguistic similarity according to typology ranging from English and German (Indo-European, closely related) to Hindi (Indo-European, distantly related) to Arabic (non-Indo-European, unrelated). This design permits a relatively wide perspective on how vocal emotions are communicated as a function of language beyond the well-studied European languages (e.g., English, German, Swedish, Dutch) to include major world languages such as Hindi and Arabic. In general, our data provide few indications that acoustic and/or perceptual patterns varied systematically as a function of linguistic similarity; rather, there appeared to be systematic tendencies which governed how well certain emotion types are recognized in the voice, and in the importance of specific acoustic parameters which seem essential for communicating emotions, but these patterns occurred irrespective of language or language typology.

The idea that linguistic similarity does not strongly predict how well vocal emotions are recognized is supported by related studies which have tested this directly through *cross-cultural* presentation of emotional utterances in a listener's native versus a foreign language (Pell et al., 2009; Thompson & Balkwill, 2006). Of special interest, when a subset of the English, German, and Arabic pseudo-utterances reported here were presented to 61 monolingual speakers of Spanish, we found that the participants could identify basic emotions equally well in all three foreign language conditions (ranging from 56% to 59% correct overall); this level of accuracy was only slightly (albeit significantly) lower than when the listeners identified vocal emotions from native, Spanish pseudo-utterances (64% correct; see Pell et al., 2009 for details). These findings allow the claim that vocal emotion expressions contain pan-cultural acoustic-perceptual properties which promote accurate recognition of emotions in a foreign language, irrespective of linguistic similarity, although the efficiency of vocal emotion processing is typically reduced in the cross-cultural setting (Pell & Skorup, 2008; Pell et al., 2009). Naturally, more comparative and cross-cultural research will be necessary before definitive claims can be made about how linguistic and/or cultural similarity influence emotional communication in the voice or through other channels (Matsumoto & Assar, 1992).

In closing, our research shows that speakers of four distinct languages exhibited many similarities in how they express and identify vocal expressions of emotion, despite important differences in the language they were speaking and in their linguistic-cultural backgrounds. These modal tendencies likely reflect properties of natural, coded signals which are used to communicate emotions in speech (Wilson & Wharton, 2006), and which are shared in large measure across languages. However, speech is unique in that it encodes vocal as well as linguistic information that can bias an emotion; our study does not speak to the *relative weight* given to prosody versus semantic cues when interpreting the emotional content of spoken language and whether this varies as a function of language. For German, there is some evidence that semantic cues have a greater influence than prosodic cues during on-line processing of emotional speech (Kotz & Paulmann, 2007), although prosody may play a stronger role than linguistic cues when interpreting emotions in certain "context-dependent" languages such as Japanese (Kitayama & Ishii, 2002). Thus, while our data highlight that discrete emotions display many cross-language similarities in their acoustic-perceptual properties, they do not inform the degree to which these attributes actually contribute to pragmatic interpretations of spoken languages which involve vocal, linguistic, facial, and other contextual parameters, and which inform a speaker's emotion or attitudes. This is an area of research that is ripe for investigation.

Acknowledgements

This work was supported by a Discovery grant from the Natural Sciences and Engineering Research Council of Canada (to M.D. Pell) and by the German Research Foundation (DFG FOR 499 to S.A. Kotz). We thank Sarah Crowder, Sarah Elgazar, Romy Leidig, and Rajashree Sen for their efforts with the recordings and for data organization.

References

Albas, D., McCluskey, K., & Albas, C. (1976). Perception of the emotional content of speech: A comparison of two Canadian groups. *Journal of Cross-Cultural Psychology*, *7*(4), 481–489.

Atkinson, A., Tipples, J., Burt, D., & Young, A. (2005). Asymmetric interference between sex and emotion in face perception. *Perception & Psychophysics*, *67*(7), 1199–1213.

Bachorowski, J., & Owren, M. J. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science*, *6*(4), 219–224.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636.

Borod, J., Cicero, B., Obler, L. K., Welkowitz, J., Erhan, H., Santschi, C., et al. (1998). Right hemisphere emotional perception: Evidence across multiple channels. *Neuropsychology*, *12*(3), 446–458.

