CONTENTS

Summary

There are several inherent problems with this account: (a) unclear origins of the majority of sounds downloaded from the Internet assumed to represent ROV, (b) lack of contrasting the data with a general model of erotic vocalization, (c) relying on a small sample of verified ROV sounds, (d) incorporating recollection into the process of scientific inquiry, (e) small sample size, (f) even smaller sample of what I termed here commercial OV, (g) lack of matching modal voicing and experimental voicing for the same sample of OV, to mention just few and very obvious drawbacks. Setting these inherent weaknesses aside, the data presented here appear to represent the first organized attempt to address this specific aspect of human vocalization. Findings are in a way surprising and seem to substantiate ancient teachings that vocalization of emotions, including those associated with lovemaking and specifically with orgasm, are brief and gentle, follow respiratory and muscular contraction rates, in a way are very intimate and do not reach the commercially presented vocal crescendos and screams, and are not really associated with or contaminated by dirty talking or in general with any talking at all.

It is mostly the voice and air (exhalatory air) that convey information and signal arrival of an orgasm, perhaps satisfaction, approval, and love. And therefore the amygdala and the hypothalamus appear to be involved perhaps more on the right than the left side as these sounds carry pleasant information.

The almost flat and unornamented F_0 pattern seems to connote peace and its prolongation signals pleasure, while the fluctuations may be attached to individual means of stressing joy, but in a gentle and not vocally violent way, as manifested by pitch elevation; pitch elevation is mostly part of joy. Although the amplitude of these vocal signals was not analyzed here, a generalized view of the signal amplitude seems to show a drop-off at the termination of each segment and with termination of the entire OV sequence. The amplitude variations are not wide.

The rate of OV was found to be rather slow and hence seems to be regulated by the respiratory rather than by the laryngeal component, which seems to have an optimal rate of 5 Hz. In this way, ROV seems to be mediated through the coordination of the sympathetic nervous system, vocalization, and ventilation and not via speech motor cortex, the cerebellum, or the left brain. The sound appears to be the by-product of the respiratory flow, onto which the voice is placed, like throwing a leaf and not a stone onto a stream of water, and making it all flow together without a splash.

In conclusion, the results seem to support the teaching of Tantra and Tao that talking may be distractive, while breathy, relaxed, and focused vocalization and air flow associated with breathing and some laryngeal (glottic) narrowing are more rewarding and reflect more the emotional state of an orgasm than screaming and yelling, “Oh God, yes, yes, yes baby yessss.” So being breathy (as in respiration) brings us closer to the sound of the laughter of apes, which in at least one experimental study was interpreted as a sound of lovemaking. It thus shows that the species differ but only by a little.

I don’t want to trivialize this subject, but I wish to admit that this chapter was intended to be a “teaser” (no pun intended) rather than a pure scientific may be associated with pleasant voice. Happiness and anger have been found to be differentiable on the basis of other vocal/acoustic characteristics and none of these include such dimensions as warmth, another characteristic associated with pleasant voice. At this point, the degree to which vocal pleasantness is associated with particular emotions or the degree to which particular emotions may be encoded in pleasant or unpleasant fashion is unclear.

Differentiating Vocal Pleasantness and Vocal Attractiveness

Researchers investigating vocal pleasantness are faced with the conceptual question of the relationship between vocal pleasantness and vocal attractiveness. Though there is some research that bears on the topic (see the Future Directions section below), it is a question that has not received sufficient attention. In addressing this issue, it may be helpful to draw some parallels to facial pleasantness and attractiveness. It is not difficult to imagine examples of attractive faces that are, in many senses, unpleasant—such as the cold and proud images of highly attractive fashion models. Conversely, one can imagine the visage of a kind and sweet-faced, but haggard and wrinkled, elderly woman. Are there vocal analogues of these two examples?

Studies may well be designed to differentiate between vocal pleasantness and attractiveness. One hypothesis is that vocal attractiveness is associated with perceptions of competence/dominance/social attractiveness, and that vocal pleasantness is associated with perceptions of benevolence/kindness/altruism.

We might well ask whether pleasantness is positively correlated with attractiveness. The answer is most likely yes, and it is probably a fairly strong correlation. However, it may be more theoretically productive to ask whether one could identify or synthesize voices that are attractive but not pleasant, or pleasant but not attractive. Such a procedure would enable us to tease apart the divergence in meaning between vocal pleasantness and vocal attractiveness and to then examine carefully the acoustic and vocal properties of such voices.

An additional distinction important to differentiating vocal pleasantness and vocal attractiveness is that between perceptions of speakers' voices and of their personalities. There is a fairly large body of research from several decades ago on personality attributions associated with particular vocal and acoustical characteristics (Apple, Streeter, & Krauss, 1979; Brown, Strong, & Rencher, 1973; 1974; 1975; Scherer, 1979). These personality attributions are made for the speaker. It is important to separate the issues of ratings of speakers from ratings of speakers' voices, as personality ratings of speakers may well be independent of ratings of their voices. A likeable and pleasant person could conceivably have a very unpleasant voice and, conversely, a dour or churlish person could have a highly pleasant voice.

Vocal Attractiveness and Vocal/Acoustical Mediators

When we make judgments about others we evaluate their actions, their looks, and even their voices. Unfavorable first impressions may be due to a "weak voice" (Zuckerman, Miyake, & Hodgins, As the interpretation of the speech signal is probabilistic, we believe that affectively marked speech patterns function as parallel codes that provide contextual information, thus helping the receiver to disambiguate the meaning of the utterance (see Chapter 15).

Vocal Manifestations of Cognitive-Affective States and Processes

Within the framework of the unified model, vocal manifestations of cognitive affective states and processes can be studied in their continuity from the states traditionally regarded as purely cognitive (doubt, certainty) and unemotional, to emotionally *colored* interpersonal stances (friendly), up to full-fledged emotional reactions. While vocal correlates of discrete emotions have been amply studied in the past two decades, those of other affectively colored states have received less attention. Exceptions are relatively recent studies in assessing the emotional tone in spontaneous dialogues (Cowie, Douglas-Cowie, & Romano, 1999), those related to the acoustic indicators of attitudes (Wichmann, 2002), and various cognitive-affective dimensions (Kehrein, 2002). Both Wichmann's and Kehrein's work emphasize the importance of social and linguistic context for the interpretation of vocally encoded cognitive-affective states. In addition to the classical vocal parameters related to F_0, Nì Chasaide and Gobl (2002) found the association between voice quality and the perceived *affective coloring* of speech. Eric Keller's study (2003) provided evidence for vocal changes related to attitude and thematic coloration of speech. Research on vocal indicators of emotional dimensions has shown that listeners can consistently rate vocal expressions of emotions on the scales of activation, valence, and potency and that each dimension is correlated with a number of vocal parameters (Laukka et al., 2005).

