In S. Hancil and D. Hirst (eds.) Prosody and Iconicity, John Benjamins Publishing Co, pp. 33-60.

Emotional expressions as communicative signals

Yi Xu, Andrew Kelly and Cameron Smillie
University College London

Abstract: It is widely assumed that emotions in speech are mainly expressed through prosody, particularly in terms of intonational contours. However, no theoretical models have been specifically developed to explain how exactly emotional meanings are conveyed by prosody. In this paper we explore the idea that emotional expressions are evolutionarily designed to elicit behaviours that are beneficial to the signaller. We show with experimental data that emotional meanings are encoded along a set of benefit-oriented *bio-informational dimensions* which involve both segmental and prosodic aspects of the vocal signal. We argue further that the proposed bio-informational dimensions allow emotional meanings to be encoded in parallel with non-emotional meanings, thus there is unlikely to be an autonomous affective prosody.

1. Introduction

While our knowledge about the linguistic aspect of human speech has reached a sophisticated level, the understanding of the emotional and attitudinal components of speech is still rather rudimentary. Although much research has been conducted in this area, the large amount of data generated have yet to lead to strong predictive models of emotional speech (Scherer, 2003). Part of the problem, as repeatedly pointed out by Scherer (1986, 2003), is a general lack of theoretical pursuit for the underlying encoding mechanisms. The most common practice in the field has been to examine as many acoustic parameters as possible and measure their correlation with various emotions (Mozziconacci, 2001; Murray & Arnott, 1993; Scherer, 2003; Shami & Verhelst, 2007; Ververidis & Kotropoulos, 2006; Williams & Stevens, 1972). As a result, the data collected are largely unconnected to each other, and difficult to be used to improve existing models. As explained by Scherer (2003:234), "As in most other areas of scientific inquiry, an atheoretical approach has serious shortcomings, particularly in not being able to account for lack of replication and in not allowing the identification of the mechanism underlying the effects found." In this paper, we will examine some of the fundamental issues about emotion and emotional expressions, and explore the possibility of developing a theory-based model of affective vocal expressions. We will present two sets of experimental data in support of the new approach. And finally, we will explore how the proposed emotion model may be linked to the recently proposed articulatory-fucntional model of speech.

1.1. Nature of emotion and emotional expressions

As human beings, we all have first-hand experiences that can be described as emotional. Thus it is easy for us to feel that we know at an intuitive level what emotion is, i.e., emotion seems to be first and foremost something we feel is unquestionably internal. But the same is also true of hunger, thirst and sexual drive,

which, though we feel just as real, are less mysterious, because it is easy to see that without hunger, thirst or sexual drive, we would not have survived thus far as a species. So, few of us would insist that we have those internal drives just for their own sake. But how about emotion? Is it also important for our survival? If it is not, why is it so common across highly diverse human communities (setting aside for the moment the issue of cultural differences in the expression of specific emotions, e.g., Ekman et al., 1987)? The survival value of emotion has been contemplated by some researchers. Ekman (1992:171) posits that emotions are mechanisms that prepare individuals to deal with fundamental life-tasks: "the primary function of emotion is to mobilise the organism to deal quickly with important interpersonal encounters, prepared to do so in part, at least, by what types of activity have been adaptive in the past." It is not yet fully clear, though, what exactly emotion prepares us to do. There has been some clues: "For example, fighting might well have been the adaptive action in anger, which is consistent with the finding that blood goes to the hands in anger. Fleeing from a predator might well have been the adaptive action in fear, which is consistent with the finding that blood goes to large skeletal muscles (Ekman, 1992:181)". Susskind et al. (2008) have reported evidence that the facial expressions of fear and disgust are to enable us to either maximize (in fear) or minimize (in disgust) sensory input. In general, however, the survival value of emotion is not yet fully clear.

Meanwhile, theories that try to model the meanings of emotion, the dimensional theories in particular, are focused mainly on the internal feelings of the emotion-bearer (Borod, 1993; Mauss, 2009; Schlosberg, 1954; Zei, 2002), which, as mentioned earlier, seem to only reflect our intuitions. The valence dimension corresponds to whether the emotion is pleasant or unpleasant, or agreeable or disagreeable, to the emotion-bearer him/herself. The activation (arousal) dimension describes the level of activation of the emotion-bearer, or whether the emotion-bearer is active or passive. And the power dimension describes power, control or attention/rejection of the emotion-bearer. Another proposed dimension is approach-withdrawal or approach-avoidance, which is again about how the emotion-bearer feels like to do him/herself (Borod, 1993; Zei, 2002). While these proposed internal feelings all seem to make intuitive sense, it is difficult to see why they need to be overtly expressed.

Of course, such difficulty would be a non-issue if we believe that emotional expressions are not intentionally made to send a signal. In fact, it has been argued that communication is not the core nature of emotional expressions (Ekman, 1997). The key argument is that the word *communication* implies that expressions are made intentionally to send a message, which contradicts the belief that the information revealed by the emotional expressions are by and large not directly intended (Ekman, 1997). Here *intentionality* is used to determine whether emotional expressions are communicational or just outer sign of internal feelings. But is it possible that a message can be sent without conscious intention of the sender? As Ekman (1992:189) has observed himself, the basic emotions have the common feature of *unbidden occurrence*. That is, "one can not simply elect when to have which emotion", and "we often experience emotions as happening to, not chosen by us". As a result, both the internal and external physiological symptoms of an emotion happen to us involuntarily. Why couldn't it be the case, then, that sending out messages is also something that happens to us without our conscious intention? In other words, is it possible that internal feelings are an evolution-engineered mechanism to quickly

mobilize *all* the reactions needed to cope with the situation, *including* effective ways to influence the behaviour of the receiver? To answer these questions, we need to take a more general perspective.

1.2. An evolutionary perspective

If, as observed by Darwin (1872), humans and nonhuman animals share much in common in terms of emotional expressions; and if, as argued by Ekman (1992), emotion is unbidden, occurring to us quickly and involuntarily; and if, as also argued by Ekman (1992), emotion has evolved to mobilise us to deal quickly with fundamental life-tasks, emotional expressions may also be the product of evolution, i.e., having been shaped by various selection pressures. But what could have been the selection pressures? Morton (1977) examined the calls of dozens of avian and mammalian species and observed that "birds and mammals use harsh, relatively low-frequency sounds when hostile and higher frequency, more pure tone like sounds when frightened, appeasing, or approaching in a friendly manner" (p. 855). He then theorized that these sound qualities are related to the body size of the animals. Low frequency and harsh sounds are related to a larger body while high pitch and pure-tone like sounds are related to a smaller body. This is because, first, a larger body has a better chance of winning a physical confrontation, thus imitating sound qualities produced by a large animal may help to deter the opponent/adversary. Secondly, selection pressure would favour developing strategies to win conflicts by creating impressions of having a large body over increasing the actual body size, as the latter is constrained by availability of resources. Morton further hypothesized that the same selection pressure has lead to calls that signal fear and appeasement by simulating sound qualities related to a small body size. But he further conjectures that a further source for fear/submission calls may come from the likelihood to elicit parental responses with infant like sounds, which are typically high pitched and pure tone like.

Are humans also subject to the kind of selection pressure Morton proposed for other animals? The answer seems to be yes, given the evidence accumulated since the early 1980s. Ohala (1984) presented rather convincing arguments that sexual dimorphism in the size and location of the human larynx is the result of mating competition among males. He showed that the increase in larynx size is mostly in the front-back dimension (rather than proportionally increasing the whole mass) which effectively lengthens the vocal folds and thus decreases F_0. He also argued that the descended male larynx is to lengthen the vocal tract, which lowers the resonances. Both changes have the effect of mimicking the vocal output of an individual with a large body. The mating-related nature of dimorphism is further evident in the fact that the enlargement and descent of the male larynx both happen at puberty, i.e., the time at which they are supposed to be preparing for the imminent mating competitions.

