DRAFT COPY ONLY. DO NOT QUOTE

The effect of manipulated appraisals on voice acoustics

Tom Johnstone1[1]
Carien van Reekum[1]
Klaus Scherer2[2]

[1] Psychology Department, University of Wisconsin-Madison

[2] Psychology Department, University of Geneva

Abstract

This study was designed to test two opposing accounts of how emotion is encoded in the voice. A simple arousal model of emotional speech holds that the changes to speech under different emotional conditions reflects a single arousal response dimension, irrespective of other aspects of emotion such as valence or potency. Such a theory is supported by most of the previous research on real or induced (as opposed to acted) emotional speech. The opposing view, the most elaborate description of which was given by Scherer (1986), is that emotions effect the acoustic characteristics of speech along a number of dimensions, not only arousal, and that these dimensional responses are the result of emotion antecedent appraisal. In contrast to most previous research, an attempt was made in this study to induce real emotional states in the laboratory, using computer game manipulations of the appraisal dimensions intrinsic pleasantness and goal conduciveness. A set of acoustic parameters selected to capture temporal, fundamental frequency, intensity and spectral vocal characteristics of the voice was extracted from speech recordings. The results indicate that a single arousal dimension cannot adequately describe the vocal changes specific to different emotions, and lend weight to a theory of multidimensional emotional response patterning as suggested by Scherer and others. Specifically, while mean energy, F0 level and speech fluency varied with goal obstructiveness, spectral measures depended upon manipulations of intrinsic pleasantness, and pitch dynamics depended upon the interaction between the two appraisal dimensions. Although the results for intrinsic pleasantness are consistent with Scherer's predictions, the effects of goal conduciveness on the voice, however, suggest a rethink of Scherer’s predictions for this appraisal dimension.

Introduction

Anyone who has felt their voice quaver and their throat constrict while nervously giving a public speech is well aware of the effects emotions can have on our speech. Often these effects are only obvious to the speaker, although sometimes the outward signs of this anxiety can be all too painfully evident to empathetic listeners, who pick up on the high-pitched, trembling voice of an anxious presenter. What, then, is known about how our emotions effect the way we produce speech? The answer is: relatively little, particularly compared with other forms of emotional expression, such as facial expression which has been the subject of close scrutiny since the pioneering work by Ekman in the late sixties and early seventies (e.g. Ekman, 1972).

Studies of emotional expression in the voice have been both fewer and less conclusive than those of facial expression. Reviews of the literature have concluded that at least for acted vocal expressions of emotion, recognition rates are comparable, though slightly lower, than for facial expressions and are also recognised extremely well across cultures (Scherer, 1989; Banse and Scherer, 1996; Johnstone and Scherer, 2000; van Bezooijen, 1984). The latter result indicates that the vocal expression of emotion reflects mechanisms that function largely independently of the mechanisms for production of a given spoken language. Only a small, albeit growing number of studies have tried to identify the emotion-specific vocal characteristics that are presumably used by listeners to infer an expressed emotion, most of them also using acted speech. Reviews of such studies (Scherer, 1986; Johnstone and Scherer, 2000) have concluded that while consistent differences in the acoustical patterns across emotions exist, they seem to indicate a single dimension of physiological arousal. Thus emotions such as anger, fear and joy were all characterised by raised fundamental frequency (F0) and high intensity, while emotions such as sadness and boredom were expressed with low F0 and low intensity. Reliable acoustic parameters that could differentiate

between two emotions of similar arousal seemed, on the basis of the empirical evidence, not to exist. Remarking, however, that the ability of judges to accurately judge expressed emotions, including those with similar arousal levels, meant that such parameters must exist, the authors suggested that a broader set of acoustic parameters would need to be analysed in future research. Supporting this claim, Johnstone and Scherer cited some of the more more recent studies (Banse and Scherer, 1996; xxrefs to others) in which a more thorough acoustic analysis of acted speech allowed better differentiation of emotions with similar arousal levels.

A question remains, however, about the use of acted speech in most of the studies of vocal emotion expression that have so far been performed. Scherer (1985) has argued that the expression of emotion in speech reflects the effects of two distinct influences on the voice. The most basic is the direct effect of an emotional response on vocal production, termed a *push effect* by Scherer. This effect might be due to interruption of cognitive processes involved in speech planning and production, perturbation of the physiological systems that not only underlie speech production but also serve to maintain the body in an optimal state for survival (e.g. the skeletal musculature, the lungs), or the activation of specialised neuromotor programs for the vocal expression of emotion. Push effects are expected to be largely involuntary. *Pull effects*, on the other hand, are those external influences that can shape vocal production independently of the actual emotional state of the speaker. Such effects include accepted sociocultural speaking styles and vocal display rules. In contrast to push effects, pull effects are expected to be largely under voluntary control, used strategically with the express purpose of sending a signal to cohorts. Such is the case with acted speech. Even those acting techniques (e.g. the Stanislovski technique) that seek to create an emotion as the basis for portraying that emotion reflect both push and pull effects. Thus the presence of emotion-specific acoustic patterns in acted speech, as have been found in previous research, might well be attributable to the adoption by the actors of culturally defined,

strategic speaking styles. The question of whether the push effects of emotions on the voice produce well defined, emotion-specific changes to acoustic patterns remains unaddressed. This is the central question of this study.

In order to examine the push effects of emotion on speech, this study adopted the approach of inducing emotional responses in the laboratory using computer game induction techniques. Since efforts to induce emotional physiological responses using passive techniques such as film watching have met with limited success in the past, an active, emotion-inducing computer game, designed to be more involving for the participants, was developed. Standard sentences, designed to be comparable across emotion conditions, were elicited from participants and recorded using high quality recording equipment.

The aims of this experiment were thus both methodological and theoretical. The methodological aim of this experiment was to develop and test the technique of studying the push effects of emotional response on the characteristics of the voice by recording the speech of players of an experimentally manipulated computer game. The principal theoretical aim was to gather evidence on whether the push effects of emotion on the voice are limited to a simple effect of arousal on the speech subsystems, or whether the effects are multidimensional and reflect factors other than arousal. The experiment also aimed to test the predictions of Scherer (1986) of the vocal characteristics resulting from appraisals along the intrinsic pleasantness and goal conduciveness dimensions.

Push effects of emotion on the voice. There is little empirical evidence that the voice changes produced by emotional responses reflect anything other than an increase or decrease in general arousal. The lack of such evidence does not necessarily exclude the existence of such non-arousal effects of emotion on the voice however, since the few studies that have been performed on real emotional speech have not examined a wide range of acoustic vocal characteristics, nor have they examined more than one or two different emotions.