Vocal Correlates of Interpersonal Stance in Medical Interviews

Zei Pollermann conducted a pilot study of vocal correlates of interpersonal stance in pre-anesthesia medical interviews. Interpersonal stance is defined as an affective stance taken toward another person in a specific interaction, coloring the interpersonal exchange in that situation as for example: distant, cold, warm, supportive, reassuring, calming, or contemptuous. As the pre-anesthesia medical interview has a well-defined structure (Wolff & Scemama-Clergue 2002), it allowed setting clear hypotheses about the type of affective stance appropriate for each of the two main phases of the interview. The aim of the first phase is to obtain anamnestic information and examine the patient, while the aim of the second phase is to decide on the type of anesthesia, to inform the patient about the risks without creating anxiety, and to obtain the patient's consent. It was hypothesized that the interpersonal stance appropriate for the examination phase could be described as encouraging, while that appropriate for the announcement of risks would be reassuring and calming. Our predictions regarding vocal correlates of such affective stances took into account discursive operations of topicalization, focalization, and comment. These operations use

Conclusions

Emotions are more appropriately interpreted intra- and cross-linguistically when both the visual and auditory signals are present than when only the auditory signal is present. Visual stimuli alone give the best interpretations. Some emotions are more easily interpreted, both intra- and cross-linguistically, prosodically and multimodally, e.g., sadness.

There is the possibility that the facial expression of emotion is more universal than prosodic expression. The prosodic production of emotional expressions per se could be universal, at least for certain emotions, but the emotional prosody is never heard in isolation, but always in combination with the speech prosody of each particular language. Another possibility, which will need further studies, is the question of whether certain emotions are more dependent on prosodic information and other emotions more dependent on facial expression.

References

Abelin, Å., & Allwood, J. (2000). Cross linguistic interpretation of emotional prosody. In R. Cowie, E. Douglas-Cowie, & M. Schröder (Eds.), *ISCA Workshop on speech and emotion* (pp. 110–113). Belfast, Northern Ireland: Textflow.

Boersma, P., & Weenink, D. (2005). Praat: Doing phonetics by computer (Version 4.3.01) [Computer program]. Retrieved September 2, 2005, from http://www.praat.org/

Darwin, C. (1872/1965). *The expression of the emotions in man and animals*. Chicago: University of Chicago Press.

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and eye. *Cognition and Emotion*, *14*, 289–311.

Laukka, P. (2004). *Vocal expression of emotion—Discrete-emotions and dimensional accounts*. Uppsala, Finland: Acta Universitatis Upsaliensis.

Massaro, D. W. (2000). Multimodal emotion perception: Analogous to speech processes. In R. Cowie, E. Douglas-Cowie, & M. Schröder (Eds.), *Proceedings of the ISCA workshop on speech and emotion* (pp. 114–121). Belfast, Northern Ireland: Textflow.

Massaro, D. W. (2002). Multimodal speech perception. In B. Granström, D. House, & I. Karlsson (Eds.), *Multimodality in language and speech systems* (pp. 45–71). Dordrecht, Netherlands: Kluwer Academic Publishers.

Matsumoto, D., & Ekman, P. (1989). American-Japanese differences in intensity ratings of facial expressions of emotion. *Motivation and Emotion*, *13*, 143–157.

Matsumoto, D., Franklin, B., Choi, J.-W., Rogers, D., & Tatani, H. (2002). Cultural influences on the expression and perception of emotion. In W. B. Gudykunst & B. Mody (Eds.), *Handbook of international and intercultural communication* (pp. 107–125). Thousand Oaks, CA: Sage.

Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1), 227–256.

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

Gender Differences in the Influence of Vocal Emotional Information on Higher Order Cognitive Processes

One might speculate that gender differences in the preattentive processing of vocal emotional expressions affect the subsequent use of vocal emotional information for higher order cognitive processes. For example, when focusing on what is said during a conversation, women may be more likely than men to integrate vocal emotional information with verbal information. Whether this is true has been investigated in a series of ERP studies that used verbal stimulus material in order to elicit an N400 (Schirmer, Kotz, & Friederici, 2002; Schirmer & Kotz, 2003; Schirmer, Kotz, & Friederici, 2005). The N400 is a negativity that peaks approximately 400 ms following word onset and that is thought to reflect the retrieval of word information from the mental lexicon (for a review see Kutas & Federmeier, 2000). To study the influence of vocal emotional information on semantic retrieval, participants were presented with words that were either congruous or incongruous to the speaker's emotional tone of voice (e.g., "success" spoken with a happy or angry voice). Congruous words elicited a smaller N400 than incongruous words, suggesting that the retrieval of word information from the mental lexicon is less effortful when a word matches a speaker's vocal emotional expression (Schirmer et al., 2002; Schirmer & Kotz, 2003; Schirmer et al., 2005). Furthermore, in accordance with the above predictions, the influence of vocal emotional information on word processing differed as a function of attention and gender. Women showed an N400 effect regardless of whether the task required them to ignore vocal emotional information (e.g., verbal emotional identification task) or to attend to both vocal and verbal emotional information (e.g., voice-word congruency judgment). In contrast, men showed an N400 effect only when asked to attend to both vocal and verbal emotional information (Schirmer et al., in 2005; Schirmer, Lui, Maess, Chan, & Penney, 2006) or when there was a longer delay between the onset of vocal and verbal information (e.g., when a spoken utterance preceded a visual target word; Schirmer et al., 2002). A more automatized use of vocal emotional information for language processing in women, as compared to men, has also been demonstrated with functional magnetic resonance imaging. Women showed larger activity in the left inferior frontal gyrus in response to words spoken with incongruous as compared to congruous emotional tone when emotional tone was task-irrelevant (Figure 5–3; Schirmer, Zysset, Kotz, & von Cramon, 2004). Given that the left inferior frontal gyrus has been implicated in semantic retrieval (Wagner, Paré-Blagoev, Clark, & Poldrack, 2001), these findings provide further evidence that, depending upon attentional focus and gender, vocal emotional information may decrease or increase semantic retrieval effort. Interestingly, a more recent study indicated that the differential effort men and women direct at encoding words spoken with congruous and incongruous emotional tone affects their memory for these words (Mecklinger, Gaebel, Schirmer, Treese, & Johansson, 2007). In accordance with the more automatized processing and use of vocal emotional expressions in women,

