EMOTIONAL SPEECH ELICITED USING COMPUTER GAMES

Tom Johnstone
johnston@fapse.unige.ch

University of Geneva

ABSTRACT

The potential of using computer games and simulations as elicitors of affective speech, particularly in research targeting current computer based speech technologies, is discussed. Initial results of acoustical analyses which have been performed on speech samples collected from 30 subjects playing an interactive computer game are presented. The game was modified to manipulate two of the stimulus evaluation checks (intrinsic pleasantness and goal conduciveness) which have been proposed by Scherer [1] as antecedents of emotion.

1. INTRODUCTION

Technical advances in signal processing have allowed the development of a large number of speech technologies, many of which are on the brink of being applied and commercialised. Automatic speaker verification, speech recognition and speech synthesis systems are examples of such technologies which could be usefully applied in areas such as security, telephone banking and computerised information retrieval. Despite the sophistication of current systems however, their reliability and usefulness is limited by the effect on speech of transient state changes to speakers, such as cognitive and physiological stress, emotional state and speaker attitude. Such changes to speaker states have been shown in past research to alter the acoustic speech signal in consistent, state-specific ways. Incorporating a knowledge of such state dependencies in speaker verification and speech recognition technology could thus lead to more robust systems, both through the selection of more appropriate acoustic parameters and processing algorithms, and through the development of speech elicitation techniques which are designed to minimise transient state changes in speakers [2]. Similarly, speech synthesis systems, which are currently limited by their unnatural sounding speech, would benefit from the addition of emotional and attitudinal information in the form of modified prosodic and suprasegmental parameters. As pointed out by Murray and Arnott [3] however, the further development of high quality, emotional synthesised speech requires a complete model of the emotion process, including effects on the nervous system and their accompanying acoustic changes, but that such a model does not yet exist.

2. A THEORETICALLY GROUNDED APPROACH

Steps towards creating such a model have been proposed by Scherer [1] in which emotion is described as consisting of a number of interacting mental and physiological processes. Scherer's component process model of emotion is based on the increasingly supported view in psychology that emotions are the result of an organism continually assessing the state of its environment in terms of its significance for the organism's goals, plans and continued well-being (i.e. so-called "appraisal theories": see [4]). Situations where events in the environment are assessed as being particularly relevant to the organism can give rise to an emotional episode in which the organism's subsystems will adapt functionally to minimise any dangers or maximise any benefits. Scherer has proposed that five subsystems are responsible for the range of mental, expressive and physiological reactions which accompany emotional episodes. Furthering this line of development, the component process model has been extended to make specific predictions of physiological changes which might occur for certain types of emotions, drawing upon both current theory and empirical evidence. These physiological predictions have subsequently been combined with knowledge of the interaction of vocal physiology, voice quality and acoustic descriptions of speech [5] to arrive at predictions of the vocal and acoustic correlates of a number of emotional states [6].

Although speculative, the predictions give some much needed structure to research on vocal expression of emotion; studies can be designed to test specific hypotheses, thus giving rise to results which are easier to interpret and combine across studies. A further benefit of Scherer's predictions is that they are not based upon common-language emotion words, which are often ambiguous (take, for example the case of "anger", which might be used to describe anything from seething, cold-blooded anger to outright rage). Instead, a speaker's state is described in terms of his or her assessment of their environment on a number of appraisal dimensions. By using such theoretically founded approach, it is possible to get closer to describing a speaker's vocal changes in terms of the most emotionally-relevant aspects of the situations in which they are typically elicited. The advantages of a situation-based research paradigm are obvious when one considers application of the knowledge gained to real-world scenarios, as with modern speech technologies.

3. SPONTANEOUS AND ACTED EMOTIONS

It has been suggested by Scherer [6] that information in the voice concerning the speaker's state reflects not only emotion-dependent, largely involuntary physiological changes to the speaker’s speech production systems (“push effects”), but also the more controlled adoption of culturally accepted speaking styles (“pull effects”). This distinction between voluntary and involuntary changes to speech production is crucial to applied research into speech technologies. In the case of speech synthesis the aim is to modify the output of synthesis systems to produce speech which sounds naturally emotional. This is essentially the same goal as that of actors. Emotional speech obtained from actors would therefore seem suitable for use in research which is aimed at adding emotional expression to synthesised speech.