In keeping with the experimental approach outlined above, this experiment aimed at measuring the acoustical changes provoked by emotional responses to computer game situations designed to be appraised in specific ways. As this was a first attempt at using such an experimental paradigm, appraisal dimensions were chosen on the basis of theoretical predictions (Scherer, 1986) that they would produce non-arousal changes to the voice as well as the ease with which they could be instantiated within the game. Scherer (1986) suggests that it is likely that the vocal characteristics of emotional speech can be quantified along the three classical emotion response dimensions: activation (or arousal), valence and potency. The little empirical evidence for the existence of three such dimensions in emotional speech (Green and Cliff, 1975) indicates that of the three, activation and valence are more easily identifiable than the potency dimension. For this reason, this experiment focussed on the valence response dimension which, according to Scherer (1986), is affected by appraisals of intrinsic pleasantness and goal conduciveness.

Arguing that appraisals of intrinsic pleasantness or unpleasantness will elicit either innate or learned approach or avoidance action tendencies respectively, Scherer draws upon past research into orofacial behaviour, particularly in response to pleasant or noxious tastes or odours. Thus unpleasant stimuli are predicted to elicit constriction of the pharynx and approximation of the faucal arches, which serve to rejecting noxious stimuli. Combining these vocal tract changes with Laver's (1980) work on voice quality, Scherer predicts strong resonance in high frequencies following appraisals of unpleasantness. In contrast, stimuli appraised as pleasant are predicted by Scherer to lead to faucal and pharyngeal expansion, which should lead to a damping of high frequency energy. Both pleasant and unpleasant appraisals are also predicted by Scherer to produce facial expressions such as the retraction of the corners of the mouth in smiles and expressions of disgust, although it is unclear how these two types of expression would differently affect vocal characteristics.

Scherer (1986) makes a theoretical distinction between appraisals of intrinsic pleasantness, which has to do with innate, or highly learned evaluation of pleasantness, and goal conduciveness, which involves an evaluation of whether a stimulus or event helps or hinders one to obtain a desired goal or need. An example of this distinction might be when one is offered a medicine that tastes terrible but will cure a bad illness. According to Scherer, the medicine will be appraised as intrinsically unpleasant but goal conducive. The opposite case might be a smoker trying desperately to give up smoking, who is offered a cigarette, an event that would be appraised as intrinsically pleasant, but goal obstructive. Despite this theoretical distinction, Scherer predicts that many of the effects of goal conduciveness appraisals on the voice will be similar to those of intrinsic pleasantness appraisals, since the neurophysiological pathways that mediate expressive responses to both appraisals are likely to be the same. Hence goal obstructive events are predicted to produce constricted pharyngeal settings and approximated faucal pillars, leading to fairly high energy at high frequencies, as opposed to less high frequency energy corresponding to wide pharyngeal and faucal settings that result from goal conducive appraisals.

Scherer's (1986) predictions of vocal changes corresponding to a hedonic valence response dimension are in direct contrast to a simple arousal theory of emotional vocal expressions. Scherer predicts that F0 level and mean overall energy of speech will not change as a function of appraisals of goal conduciveness or intrinsic pleasantness. An arousal theory of emotional expression in the voice would predict at least F0 level and mean energy to increase with increasing arousal. Indeed, studies that have been carried out on both real and acted emotional speech have almost always measured an increase in F0 and energy for anger, joy and fear and a decrease for boredom and sadness and attributed it to changes in physiological arousal. Other acoustic variables, such as the spectral distribution of energy,

might also be expected to change according to an arousal theory, but they would be expected to covary with F0 and energy in a consistent manner.

In summary, by measuring the acoustical properties of speech immediately following game events that are designed to elicit appraisals of intrinsic pleasantness and goal conduciveness, this experiment sought to test the hypothesis that such events provoke changes to the voice that are inconsistent with a simple arousal model of emotional vocal expression. To this end, changes were predicted to the spectral distribution of energy of elicited speech due to the different experimental conditions that would not covary with any changes observed in speech F0 or speech energy. In addition, the specific predictions of Scherer of more high frequency energy under conditions of negative hedonic valence (unpleasant or obstructive) than under conditions of positive hedonic valence (pleasant or conducive), and no valence-dependencies for F0, were also tested.

Method

Participants. Thirty-three volunteer adolescents between the ages of 13 and 15 (27 males and 6 females) were recruited from nearby high schools. Both the schools and parents of all children gave full written consent for the participation of their children based upon a full explanation of the aims and procedure of the experiment, who were reimbursed SFr.15 for their participation in the experiment. Adolescents were chosen because they were considered most likely to be familiar with, and get involved in video games, thus making it more likely that emotional responses to the video games would be elicited.

Description of the game. The game XQuest situates the player in a fictional galaxy, filled with crystals, mines and enemies. The general assignment is to gather the crystals which are present in each galaxy. Acceleration of the player's space ship is controlled by moving the mouse, thus leading to a ballistic movement of the ship. This feature makes the game particularly interesting to play. Pressing the left button of the mouse launches bullets in

the direction in which the ship is going, which destroy any enemies which are hit. Pressing the right button launches a bomb, if available, which destroys every enemy and mine in the galaxy. Once the player has picked up all the crystals in the galaxy, a gate opens through which the player proceeds to the next galaxy (or game level). Points are awarded for every crystal gathered, completion of a game level within a certain time range, and every enemy destroyed. Depending on the amount of points gained, extra ships are given. The difficulty increases in successive game levels since there are more crystals to pick up, the number of mines increases, the enemies become more numerous and difficult, and the exit to the next game level becomes smaller. The game ends when the player loses all the ships, after which the player starts a new game at the first game level.

Procedure. Upon their arrival in the laboratory, participants were fully informed of the nature of the computer game and how long they would be asked to play for. Participants were then given a demonstration of how to play the game, during which the experimenter played the game and explained the appearance of the different objects (e.g. mines, crystals, player's space ship, enemies), movement of the player's space ship, how to fire bullets and to use the bomb, how to pick up the crystals and how to exit the galaxy when all the crystals had been collected. Players were also shown the emotion rating screen and verbal report screen and how to use them properly. Then they played the game for a 20 minute practice session. During this time, they were given extra instruction and reminders whenever necessary. In order to ensure that all players were sufficiently involved in the game, and to establish a minimal performance level for all players, a selection criterion of reaching at least the fourth game level by the end of 20 minutes practice was used. All players met this criterion. After the practice session, sensors for physiological measures and the microphone were attached. The players were asked to speak aloud in a normal voice while the microphone recording level was adjusted. The experiment started with a 2.5 minute relaxation phase, to establish a

resting baseline. Then the game started and participants played for 45 minutes, after which the game halted automatically.