Vocal Perception of Emotion from Speech

Theoretical approaches and empirical outcomes associated with perception of emotion from speech acoustics generally parallel those we have reviewed for production. Specifically, some researchers adopt a cue-configuration perspective whereas others emphasize a dimensional view. When listeners are asked to identify the intended emotion in utterances produced by actors, accuracy is again relatively modest—about 55% across studies (reviewed by Johnstone & Scherer, 2000). Similar outcomes and confusions are observed across different cultures and language groups, although identification errors also reflect the degree of disparity between vocalizer and listener language (Scherer, Banse, & Wallbott, 2001). The standard strategy in these experiments has been to use a forced-choice identification paradigm in which listeners select a single emotion word deemed to best describe the affect being conveyed. Stimulus sets usually include only a small number of talkers and emotions, and are often selected to include presumably prototypical instances of the emotions in question (e.g., Banse & Scherer, 1996; Leinonen, Hiltunen, Linnankoski, & Laakso, 1997; Sobin & Alpert, 1999). As noted earlier, this strategy of using acted emotional samples and then testing only a screened subset may in and of itself be accounting for some of the evidence of differentiated perception of emotion.

When Pereira (2000) had listeners rate vocal samples of various discrete emotions using dimensional scales, by far the strongest associations between the ratings and vocal acoustics were arousal-related (see also Green & Cliff, 1975). F_0 and amplitude were again primary, as they were in a study by Streeter and colleagues (1983) in which participants evaluated talker stress levels from speech. Here, listeners reported that vocalizers were stressed when hearing significant variation in talker F_0 and amplitude, but otherwise usually failed to perceive stress. The link between speech acoustics and perceived valence is generally weaker, which again parallels outcomes observed on the production side. Ladd and colleagues (1985), for example, systematically varied several acoustic parameters as listeners rated vocalizer affect and attitude. A central finding in this study was that listeners' responses did not reveal categorical responses associated with discrete emotions, but rather varied continuously in accordance with continuous changes in F_0. Results are similar when cue-configuration and dimensional perspectives are compared within the same experimental setting, although only a few such studies have been conducted. Overall, listeners are found to be most likely to make errors when stimuli reflect similar arousal levels (e.g., Pakosz, 1983; Pereira, 2000) and among similarly valenced members of emotion families (e.g., Banse & Scherer, 1996; see also Breitenstein, Van Lancker, & Daum, 2001; Ladd, Silverman, Tolkmitt, Bergmann, & Scherer, 1985).

Taken together, the perceptual evidence shows that overall listener accuracy is quite moderate. Many investigators argue that the generally observed outcome of about 55% correct indicates both that vocalizer emotion is associated with differentiated acoustics and that listeners can in turn perceive these cues (e.g., Banse & Scherer, 1996; Johnstone & Scherer, tures, which the classifier must be trained to recognize. Eventually, in the classification procedure, the constantly varying prosodic features and the more abstract features must be combined.

Conclusion

Our research project on the human and automatic classification of emotion in spoken Finnish, which is the first such research on this minority language, has yielded very interesting results. First, features of F_0 and intensity have been found to accompany emotional Finnish speech—this is probably a universal phenomenon in the expression of emotion. Second, the performance level of the human emotion classification exceeds that of the automatic classification. Although this is not surprising in itself, we suggest that phonological features of F_0 variation, especially rising F_0, are emotion-carrying features in spoken Finnish, in addition to the global constantly varying average features of F_0, intensity, duration, etc. Also in this respect, it can be argued that Finnish, a minority language in a minority language group, is not qualitatively different from major Indo-European languages. This finding contradicts prior notions of "emotionality" of the Finnish language. In languages in general, prosodic parameters are hierarchically organized as concrete (phonetic or paralinguistic) and as more abstract (phonological or linguistic) phenomena, and there is no reason to assume that some of these levels would be irrelevant from the viewpoint of the vocal communication of emotion. Finally, the results suggest that contrastive research on human vs. computer categorization of emotions is promising, and that, in the near future, computer recognition of human vocal emotions may approach a natural state, yielding access to new exciting product applications.

References

Hakulinen, L. (1979). *Suomen kielen rakenne ja kehitys* (*The structure and development of the Finnish language*). Helsinki: Otava.

Iivonen, A. (1998). Intonation in Finnish. In D. Hirst & A. Di Cristo (Eds.), *Intonation systems*: *A survey of twenty languages* (pp. 311–327). Cambridge, UK: Cambridge University Press.

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

Laukkanen, A.-M., Vilkman, E., Alku, P., & Oksanen, H. (1997). On the perception of emotions in speech: The role of voice quality. *Logopedics Phoniatrics Vocology*, *22*, 157–168.

Niemi, J. (1984). *Word level stress and prominence in Finnish and English*. *Acoustic experiments on production and perception*. Joensuu, Finland: Publications in the Humanities 1, University of Joensuu.

Seppänen, T., Toivanen, J., & Väyrynen, E. (2003). MediaTeam Speech Corpus: A first large Finnish emotional speech database. In (Ed.), *Proceedings of the 15th International Congress of Phonetic Sciences* (Barcelona)*: Vol. 3* (pp. 2469–2472).

Suomi, K., Toivanen, J., & Ylitalo, R. (2003). Durational and tonal correlates of accent in Finnish. *Journal of Phonetics*, *31*, 113–138.

ten Bosch, L. (2003). Emotions, speech and the ASR framework. *Speech Communication*, *40*, 213–225.

there was no effect of the encoding situation on judgments or on the impact of the context.

Discussion

Our results clearly demonstrate that judgments of vocal stimuli were affected by decoders' beliefs about the situational context in which they were produced. This is the first time, to our knowledge, that such an effect has been demonstrated with spontaneous vocalizations and the first time that a context effect has been shown for adults' decoding of any type of affective vocalizations. We are not surprised by these findings and think it only logical that evaluations of affective vocalizations are influenced by the same type of information that has been shown to influence judgments of facial displays. Yet, this finding is far from trivial and underlines the necessity that models dealing with decoding processes in nonverbal communication are not specific to a particular channel of communication, but are general in nature.