Lowering vocal resonances by lengthening the vocal tract is a size-related acoustic feature not considered by Morton (1977). But its importance is no less than that of F_0 and voice quality for several reasons. First, as has been demonstrated, at least in mammals, the length of the vocal tract is limited by the sternum, whose position is proportional to actual body size (Fitch, 1997; Reby & McComb, 2003). Thus vocal tract length may provide reliable information about the body size of the vocalizer. In contrast, the size of the larynx is less limited by the actual body size. Support for this

argument can be found in the finding that male human F_0 is not correlated with body size, but male vocal tract length is (González, 2004). Second, as will be discussed in more detail later, other things being equal, F_0 is positively related to vocal effort/subglottal pressure/loudness, which in turn is related to activation/arousal level. Thus, unlike F_0, vocal tract length more purely signals size information. Third, vocal tract elongation for the sake of exaggerating body size has been found in many animals. The most remarkable cases are birds whose vocal tracts (which are unrestricted by the sternum) have become so long that they form loops and coils within the thorax (Fitch, 1999).

The importance of manipulating vocal tract length to signal size is further highlighted by Ohala's (1984) hypothesis about the origin of the smile. Though showing more teeth, which are potential weapons, smiling has the acoustic effect of shortening the vocal tract. Ohala thus suggested that the smile is for the sake of mimicking a smaller body size during vocalization, thus signalling appeasement and sociability, just like increasing frequency and tone-like quality to signal submission as hypothesized by Morton (1977).

1.3. A bio-informational dimensions theory

The theorizations by Morton and Ohala could be further extended to a more comprehensive theory of human emotional vocalizations, and this is one of the objectives of this paper. Following Morton (1977), Ohala (1997) and Russell, Bachorowski and Fernández-Dols (2003), we believe that *human vocal expressions of emotions are evolutionarily shaped to elicit behaviours that may benefit the vocalizer*. As such they are not arbitrary signals, although their meanings are often not intuitively transparent, thanks to the deep evolutionary root that makes them highly automatic. More specifically, we propose that vocal emotional expressions are to influence the behaviour of the receivers by manipulating the vocal signal along a set of bio-informational dimensions, namely, *size projection*, *dynamicity*, *audibility* and *association*.

The *size projection* dimension, which is equivalent to what has been termed the frequency code (Ohala, 1984) or the size code (Chuenwattanapranithi et al., 2008; Gussenhoven, 2002), is to project either a large body size to create an effect of repelling or dominating the receiver, so as to express threat or assertiveness, or a small body size to create an effect of attracting or appeasing the receiver, so as to express friendliness, subordination or sympathy-pleading. At least three parameters are likely to be involved in this dimension — vocal tract length (VTL), which directly controls spectral density, F_0 and voice quality.

The *dynamicity* dimension controls how vigorous the vocalization sounds, depending on whether it is beneficial for the vocalizer to appear strong or weak. A vigorous vocalization has a large movement range with high velocity, in terms of both F_0 and formant movements, whereas a less vigorous vocalization has a narrow movement range with low velocity.

The *audibility* dimension controls how far a vocalization can be transmitted from the vocalizer, depending on whether and how much it is beneficial for the vocalizer to be heard over long distance. The control of audibility is mainly through intensity. But it

may have a significant effect on voice quality (Stevens, 1998).

The *association* dimension controls associative use of sounds typically accompanying a non-emotional biological function in circumstances beyond the original ones. For example, the disgust vocalization seems to mirror the sounds made when a person orally rejects unpleasant food (Darwin, 1872). Articulating this kind of sounds involves tightening the pharynx, which would result in raised F1 (Stevens, 1998) as well as devoicing.

The advantage of this bio-informational dimensions theory (henceforth BID) is that it allows us to construct testable hypotheses about specific emotions, moods and attitudes, and it allows us to connect findings that otherwise seem to be unrelated. In the following, we will discuss hypotheses about a number of commonly recognized basic emotions and how they are linked to some previous findings.

2. Preliminary BID interpretation of existing data

Despite being largely atheoretical as discussed earlier, past research has generated a large amount of data, some of which have actually provided initial, though retrospective, evidence for BID. In the following we will briefly overview the data on individual emotions and try to offer interpretations from a BID perspective. We will also briefly discuss some recent findings that seem to provide direct support for the size code hypothesis, a major precursor to BID.

2.1. Anger/happiness

These two emotions are discussed together because they are supposed to be the direct opposite of each other. They are also two of the most studied emotions, presumably because they are the most frequently encountered beside neutral emotion. According to Morrison, Wang and De Silva (2007), anger and happiness are by far the largest emotion categories (3.1% and 1.8%, respectively) communicated through a call centre system following neutral speech (93.3%), while the other emotional classes such as sadness, fear, surprise, and disgust have much lower percentages of occurrence (the highest being sadness, at only 0.1%). Interestingly, however, anger and happiness are also among the least distinguishable emotion pairs both in terms of identifiable acoustic parameters and rate of automatic recognition. Of the seven acoustic patterns summarized in Scherer (2003: Table 1), only one is different between the two emotions (*Sentence contour*: fall for anger/rage; undefined for joy/elation). In Murray & Arnott, (1993) of the seven parameters shown in Table I, three have identical or nearly identical features (*Speech rate*, *Pitch range*, *Intensity*). Of the eight parameters examined by Ververidis & Kotropoulos (2006:Table 2), three are identical for anger and joy (*Pitch range*, *Intensity range*, *Transmission duration*), two have missing values for one of the emotions (*Pitch contour*, *Speech rate*), and the rest three have the same direction of changes. In terms of automatic recognition, Kwon et al. (2003) found that angry samples were frequently categorized as happy, and happy samples as angry, when using acoustic parameters such as F_0, formant, energy and mel-frequency cepstral coefficients. Shami and Verhelst (2007: Table 21) reported that 20.3% of the angry speech was classified as happy, and 35.6% of the happy speech classified as angry, using a presumably state of the art recognition algorithm.

From a communicative perspective, however, anger and happiness should be among

the most easily distinguished emotions. The question is, what are the acoustic parameters used to encode them? BID would predict that the size projection dimension best separates these two emotions. That is, angry expressions project a large body size to repel or dominate the receiver, while happy expressions project a small size to attract or appease the receiver. It can also be predicted that the two emotions are likely to be similarly located near the high end of the dynamicity and audibility dimensions, because it is likely beneficial for an angry or happy vocalization to sound highly vigorous, although anger, especially hot anger, may have slightly greater dynamicity and audibility. Both anger and happiness are probably neutrally located along the association dimension.

2.1.1. Preliminary evidence

First, there is evidence that listeners are highly sensitive to acoustic manipulations along the size-projection dimension. They use spectral density and mean F_0 to judge speakers' body size and shape, and male speakers with deeper voice (lower F_0 and denser spectrum) are judged, although often wrongly, as taller and heavier (van Dommelen & Moxness, 1995). Female listeners judge men with voices with closely spaced low-frequency harmonics as more attractive, heavier, more likely to have a hairy chest and more muscular. Interestingly, their judgments are often wrong, but they nevertheless agree with each other closely (Collins, 2000). These findings are consistent with the hypothesis that males try to win females by projecting a larger body size to show that they have a better chance of wining a physical contest, and listeners are sensitive to the size-projection signals.

Second, there is evidence that smiles during speech are clearly audible (Aubergé and Cathiard, 2003; Drahota, Costall & Reddy, 2008; Robson & MackenzieBeck, 1999) and listeners can perceive happiness and unhappiness from speech spoken with a smiling or frowning face (Tartter and Braun, 1994).

Third, in terms of the acoustic parameters, most studied have found that both anger and happiness involve higher than neutral pitch, but there is great inconsistency as to which of the two has higher pitch. The only consistency with the prediction of size-projection dimension is the finding that anger involves steeper F_0 fall than happiness (Scherer, 2003). Also consistent with the prediction is the finding by Gobl & Ní Chasaidi (2003) that the best correlation of voice quality with emotion is between tense voice and anger. In general, therefore, previous evidence for these two emotions consistent with BID is somewhat scattered.

The clearest evidence is seen in Chuenwattanapranithi et al. (2008), which directly tested the hypothesis that anger and happiness are encoded by projecting body size along a large-small continuum. Human listeners were asked to judge the size and emotion of the speaker from vowels synthesized with different vocal tract lengths (VTL) and F_0. The results were consistent with the prediction of BID, i.e., vowels with longer VTL and lower F_0 were heard both as produced by a larger person and as by an angry person, and those with a shorter VTL and higher F_0 were heard as produced by a smaller and a happier person.

