Researching Emotions in Speech

Gordon McIntyre[1] **and Roland Göcke**[1,2]

[1]Research School of Information Sciences and Engineering, Australian National University
[2]Vision Science, Technology and Application, NICTA Canberra Laboratory *
gordon.mcintyre@anu.edu.au, roland.goecke@nicta.com.au

Abstract

Many studies into affective computing cover similar ground or overlap, pointing to much duplication of effort. Two problem areas exist in the research. The first relates to the way in which emotions are defined and described. The second is in the techniques used to elicit samples of speech. Further, few studies are based on natural speech, ignoring social, contextual, cultural and agent characteristics and few take account of the affective state of the receiver. This paper presents a survey of the problem areas and proposes a comprehensive model and a set of ontologies for use in affective communication.

1. Introduction

The December 2005 edition of New Scientist published an article titled, "Tell Laura I love her..." (Daviss, 2005), featuring an insight into the work of esteemed, affective-computing pioneer, Ros Picard. The article, whilst positive, did air the view of one critic who complained,

> "...that there is no valid theory, or even commonly accepted definition of 'emotion' to guide the work. These are smart people and they're producing useful things, but I think that's despite the framework they work in, not because of it...If the banner concept of your research is something that you can't even define, ...then you run the risk of producing things that do not advance human knowledge, but can only be described as 'cute' a digital motivator for a fitness programme, for example."

Overall, the article presented Picard and affective computing in a positive light. However, the criticism is understandable. Notions of *systematic*, *verifiable means* and *acquiring knowledge through empiricism*, are important in science. It is quite reasonable then for scientists from other areas of interest to expect that published results are built on a foundation that is systematic and verifiable.

Research in the field of affective computing is not consistent and, arguably, not verifiable. There is a lack of consistency in the way in which emotion is defined, measured and catergorised. This in turn makes verification and reuse of results difficult.

This paper presents a survey of two of the major areas of contention in affective computing and proposes a more complete model that could be used to progress research into emotions in speech.

The next section explains the motivation behind this study. Section 3 provides some background to some fundamental problems in defining, categorizing and measuring emotions. Section 4 summarises the most common techniques for eliciting data for research. Section 5 presents a new, more comprehensive model of affective communication and a set of ontologies for use when researching the field. Section 6 concludes the paper.

The terms affect, affective state and emotion, although not strictly the same, are used interchangeably in this paper.

2. Motivation

The human computer interface is, for other than basic operations, limited without some level of incorporation of the affective human state. An interface that responds and adapts to emotions can be more satisfying, much more productive and can provide many opportunities. One of the most natural interfaces is speech and we know that evidence of emotions is often present in speech. Emotion recognition in speech by computers could provide many benefits.

However, the field would benefit from a more consistent modelling, reporting and research framework for affective communication.

3. Problem 1 - Describing emotions

Emotion theory is a large amalgam of approaches to the study of emotion, mostly based on cognitive psychology but with contributions from learning theory, physiological psychology, clinical psychology, and other disciplines including philosophy.

There are many reviews on emotional classification, e.g. (Cowie & Cornelius, 2003). Two approaches to describing emotions seem to be dominant. The first is to define categories and is the more common technique. The second approach is through the use of dimensions. Another approach, discussed briefly, is through the use of appraisal theory.

3.1. Category approach

The most popular grouping, referred to as the "big six", comprises fear, anger, happiness, sadness, surprise and disgust (Cornelius, 1996; Ekman, 1999). These are regarded as "full-blown" emotions (Scherer, 1999) and are evolutionary, developmental and cross-cultural in nature (Ekman & Oster, 1982). However, there are many alternative groupings both across disciplines and within disciplines. Some

* National ICT Australia is funded by the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

studies concentrate on only one or two select categories. Other employ schemes using more than twenty emotional archetypes.

One of the difficulties then in comparing results from studies into emotions is that the choice of categories used between studies is not consistent and will often depend on the application that the researcher has in mind. If the focus of the research is to understand full-blown emotions then the big-six or a subset might be adequate. However, if the aim is to study the less dramatic emotional states in everyday life, with all the shades and nuances that we know distinguish them, then the choice of categories is much more difficult (Schröder & Cowie, 2005).

Thus, there are many problems with the category approach such as:

- no agreed number of categories or definitions
- the large number of descriptors with overlapping meanings
- lack of consistency of words across languages
- inconsistency in the use of categories in research

3.2. Dimensional approach

Another way to label the affective state is to use multiple dimensions. Instead of choosing discrete labels, one or more continuous scales, such as pleasant/unpleasant, attention/rejection or simple/complicated are used. Two common scales are valence (negative/positive) and arousal (calm/excited). Valence describes the degree of positivity or negativity of an emotion or mood; arousal describes the level of activation or emotional excitement. Sometimes a third dimension, control or attention, is used to address the internal or external source of emotion. A software application has been developed at (Cowie et al., 2000) to assist in the continuous tracking of emotional state on two dimensions. The package, called FEELTRACE, has been used in a number of studies.