Manipulation of appraisal dimensions. Two appraisal dimensions were operationalised by either manipulating or selecting specific game events. From all possible events in the game which may elicit emotion-antecedent appraisal, two types of events were selected a-priori since it is highly likely that these events would be appraised in a similar way by all players. These events are loosing a ship (by hitting an obstacle or being shot by an enemy) and passing to the next game level after successful completion of a galaxy. In the context of the game, the first type of event is obstructive and the latter conducive in the pursuit of gaining points and progressing to as high a game level as possible. The independent variable goal conduciveness was thus operationalised by selecting situations in which the player's ship was destroyed (low goal conduciveness) or a game level was successfully completed (high goal conduciveness). The other appraisal dimension which was studied was the intrinsic pleasantness of critical game events. As opposed to goal conduciveness, this appraisal dimension was directly manipulated by playing valenced (i.e. pleasant and unpleasant) sounds. The pleasantness of these sounds, which were equal in duration and average intensity, had been established in an independent pre-test of 15 judges who were asked to rate the sounds on a seven point scale from -3 (very unpleasant) to +3 (very pleasant). The mean ratings are given in table 1.

Table 1 About here

The appraisal dimensions were manipulated in a 2 (intrinsic pleasantness) x 2 (goal conduciveness) within-subjects design. Thus concurrent with the player reaching the next level or losing a life, either a pleasant or an unpleasant sound was presented.

Vocal reports. Speech was elicited by means of a vocal report pop-up screen, which requested a vocal report of the immediately preceding game events. The report screen, which was designed to seem to the player like part of the game, rather than an intrusion on it, was displayed whenever an experiment-relevant event (i.e. loss of ship or new level) occurred, with the constraint that no more than one screen appeared every two minutes, so that the continuity of the game was not unduly interrupted (see figure 1). The players were requested to respond to the screen by pronouncing aloud the identification number, choosing one of the three given reasons for the preceding event, and estimating the percentage chance that they would be successful in the following game level. The pop-up screen provided both strings of isolated letters and connected phrases to be pronounced by the subject. For each presentation the identification number changed, but the first six characters remained constant across all presentations.

Figure 1 about here

Emotion self-report. An emotion self-report was obtained using a pop-up screen which displayed a popular French comic strip character (Gaston Lagaffe), expressing eight different emotions (interest, joy, surprise, anger, shame, pride, tenseness and helplessness; see Figure 2). The images of the characters, which were used to make the rating screen clearer and easier for the adolescents, were accompanied by the corresponding emotion-labels and a continuous graphic scale on which the felt intensity of each emotion could be indicated by means of clicking and dragging with the mouse. The ratings were converted to 100 decimal values ranging from 0 to 1. The pop-up screen was presented immediately after a random sample of critical game events, but not more often than once every four minutes, so that the continuity of the game was not unduly interrupted.

Figure 2 about here

Results

Emotion reports.

For each subject, each experimental condition contained on average two observations, the data for which were averaged. Mean reported emotion intensities are shown in figure 3. Since the data for all emotions were heavily skewed, and had distributions that could not be corrected with mathematical transformations, they were then analysed with a Friedman test (this is an appropriate non-parametric within-subject test of differences between mean ranks) to determine for each emotion if the reported intensities differed across the four experimental conditions. Reported joy ($\chi^2_{(3, 32)} = 10.0$, $p = 0.02$), pride ($\chi^2_{(3, 32)} = 29.9$, $p < .0005$), anger ($\chi^2_{(3, 32)} = 29.2$, $p < .0005$) and surprise ($\chi^2_{(3, 32)} = 12.3$, $p = 0.007$) all differed across the four conditions. For these emotions, posthoc two-way Wilcoxon signed ranks tests were used to compare the reported intensities between conducive and obstructive, and pleasant and unpleasant conditions respectively. Joy ($Z = 2.0$, $p = 0.004$) and Pride ($Z = 4.1$, $p < 0.0005$) were both higher in conducive conditions than in obstructive conditions. Anger ($Z = 4.0$, $p < 0.0005$) and Surprise ($Z = 3.1$, $p = 0.002$) were both higher in obstructive conditions than in conducive conditions. Surprise was also higher following games events accompanied by pleasant sounds than following events accompanied by unpleasant sounds ($Z = 2.4$, $p = 0.02$).

Figure 3 about here

Acoustic Analyses.

Sections of the DAT speech recordings corresponding to the prompted vocal reports were identified using timing data recorded in the participants' data log files. These sections were then digitised using 16-bit Kay Computer Speech Laboratory (CSL) 5300B speech analysis hardware and software at 20000 samples/second and stored as separate digital PC sound files. A number of acoustic analyses were then performed on each speech file, using CSL speech analysis software. These are detailed below for each type of analysis.

Fundamental frequency (F0). For each speech file, a three stage procedure was used to extract the F0 contour. The CSL software was first used to mark the onset of each pitch impulse. For each participant, all their speech files were analysed with a set minimum allowed F0 of 150 Hz and a set maximum allowed F0 of 400 Hz. These values are used by the CSL routine to limit the search for pitch impulse peaks in the speech waveform to within the expected range of F0 values for the adolescent participants. The positions of the impulse markers were then visually compared with the speech waveform, and obvious errors were manually corrected. For some participants for which there were many errors, due to F0 being either above the maximum or below the minimum allowed values, the minimum and maximum allowed F0 values were adjusted appropriately and all the participant's speech files were reanalysed and pitch impulses re-inspected. For all participants, a single adjustment of the minimum and maximum allowed F0 values was sufficient to ensure an accurate calculation of pitch impulse markers. Finally, the CSL software was used to calculate the F0 contour for each speech files based upon the pitch impulse markers.

From each calculated F0 contour, the following statistical measures of F0 were calculated: mean F0, standard deviation of F0, F0 5th percentile value and F0 95th percentile value. The two percentile values were calculated as measures of F0 floor and F0 ceiling respectively as reported in Banse and Scherer (1996).

Energy. The mean voiced energy of speech in each speech file was measured by calculating the root mean square (RMS) value of 15 millisecond frames of the speech signal, centred about each pitch impulse marker. The RMS value has advantages over other energy measures since it reflects more accurately the perceived intensity of speech (Deller, Proakis and Hansen, 1993). A 15 millisecond calculation window is long enough to ensure that the energy is averaged over 2-3 fundamental periods.