2.2. Fear

Fear is probably the most important emotion after anger and happiness. Among the

measurements found to be relevant for fear the most consistent is pitch. Most studies have found that fear is associated with high or very high pitch (Burkhardt & Sendlmeier, 2000; Cowie et al., 2001; Protopapas & Lieberman, 1997; Mozziconacci, 2001; Murray & Arnott, 1993; Ververidis & Kotropoulos, 2006). But some studies also reported lower F_0 than neutral emotion (Williams & Stevens, 1972). The other measurements are less consistent. Pitch range has been found to be either wider (Murray & Arnott, 1993; Mozziconacci, 2001; Cowie et al., 2001) or narrower (Fónagy & Magdics, 1963; Fónagy, 1978) than neutral emotion. Intensity has been found to be either higher (Scherer, 2003) than neutral emotion, or the same (Murray & Arnott, 1993). And speech rate has been found to be either faster (Mozziconacci, 2001) or slower (Sulc 1977; Williams & Stevens, 1972) than neutral emotion. Also, it is often believed that fear involves unintentional tremor, which is audible in fear vocalization. But no audible tremor was not found by Protopapas & Lieberman (1997). Also striking is that all these characteristics seem to be shared with anger and happiness. This can be best seen in the summary table in Scherer (2003:233), where none of the acoustic cues for fear can distinguish it from anger or joy with the only exception of its pitch range, which, is optionally narrower than that of anger and joy.

According to Morton (1977), fear is the opposite of hostility in animal calls, and is functionally similar to submission and appeasement. It is therefore reasonable to assume that in humans fear is also the opposite of anger which is presumably equivalent to hostility. Thus fear expressions should project a small body size — with high pitch, low spectral density and tone-like voice quality. On the dynamicity dimension, fear should be located toward the low end, because the vocalizer would want to give the receiver, who is likely to be an adversary, the impression that it is not a threat. On the audibility dimension, fear is also likely to be located near the low end, because it is likely to have evolved in situations where an aggressor is approaching, and so it is beneficial for the vocalizer to be heard as late as possible.

The generally reported high mean F_0 in fear seems to be consistent with the above prediction. For the tone-like voice quality, there has been some evidence. Burkhardt and Sendlmeier (2000) report that falsetto voice can be heard as fear, which is consistent with both very high F_0 and tone-like voice quality. Gobl & Ni Chasaidi (2003) found that synthetic whispery voice had the strongest response for fear, although in general fear was one of the least recognized emotion when only voice quality was manipulated. There has been no prior evidence for the above prediction that fear vocalization would project a mall body size also in terms of spectral density. Interestingly, Fónagy (1978) reported a high rate of human recognition of fear vocalization as reproach or suppressed anger.

2.3. Sadness

The case of sadness is probably more complicated than is usually recognized. Scherer (1979) suggests that there may be two kinds of sad vocalization, a quiet and passive type, and an active grief often seen in mourning. The word “sad” in English and some other languages actually refer to two rather different emotional states: a grieving type with sobbing voice, and a depressed type characterized by very low energy.[1] The problem can be seen in the typical facial expression of sadness, which is apparently that of the sobbing/grieving type (Ekman, 1998), whereas the acoustic parameters

typically found associated with sadness suggest that the vocal sadness being studied is often of the depressed type. As summarized by Banse & Scherer (1996), sadness is characterized by reduced F_0, pitch range, F_0 variability, intensity and speech rate. All these characteristics seems to suggest the depressed type of sadness. It seems that in most studies when asked to act out sadness, speakers typically produce the depressed type of vocalization.

Some studies have found rather different acoustic characters for sadness. Costanzo, Markel & Costanzo (1969) found that utterances with higher perceived pitch were heard as grief sounding. Interestingly, the paragraph they used to induce grief is a person expressing the feeling of losing a close relative. Likewise, Erickson et al. (2006) studied the voice of two speakers who were grieving for the loss of someone very close to them, and they also found higher F_0 associated with sad speech. Burkhardt & Sendlmeier (2000) found that perceived sadness was associated with raised pitch contour and falsetto voice as well as narrow pitch range, slow speech rate and breathy articulation. They also suggested that sadness should be split into two categories, *crying despair* with high arousal and *quiet sorrow* with low arousal.

From an BID perspective, since the two types of sadness have rather different communicative functions, they should be located well apart on the bio-informaitonal dimensions. For grieving sadness, its location on the size projection dimension should be toward the small-size end if the function of the vocalization is mainly to plead for sympathy, but toward the large-size end if the function is mainly to make a demand. For depressed sadness, the localization on the size-projection dimension should be neutral, as it is inconceivable why it should be located toward either the large- or small-size end. For depressed sadness, its location on the dynamicity and audibility dimensions should be rather low, mainly because of the low or even compressed activation level. For grieving sadness, the location on the audibility dimension should be high because it would be beneficial for vocalization to be easily heard. Its location on the dynamicity dimension cannot be straightforwardly predicted, although based on Erickson et al. (2006) it should be located toward the lower end.

2.4. Disgust

Although disgust is not as much researched as the emotions discussed thus far, the limited report has revealed some curious facts. First, recognition of disgust from facial expressions is much easier than from speech (Scherer 2003). Second, disgust is one of the most easily recognized emotions from non-verbal vocalizations (Sauter et al., 2009), but its recognition from speech is among the most difficult (Juslin, 2001; Scherer 2003). The recognition difficulty may be seen in the fact that the reported acoustic properties for disgust are not highly distinct from other emotions. According to Ververidis & Kotropoulos (2006), "Disgust is expressed with a low mean pitch level, a low intensity level, and a slower speech rate than the neutral state does." This description is very similar to that of the depressed type of sadness. According to Murray and Arnott (1993), compared to neutral speech, disgust has slower speech rate, lower mean pitch, lower intensity, which are again all similar to sadness. The only difference is that disgust has slightly wider pitch range than neutral speech, and wide, downward terminal inflections and grumbled voice quality. But these characteristics are somewhat similar to those of anger. There have also been report of increased F_0

(Scherer, 1986).

Darwin (1872) proposes, based on his principle of "serviceable habits", that the expression of disgust is derived from warnings to conspecifics of toxic or rotten food. If so, from the principle of "inclusive fitness" (also known as kin selection) (Hamilton, 1964), an expression for such a purpose is most likely to be selected if it is most effectively delivered to conspecifics feeding near by, who are likely to close relatives. As a result, facial expression, which is best viewed at a close distance, is probably more important than vocal expression of disgust (Scherer, 2003). Also conceivably, short vocal bursts are more effective than modifications of whole sentences in signalling the toxicity or foul taste/smell of food.

From the BID perspective, the most distinctive characteristics of disgust should be seen in the association dimension, on which it should be located toward the end that best emulate the sound of vomit or regurgitation. Acoustically, this would mean raised F1, increased noisiness or even devoicing. On the size-projection dimension, disgust may be located toward the large-size end, as there may be a need to sound assertive when warning about the danger of toxic food. But this may be counteracted by the retraction of the lip corners seen in the typical facial expression of disgust (Sherer, 1986). The audibility of disgust is likely to be low because it has presumably evolved from warnings to close relatives who are nearby during feeding time as discussed above. Its location on the dynamicity dimension is likely to be neutral except that speech rate may be reduced.

3. New data

In this section we will report preliminary data from two experiments. The first experiment is a partial replication of Chuenwattanapranithi et al. (2008) with a different method of acoustic manipulation. The second experiment is an initial test of BID.

3.1. Experiment 1

The goal of this experiment is to replicate the basic findings of Chuenwattanapranithi et al. (2008), but with a different method. Instead of generating vowels with an articulaory synthesizer, we resynthesized real human speech while manipulating spectral dispersion (inverse of spectral density) and F_0 along the size-projection dimension. Altering spectral dispersion had the equivalent effect of altering the length of the vocal tract. The use of real speech has the advantage of keeping all the other aspects of the acoustic signal as natural as possible.