It might be argued that we observed context effects because the vocal stimuli were not very clear. Indeed, the ratings of the stimuli in the first experiment were rather low and we would be surprised if the impact of context would have been less for stimuli that were evaluated as being very positive or very negative in the absence of contextual information or more general beliefs about context. Contrariwise, one might also argue that the fact that our stimuli were only mildly positive and negative made our manipulation more plausible. We had no reason to assume that the participants suspected that vocalizations were presented with a discordant context. As we stated in our introduction, we do not intend to make any claims as to the relative importance of voice vs. context. However, we do find it important to demonstrate that, at least under some circumstances, beliefs about the context clearly influence affective judgments.

We are also intrigued that we could show context influences despite the fact that we did not use categorical judgments, but a simple valence dimension. If our reading of Russell's model (e.g., 1997) of the interaction of dimensional judgments and attribution of emotion labels is correct, then we should not have been able to demonstrate an impact of context on the valence dimension itself. However, it is possible that the explicit situational context information we provided impacts the valence estimation via affective bias influences on nonverbal decoding. As Pell states:

> The past decade has seen burgeoning evidence of affective bias or priming in cognitive processing, where the encoding of a "target" event—typically, a written word, picture, or facial stimulus—is systematically influenced by shared affective evaluations of a spatially or temporally contiguous stimulus "prime" (Pell, 2005, p. 46).

In the present study the prime would be a mental activation of the game-event that supposedly led to the vocalization evaluated. Even if the evaluation of valence was automatic, as Russell (1997) argued for the context of facial expressions, an affective bias could moderate this automatic process. Clearly, more research is needed to test whether a vignette, as used in our study, can cause such priming effects.

We have also shown that the impact of beliefs regarding the situational context

CHAPTER 9

Modification of Emotional Speech and Voice Quality Based on Changes to the Vocal Tract Structure

Brad H. Story

Abstract

Speech production can be represented as the combination of a sound source and an acoustic filter. For vowels and vowel-like sounds, the sound source is the periodic airflow signal generated by the vibration of the vocal folds, whereas the filter is formed by the vocal tract airspace. Although the goal of speech production is to transmit a linguistic message to a listener, the structure and idiosyncratic use of the speech organs impose unique variations on both the source signal and filtering properties (formant frequencies) provided by the vocal tract. These variations create the acoustic characteristics that give rise to the "emotional quality" of the speech. In this chapter, a computational speech production model is used to demonstrate how structural modifications of the vocal tract can generate changes in the quality of the resultant speech. The results suggest that combining this type of modeling with psychoacoustic experiments could eventually provide a powerful means for learning about the emotional load carried in natural speech.

1997; Schmitt, Hartje, & Willmes, 1997; Borod et al., 1998). In healthy subjects, behavioral studies of dichotic listening of prosodic stimuli (e.g., Schmitt et al., 1997) and nonlinguistic vocalizations (Carmon & Nachshon, 1973) have shown a left ear advantage for processing emotional stimuli, thus suggesting a right hemisphere dominance. Also, some neuroimaging studies in healthy subjects supported the right hemispheric dominance in processing emotional prosody (Kawashima et al., 1993; George et al., 1996; Buchanan et al., 2000; Kotz, Alter, Besson, Friederici, & Schirmer, 2000; Pihan et al., 2000; Rama et al., 2001; Wildgruber et al., 2002; Mitchell et al., 2003), as well as emotional nonlinguistic vocalizations (Sander & Scheich, 2001; Meyer et al., 2005).

However, some results did not support the right hemispheric dominance hypothesis in processing vocal emotions. Some studies reported bilateral responses to emotional prosody, and even stronger activations in the left hemisphere have been observed (e.g., Kotz et al., 2003). Moreover, the right hemispheric dominance hypothesis has not been clearly established in lesion studies. Although some patients with difficulties in identifying and discriminating emotional prosodic patterns predominantly presented lesions to the right hemisphere (Starkstein et al., 1994), deficits in perception of emotional prosody following lesions to the left hemisphere have also been reported (e.g., Cancelliere & Kertesz, 1990; van Lancker & Sidtis, 1992; Pell & Baum, 1997). Some studies with patients with left-hemisphere infarcts have even suggested a left hemispheric dominance in perception of emotional prosody (e.g., Pell, 1998).

Valence Hypothesis

The valence hypothesis suggests a left hemispheric dominance for processing positive emotions and a right hemispheric dominance for processing negative emotions (e.g., Sackeim et al., 1982; Ross, Homan, & Buck, 1994; Pell, 1998; Davidson & Irwin, 1999). However, some neuroimaging studies of emotional prosody did not support this hypothesis (George et al., 1996; Buchanan et al., 2000; Wildgruber et al., 2002; Kotz et al., 2003).

Some studies have supported both of these hypotheses according to the brain regions studied. For instance, findings from Sander, Roth, and Scheich (2003), studying neural substrates associated with perception of prosodic patterns of happiness and sadness, supported both hypotheses. Activations for emotional stimuli (versus vowel detection) in a temporo-parietal region of interest, although bilateral, were greater in the right hemisphere, thus supporting the right hemispheric dominance hypothesis. The authors also reported lateralized activations in support of the valence hypothesis in an occipital region of interest: stronger activations were observed for prosody expressing sadness in the right hemisphere than in the left one, whereas stimuli depicting happiness elicited greater responses in this region in the left hemisphere than in the right one.

Hypothesis of the Type of Emotion

Another hypothesis suggests that brain activations for primary emotions (e.g., happiness, sadness) are lateralized to the right hemisphere, whereas social emotions (e.g., culpability, shyness, envy) are lateralized to the left hemisphere (Ross et al., 1994). The study of social emotions is relatively recent and this hypothesis needs further investigation (for more details on social emotions, see Eisenberg, 2000).