Duration and fluency measures. Using the calculated pitch impulse markers, an estimation was made of the length of each utterance, by measuring the time from the first pitch impulse marker to the last pitch impulse marker in each speech file. This technique is inaccurate in so far as it ignores unvoiced sounds at the beginning and end of each utterance, but was nevertheless used since such unvoiced sounds were not expected to vary greatly between experimental conditions (at least compared with the variation expected for voiced parts of the utterance), and since no CSL algorithm could be used for the automatic determination of the onset and offset of unvoiced sounds. A measure of the proportion of each speech utterance that was voiced was made by using the pitch impulse markers to estimate the voiced portions of each utterance and dividing the summed duration of these portions by the estimated total utterance duration. Both measures (i.e. utterance length and proportion voiced) were thus used as an indicator of speech fluency.

Spectral measures. The pitch impulse markers were used to separate each speech file into voiced and unvoiced parts. The average power spectrum of voiced parts of each utterance was then calculated using the CSL software, with a frame size of 512 samples. This thus yielded a power spectrum with 256 frequency bins, each one of width 39.06 hertz. The proportion of energy under 500 Hz, which is a measure of low frequency energy that has been found to vary with different emotions (Banse and Scherer, 1996), was calculated by summing all the frequency bins of the power spectrum below 500Hz, and dividing by the

sum of all the frequency bins across the entire spectral range. An equivalent calculation was made of the proportion of energy under 1000 Hz.

Statistical analyses.

To determine the effects of the experimental manipulations on the acoustic parameters, each parameter was separately analysed with a univariate mixed-model ANOVA, with conduciveness and pleasantness as two-level fixed factors, and participant as a 30-level random factor. A univariate mixed-model approach was chosen because a number of the acoustic variables measured are known to be highly interdependent, and the aim of these analyses was to identify all the parameters that varied across experimental conditions, rather than only those that contributed uniquely to a single composite dependent variate. Because of the high interdependence, no Bonferroni corrections were performed – rather, the intention was to use replication in following studies to verify the results of these analyses.

Pleasantness x Conduciveness interaction. There was very little interaction between the two independent variables. A weak interaction was measured for the F0 ceiling ($F(1,30)=2.9$, $p=0.10$). Post-hoc comparisons showed that this was due to F0 ceiling being higher in response to unpleasant than to pleasant sounds that accompanied obstructive events ($F(1,30)=4.4$, $p=0.04$), but the lack of such a difference for conducive events ($F(1,35)<1$). An interaction for F0 standard deviation ($F(1,30)=5.0$, $p=.03$) was also due to higher F0 standard deviation in response to unpleasant than to pleasant sounds that accompanied obstructive events ($F(1,32)=5.6$, $p=0.02$), but the lack of such a difference for conducive events ($F(1,33)<1$). These two interactions are illustrated in figure 4. No other pleasantness x conduciveness interactions were observed.

Figure 4 about here

Pleasantness. No effects of pleasantness on mean energy ($F(1,29)<1$), F0 floor ($F(1,29)<1$), F0 ceiling ($F(1,30)=2.1$, $p=0.15$), mean F0 ($F(1,29)<1$), F0 standard deviation ($F(1,30)=1.2$, $p=0.28$) were observed. The proportion of energy below 500 hertz was significantly lower for unpleasant than for pleasant sounds ($F(1,31)=7.3$, $p=0.01$). A similar result was found for the proportion of energy under 1000 hertz, which was also lower in response to unpleasant sounds than to pleasant sounds ($F(1,29)=4.2$, $p=0.05$).

Figure 5 about here

Conduciveness. The effects of conduciveness on mean energy and F0 floor are shown in figure 5. Mean energy was lower for conducive than for obstructive events ($F(1,29)=6.4$, $p=0.02$). F0 floor was lower for conducive events than for obstructive events ($F(1,30)=4.6$, $p=0.04$), although no effects of conduciveness on F0 ceiling ($F(1,30)<1$), mean F0 ($F(1,29)<1$) or F0 standard deviation ($F(1,30=1.2$, $p=0.27$) were measured. The effects of conduciveness on the fluency parameters is shown in figure 6. The percentage of each utterance that was voiced was lower for conducive than for obstructive events ($F(1, 30)=23.4$, $p<0.0001$). Utterance duration was higher for conducive events than for obstructive events ($F(1,30)=22.0$, $p<0.0001$). No significant differences due to the conduciveness manipulation were found for the proportion of energy under 500 hertz ($F(1,30)=1.8$, $p=0.19$) nor for the proportion of energy under 1000 hertz ($F(1,29)=1.3$, $p=0.27$).

Figure 6 about here

Interactions with participant. As expected, all of the acoustic parameters differed significantly across participants. More relevant to the current study is whether the ways in which acoustic parameters varied across experimental conditions were participant-dependent. In other words, how stable across participants were the effects of the experimental conditions? The interactions between the random participant factors and the two experimental factors can be used as an indicator of such a dependency. Weak conduciveness x participant ($F(29,28)=1.7$, $p=0.07$) and pleasantness x participant ($F(29,26)=1.8$, $p=0.06$) interactions were measured for the percentage of each utterance that was voiced. A significant pleasantness x participant interaction was observed for utterance duration ($F(29,26)=2.1$, $p=0.03$). No such interactions were measured for the other acoustic parameters.

Data reduction. As mentioned above, the set of acoustic parameters measured in this experiment are not indicators of distinct and independent voice production characteristics, nor even indicators of distinct characteristics of the acoustic signal. There is a large amount of overlap between certain parameters, for example the measures of F0, which is unavoidable since parameters that independently measure distinct acoustic properties are not yet well defined, particularly with respect to the acoustic properties of *emotional* speech. Because of this lack of independence, one can expect the acoustic parameters to be correlated to varying degrees. Table 2 lists the pairwise Pearson correlations between all the acoustic parameters measured in this experiment. The parameter values used for the calculation of the correlations were first zero-meaned for each participant, by fitting a linear model with an intercept and participant as the single predictor variable to the original acoustic parameters and taking the non-standardised residuals. This ensures that the correlations between parameters reflect within-subject variability (i.e. variability due to experimental manipulations) rather than between-subject variability. As can be seen, parameters that are primarily associated with the same specific aspect of the acoustic signal, such as the F0-related measures, tend to be highly

correlated. In addition, there are a number of substantial correlations between parameters that are based on different acoustic features, such as the correlations between the F0-related parameters and the measures of spectral energy distribution. It is possible that the latter correlations reflect underlying common characteristics of vocal production, which might be directly affected by emotional responses.

Table 2 about here

To address this possibility, a principal components analysis (PCA) was used to extract a small number of orthogonal factors that could be linked to specific voice production characteristics. Three, four and five factor solutions were calculated on the basis of the zero-meaned acoustic parameters, which accounted for 73%, 83% and 91% of the within-subject variance respectively. The factors were Quartimax rotated, in order to minimise the number of factors needed to explain the variables and to simplify interpretation. The rotated factor weightings for the three solutions were then examined to determine which of the solutions was most clearly interpretable with respect to current understanding of voice production and speech acoustics. Of the three solutions, the four factor solution provides the most parsimonious explanation of factors. Factor weightings for the four factor solution are shown in table 3.