3.1.1. Stimuli

The stimuli were the English digits 1, 2, 3 … 10, spoken by a male speaker of South British English, age 20, recorded in an anechoic chamber at University College London, in an emotionally 'neutral' voice. The spoken digits were then modified in terms of F_0 and spectral dispersion using the program Speech Filing System (Huckvale, 2008). Three factors were controlled in modifying the digits: *acoustic parameter* (F_0, spectral dispersion, both), *direction of modification* (up, down) and *manner of modification* (static, dynamic). Thus the total number of stimuli were 3 parameters x 2 directions x 2 manners x 10 digits = 120. Such a design avoids

combinations of parameter changes that are ambiguous in terms of size projection, e.g., increasing F_0 but decreasing spectral dispersion.

The [mean/median?] fundamental frequency of all the spoken digits was first set to 106 Hz and then the F_0 of each digit is either raised or lowered by 10 Hz. Also the change is applied either statically, i.e., by the same amount throughout a digit, or dynamically, i.e., increasing the amount of change from 0 to 10 Hz from the onset to the offset of the digit. Spectral dispersion was altered by either compressing or expanding the entire spectrum by 10%. Like F_0 modification, the spectral changes were applied either statically or dynamically throughout each digit.

3.1.2. Subjects and Procedure

Seven native speakers of British English participated as subjects. They were university students aged 20-22, 4 males and 3 females with no self-reported hearing problems.

The perceptual tests were carried out in a quiet room. The tests were run by the ExperimentMFC module of the Praat program (Boersma, 2001) on a laptop computer. Subjects listened to the stimuli through a set of BOSE Quiet Comfort 2 Acoustic Noise Cancelling headphones and performed two forced choice tasks. The first was to determine whether the speaker was large or small in body size, and the second was to determine whether the speaker was angry or happy. During each trial, a resynthesized digit was played once, and the subject indicated his/her decision by pressing a button on the screen.

The tokens were presented in random order and repeated in three blocks. Thus each subject made 360 judgments in each task. They did the emotion judgment task before doing the size judgment task. The subjects carried out the experiment individually and were given a practice round to customize themselves to the voice and nature of the experiment. They were instructed to make judgments instinctively without thinking too hard.

3.1.3. Results

Size perception

Each of subjects' responses is coded as 1 for judging the speaker as happy or small, and 0 for judging the speaker as angry or large, and the average of the three repetitions for each combination of parameter changes was computed as the response score. Figure 1a displays response scores for body size as a function of parameter and direction of manipulation. Digits with increased F_0, increased spectral dispersion or both led to higher scores for smaller body size judgment, while those with decreased F_0, decreased spectral dispersion or both led to lower scores. A three-way repeated measures ANOVA shows that the effect of manipulation direction is highly significant ($F[1,6] = 166.21$, $p < 0.001$). Figure 1a also shows that the scores differed across the three parameter conditions, and the differences are significant ($F[1,6] = 3.99$, $p < 0.05$). Also the effect of direction becomes larger as the parameter condition changes from F_0 to spectrum to both F_0 and spectrum, as is shown by the significant interaction between direction and parameter of manipulation ($F[2,12] = 18.25$, $p < 0.001$).

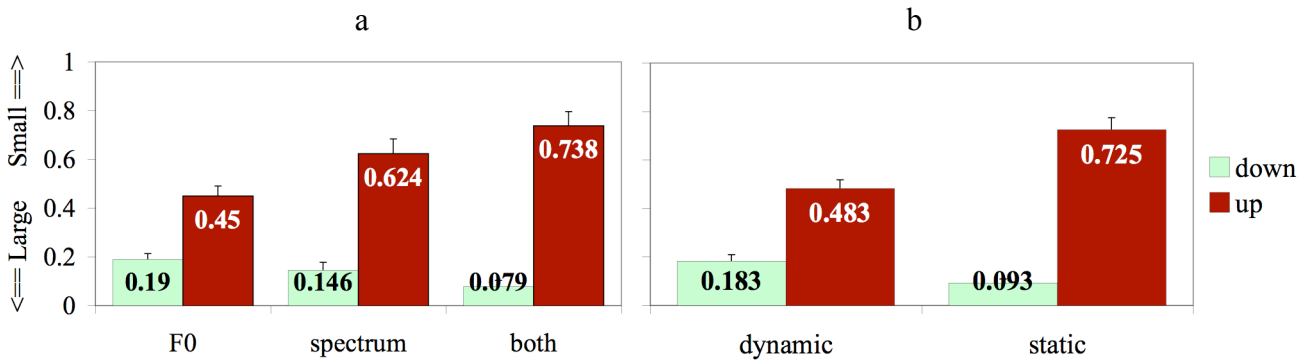

Figure 1. a) Response scores for body size as a function of acoustic parameter and direction of manipulation. b) Response scores for body size as a function of manner and direction of parameter manipulation.

Figure 1b shows that size judgment scores are also affected by manner of parameter manipulation ($F[1,6] = 15.71$, $p < 0.01$). The scores become more extreme when the parameter change is static than when it is dynamic. There is a significant interaction between manner and direction of parameter change ($F[2,12] = 128.99$, $p < 0.0001$).

These results show that listeners are highly sensitive to the parameter manipulations performed on the spoken digits when judging the body size of the speaker. They judged digits with higher F_0, greater spectral dispersion or both as spoken by a smaller person, and they judged digits with lower F_0, smaller spectral dispersion or both as spoken by a larger person. Also they were more sensitive to static than dynamic parameter changes.

Emotion perception

Figure 2a displays response scores for emotion as a function of parameter and direction of manipulation. Digits with increased F_0, increased spectral dispersion or both led to higher happiness scores, while those with decreased F_0, decreased spectral dispersion or both led to lower happiness scores. A three-way repeated measures ANOVA shows that the effect of manipulation direction is highly significant ($F[1,6] = 79.17$, $p < 0.001$). The effect of parameter conditions is not significant, although differences in the means can be seen in the Figure. There is, however, a significant interaction between direction and parameter of manipulation ($F[2,12] = 32.64$, $p < 0.001$), this is because the direction effect becomes larger as the parameter condition changes from F_0 to spectrum to both F_0 and spectrum.

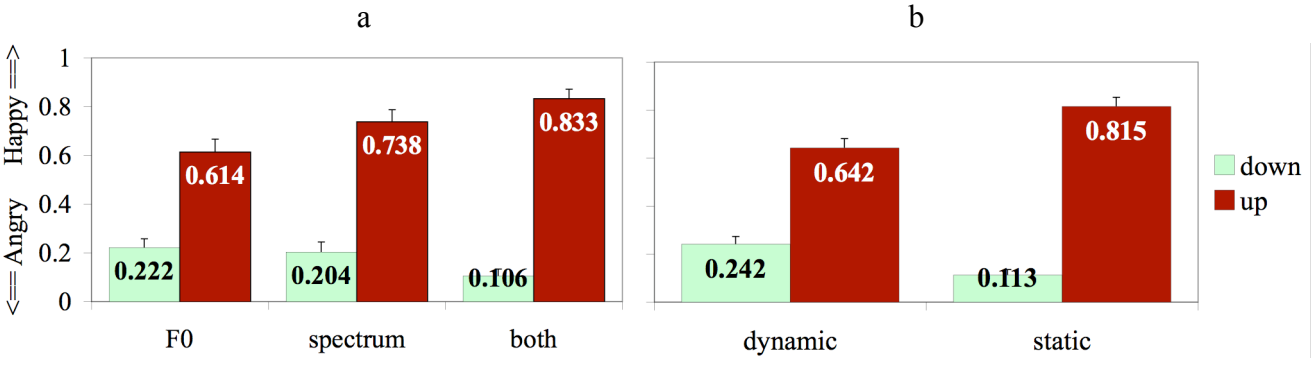

Figure 2. a) Response scores for emotions as a function of acoustic parameter and direction of manipulation. b) Response scores for emotions as a function of manner and direction of parameter manipulation.

There is no main effect of manner of manipulation, but there is a significant interaction between manner and direction of parameter change ($F[2,12] = 142.64$, $p < 0.0001$). The scores become more extreme when the parameter change is static than when it is dynamic.