Lateralization in Function of the Acoustic Parameters

Alternatively, it has been suggested that lateralization of prosody processing varies with the acoustic parameters. Most studies conceptualize emotional prosody as being a distinct entity instead of a manipulation of prosodic parameters to express an emotion. Processing a vocal emotion entails an acoustic analysis (Ethofer et al., 2005); and processing acoustic properties appears to involve lateralized brain areas (Zatorre & Belin, 2001; Boemio, Fromm, Braun, & Poeppel, 2005; Ethofer et al., 2005). According to a model proposed by van Lancker and Sidtis (1992), the left and right hemispheres are responsible for processing different aspects of emotional prosody. The right hemisphere is superior to the left one with regards to the extraction of fundamental frequency (F_0) information, whereas the left hemisphere is more involved in the extraction of temporal information. For instance, perception of the F_0 has been shown to involve the right hemisphere (e.g., Zatorre, Evans, Meyer, & Gjedde, 1992; Zatorre & Belin, 2001; Zatorre, Belin, & Penhune, 2002; but see Wildgruber et al., 2002), whereas the left hemisphere has been involved in temporal cues processing (e.g., Carmon & Nachshon, 1973; Zatorre et al., 1992; Belin et al., 1998; Zatorre et al., 2002). Moreover, hemispheric differences have been reported according to the processed feature. For instance, in Ladd (1996) the lateral temporal lobe of the right hemisphere processed tonal direction, but not tonal range or height. Also, the right hemisphere has shown superiority in processing lower frequencies, such as the F_0, compared to higher frequencies (Ivry & Lebby, 1993, but see Wolf, 1977). However, in an elegant study from Wildgruber et al. (2002), superiority of the right hemisphere was observed during processing emotional prosody that was not explained by the acoustic structures of the F_0, nor the duration.

Some prosodic cues used to perceive emotional states are relatively specific and stable; for example, longer syllable length is strongly associated with prosody expressing sadness (e.g., Cosmides, 1983; Wallbott & Scherer, 1986; Banse & Scherer, 1996; Wildgruber et al., 2002). Moreover, relations between physiological changes underlying emotional states and changes in acoustic parameters have been reported (Scherer, 1986). Therefore, superiority of the right or the left hemisphere in emotional processing may be due in part to the fact that some acoustic cues play a greater role in the perception of a given emotion. This suggests that some cues and/or some aspects of these acoustic cues are processed in a lateralized fashion.

Several levels of processing are involved in the perception of vocal emotions and each of them modulates brain activity and can contribute to lateralized activity. For instance, attention modulates early acoustic processing (Rinne et al., 2005) as well as vocal emotional processing (Schirmer, Kotz, & Friederici, 2005; Sander et al., 2005), and attentional modulation seems to elicit greater activity in the right auditory cortex than the left one (e.g., Pugh et al., 1996). The type of task also presumably elicits differential patterns of brain activity. For instance, Peper and Irle (1997) showed that frontal regions are involved in decoding the emotional valence of prosodic stimuli, the dorsolateral, parietal, and temporal regions in categorization and matching tasks, whereas decoding arousal seems to involve all of these regions within the right hemisphere. Moreover, increases in cognitive demands when tasks require complex processes have been correlated with enhanced activity in the left hemisphere (frontal and temporal) in healthy subjects (e.g., Kotz et al., 2000; Pihan et al., 2000), as well as in patients with brain lesions (Tompkins & Flowers, 1985). Another important factor is that semantic information influences brain responses and greater activations have been observed in the left hemisphere (Vikingstad, George, Johnson, & Cao, 2000). Indeed, differential patterns of activation for prosodic stimuli with regards to availability of semantic content have been reported. Positive and negative emotions expressed through prosody when semantic information was not available (using acoustic filters) elicited enhanced activity in the inferior frontal gyrus, whereas activations in these regions when semantic content was available were observed only for positive emotions (Kotz et al., 2003). Neuroimaging techniques with better temporal resolution, such as event-related potentials, may describe differences between neural activity underlying emotional processing of prosodic and semantic information by distinguishing temporal patterns (e.g., Schirmer et al., 2005). Linguistic prosody (i.e., modulation of acoustic parameters to communicate the phrasing and accentuation of words or sentences) is yet another candidate that influences neural activity associated with emotional prosody processing (as discussed by Ross et al., 1997; Pell, 1998; Baum & Pell, 1999). Weintraub, Mesulam, and Kramer (1981) and Brådvik et al. (1991) have even suggested that solely linguistic prosody is processed in the right hemisphere (studies from unilateral lesion to the right hemisphere). It is thus essential to carefully isolate these two types of prosody in order to study brain regions involved in processing emotional prosody.

Another important factor to consider in vocal emotion processing is that emotional stimuli can easily induce an emotional state, especially when a block design is used or when duration of the stimuli is long. In such cases, significant changes in gestures, facial expressions, and other physiological cues, such as arterial pressure and body temperature, are observed. Most studies of vocal emotions did not use electrophysiological measurements or questionnaires assessing the arousal of participants, and thus it is difficult to differentiate neural activity related to the perception of the stimuli from that related to mood induction effects. It may also be important to include different cognitive processes within a same study. For instance, it seems that the identification of the emotion and hedonic judgment are sometimes in *opposition*. In Wallbott and Scherer (1986), even if participants correctly identified sadness in prosodic stimuli, they judged these stimuli as being pleasant. These dichotomies remain to be explored by behavioral, electrophysiological, and neuroimaging studies.

Future studies need also to explore the distinction between emotion recognition and emotion understanding. It has been suggested that the mirror neuron system is one of the essential mechanisms involved in how humans understand emotions. According to this hypothesis, we understand emotions as we understand any other actions (e.g., Gallese, Keysers, & Rizzolatti, 2004). Specifically, it is hypothesized that we understand a given emotion expressed by a peer because part of the neural network that represents that given emotion is activated. For instance, the observation of a face expressing fear may engage part of the amygdala (Carr, Iacoboni, Dubeau, Mazziotta, & Lenzi, 2003), whereas observation of a face expressing disgust may recruit part of the insula (Wicker et al., 2003), in conjunction with the classical mirror neuron system neural networks, which include the rostral part of the inferior parietal lobe and the pars opercularis of the inferior frontal gyrus (Broca's area) (for more details see Rizzolatti, Fogassi, & Gallese, 2001). As far as we know, no studies have explored the mirror neuron system using vocal emotions; neuroimaging studies are needed to characterize the possible contribution of this mirror system in vocal emotion processing. (Editor: in agreement February 2006, but for a more popular discussion of valence, feelings, emotions and specific neuronal cells responses to selective visual (sic!) stimuli. See Scientific American: Mind, February/March 2006).