With speech recognition and speaker verification systems however, the problem exists of emotion induced changes to speech which the speaker is largely unable to control. Real affective states probably lead to changes in the vocal signal which serve no intentional communication purpose, but instead reflect changes to the speaker's underlying physiological systems. In contrast, actors are able to achieve a desired change in speech quality by using largely voluntary vocal settings. Thus acoustic analyses of actor portrayed speech might not provide an accurate description of spontaneous affective speech modulation, which is likely to differ both qualitatively and quantitatively, and be less perceptually obvious than those which accompany the full blown voluntary expression of emotion. In order to better understand speech variability for the purpose of improving speech and speaker recognition systems, it would seem necessary to examine not only acted emotional speech, but also speech elicited from subjects under a variety of real or induced emotional conditions.

Unfortunately, this type of research has been scarce due to practical problems. Field-studies of "real-life" emotions are complicated by many possibly confounding factors which make it difficult to accurately specify the emotional state of the speaker. The more controlled elicitation of emotional states in experimental studies has also proved troublesome. Martin [7] provides an overview of the various techniques which have been commonly used in the experimental induction of emotions. These typically include mental imagery techniques, presentation of emotionally valenced films and stories and techniques based upon the recall of past emotional experiences. Essentially, the problem with induction techniques is that ethical constraints prevent anything but the induction of a limited number of low magnitude emotions, which may not provide enough power for their differentiation on the basis of vocal cues. One thorough review of previous studies on the vocal expression of emotion ([6], p. 160) listed 23 studies which used induced or natural emotions, of which only 4 studies examined more than 2 different emotions. Furthermore, there was little overlap between studies of the particular emotions elicited.

4. COMPUTER INDUCTION OF EMOTION

Recently the promise of using computer games and simulations for the induction of moderate to high intensity real emotions has been demonstrated [8]. The ability of computer simulation to induce stronger, more varied emotions than other induction techniques could be expected from their closeness in many ways to "real-life". Games and simulations in general allow players to be submerged in scenarios which can be modelled on everyday situations. Furthermore, games allow players to become involved in the types of events which might only happen infrequently (or never) in normal life, but nevertheless produce strong emotional reactions when they do occur. The high level of interactivity involved in game playing is more conducive to emotion elicitation than more passive activities such as watching films. Modern computer games, with their realistic graphics and sound effects, add to this realism and player involvement.

From an experimental viewpoint, computer games have the advantage that they can be played in a controlled laboratory environment, enabling high quality digital recording of speech. Another major advantage of using computer games as emotion inducers is the ability to change the game in order to systematically manipulate emotion eliciting situations, in accordance with theoretical hypotheses. Moreover, many of the possibly confounding factors which would be uncontrollable in "real-life" situations can be manipulated or controlled in a computer game experiment.

Our research group in Geneva is currently taking some initial steps in using various computer games and simulations to elicit a range of emotions, which are then studied in terms of their subjective, physiological and expressive manifestations [9]. It is hoped that the inclusion of vocal measurements in this effort will contribute towards understanding the interaction between psychological state, physiology and speech acoustics. Specifically, such studies are aimed at discovering to what extent spontaneous, involuntary effects of emotion on speech overlap or differ from the more culturally determined, voluntary aspects of vocal communication of emotion which have been previously studied.

5. THE RESEARCH DESIGN

A series of studies is being conducted to examine the acoustic differences in subject vocal responses under computer-based manipulations of expected outcomes of stimulus evaluation checks (SECs) in Scherer's component process theory of emotion. Eventually it is proposed to create a suite of computer games and simulations for the induction of a variety of emotional states. Programming of the first of these is in process, but for a primary study a readily available and popular arcade-style computer game was chosen (XQuest, programmed by Mark Mackay - see acknowledgements).

Throughout this game the player encounters various active obstacles (enemy space ships), passive obstacles (enemy mines) as well as useful objects which augment the player's power and defensive ability. Players control their space ship using a mouse. The game is chosen for its clear goals (completing each level and accumulating points), as well as the possibility to manipulate game objects in accordance with the SECs of the component process theory. A number of modifications have been made to the game in order to manipulate the subject's probable SEC outcomes at times in the game when subjective or vocal data are collected. A summary of appraisal manipulations within the XQuest game is given in Table 1. The preliminary nature of these manipulations should be emphasised here. In future studies more sophisticated designs using purpose-built computer simulations are envisaged.

In each experimental session the subject is fitted with a headset microphone connected to a DAT recorder, and proceeds to play the game for one hour. At certain times in the game, the subject is requested to verbally report an assessment of the current state of play. In subsequent studies the game will include a limited set of standard vocal commands which the player will use to perform specific actions, as well as vocal interaction with another intelligent agent (i.e. friendly or enemy ship), played by a cohort. This will allow more seamless integration of subject speech into the game, which is expected to provide more natural, emotional utterances. The report and commands are used to collect speech data on a limited set of standard

phrases and isolated words and digits. Interaction with the cohort will provide samples of spontaneous speech. The subject is also required to answer a computerised subjective questionnaire which gives details about the subject's emotional state at key points during the game.