These results show that listeners are highly sensitive to the parameter manipulations performed on the spoken digits when judging the emotion of the speaker. They judged digits with higher F_0, greater spectral dispersion or both as spoken by a happy person, and digits with lower F_0, smaller spectral dispersion or both as spoken by an angry person. Also they were more sensitive to static than dynamic parameter changes.

Overall, there is a bias toward hearing a large body size and angry voice, as can be seen in Tables 1 and 2, in which the scores for the down stimuli have been transformed by applying the following equation:

$$S' = 1 - S \qquad [1]$$

where S' is the new score and S the original score.

Table 1. Mean size judgment scores computed with equation (1). Standard errors are shown in parentheses.

Parameter / Direction	F_0	Spectrum	Both
down	0.81 (0.024)	0.854 (0.032)	0.921 (0.025)
up	0.45 (0.041)	0.624 (0.06)	0.738 (0.058)
Manner	dynamic	static	
down	0.817 (0.026)	0.907 (0.016)	
up	0.483 (0.035)	0.725 (0.05)	

Table 1. Mean emotion judgment scores computed with equation (1). Standard errors are shown in parentheses.

Parameter / Direction	F_0	Spectrum	Both
down	0.778 (0.036)	0.796 (0.041)	0.894 (0.028)
up	0.614 (0.053)	0.738 (0.05)	0.833 (0.039)
Manner	dynamic	static	
down	0.758 (0.031)	0.887 (0.023)	
up	0.642 (0.039)	0.815 (0.038)	

3.1.4. Findings of Experiment 1

The results of experiment 1 show that listeners are highly sensitive to variations in F_0 and spectral dispersion both in judging body size and in judging emotion even when the manipulations are performed on naturally spoken words. Increased F_0 and spectral dispersion lead to perception of smaller body size and happiness, and decreased F_0 and spectral dispersion lead to perception of larger body size and anger. The perceptual sensitivity in the case of body size judgment agrees well with the finding of Ives et al., (2005), Smith et al. (2005) and Turner & Patterson (2003). The sensitivity in the case of emotion judgment is consistent with the findings of Chuenwattanapranithi et al. (2008). This offers further support for the size projection dimension of the BID theory.

One finding of Chuenwattanapranithi et al. (2008) not replicated here is that temporally dynamic parameter changes did not lead to more consistent emotion judgment. Rather, it is the stimuli with fixed parameter changes that elicited more consistent judgments. A likely explanation is that the acoustic parameters in question — F_0 and spectral properties — were already dynamic in the spoken digits, whereas in Chuenwattanapranithi et al. (2008) the manipulated parameters in the steady-state vowels were genuinely static. It is possible that the sensitivity of emotional perception is subject to the presence/absence rather than the magnitude of dynamic movements. Another possibility is that the dynamic manipulation performed in the present experiment generated smaller overall differences in F_0 and spectral property than in Chuenwattanapranithi et al. (2008), judging from the fact that in Figure 1, the judgment difference is much smaller in the dynamic condition than in the static condition. This is rather different from the very similar size judgment difference between static and dynamic conditions in Chuenwattanapranithi et al. (2008).

3.2. Experiment 2

The goal of this experiment is two-fold. This first is to perform a preliminary test of BID, and the second is to test the idea that emotional encoding is parallel to speech prosody. That is, given an emotionally neutral utterance with proper intonation, it is possible to make it sound "emotional" by imposing global (rather than local) modifications of certain acoustic parameters along the bio-informational dimensions. In doing so the normal intonation carrying "linguistic" meanings remains largely intact although emotional information has been added.

3.2.1. Stimuli

The base stimuli is a complete sentence recorded by a male speaker of Southern British English. The sentence is "I owe you a yoyo", which was chosen because it is emotionally neutral, and because it consists of only vowels and glides, which makes the manipulations of voice quality and F_0 maximally audible. The sentence was recorded at normal speech rate with prosodic focus on the word "owe". Thus the intonation of the sentence carries the non-emotional, pragmatic meaning of focus (Gussenhoven, 2007; Xu, 2005; Xu & Xu, 2005).

The recording was made in a sound-proof room onto the CoolEdit 2000 computer program (via a Mono RS microphone AKG 249-946). The manipulation of the parameters was done using the "Change gender" function of Praat (Boersma, 2001).

This function allows users to independently change formant shift ratio (larger ratio = greater spectral dispersion), pitch median, pitch range factor and duration factor (larger value = longer duration). A Praat script was written to apply the parameter values shown in Table 3, resulting in 4 x 4 x 4 x 2 = 128 unique stimuli.

Table 3. Parameters and their values used in the resynthesis of the original speech utterance.

Formant shift ratio	Pitch median (Hz)	Pitch range factor	Duration factor
1.2	400	4	1.1
1.078	200	1.170	
0.956	100	0.341	0.9
0.833	50	0.1	

3.2.2. Subjects and procedure

Fifteen speakers of British English, age 20-22 (4 males and 3 females) participated as subjects. Like in experiment 1, the perception tests were run by a MFC script in Praat. The subjects were instructed to listen to the resynthesized sentences, which were presented to them in random order and determined whether the speaker sounded happy, angry, scared, grief-stricken, or depressed. The emotion 'grief-stricken' was described to the subjects as experienced by someone who could well be on the verge of tears. The subjects were then played 3 cycles of the 128 utterances created (128 x 3 = 384 utterances in total) in a randomised order via headphones (Sennheiser HD 265 linear).

3.2.3. Results

Table 4 show the best recognition score (% recognition) for the five emotions and the parameter values (cf. Table 3) for the best scores.

Table 4. Best recognition scores for the five emotions and their corresponding parameter values used in the resynthesis. A value with ** or *** indicates that a 4-way repeated measures ANOVA found the effect of that parameter to be significant at the level of $p < 0.01$ or $p < 0.001$ for the corresponding emotion.

Emotion	Best score (%)	Formant shift (df = 3,42)	Pitch median (df = 3,42)	Pitch range (df = 3,42)	Duration (df = 1,14)
Happy	73.3	1.20 ***	200	4.00 ***	0.9 ***
Depressed	71.1	0.96	100 ***	0.10 ***	1.1 ***
Grief-stricken	60.0	0.83 ***	400 **	0.34 ***	1.1 **
Scared	48.9	0.83 **	400 ***	1.17	0.9
Angry	48.9	0.83	50 ***	0.10	1.1

As can be seen in Table 4, all the five emotions had recognition rates well above

chance (~~25~~%). And the rates are especially high for happy, depressed and grief-stricken. For happiness, formant shift, pitch range and duration all had highly significant effects, and the direction of the parameter values are consistent with the predictions of BID: Small body size (large spectral dispersion) and high dynamicity (large pitch range and fast speech rate). The pitch median for the best score, 200 Hz, is not very high, but when combined with a large pitch range, the peak F_0 is as high as 700 Hz.

At its best recognition rate (71.1%), depressed is associated with low median pitch, very small pitch range and slow speech rate. These correspond well with previous findings about sadness as discussed earlier. It also agrees well with our earlier prediction that the typically reported sadness is of the depressed type.

Grief-stricken, which was also well recognized (60.0%), was associated with very low spectral dispersion, very high median pitch, small pitch range and slow speech rate. Only the last two parameter values are similar to those of depressed. The high pitch median is diametrically opposed to that of depressed, but consistent with findings of studies that specifically examined the grieving type of sadness (Costanzo et al., 1969; Erickson et al., 2006).