In summary, despite the distributed nature of the perceptual processing of vocal emotions, the right hemisphere, in particular the inferior frontal regions, seems to be a critical component of the system, which appears to work in collaboration with more posterior regions of the right hemisphere (such as the medial temporal gyrus), frontal regions of the left hemisphere, as well as subcortical structures. Ethofer et al. (2005) studied the connectivity of brain areas associated with emotional prosody and showed that processing emotional prosody first requires an acoustic analysis involving the right temporal cortex, after which the information is processed within the bilateral inferior frontal cortices.

Conclusion

Overall, the goal of this chapter was to discuss emotional processing through prosody and nonlinguistic vocalizations in an attempt to show that it is essential to further explore emotional nonlinguistic vocalizations. It is important to relate the literature on speech, facial expressions, and body gestures to the one on nonlinguistic vocalizations to have a complete view of how humans process emotions. The study of vocalizations will contribute to the understanding of other aspects of emotional processing that may not be possible (or different) using prosody. For instance, prosody and nonlinguistic vocalizations do not cover the same emotional spectrum. Indeed, the range of acoustic parameters used in prosodic patterns does not allow the expression of a set of emotions as large as that in vocalizations and limits the emotional intensity as well as spontaneity of expression (e.g., Scherer, 1981a; Barr, Hopkins, & Green, 2000). Moreover, language and culture represent confounding factors in emotional perception of prosodic stimuli. Cultural and language differences between speakers and decoders are at the center of emotional response. Pull effects, on the other hand, involve external conditions such as social norms. In any given case of emotion expression, both push and pull effects can be present and affect the resulting expression.

A consequence of the coexistence of push and pull effects is that there is no one-to-one relationship between expression and other components of emotion (e.g., subjective feeling). Individuals are most likely to report an emotion, and theorists are most likely to claim that an emotion has occurred, to the extent that many components of emotion co-occur (such as cognitive appraisal, subjective feeling, physiological arousal, expression; see Ekman, 1993).

Studies on Vocal Expression

Most studies have considered vocal expression as a means to communication. Hence, fundamental issues include (a) the *content* (What is communicated?), (b) the *accuracy* (How accurately is it communicated?), and (c) the *code* (How is it communicated?). The following sections first review the methods that have been used to address these questions, and then review the literature on decoding and encoding of emotions, respectively.

Methods of Collecting Vocal Expressions

A majority of studies on vocal expression have used some variant of the *standard content paradigm*. That is, someone (e.g., an actor) is instructed to read some verbal material aloud, while simultaneously portraying particular emotions chosen by the investigator. The emotion portrayals are first recorded and then evaluated in listening experiments to see whether listeners are able to decode the intended emotions. The same verbal material is used in portrayals of different emotions, and most typically has consisted of single words or short phrases. The assumption is that because the verbal material remains the same in the different portrayals, whatever effects appear in listeners' judgments should mainly be the result of the voice cues produced by the speaker. Other common methods include the use of emotional speech from real conversations (Eldred & Price, 1958; Greasley, Sherrard, & Waterman, 2000; Huttar, 1968), induction of moods in the speaker using various methods (Bachorowski & Owren, 1995; Bonner, 1943; Millot & Brand, 2001), and the use of speech synthesis to create emotional speech stimuli (Burkhardt, 2001; Cahn, 1990; Murray & Arnott, 1995).

Listeners' responses have most often been collected through forced-choice procedures, where the listener is asked to select one among several emotion labels. Another fixed-alternative method is to ask listeners to rate the stimuli on scales representing either emotion labels or emotion dimensions (Scherer, Banse, Wallbott, & Goldbeck, 1991). Free descriptions have also been used, though more sparsely.

The use of forced-choice methodology produces an ecologically valid task, but the fixed number of alternatives may produce artifacts. Some of the problems with the forced-choice method are alleviated by the use of rating scales, but the listeners' responses are still being influenced by the alternatives present. There have been several suggestions as to how one can improve the validity of the fixed-choice methodology, for instance by correcting for guessing (Wagner, 1993), or including "other emotion" as a response alternative (Frank & Stennet, 2001). The use of free descriptions is the least biasing task, but free descriptions are difficult to classify. It has been reported that free descriptions and the forced-choice task yield similar results, though free descriptions give more detailed information (Greasley et al., 2000).

Decoding of Vocal Expressions

It was early agreed that emotions can be communicated accurately through vocal expressions, a finding that is supported by common, everyday experience (Kramer, 1963). In the most comprehensive review to date, Juslin and Laukka (2003) conducted a meta-analysis of the literature on decoding accuracy of discrete emotions in both within-cultural and cross-cultural communication. Included in the analysis were studies that presented forced-choice decoding data relative to some independent criterion of encoding intention. To be able to compare accuracy scores from different studies with different numbers of response alternatives in the decoding task, the accuracy scores were transformed to Rosenthal and Rubin's (1989) effect size index for one-sample, multiple-choice-type data, *pi*. This index transforms accuracy scores involving any number of response alternatives to a standard scale of dichotomous choice, on which .50 is the null value and 1.00 corresponds to 100% correct decoding. The results of the meta-analysis, in terms of the pi index, are shown in Table 11–1. Also shown in the table are additional data regarding surprise and disgust taken from Laukka and Juslin (2002).

The means and confidence intervals presented in Table 11–1 suggest that the decoding accuracy is typically significantly higher than what would be expected by chance alone for both within-cultural and cross-cultural vocal expression. However, the accuracy was significantly higher (*t*-test, $p < .01$) for within-cultural expression (pi = .90) than for cross-cultural expression (pi = .85). Among the individual emotions, anger and sadness were best decoded, followed by fear, happiness, and surprise. Disgust and tenderness received the lowest accuracy although it must be noted that the estimates for these emotions were based on fewer data points. The pattern of results visible in Table 11–1 is consistent with previous reviews of vocal expression featuring fewer studies, but differs from the pattern found in studies of facial expression, in which happiness is usually better decoded than other emotions (Elfenbein & Ambady, 2002).

Conclusions (decoding): (a) The communication of emotions may reach an accuracy well above chance level, at least for broad emotion categories corresponding to basic emotions (that is, anger, disgust, fear, happiness, sadness, surprise, tenderness). (b) Vocal expressions of emotion are accurately decoded cross-culturally, although the accuracy is somewhat lower than for within-cultural vocal expression.