Table 1: Manipulations of SECs in a computer game.

1. Novelty:	High. Previously unencountered game object appears.
	Low: Commonly encountered game object appears.
2. Intrinsic Pleasantness:	High. Pleasant sound accompanies event.
	Low. Unpleasant sound accompanies event.
3. Goal/Need Significance:	
a) Relevance:	High. Object appears which is vital to player's well-being.
	Low. Object appears which has no relevance to player.
b) Conduciveness:	High. Object is friendly or worth points.
	Low. Object is an enemy or destructive.
c) Urgency:	High. Must respond quickly in order to defend (enemy) or benefit (friend).
4. Coping Potential.	
a) Causation:	This subcheck is analysed in terms of a vocal assessment by the subject of the cause of an event. Thus at certain events the subject is required to assess whether the event had an internal (i.e. self) or external (i.e. other) cause, or happened by chance.
b) Control:	High. Player has complete control of ship.
	Low. Ship control disturbed by random perturbation.
c) Power:	High. Player has high shooting power.
	Low. Player has low shooting power.
d) Adjustment:	High. Player has reserve ships in case of damage.
	Low. Player has no reserve ships.
5. Norm/Self-compatibility:	
a) External standards:	High. Player rescues a friendly object.
	Low. Player destroys a friendly object.
b) Internal standards:	High. Player succeeds against difficult/strong enemy.
	Low. Player is defeated by easy/weak enemy.

6. INITIAL RESULTS

A preliminary study has been conducted to examine the acoustic differences in subject vocal responses under manipulations of expected outcomes of the intrinsic pleasantness and goal conduciveness checks. Intrinsic pleasantness was manipulated using valenced (i.e. pleasant and unpleasant) sounds. Goal conduciveness was examined using points in the game which were either highly goal conducive (player reached the next highest level) or highly goal obstructive (player's ship was destroyed). The checks were manipulated in a 2 (intrinsic pleasantness) x 2 (goal significance) within-subjects design. Thus concurrent with the player reaching the next level or losing a life, either a pleasant or unpleasant sound was presented. Speech was elicited by means of a vocal report pop-up screen, which requested a vocal report of the immediately preceding game events. The screen 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. The pop-up screen provided both strings of isolated letters and connected phrases to be pronounced by the subject.

Preliminary results of primary acoustical analyses of the speech samples for 30 subjects are presented in Figure 1, and will be reported in more detail in a future paper. Goal obstructive events were characterised by significantly higher energy speech and more rapid speech (indicated by shorter overall durations with an increased voiced proportion) than goal conducive events. This matches predictions [6] and previous findings [10] for quiet forms of happiness and interest, which were indeed reported by subjects in this condition. Measures of F0 are more ambiguous, with an interaction between pleasantness and goal conduciveness for mean F0 and F0 variability. This was the only effect found for the intrinsic pleasantness manipulation, and might be explained by a perceived discrepancy between the sound and event, although the effect is too small to place much confidence in its interpretation. Finally, there was significantly less energy at low frequencies (i.e. less than 1000 Hertz) as a proportion of total energy, for goal conducive events than for goal obstructive events. This runs somewhat counter to the observed decrease of overall energy for goal conducive events, which one would expect to lead to increased low frequency energy. However, it is difficult to interpret such global spectral parameters without a closer look at the shape of the glottal wave form and the positions and bandwidths of formants. This analysis is currently in progress.

These data presented here are only preliminary. Before making generalisations one would have to replicate the findings with a wider range of game events for each appraisal dimension, to ensure that the findings are not due to idiosyncrasies in the events. However, the data do seem to indicate the potential of computer games to induce psychological state changes which in turn affect speech. This is despite the rather non-spontaneous nature of the vocal reports, in which the game was paused to display the report screen, thus introducing a delay between the emotion-inducing event and the onset of speech. It might be expected that with future integration of speech commands into computer games will lead to clearer and larger effects.