Scared had somewhat lower recognition rate (48.9%) than the three emotions just mentioned, although still well above chance. The most significant parameter for scared is a very high pitch median, which agrees well with the general findings discussed earlier. There was no effect of pitch range or speech rate, and this goes against our earlier hypothesis that fear is situated low on the dynamicity dimension. Perhaps the most surprising is the significant effect of formant shift which puts fear on the very low end of the size-projection dimension. This goes against not only our own hypothesis outlined earlier, but also Morton's (1977) grouping of fear with submission as the opposite of hostility/aggression. Mozziconcci (2001): Fear appeared to be confused quite frequently with indignation"

Finally, the perception pattern of angry is somewhat unexpected as it seems to agree with our own previous findings (Chuenwattanapranithi et al., 2008 and experiment 1) only in terms of low pitch median. And it also disagrees with the general finding that anger, especially hot anger, is usually associated with high pitch. A likely explanation can be found in the highly significant interaction between pitch median and pitch range ($F[9,126] = 4.064$, $p < 0.001$) as shown in Figure 3. The scores for all the pitch ranges are relatively high when pitch median is 50, but the trend is in favour of the smallest pitch range. But with other pitch medians, is the highest pitch range that is the most favoured. When listening to the stimuli ourselves, we noticed that when pitch median was 50 Hz, the voice sounded very rough, with clearly audible glottalizations, especially when the pitch range was small. It seems that such rough voice quality is associated by the listeners to anger, and this agrees with the finding of Gobl & Ní Chasaidi (2003), and with Morton's (1977) hypothesis that hostile vocalizations tend to have rough sound quality. Also unexpected is that formant shift had no significant effect on anger perception. This could have been due to the fact that listeners were biased by the rough voice quality at the 50 Hz pitch median and allocated most of the anger responses to those sentences.

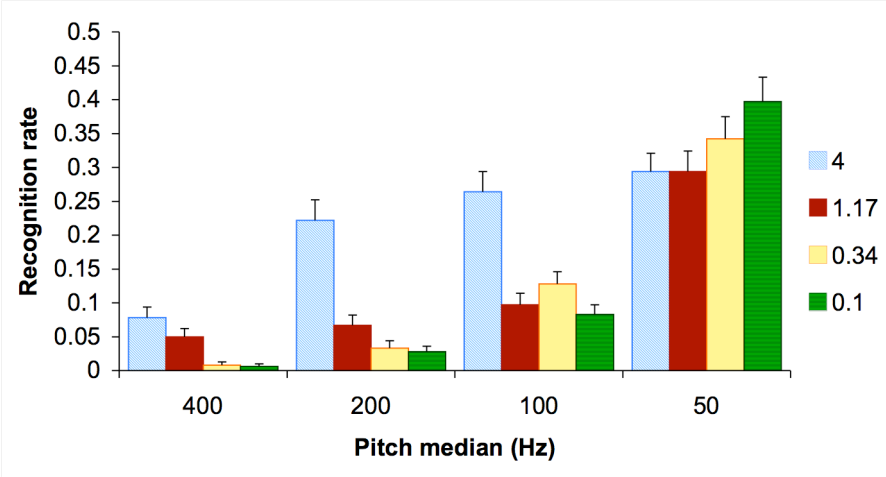

Figure 3. Recognition rate of anger as a function of pitch median and pitch range.

3.2.4. Findings of Experiment 2 and further implications

The results of experiment 2 have provided support for some of the predictions based on BID, but they have also suggested a need to change the assumptions behind some of the predictions. That happiness is located toward the small-size end of the size-projection dimension and high end of dynamicity dimension is clearly supported. The separation of sadness into two rather different types of expressions is also well supported, with the depressed type corresponding to the most commonly reported sadness, but grief-stricken to a rather different type. As discussed in 2.3, the question is whether this expression is to *beg* or *demand* for sympathy. The lengthened vocal tract (small formant shift ratio) suggests that it is more for demanding than for begging. Perhaps the biggest surprise is the finding that fear vocalization, too, is associated with lengthened vocal tract high pitch. When combined with high pitch, it seems that the expression is sending a mixed signal: I may be small (high pitch), but I am willing to fight (long vocal tract). This is further supported by the high dynamicity indicated by relatively large pitch range (1.17). Thus there seems to be a need to revise the grouping together of fear and submissiveness as suggested by Morton (1977). While submissive expression probably indeed signals total surrender, a fear expression still signals a demand for the aggressor to back off. This seems to make evolutionary sense, because a total surrender to a predator can only mean one thing: to be eaten.

The results of anger perception suggest that the use of 50 Hz as the lowest pitch median may have inadvertently introduced a confound: voice quality, which in this experiment is supposed to remain constant. But this finding may actually point to a critical role of voice quality as suggested by Morton (1977). Further exploration of BID therefore should try to control voice quality more directly.

The findings of this experiment also demonstrate more clearly than by previous investigations that emotional encoding is likely to be parallel to speech prosody. That is, given an emotionally neutral utterance with proper intonation, it is possible to make it sound “emotional” by imposing global (rather than local) modifications of certain acoustic parameters along the bio-informational dimensions. In doing so the

normal intonation carrying "linguistic" meanings remains largely intact although the emotional information has been added. In the following discussion, we will take a closer look at the relationship between emotion and prosody.

4. Parallel encoding of emotional and linguistic information

Much of the research effort on vocal expression of emotions to date has been devoted to searching for characteristic affective prosody, in particular, emotional intonation, but little clear evidence has been demonstrated so far (Hirschberg, 2002). As concluded by Scherer & Bänziger (2004:365) based on results of a systematic production study, "there is little evidence for emotion-specific intonation contours." To understand this difficulty, we need to ask a question that is not often considered. That is, suppose that emotion is indeed encoded through prosody, what happens to the prosodic patterns that carry linguistic information? Does the emotional prosody replace linguistic prosody, or does it largely leave it intact? Scherer, Ladd and Silverman (1984) and Ladd et al. (1985) made an attempt to ask a similar question by contrasting "configuration" and "covariance" as two alternative strategies of producing emotional prosody, and their conclusion was that both strategies are used. Our experiment 2 discussed above seems to have provided a more direct answer to this question. The base sentence in the experiment was a natural human utterance with prosodic focus on the word "owe", as shown in the top panel of Figure 4. Consistent with empirical findings about focus in English (Cooper, Eady and Mueller, 1985; Xu & Xu, 2005), the pitch range of the focused word is expanded (raised in this case because it is a statement rather than a question, cf. Liu & Xu, 2007), but the post-focus pitch range is compressed. Through global manipulation of pitch range and pitch median, dramatic changes in pitch can be introduced into the F_0 contours, as can be seen in the rest of the panels in Figure 4, which show the spectrogram and pitch tracks of the sentences resynthesized with the parameter combinations (Table 4) that had the best perceptual scores for each of the emotions studies in Experiment 2. However, it can be also seen that the effect of these manipulations is to exaggerate, compress or vertically shift the original contours without eradicating them. The same is true, of course, for what happens to the original spectral movements when global spectral manipulations are applied, although the integrity of the formant patterns are not as obviously seen as that of F_0 contours.