Borod, J., Pick, L., Hall, S., Sliwinski, M., Madigan, N., Obler, L., et al. (2000). Relationships among facial, prosodic, and lexical channels of emotional perceptual processing. *Cognition and Emotion*, *14*(2), 193–211.

Burkhardt, F., Audibert, N., Malatesta, L., Türk, O., Arslan, L., & Auberge, V. (2006). Emotional prosody—Does culture make a difference? In: R. Hoffmann, & H. Mixdorff (Eds.), *Proceedings of the speech prosody 3rd international conference Dresden*, May 2–5, 2006, Dresden, Germany.

Cosmides, L. (1983). Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(6), 864–881.

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*, 5–32.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*, 169–200.

Ekman, P., & Friesen, W. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, *1*, 49–98.

Ekman, P., & Friesen, W. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124–129.

Ekman, P., Friesen, W., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., et al. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, *53*(4), 712–717.

Ekman, P., Sorenson, E. R., & Friesen, W. (1969). Pan-cultural elements in facial displays of emotion. *Science*, *164*, 86–88.

Elfenbein, H., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, *128*(2), 203–235.

Elfenbein, H., Beaupré, M., Lévesque, M., & Hess, U. (2007). Toward a dialect theory: Cultural differences in the expression and recognition of posed facial expressions. *Emotion*, *7*(1), 131–146.

Fónagy, I., & Magdics, K. (1963). Emotional patterns in intonation and music. *Zeitschrift fur Phonetik*, *16*, 293–326.

Frick, R. (1986). The prosodic expression of anger: Differentiating threat and frustration. *Aggressive Behaviour*, *12*, 121–128.

Gobl, C., & Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*(1-2), 189–212.

Goos, L., & Silverman, I. (2002). Sex related factors in the perception of threatening facial expressions. *Journal of Nonverbal Behavior*, *26*(1), 27–41.

Hess, U., Adams, R., & Kleck, R. (2005). Who may frown and who should smile? Dominance, affliction, and the display of happiness and anger. *Cognition and Emotion*, *19*(4), 515–536.

Hofmann, S., Suvak, M., & Litz, B. (2006). Sex differences in face recognition and influence of facial affect. *Personality and Individual Differences*, *40*, 1683–1690.

Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin*, *115*(2), 288–299.

Johnson, W. F., Emde, R. N., Scherer, K. R., & Klinnert, M. D. (1986). Recognition of emotion from vocal cues. *Archives of General Psychiatry*, *43*, 280–283.

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion*, *1*(4), 381–412.

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.

Juth, P., Lundqvist, D., Karlsson, A., & Ohman, A. (2005). Looking for foes and friends: Perceptual and emotional factors when finding a face in the crowd. *Emotion*, *5*(4), 379–395.

Kitayama, S., & Ishii, K. (2002). Word and voice: Spontaneous attention to emotional utterances in two languages. *Cognition and Emotion*, *16*(1), 29–59.

Kotz, S. A., Meyer, M., Alter, K., Besson, M., von Cramon, Y., & Friederici, A. (2003). On the lateralization of emotional prosody: An event-related functional MR investigation. *Brain and Language*, *86*, 366–376.

Kotz, S. A., & Paulmann, S. (2007). When emotional prosody and semantics dance cheek to cheek: ERP evidence. *Brain Research*, *115*, 107–118.

Kramer, E. (1964). Elimination of verbal cues in judgments of emotion from voice. *Journal of Abnormal and Social Psychology*, *68*(4), 390–396.

Laukkanen, A.-M., Vilkman, E., Alku, P., & Okanen, H. (1996). Physical variations related to stress and emotional state: A preliminary study. *Journal of Phonetics*, *24*, 313–335.

Levitt, E. A. (1964). The relationship between abilities to express emotional meanings vocally and facially. In J. R. Davitz (Ed.), *The communication of emotional meaning* (pp. 87–100). New York: McGraw-Hill.

Matsumoto, D. (2006). Are cultural differences in emotion regulation mediated by personality traits? *Journal of Cross-Cultural Psychology*, *37*(4), 421–437.

Matsumoto, D., & Assar, M. (1992). The effects of language on judgments of universal facial expressions of emotion. *Journal of Nonverbal Behavior*, *16*(2), 85–99.