Encoding Studies of Vocal Expression

Almost from the beginning of empirical research on vocal expressions, researchers started to acoustically analyze the emotional speech, hoping to find acoustic voice cues that signal various emotional states sadness and tenderness. F4 is slightly higher in the sample of tenderness and the overall spectral slope is less steep, which is prone to make the voice quality brighter in general.

Discussion

On the basis of the studies reviewed here (Laukkanen et al. 1995; 1996; 1997; Airas & Alku, 2006; Waaramaa et al., 2006, Waaramaa et al., 2007), the glottal waveform seems to play a remarkable role in conveying emotional content of speech. After elimination of differences in F_0, SPL, and duration, samples were still perceived to signal some emotional states based on the characteristics of the glottal waveform and formant frequencies. Samples with low SQ and high QOQ reflecting low vocal effort level tended to be perceived as expressions of emotional states with low psychophysiological activity level, such as sadness, tenderness, or surprise (Laukkanen et al., 1997). Samples with high SQ and low QOQ, on the other hand, tended to be perceived as expressions of enthusiasm or anger.

These results of the expressive role of voice source are in accordance with the theory on *glottal mimicry* presented by Fónagy (1962). The basic dualism was not perfect, though. The subjects of the study by Laukkanen et al., 1997 perceived joy, fear, and sadness in samples with opposite glottal waveform characteristics. This is not likely to be just a sign of the listeners' insecurity in their judgments. Instead, it may illustrate the fact that emotional states and their expressions may take different forms depending on the vocalizer's evaluation of the situation. For example, the subject may be depressive and acquiescent in sadness and fear and therefore use low activity level, or he or she may be full of grief and desire to change the negative situation, thus exploiting a high activity level.

Strength and activity level related to an emotion has been suggested to be mainly expressed by pitch, loudness, and duration, while valence of an emotion is assumed to be communicated by rhythm and voice quality (Murray & Arnott, 1993). In the study by Laukkanen et al., 1997, there was reasonable consistency between the intended and the perceived valence, suggesting that perception of valence has still been possible regardless of the elimination of F_0, SPL, and temporal differences between the samples, and the regression analysis suggested that only F1 and F4 (in vowel [a:]) had significant effect on the perception of valence. In further work by Waaramaa et al. (2007), F3 was found to have relevance in valence coding for specific vocalic back open unrounded to semiclosed rounded segments [a:, o:], while F2 seemed to play a role in open back-to-front semiopen unrounded [a:, e:] vocalic segments.

The experiments with semisynthetic material (Waaramaa et al., 2006) also suggested that F3 and spectral slope have some role in coding the valence. Similarly, Laukkanen et al., 1997 concluded that the voice source type seemed to affect the perception of valence. This is naturally to be expected because the voice source type, in turn, relates to the spectral slope. The findings of for instance Gobl and Ní Chasaide (2003) and Alku and Airas (in press) suggest that the pressed vocal quality (and thus also the steepness of the spectral slope) correlates with the emotion activation rather than valence.

are universal because of phylogenetic continuity (e.g., Scherer, 1999b). This would mean that vocal expression of emotion must be, and must have been, an adaptive behavior, that is, it must be functional. Its *functionality* determines the properties of an emotional expression (Darwin, 1872). However, in the sequence model of vocal expression discussed in the second section, the local display rules also *render* vocal expression functional. A more positive view of vocal communication of emotion would suggest that the community, the culture of the speaker, is itself one of the environmental systems that influence ongoing social microevolution. In this perspective, then, there is mutuality, rather than a unidirectional influence, between the two systems discussed here: the emotional-affective action systems and the cognitive communicative systems.

Conclusions

Raw affective experiences are a gift of nature; they are not cognitively penetrable (Panksepp, 2005). However, the percepts that result from vocalization *are* penetrable for humans: they can be attended to, repeated, reflected upon, dramatized, etc. Diversity develops in communication systems because expressions of emotion involve multiple channels, including various cognitive ones, but intentional vocalization might utilize only a subsection of these. In situations of fear the voice might become tense, it might tremble, and so on. In attending to the vocalization, a listener might pick out one of the vocal qualities, and ultimately a tense voice can become associated with fear in one group, whereas in another group a trembling voice is the more typical expression of fear. Communication systems undergo processes of conventionalization and change that are in part independent from the universal constraints of the affective system. Communication is a way of engaging the world that is distinct in principle from such evolutionarily more basic systems (Panksepp, 2005, p. 162).

The human voice comes to convey many subtle changes in emotional state —at least to somebody who has a history of engagement with the speaker. It is misleading to narrow down the universal to only a handful of emotions, in the hope that they will be the major way in which emotionality is communicated in the voice. Rather, it seems to us that there is a lot that is universal in the communication of affective states, and at the same time there is a lot that is diverse: many affective states (tiredness, playfulness, exhilaration, fear, many shades of distress, and many shades of joy) are surely felt by all people (Panksepp, 2005). However, as soon as such states are communicated, not to mention given names and thus ordered into categories, it is increasingly less fruitful to look at these vocalizations for a universal core of emotionality. Emotions become semiotically formed in communication, and emotions are easily recognizable from the voice if hearer and speaker share a history of using the same communication system. For example, it is easy to recognize contempt from the voice of a person that one knows well. However, the recognition of contempt has been very low in recognition studies (Elfenbein & Ambady, 2002).

What follows is that universality in the vocal expression of emotions can be studied in two ways. Studies of biologically entrenched action systems might be more successful with nonhuman animals and very early stages of communication development in our own species. Studies of the vocal communication of emotions in humans that have been or are becoming enculturated might more fruitfully focus on the development of universality in histories of social engagement.

Models trying to incorporate both universality and diversity have taken the form of unidirectional sequence models: An emotional state triggers a universal mechanism (a motor program for vocal expression, for example), which is then modulated according to local display rules. Fundamentally, these models use conduit-metaphors of communication (Reddy, 1979/1993), where the basic emotions are the substance that itself remains untouched, and just becomes wrapped into different cognitive-communicative packages by the local code that needs to be decoded by the listener. We have suggested an alternative model that accommodates the ambiguous results of empirical research and motivates new questions about the universality and diversity of vocal expressions of emotions.

Vocal expression of emotion is usually studied as a natural sign, with a self-evident meaning. However, the fact that we can study the acoustic properties of an auditory signal and establish acoustic regularities does not mean that the meaning of such a natural sign is self-evident. The fact that a certain percept (such as a particular pitch contour) is a natural symptom of a state does not mean that it is understood *as* a sign of this state. As long as it is not clear what the relation is between auditory perception and conscious states, the description of acoustic properties cannot be used as a consistent description of a meaning.