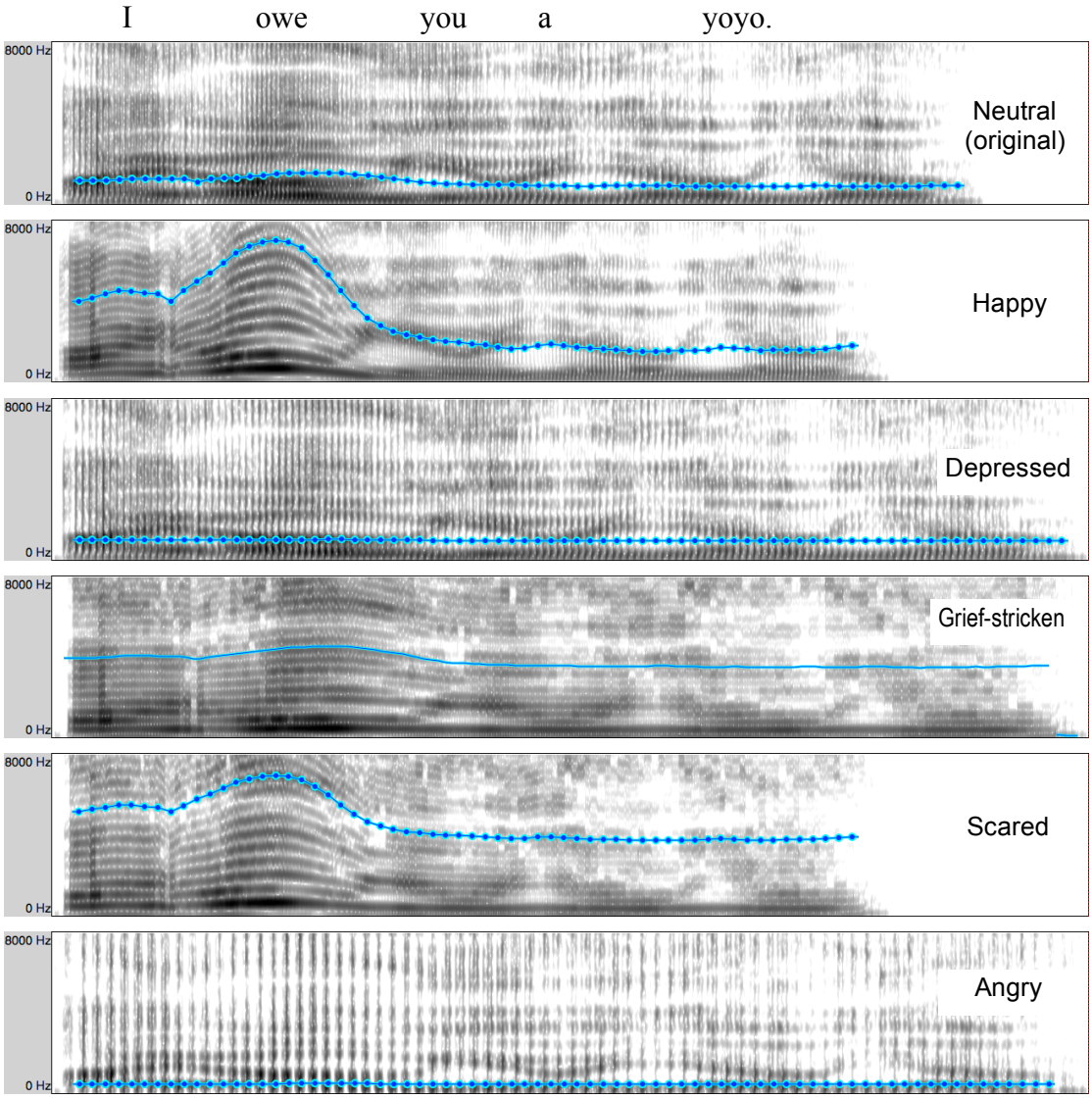

Figure 4. Spectrograms and F_0 tracks of the original and resynthesized speech utterance "I owe you a yoyo" that had the best recognition scores for each of the five emotions. In all the renditions the prosodic focus is on the word "owe".

Such global manipulations of F_0 and spectral properties without obliterating the linguistic intonational components is consistent with the conceptualization of the Parallel Encoding and Target Approximation model (PENTA) for speech prosody as shown in Figure 5 (Xu, 2005). PENTA assumes that various linguistic as well as para-linguistic functions are encoded in parallel, each with a unique encoding scheme that specifies the control parameters of the articulatory process of target approximation (TA). The encoding schemes differ from each other not only in terms of individual parameters, but also in terms of the temporal scope of application, and such temporal scope is determined by the nature of the function. For example, for lexical tones and lexical stress, the parameter specifications are largely local to individual syllables, as has been successfully tested in Prom-on, Xu and Thipakorn (2009). For focus, the scope is divided into pre-focus (if any), on-focus and post-focus (if any) regions, as

also has been successfully tested (Prom-on et al., 2009). For sentence type (statement vs. question), the scope is likely to be nearly the entire sentence, i.e., excluding the initial unstressed syllables (Thorsen, 1980; Liu & Xu, 2005, 2007). Functions with non-local encoding schemes sometimes also change the local parameters, e.g., changing the [high] target of stressed syllables in English to [rise] in question intonation (Liu & Xu, 2007), but more often than not the local targets are left intact in terms of target slope, and relative target height (Liu & Xu, 2005). The results of both experiments presented in the present paper, especially those of Experiment 2, suggest that the temporal scope of parameter control for encoding emotional meanings is likely to be at least as broad as an entire utterance.

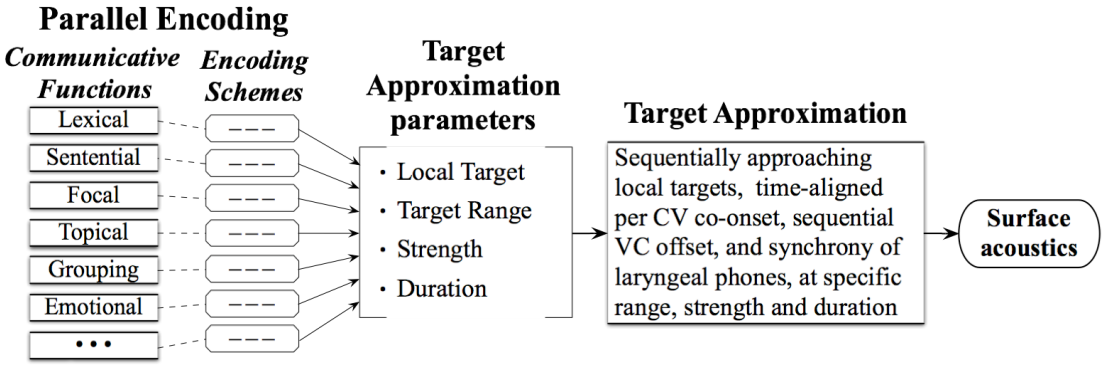

Figure 5. A sketch of the PENTA model. Modified from Xu (2005).

The PENTA account of emotional expressions is also in contrast with some of the previous attempts at demonstrating the use of the size projection principle (Morton, 1977; Ohala, 1984) or other potential "biological codes" (Gussenhoven, 2002). Many of these accounts focus on the grammaticalization of these codes, e.g., the morphological use of high-front vowels to denote smallness and low-back vowels to denote largeness, or the intonational use of the final rise to indicate questions (Gussenhoven, 2002; Fitch, 1994; Ohala, 1997). However, the correlation between vowel shape and size information can hardly be said to be very high, as exceptions are easily found in any language. As for the rising question intonation, a recent study has demonstrated that it is entirely missing in a group of languages located in the Sudanic belt of Africa, which use instead "lax prosody" (consisting of lengthening and/or breathy voice) to indicate the interrogative meaning (Rialland, 2009). Thus there may not be universal or pervasive grammaticalization of the biological codes. In contrast, it is very likely that the bio-informational dimensions outlined earlier are used in encoding emotional meanings *all the time by all the languages*, and the cases of grammaticalizations are just occasional byproducts of the constant use of these dimensions.

5. Conclusions

From an evolutionary point of view, emotion, like everything else about ourselves, is the result of adaptation to our ancestral conditions in the long past (Tooby & Cosmides, 1990). Affective expressions as well as corresponding internal neurophysiological states, as correlates of emotion, are both part of such adaptation. It

is therefore unlikely that we can explain emotional expressions on the basis of our internal feelings, as most current approaches have been trying to do. In this paper we have explored, instead, the idea that emotional expressions are evolutionarily designed to elicit behaviours that may benefit the emotion bearer. Extending the work of Morton (1977) and Ohala (1984), we have proposed that emotional meanings are encoded along a set of behaviour-eliciting *bio-informational dimensions* (BID), which involve both segmental and prosodic aspects of the vocal signal. Initial evidence for BID can be seen not only scattered around in the literature, but also in data from two new experiments that we have presented. The new data, especially those of experiment 2, also demonstrate that there need to be changes in the previous assumptions about certain emotions, such as grouping fear with submission (Morton, 1977), treating sadness as a single emotion, or what exactly a grief-stricken expression means. Such findings, though still preliminary, demonstrate the potential effectiveness of an evolution-based and theory-driven approach. Finally, we have explored how BID can be seamlessly incorporated into the PENTA model of speech prosody. Within such a model, emotional meanings can be encoded in parallel with non-emotional meanings, rather than forming their own autonomous prosody as often assumed previously.

References

Auberge, V. and Cathiard, M. (2003). Can we hear the prosody of smile. *Speech Communication* **40**: 87-97.

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

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10: 341-345.

Borod, J. C. (1993). Emotion and the brain -- Anatomy and theory: An introduction to the special section. *Neuropsychology* 7: 427-432.

Burkhardt, F., and Sendlmeier, W. F. (2000). "Verification of acoustical correlates of emotional speech using formant-synthesis," in *ISCA Workshop on Speech and Emotion: A conceptual framework for research* (Belfast).