Table 3 about here

The first factor loads most heavily on mean energy, F0 floor and mean F0. These three parameters are all indicators of physiological arousal or excitation, hence the first factor

would most appropriately be interpreted as an excitation factor. The second factor loads highly on F0 ceiling and F0 variability, both indicators of F0 variability and range, and would thus be appropriately labeled pitch dynamics. The third factor is most clearly an indicator of fluency, as indicated by a high loading on percentage voiced and a high negative loading on utterance duration. The fourth factor loads highly on the two measures of spectral energy distribution, which have most commonly been linked to vocal fold dynamics and the profile of the vocal tract.

The factor scores were entered into a univariate mixed-model ANOVA, with conduciveness and pleasantness as two-level fixed factors and participant as a 30-level random factor. Scores on the excitation factor were higher for obstructive events than for conducive events ($F(1,30)=4.1$, $p=0.05$). There was no difference in the excitation factor between pleasant and unpleasant events ($F(1,32)<1$), nor were there any significant interactions. For the pitch dynamics factor, no significant difference was observed between events paired with pleasant sounds and those paired with unpleasant sounds ($F(1,31)<1$), nor between conducive events and obstructive events ($F(1,30)<1$). A significant interaction between conduciveness and pleasantness ($F(1,30)=4.0$, $p=0.05$) indicated that the pitch was less dynamic following pleasant sounds than following unpleasant sounds when accompanied by obstructive events. The fluency factor scores were significantly higher for conducive than for obstructive events ($F(1,31)=27.6$, $p<0.000$). The fluency factor did not vary significantly across levels of pleasantness ($F(1,31)=1.9$, $p=0.18$), nor was the pleasantness x conduciveness interaction significant ($F(1,31)<1$), but there was a significant interaction between pleasantness and participant ($F(29,24)=2.3$, $p=0.02$), indicating that fluency varied with pleasantness in a participant-dependent manner. The spectral factor scores were significantly lower following unpleasant sounds than following pleasant sounds ($F(1,32)=6.9$, $p=0.01$), with no other significant main effects or interactions for this factor.

Discussion

The measured effects of the intrinsic pleasantness manipulation on the acoustic speech signal were limited to the distribution of energy in the spectrum, with a greater proportion of energy in higher frequencies being measured after unpleasant sounds than after pleasant sounds. This result is consistent with the predictions of Scherer (1986) which were based upon a presumed constriction of the faucal pillars and pharynx in response to the appraisal of a stimulus or event as intrinsically unpleasant. Scherer did not predict changes in speech intensity nor F0 due to pleasant or unpleasant appraisals, and indeed no main effects of pleasantness on energy or F0 parameters were measured in this experiment. The measured interaction between pleasantness and conduciveness for F0 ceiling and F0 standard deviation, indicating that for obstructive events only, F0 had a reduced dynamic for pleasant than for unpleasant sounds, was, however, unexpected and is difficult to explain. The predictions of Scherer have little to say about the interaction between different appraisal outcomes, other than that the predicted effects of two appraisal outcomes might reinforce or negate each other and thus produce a large or small change to speech acoustics accordingly.

Scherer (1986) predicted that the vocal changes caused by appraisals of goal conduciveness would parallel those of intrinsic pleasantness appraisals. Thus for events appraised as goal obstructive, a voice described as “narrow”, with more high frequency energy is expected, whereas for goal conducive events, a “wide” voice, with greater low frequency energy was predicted. The results of the current experiment do not support these predictions, with no significant spectral differences measured between conducive and obstructive events. Also contrary to the predictions of Scherer, conducive events were lower in energy and had a lower F0 level, as indicated by F0 floor, than obstructive events. These latter results, in combination to the higher scores on the excitation PCA factor for obstructive

events than for conducive events, suggest that physiological arousal was higher following the destruction of a ship than following the completion of a game level.

Such an arousal dimension is equivalent to the distinction used by Scherer to predict the difference between vocal changes in response to goal discrepant appraisals and those resulting from goal consistent appraisals. According to Scherer, appraisals of goal discrepancy should result in an *ergotropic* shift, whereby the sympathetic branch of the ANS is activated in preparation for probable action, leading to an overall tensing of the skeletal musculature, including the vocal folds, and increased depth of respiration. Such changes are expected to produce increases in F0 level and speech intensity. In situations appraised as goal consistent, a shift towards *ergotropic-trophotropic balance* is expected, in which sympathetic ANS activity is decreased and activity in the parasympathetic branch of the ANS is increased for the purposes of energy conservation and recovery, resulting in a "relaxed" voice with low to moderate F0 and intensity. Indeed, it is probable that the ship destroyed events were more discrepant with expectations than were the new level events, since when passing to a new level in XQuest, players see their ship approaching the exit gate, whereas the enemies that collide with the player's ship often behave in unpredictable ways. It is thus quite conceivable that the acoustic differences observed between ship destroyed and new level events were due to differences in sympathetic ANS activity produced by appraisals of discrepancy and consistency respectively. Support for this explanation comes from the reports of surprise, which were higher for obstructive events than for conducive events.

Unfortunately, it is not possible to say whether the observed differences between ship destroyed events and level completion events in this experiment were due to a failure on the part of Scherer's predictions, or a failure of the experiment to isolate the desired appraisal dimension. Although the two events were clearly objectively obstructive and conducive to the main goals of the game, it is not clear that they were appraised as such by participants, nor

that they were only appraised along that dimension, and not also appraised along other dimensions such as discrepancy.

The practical problems with asking a player to report their appraisals as they play a computer game include interrupting the flow of game play, as well as changing the very way in which game events are perceived and evaluated (and hence affecting their emotional responses). In addition, most appraisal theorists hold that appraisals often occur unconsciously (e.g. Leventhal & Scherer, 1987; see van Reekum and Scherer, 1997, for a discussion of levels of processing in appraisal) and are thus not accessible for reporting, a problem analogous to that of extracting problem solving techniques from experts in the domain of expert systems research (Barry, 1987). Hence even if subjective appraisal reports had been collected from players, they would not have provided a valid indication of the way players actually appraised events during the game. It would be desirable in future experiments to use a method of estimating participants' appraisals to events which provides a suitable compromise between validity and avoiding interference with the ongoing experiment. Moreover, an effort needs to be made to avoid unwittingly using experimental manipulations that lead to appraisals other than those intended.