In sum, we have been arguing that universals of iconic vocal emotional communication develop in histories of engagement. If such a process of developing universals exists, it must necessarily build on yet other universal potentials of the brain and mind. One of these potentials is accessible in the acoustics of affect bursts. Others have to be found in consistent patterns of social cognition, which may become increasingly complex with increasingly complex manmade environments. Presumably, the universal tendencies that govern social cognition would have less variability in humans still living in natural environments, but they may vary substantially depending upon the ecological factors of those environments. Future research might attempt to address the development of universals in the intersubjective engagements that emerge in different sociocultural contexts and world environments, and the relation between such developing universals and evolved vocal action patterns.

Acknowledgments. We are grateful to Alan Costall for discussions on an earlier draft of this chapter. Jörg Zinken's research is supported by the European Union's sixth framework project "Stages in the Evolution and Development of Sign Use."

References

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

Bachorowski, J. A., & Owren, M. J. (1995). Vocal expression of emotion: Acoustic

against American English. The order of speaking foreign languages for all participants was at random, and the speakers self-judged the fluency of the nonnative language.

Results were analyzed for fundamental frequency only as opposed to the Finnish experiment, where F_0 and sound pressure level (SPL), voiced-voiceless ratio, and vocal loading index for Finnish females reading in Finnish and in one to three foreign languages were analyzed.

Results

As in the Finnish experiment, the F_0 was higher for all speakers (male and female) when the nonnative language was produced. The greatest F_0 distance was found for the least familiar or the imitated language.

For the male speakers producing familiar foreign languages the maximum F_0 difference was nearly 20 Hz, with average difference not exceeding 10 Hz. For the female the maximum difference was 30 Hz, with the average difference not exceeding 15 Hz. The F_0 distance reached 70 Hz when imitated language was produced by American English native speakers imitating Cantonese Chinese.

Discussion

The results showed that all native speakers, be it of Finnish, Polish, American English, or Tagalog and irrespectively of gender, used higher mean F_0 when speaking foreign languages as compared to their native language. The widest F_0 distance was found for the imitation condition. For the non-Finnish group a trend of mode pitch distance between the native and "the least fluent or the completely unknown (imitation condition) language" appeared. This was interpreted as indicative of added stress factor from the idea of pitch matching to the target speaker, no matter the gender of the target speaker and the experimental speaker. In that sense, the adult speakers behaved in way youngsters behave when imitating adult prosody.

This rise in F_0 most likely can be interpreted on many grounds, but it may reflect a higher psycho-physiological activity (Orlikoff & Baken, 1988; 1989) when speaking in a foreign language, which supposedly is a more demanding and hence emotionally loading task than speaking in one's mother tongue. Elevated F_0 has also been explained as a sign of submissiveness and willingness to cooperate, as in interrogative prosody, indicating a willingness to participate in a discourse (Grabe & Karpinski, 2003).

The use of pitch and its variation are also culture and gender dependent (Lewis, 2002). It is possible that the subjects of the present study tried to some extent to adapt their voices to a certain level they imagined the native speakers of the foreign language would use. This adaptation phenomenon appears to be a useful tool in reducing the so-called foreign accent effect.

According to previous studies the average fundamental frequency in Finnish-speaking females is lower than for example in Swedish or English-speaking women (Laukkanen & Leino, 1999; Pegoraro Krook, 1988; Rantala 2000). Some subjects of the present study reported that they felt that they use a higher pitch in foreign languages but they could not explain the reason for it. Ohara has found in her studies (1992, 1999) that

MELISM—An Automatic Method of Segmentation and Melodic Coding of Speech

To allow accurate descriptions of melodic salience, an automatic analysis tool, MELISM, was developed by Caelen-Haumont and Auran (2004; 2005). MELISM procedure is based on Praat software (Boersma & Weenink, 1996) and allows automatic detection of melisms, their segmentation into *tonal syllables*, and their positioning on a nine-level scale based on a stylized F_0 curve generated by MOMEL[2] (Hirst & Espesser, 1993). The tonal syllables are mono- or bitonal sequences obtained at the points of change of melodic slopes. MELISM requires (a) a preliminary segmentation of the signal into linguistic units considered as relevant (e.g., single words or prosodic words), and (b) a stylization of the F_0 contour by determining target points with MOMEL algorithm.

Table 15–2 illustrates the nine-level scale based on Delattre's four-level model (Delattre, 1966) and obtained by dividing the space between each of the four levels into three segments. The nine-level scale is expressed by the following symbols: a = Acute; s = Supra; h = High; s = Elevated; c = Central; b = Bottom; I = Infra; g = Grave. The more acute ones (A, S or H) are involved in the definition of melisms.

To illustrate the MELISM procedure, two spontaneous speech samples are presented here, one expressing the emotion of joy and the other two attitudes.

Table 15–2. Matrix of tonal sequences describing the melodic configurations of words

	Melisms								
Tone	**Acute** **a**	**Supra** **s**	**High** **h**	**Elevated** **e**	**Middle** **m**	**Central** **c**	**Bottom** **b**	**Infra** **i**	**Grave** **g**
a	*aa*	*as*	*ah*	*ae*	*am*	*ac*	*ab*	*ai*	*ag*
s	*sa*	*ss*	*sh*	*se*	*sm*	*sc*	*sb*	*si*	*sg*
h	*ha*	*hs*	hh	he	hm	hc	hb	hi	hg
e	*ea*	*es*	eh	ee	em	ec	eb	ei	eg
m	*ma*	*ms*	mh	me	mm	mc	mb	mi	mg
c	*ca*	*cs*	*ch*	ce	cm	cc	cb	ci	cg
b	*ba*	*bs*	*bh*	be	bm	bc	bb	bi	bg
i	*ia*	*is*	*ih*	ie	im	ic	ib	ii	ig
g	*ga*	*gs*	*gh*	ge	gm	gc	gb	gi	gg

[2] MOMEL (Hirst & Espesser, 1993) allows stylization of fundamental frequency contours as a combination of their macromelodic and a micromelodic components. This is assumed to correspond to the global pitch contour of the utterance, which is continuous and independent of the nature of the constituent phonemes. It corresponds approximately to what we produce if we hum an utterance instead of speaking it.