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B. and Maneewongvatana, S. (2008). Encoding emotions in speech with the size code — A perceptual investigation. *Phonetica* 65: 210-230.

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour* **60**: 773–780.

Cooper, W. E., Eady, S. J. and Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America* 77: 2142-2156.

Costanzo, F. S., Markel, N. N. and Costanzo, P. R. (1969). Voice quality profile and perceived emotion. *Journal of Counseling Psychology* 16(3): 267-270.

Cowie, R. and Cornelius, R. R. (2003). Describing the emotional states that are expressed

in speech. *Speech Communication* 40: 5-32.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. London, England: John Murray.

Drahota, A., Costall, A. and Reddy, V. (2008). The vocal communication of different kinds of smile. *Speech Communication* 50(4): 278-287.

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

Ekman, P. (1997). Should we call it expression or communication? *Innovations in Social Science Research* 10: 333-344.

Ekman, P. (1998). Universality of emotional expression? A personal history of the dispute. In *Third Edition of Charles Darwin's The Expression of The Emotions in Man and Animals, with introduction, afterwords, and commentaries*. P. Ekman. London: HarperCollins pp. 363-393.

Ekman, P. (1999). Basic Emotions. In *The Handbook of Cognition and Emotion*. T. Dalgleish and T. Power. Sussex, U.K: John Wiley & Sons, Ltd. pp. 45-60.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M. and Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology* 53(4): 712-717.

Erickson, D., Yoshida, K., Menezes, C., Fujino, A., Mochida, T. and Shibuya, Y. (2006). Exploratory Study of Some Acoustic and Articulatory Characteristics of Sad Speech. *Phonetica* 63: 1-25.

Fitch, W. T. (1994). *Vocal tract length perception and the evolution of language*. Ph. D. Dissertation, Brown University.

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *Journal of the Acoustical Society of America* 102: 1213-1222.

Fitch, W. T. (1999). Acoustic exaggeration of size in birds by tracheal elongation: Comparative and theoretical analyses. *Journal of Zoology (London)* 248: 31-49.

Fónagy, I. (1978). A New Method of Investigating the Perception of Prosodic Features. *Language and Speech* 21: 34-49.

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

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

González, J. (2004). Formant frequencies and body size of speaker: a weak relationship in adult humans. *Journal of Phonetics* 32: 277–287.

Gussenhoven, C. (2002). Intonation and interpretation: Phonetics and Phonology. In *Proceedings of The 1st International Conference on Speech Prosody*, Aix-en-Provence, France: 47-57.

Gussenhoven, C. (2007). Types of focus in English. In *Topic and Focus: Cross-linguistic*

Perspectives on Meaning and Intonation. C. Lee, M. Gordon and D. Büring. New York: Springer pp. 83-100.

Hamilton, W. D. (1964). The genetical evolution of social behaviour I & II. *Journal of Theoretical Biology* **7**: 1-52

Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication* **36**: 31-43.

Huckvale, M. (2008). SFS Speech Filing System 4.7, http://www.phon.ucl.ac.uk/resource/sfs/, University College London.

Ives, D. T., Smith, D. R. R. and Patterson, R. D. (2005). Discrimination of speaker size from syllable phrases. *Journal of the Acoustical Society of America* 118: 3816-3822.

Juslin, P. N. and Laukka, P. (2001). Impact of Intended Emotion Intensity on Cue Utilization and Decoding Accuracy in Vocal Expression of Emotion. *Emotion* **1**: 381-412.

Kwon, O. W., Chan, K., Hao, J. and Lee, T. W., . (2003). Emotion Recognition by Speech Signals. In *Proceedings of Eurospeech*, Geneva, Switzerland**:** 125-128.

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G. and Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *Journal of the Acoustical Society of America* 78: 435-444.

Liu, F. and Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica* **62**: 70-87.

Liu, F. and Xu, Y. (2007). Question intonation as affected by word stress and focus in English. In *Proceedings of The 16th International Congress of Phonetic Sciences*, Saarbrücken: 1189-1192.

Mauss, I. B. and Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion* **23**(2): 209-237.

Morton, E. W. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist* 111: 855-869.

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

Mozziconacci, S. (2002). Prosody and Emotions. In *Proceedings of The 1st International Conference on Speech Prosody*, Aix-en-Provence, France**:** 1-9.

Murray, I. R. and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America* 93: 1097-1108.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica* 41: 1-16.

Ohala, J. J. (1997). Sound symbolism. In *Proceedings of 4th Seoul International Conference on Linguistics*: 98-103.

Prom-on, S., Xu, Y. and Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America* 125: 405-424.

Protopapas, A., and Lieberman, P. (1997). "Fundamental frequency of phonation and perceived emotional stress," *Journal of the Acoustical Society of America* 101, 2267-2277.

Reby, D. and McComb, K. (2003). Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags. *Animal Behaviour* 65: 519-530.

Rialland, A. (2009). African "lax" question prosody: its realisations and its geographical distribution, *Lingua* 119, 928-949.

Robson, J. and MackenzieBeck., J. (1999). Hearing Smiles - Perceptual, Acoustic And Production Aspects Of Labial Spreading. In *Proceedings of The 14th International Conference of Phonetic Sciences*, San Francisco: 219-222.

Russell, J. A., Bachorowski, J.-A. and Fernández-Dols, J.-M. (2003). Facial and Vocal Expressions of Emotion. *Annual Review of Psychology* 54(1): 329-349.

Sauter, D., Eisner, F., Ekman, P. and Scott, S. K. (2009). Universal vocal signals of emotion. In *Proceedings of The 31st Annual Meeting of the Cognitive Science Society*, Amsterdam, The Netherlands

Scherer, K. R. (1979). Nonlinguistic vocal indicators of emotion and psychopathology. In *Emotions in personality and psychopathology*. C. E. Izard. New York: Plenum Press pp. 493-529.

Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin* 99: 143-165.

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

Scherer, K. R. and Bänziger, T. (2004). Emotional expression in prosody: a review and an agenda for future research. In *Proceedings of Speech Prosody 2004*: 359-366.

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

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review* 61(2): 81-88.

Shami, M. and Verhelst, W. (2007). An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication* 49: 201-212.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H. and Irino, T. (2005). The processing and perception of size information in speech sounds. *Journal of the Acoustical Society of America* 117: 305-318.

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA: The MIT Press.

Sulc, J. (1977). To the problem of emotional changes in the human voice. *Activitas Nervosa Superior* 19: 215-216.

Susskind, J. M., Lee, D. H., Cusi, A., Feiman, R., Grabski, W. and Anderson, A. K. (2008). Expressing fear enhances sensory acquisition. *Nat Neurosci* 11(7): 843-

850.

Tartter, V. C. and Braun, D. (1994). Hearing smiles and frowns in normal and whisper registers. *Journal of the Acoustical Society of America* **96**: 2101-2107.

Thorsen, N. G. (1980). A study of the perception of sentence intonation — Evidence from Danish. *Journal of the Acoustical Society of America* 67: 1014-1030.

Tooby, J. and Cosmides, L. (1990). The past explains the present : Emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology* **11**(4-5): 375-424.

Turner, R. E. and Patterson, R. D. (2003). An analysis of the size information in classical formant data: Peterson and Barney (1952) revisited. *Journal of the Acoustical Society of Japan* **33**: 585–589.

van Dommelen, W. A. and Moxness, B. H. (1995). Acoustic Parameters in Speaker Height and Weight Identification: Sex-Specific Behaviour. *Language and Speech* **38**: 267-287.

Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication* 48: 1162-1181.

Williams, C. E. and Stevens, K. N. (1972). Emotion and speech: Some acoustical correlates. *Journal of the Acoustical Society of America* **52**: 1238-1250.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication* 46: 220-251.

Xu, Y. and Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics* **33**: 159-197.

Zei Pollermann, B. (2002). A Place for Prosody in a Unified Model of Cognition and Emotion. In *Proceedings of The 1st International Conference on Speech Prosody*, Aix-en-Provence, France: 17-22.

---

[1] In Chinese and Japanese, the character 悲 is used to refer to the grieving type of sadness. In Chinese there are also synonyms like 伤心, 悲伤, 悲痛, all referring to grieving sadness.