Push effects as simple arousal. The hypothesis that push effects on the voice are limited to an arousal dimension was not supported by the results. The fact that manipulations of intrinsic pleasantness produced changes in certain acoustic parameters, while the conduciveness of events produced changes in completely different acoustic parameters, is very difficult to explain within a simple arousal model of emotional vocal response. Such a model would predict that those acoustic parameters affected by general physiological arousal would covary – that a change in the value of one such variable would be accompanied by a concomitant change in the values of the others, and that for a given pair of variables, the relative direction of change would be constant. Although it is possible that not all such

changes would be measurable, since some acoustic parameters might be more sensitive than others, a purely arousal-based model would hold that their relative sensitivity would be constant. In other words, if one experimental condition were to produce a change in parameter A but not in parameter B, one would not expect that in another experimental condition, a change would be observed in parameter B but not in parameter A. This is, however, what the results of this experiment reveal: That variations in the intrinsic pleasantness of an event cause changes to spectral energy distribution, but not to overall energy, F0 level nor fluency, but that changes to the conduciveness of an event produce changes to the latter set of variables, but not to spectral energy distribution. Although a single-dimension arousal model could be modified to fit such data, a more parsimonious explanation is that emotional changes to the voice reflect two or more dimensions, presumably reflecting two or more underlying mechanisms. This does not come as a surprise, given the evidence and theoretical justification that exists for the existence of at least three dimensions that seem to characterise emotional responses in general: activation, valence and potency. Scherer (1984) has suggested that the three dimensions can be useful in describing the structure of emotional response, although they are of more limited use in explaining the elicitation of emotion. In making his predictions of the acoustic characteristics of speech corresponding to different emotions, Scherer uses such a three-dimensional categorisation, which he proposes corresponds to three dimensions of voice type. Thus Scherer proposes that hedonic valence corresponds to variation in the voice from “wide” to “narrow”, activation corresponds to vocal variation from “lax”, through “relaxed” to “tense, and that potency effects voice along a “thin” to “full” dimension. The different vocal types, which are largely based upon the work of Laver (1980) on vocal settings and voice quality, are also described by Scherer in purely acoustic terms. Despite the widespread acceptance of a three dimensional description of emotional responses, there is, however, very little empirical

evidence supporting such a view with respect to push effects on the voice. Green and Cliff (1975) arrived at a three dimensional description of vocal changes in acted emotional speech, labeling the factors “pleasant-unpleasant”, “excitement” and “yielding-resisting”. The results from the factor analysis in this experiment also provide some support for a three dimensional view of vocal response patterning. A factor that could clearly be related to activation, which consisted of F0 level and mean intensity, was identified. These two acoustic parameters were indeed posited as indicators of the activation dimension by Scherer (1986). In addition, the spectral energy factor, which varied significantly with manipulations of intrinsic pleasantness, seems to match the description given by Scherer for a hedonic valence dimension. The two other factors, F0 dynamics and fluency, do not, however, easily map on to the remaining factor suggested by Scherer, the one of potency. The lack of emergence of a factor that clearly corresponds to the potency dimension is perhaps not surprising, since power, the appraisal dimension most implicated by Scherer in the potency dimension, was not manipulated in the current experiment, and thus might not have varied sufficiently to produce measurable effects in the acoustic data. In addition, Scherer did not include parameters related to speech fluency in his predictions. Nevertheless, at least the fluency factor that was measured in this experiment deserves an explanation, since it varied significantly with goal conduciveness. One possibility is that fluency is related to the cognitive factors involved in emotional situations, such that situations requiring greater use of cognitive resources “steal” cognitive resources away from speech planning and execution, leading to speech that has more pauses and errors. Such situations are likely to be those appraised as obstructive and needing an urgent response – the type of situations that typically provoke anxiety and fear. The fluency measures from this experiment do not support such a hypothesis however, since obstructive situations provoked speech that was more fluent, as indicated by shorter overall duration and greater voiced percentage. Alternatively, the

increase in speech fluency observed for obstructive conditions over conducive conditions could also be the result of increased arousal and excitation, although the low correlations between both fluency measures and the mean energy and F0 measures would seem to cast doubts on this explanation as well.

Conclusions

The results from this experiment indicate the potential of computer games to induce emotional responses which in turn affect the acoustic properties of speech. The main hypothesis of the experiment, that such changes reflect more than the unidimensional effects of physiological arousal on the voice, was supported by the data. In addition, acoustic analyses supported the specific hypotheses put forward by Scherer (1986) concerning the acoustic changes to speech due to appraisals of intrinsic pleasantness. Scherer's predictions of changes to speech due to goal conduciveness appraisals were not, however, supported in this experiment. A possible explanation for the discrepancy is that players did not only appraise events along the two intended dimensions, but also in terms of goal discrepancy.

This experiment has also highlighted a number of methodological and theoretical issues that need to be addressed in future studies. The experiment was successful in eliciting measurable vocal differences between experimental conditions despite the non-negligible delay between the emotion-inducing event and the onset of speech, a result of pausing the game to display the report screen. The measured acoustic differences in speech between the experimental conditions, which in this experiment were small, were presumably what remained of more immediate, possibly larger, acoustic effects. Alternatively, it is possible that the measured effects did not reflect the immediate emotional response to appraisal of the game event, but rather a secondary response, perhaps produced by reappraisal of the event outcome, or even appraisal of the initial emotional response or efforts to control such a response. Given the lack of knowledge of the temporal course of emotional responses (an

issue discussed in some depth by Edwards, 1998), it is impossible to say with certainty which of these alternatives is correct. Clearly, it would be preferable in future computer game studies to measure the acoustic changes to speech more immediately after, or even during, appraised game events, although how this can be done without unduly interrupting the game (and thus interfering with the experimental manipulations) is not obvious.

A further issue concerns the difficulty of interpreting acoustic measurements, in particular those that are inconsistent with theoretical predictions, in terms of the their supposed physiological causes. It is difficult to interpret such global energy, F0 and spectral parameters as were measured in this experiment, without any corresponding data on the respiratory and muscular changes that are thought to affect them. Speech production is highly redundant, so that the same or similar acoustic effects can be produced using a variety of vocal settings. Thus acoustic measurements that are consistent with theoretical predictions provide only indirect support for the theory of how those acoustic changes are produced. More problematic still is when acoustic measurements are inconsistent with the predictions, since one can only speculate as to the causes of the discrepancies, as was the case with the goal conduciveness factor in this experiment. The obvious way to solve this problem is by measuring theoretically relevant physiological variables, such as respiration rate and depth, and muscular tension, concurrently with vocal measurements, although such an approach brings with it many new methodological problems.