Menn, L., & Boyce, S. (1982). Fundamental frequency and discourse structure. *Language and Speech*, *25*(4), 341–383.

Mesquita, B., & Frijda, N. (1992). Cultural variations in emotions: A review. *Psychological Bulletin*, *112*(2), 179–204.

Mitchell, R. L. C., Elliott, R., Barry, M., Cruttenden, A., & Woodruff, P. W. R. (2003). The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. *Neuropsychologia*, *41*(10), 1410–1421.

Mozziconacci, S. (2001). Modeling emotion and attitude in speech by means of perceptually based parameter values. *User Modeling and User-Adapted Interaction*, *11*, 297–326.

Murray, I. R., & Arnott, J. L. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, *16*, 369–390.

Paulmann, S., & Kotz, S. A. (2008). An ERP investigation on the temporal dynamics of emotional prosody and emotional semantics in pseudo- and lexical-sentence context. *Brain and Language*, *105*(1), 59–69.

Pell, M. D. (2001). Influence of emotion and focus location on prosody in matched statements and questions. *Journal of the Acoustical Society of America*, *109*(4), 1668–1680.

Pell, M. D. (2002). Evaluation of nonverbal emotion in face and voice: Some preliminary findings on a new battery of tests. *Brain and Cognition*, *48*, 499–504.

Pell, M. D. (2005). Nonverbal emotion priming: Evidence from the 'facial affect decision task'. *Journal of Nonverbal Behavior*, *29*(1), 45–73.

Pell, M. D., & Baum, S. R. (1997). The ability to perceive and comprehend intonation in linguistic and affective contexts by brain-damaged adults. *Brain and Language*, *57*(1), 80–99.

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, *33*(2), 107–120.

Pell, M. D., & Skorup, V. (2008). Implicit processing of emotional prosody in a foreign versus native language. *Speech Communication*, *50*, 519–530.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, *115*(1), 102–141.

Scherer, K. R. (1997). The role of culture in emotion-antecedent appraisal. *Journal of Personality and Social Psychology*, *73*(5), 902–922.

Scherer, K. R., Banse, R., & Wallbott, H. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, *32*(1), 76–92.

Scherer, K. R., Banse, R., Wallbott, H. G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, *15*(2), 123–148.

Scherer, K. R., London, H., & Wolf, J. J. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research and Personality*, *7*, 31–44.

Schröder, M. (1999). Can emotions be synthesized without controlling voice quality? *Phonus*, *4*, 35–50.

Shimamura, A., Ross, J., & Bennett, H. (2006). Memory for facial expressions: The power of a smile. *Psychonomic Bulletin & Review*, *13*(2), 217–222.

Siegman, A. W., & Boyl, S. (1993). Voices of fear and anxiety and sadness and depression: The effects of speech rate and loudness on fear and anxiety and sadness and depression. *Journal of Abnormal Psychology*, *102*(3), 430–437.

Sobin, C., & Alpert, M. (1999). Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy. *Journal of Psycholinguistic Research*, *23*(4), 347–365.

Thompson, W., & Balkwill, L.-L. (2006). Decoding speech prosody in five languages. *Semiotica*, *158*(1/4), 407–424.

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

Wallbott, H. G. (1988). Big girls don't frown, big boys don't cry—Gender differences of professional actors in communicating emotion via facial expression. *Journal of Nonverbal Behavior*, *12*(2), 98–106.

Wallbott, H. G., & Scherer, K. R. (1986). Cues and channels in emotion recognition. *Journal of Personality and Social Psychology*, *51*(4), 690–699.

Whiteside, S. (1999). Acoustic characteristics of vocal emotions simulated by actors. *Perceptual and Motor Skills*, *89*, 1195–1208.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, *52*(4(2)), 1238–1250.

Wilson, D., & Wharton, T. (2006). Relevance and prosody. *Journal of Pragmatics*, *38*, 1559–1579.

Zuckerman, M., Lipets, M. S., Koivumaki, J. H., & Rosenthal, R. (1975). Encoding and decoding nonverbal cues of emotion. *Journal of Personality and Social Psychology*, *32*, 1068–1076.