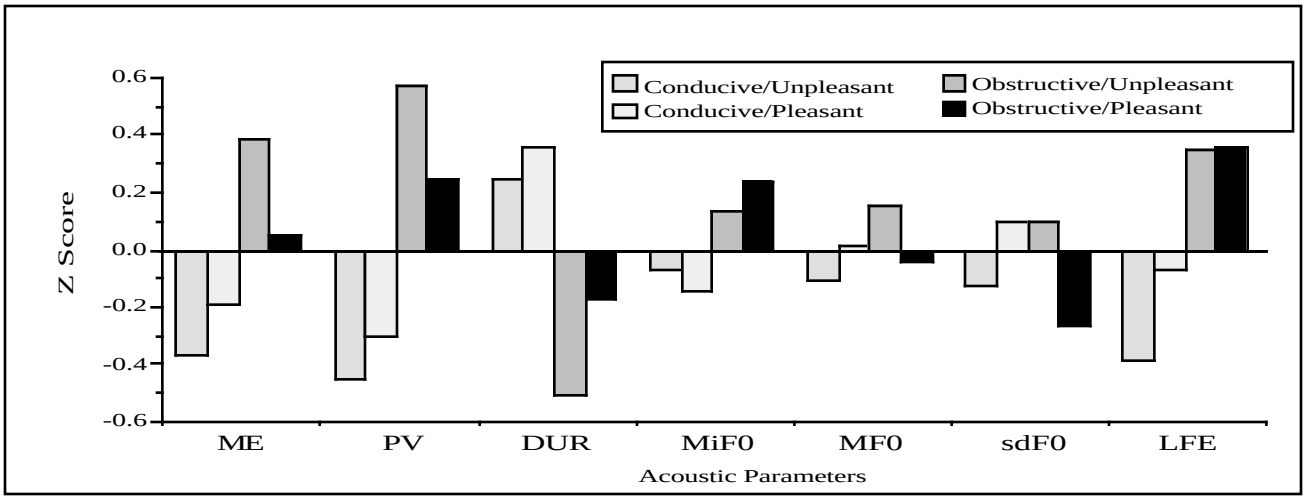

Figure 1: Change of acoustic parameters with manipulations of intrinsic pleasantness and goal conduciveness. Acoustic Parameters: ME: Mean Energy; PV: Percentage voiced; DUR: Mean Duration; MiF0: Minimum F0; MF0: Mean F0; sdF0: Standard Deviation F0; LFE: Low Frequency (<1000 Hz.) Energy.

7. SUMMARY

This paper has discussed the importance of research into affective speech to the effort of improving the robustness of speech recognition and speaker verification technologies. The requirement of studies involving real, spontaneous affective speech has been highlighted, together with a brief review of the problems such studies present. To overcome some of these problems, an experimental plan has been outlined in which computer games are used to induce emotions and elicit affective speech, based upon current psychological knowledge of emotion processes. Initial results have been presented which indicate the promise of such an approach. It is hoped that the continuation of this research will result in greater knowledge of the spontaneous effects of emotion on speech, which can be applied to the problem of improving reliability and robustness in a range of current speech technologies.

5. ACKNOWLEDGEMENTS

This research has been supported by the Swiss National Scientific Research Fund (project FNRS 1114-037504.93). The author would like to thank Mark Mackay for making the source code of XQuest available for this research.

6. REFERENCES

1. Scherer, K. R. "Emotion as a multicomponent process: A model and some cross-cultural data," *Review of Personality and Social Psychology, 5, 1984, 37-63.*

2. Furui, S. "Speaker Recognition," in R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, V. Zue (Eds.)**,** *Survey of the State of the Art in Human Language Technology, Ch. 1.7.* CSLU, Oregon Graduate Institute Of Science & Technology.

3. Murray, I. R., and Arnott, J. L. "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, 1993, 1097-1108.*

4. Gehm, T. L., and Scherer, K. R. "Relating Situation Evaluation to Emotion Differentiation: Nonmetric Analysis of Cross-cultural Questionaire Data," In K.R. Scherer (Ed.) *Facets of Emotion - Recent Research (pp. 61-77).* Hillsdale, N.J.: Lawrence Erlbaum Associates, 1988.

5. Laver, J., and Hansen, R. "Describing the normal voice," in J.K. Darby (Ed.), *Speech Evaluation in Psychiatry (pp. 51-78).* New York: Grune & Stratton, 1981.

6. Scherer, K. R. "Vocal Affect Expression: A review and a Model for Future Research," *Psychological Bulletin, 99, 1986, 143-165.*

7. Martin, M. "On the induction of mood," *Clinical Psychology Review, 10, 1990, 669-697.*

8. Kaiser, S., Wehrle, T. and Edwards, P. "Multi-modal emotion measurement in an interactive computer game: A pilot study," *Proceedings of the 8th Conference of the International Society for Research on Emotions,* Storrs, CT: ISRE Publications, 1993.

9. Banse, R., Etter, A., van Reekum, C. M., and Scherer, K.R. "Psychophysiological responses to emotion-antecedent appraisal of critical events in a videogame," *in preparation.*

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