One of the advantages of using a computer game to test the predictions made by appraisal theories of emotion is that it can highlight ambiguities in the constructs of the theories themselves. Such was the case in this experiment with goal conduciveness, the manipulation of which was questionable on the grounds that goal discrepancy might have also been manipulated. Although this could be considered a purely methodological problem concerning the design of the goal conduciveness manipulation, a closer look at the theory

itself reveals a certain amount of confusion between the two appraisal dimensions. Thus while goal discrepancy is sometimes described as an evaluation of discrepancy with the *expected* state of proceedings (Scherer, 1986, p. 147), it is also described as an evaluation of discrepancy with the *desired* state of proceedings(Scherer, 1986, p.153)[1], which seems to overlap with the goal conduciveness appraisal dimension. In addition to such definitional issues, there is also the possibility that the different appraisal dimensions are interdependent, that is, that the outcome of one appraisal will influence one or more of the others. This possibility is indeed consistent with the outline of Scherer's theory, in which appraisals are postulated to function in a sequential manner, the outcome of one appraisal check feeding forward into the next check in the sequence (Scherer, 2001). Of course, if the different appraisal dimensions do influence one another, it is difficult to see how they can be independently manipulated in an experiment, an issue that will need to be addressed in future studies.

[1] In the most recent version of Scherer's theory (2001), a clearer distinction is made between discrepancy and goal conduciveness.

References

Banse, R. & Scherer, K. R. (1996).

Barry (1987).

Van Bezooijen (1984).

Cacioppo, Bernston, Larsen, Poehlmann & Ito (2000).

Darwin, C. (1872).

Davidson et al. (1995).

Deller, Proakis & Hansen (1993).

Edwards (1998).

Ekman, P. (1972,1973,1982,1984).

Green & Cliff (1975).

Johnstone, T. & Scherer, K. R. (2000).

Laver (1980).

Leventhal & Scherer (1987).

van Reekum & Scherer (1997).

Scherer, K. R. (1984, 1985, 1986, 1989).

Table 1. Ratings of the two sounds used in the experiment as a manipulation of intrinsic pleasantness.

	Mean rating	Std. Dev.
Unpleasant sound	-2.3	1.1
Pleasant sound	2.2	0.8

Table 2. Pearson correlations between acoustic parameters.

	Mean Energy	Percentage voiced	Duration	F0 floor	F0 ceiling	mean F0	F0 std. dev.	E < 500 Hz	E < 1000 Hz
Mean energy	1.00	.22**	-.06	.33**	.30**	.60**	.14*	-.23**	-.01
Percentage voiced	.22**	1.00	-.69**	.19**	-.03	.09	-.11	.13*	.09
Duration	-.06	-.69**	1.00	-.18**	.06	-.05	.17**	.01	.00
F0 floor	.33**	.19**	-.18**	1.00	.08	.55**	-.41**	.10	.27**
F0 ceiling	.30**	-.03	.06	.08	1.00	.66**	.73**	-.28**	-.44**
mean F0	.60**	.09	-.05	.55**	.66**	1.00	.30**	-.23**	-.13
F0 std. dev.	.14*	-.11	.17**	-.41**	.73**	.30**	1.00	-.30**	-.45**
E < 500 Hz	-.23**	.13*	.01	.10	-.28**	-.23**	-.30**	1.00	.60**
E < 1000 Hz	-.01	.09	.00	.27**	-.44**	-.13	-.45**	.60**	1.00

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 3. Factor weightings for a four factor, Quartimax rotated PCA of acoustic variables.

	Component			
	1	2	3	4
mean energy	.74	.13	.11	-.12
Percentage voiced	.13	.01	.91	.13
Duration	-.06	.12	-.91	.08
F0 floor	.81	-.36	.09	.18
F0 ceiling	.41	.82	-.04	-.17
F0 mean	.87	.37	.01	-.08
F0 std. dev.	-.04	.93	-.09	-.20
energy < 500 Hz	-.16	-.08	.06	.92
energy < 1000 Hz	.10	-.38	-.02	.80

Figure Captions

Figure 1. Vocal report screen used in the experiment (left) and English translation (right). Italics indicate phrases that were displayed in place of the preceding phrases for ship destroyed events.

Figure 2. Subjective emotion rating screen.

Figure 3. Mean reported emotion intensity as a function of experimental condition.

Figure 4. Interaction of pleasantness and conduciveness for F0 ceiling (left) and F0 standard deviation (right; values are participant-standardised scores). Solid lines indicate conducive events, broken lines indicate obstructive events. Error bars represent +/- 1 standard error.

Figure 5. Mean voiced energy in decibels (left) and F0 floor (right) as a function of conduciveness (values are participant-standardised scores). Error bars represent +/- 1 standard error.

Figure 6. Mean utterance duration in seconds (left) and the percentage of each utterance that was voiced (right) as a function of conduciveness (values are participant-standardised scores). Error bars represent +/- 1 standard error.

Figure 1

Galaxie Franchie! **Votre rapport s.v.p.** Identification du vaisseau: **AG01813** Votre compte rendu: **Ennemis peu efficaces** ***Attaque ennemie*** **Navigation correcte** ***Mauvais navigation*** **Chance** Probabilite de franchir la galaxie suivante? **X pourcent**	Galaxy Completed! **Your report please** Identification of ship: **AG01813** Your account: **Enemies not very good** ***Enemy attack*** **Good navigation** ***Bad navigation*** **Luck** Probability to complete the next galaxy? **X percent**

Figure 2


Comment vous sentez-vous en ce moment ?
Interessé / Excité
Etonné / Surpris
Amusé / Joyeux
Content de soi / Fier
Impuissant / Dépassé
Tendu / Nerveux
Enervé / En colére
Mal à l'aise / Honteux
RIEN
OK

Figure 3

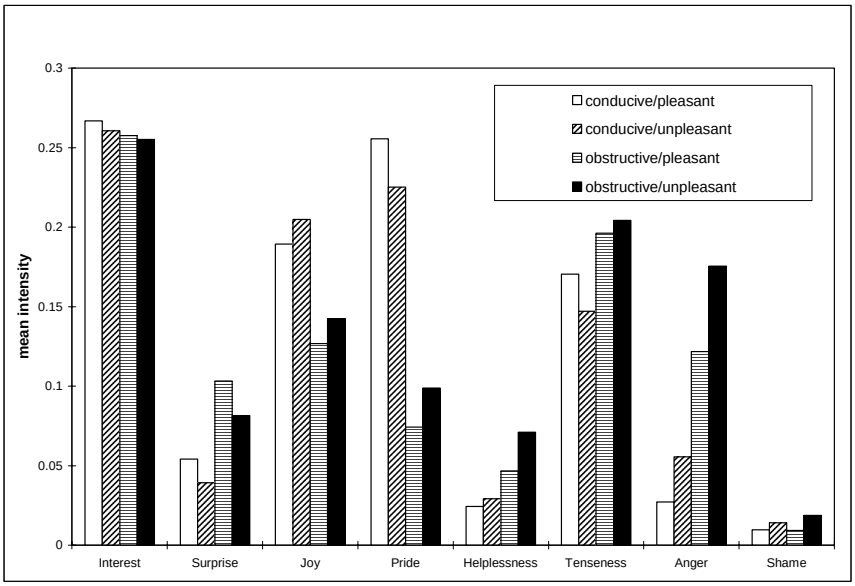

0.3
0.25
0.2
0.15
0.1
0.05
0
mean intensity
conducive/pleasant
conducive/unpleasant
obstructive/pleasant
obstructive/unpleasant
Interest
Surprise
Joy
Pride
Helplessness
Tenseness
Anger
Shame

Figure 4

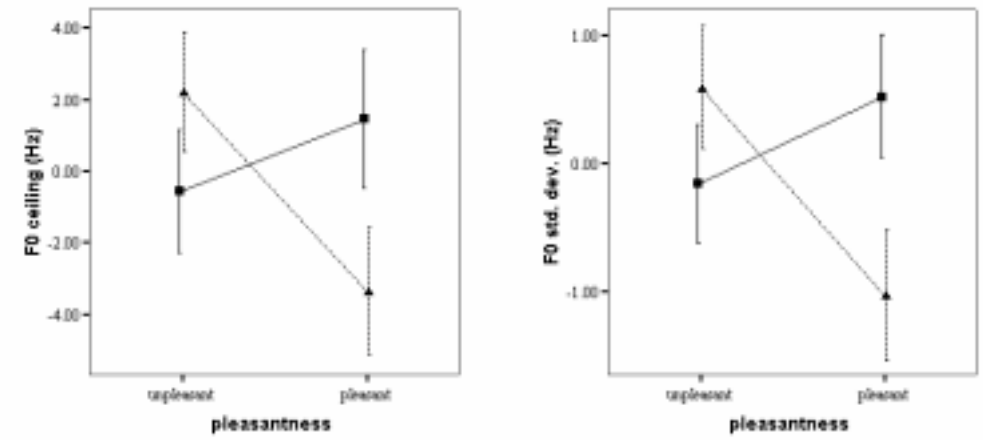

4.00
2.00
0.00
-2.00
-4.00
F0 ceiling (Hz)
unpleasant
pleasant
pleasantness
1.00
0.00
-1.00
F0 std. dev. (Hz)
unpleasant
pleasant
pleasantness

Figure 5

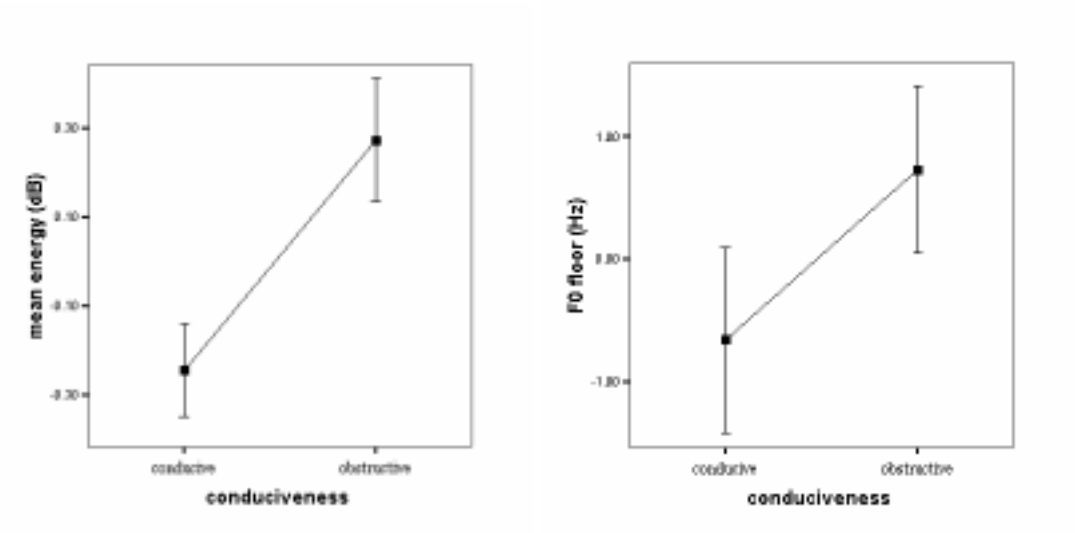
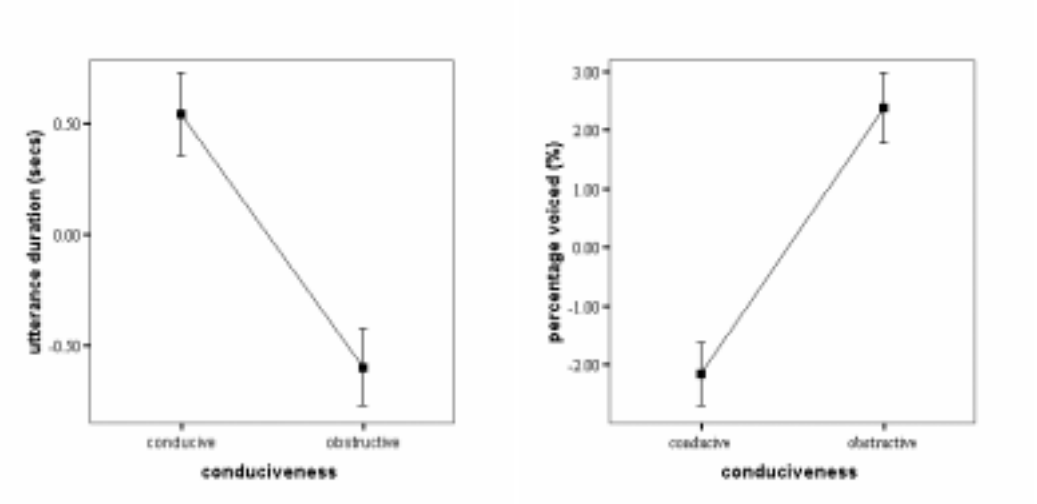

mean energy (dB)
0.30
0.10
-0.10
-0.30
conductive
obstructive
conduciveness
F0 floor (Hz)
1.80
0.80
-1.80
conductive
obstructive
conduciveness

Figure 6


utterance duration (secs)
0.50
0.00
-0.50
conducive
obstructive
conduciveness
percentage voiced (%)
3.00
2.00
1.00
0.00
-1.00
-2.00
conducive
obstructive
conduciveness