3.3. Appraisal approach

Scherer has studied the assessment process in humans and suggests that people affectively appraise events with respect to novelty, intrinsic pleasantness, goal/need significance, coping, and norm/self compatibility (Scherer, 1999). It is not yet clear how to implement this this approach in practice.

3.4. Discussion

Both the categorical and dimensional approaches suffer from being highly subjective. Recent studies have shown that this is not only because of complexity but also because of the differences in the efficiency of listeners' physiology and the fact that the listener's own affective state influences their judgment of the speaker's affective state. Hence, there is a need for a model that includes listener attributes.

Analysis of Belfast Naturalistic Database has shown that full-blown emotions occur more rarely in natural speech than was thought (Cowie & Cornelius, 2003). A study by the Human-Machine Interaction Network on Emotion (HUMAINE) group proposes that emotions be dichotomised into *episodic* and *pervasive* categories. *Episodic* emotions are more like the traditional set of "full-blown" emotions or what Scherer would describe as emotions. *Pervasive* are the everyday emotions such as grief/sorrow, sarcasm/irony and surprise/astonishment. Regardless of the method employed to describe emotions, this greatly complicates the task of describing emotional content in speech samples.

A recent study by (Devillers, Vidrascu & Lamel, 2005) has included labelling of *blended* and *secondary* emotions in a corpus of medical emergency call centre dialogues, as well as including task-specific context annotation to one of the corpora. Introducing the concept of multiple, concurrent emotions further compounds the difficulty of describing emotional content in speech samples.

4. Problem 2 - Eliciting emotional speech

Most studies into affective communication begin with the collection of audio and/or video communication samples. The topic of collection of emotional speech has been well covered by other reviews (Scherer, 2003; Cowie & Cornelius, 2003; Cowie, Douglas-Cowie & Cox, 2005) so it is only briefly covered here.

4.1. Naturally occurring speech

To date, call centre recordings, recordings of pilot conversations, and news readings have provided sensible sources of data to research emotions in speech. These type of samples have the highest ecological validity. However, aside from the copyright and privacy issues, it is very difficult to construct a database of emotional speech from this sort of naturally occurring emotional data sources. In audio samples there are the complications of background noise and overlapping utterances. In video there are difficulties in detecting moving faces and facial expressions. A further complication is the suppression of emotional behaviour by the speaker who is aware of being recorded.

4.2. Induced emotional speech

One technique introduced by Velten, is to have subjects read emotive texts and passages which, in turn, induce emotional states in the speaker (Velten, 1968). Other techniques include the use of Wizard of Oz setups where, for example, a dialog between a human and a computer is controlled without the knowledge of the human. This method has the benefit of providing a degree of control over the dialogue and can simulate a natural setting.

The principal shortcoming of these methods is that the response to stimuli may induce different emotional states in different people.

4.3. Acted emotional speech

A popular method is to engage actors to portray emotions. This technique provides for a lot of experimental control over a range of emotions and like the previous method provides for a degree of control over the ambient conditions.

One problem with this approach is that acted speech elicits how emotions should be portrayed, not necessarily

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

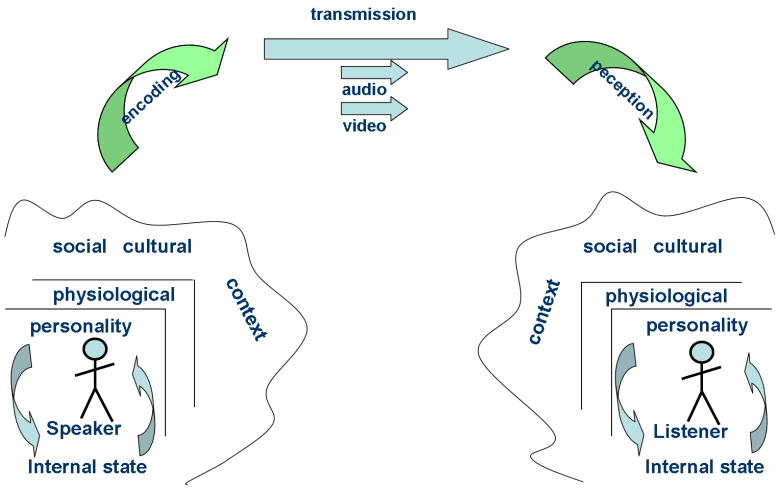

Figure 1: A model of affective communication

how they are portrayed. The other serious drawback is that acted emotions are unlikely to derive from emotions in the way that Scherer et al. describe them, i.e. "episodes of massive, synchronised recruitment of mental and somatic resources to adapt or cope with a stimulus event subjectively appraised as being highly pertinent to the needs, goals and values of the individual (Scherer, 2004)".

4.4. Discussion

Relatively little research into affect has been based on natural speech. In many cases, the approach to affect recognition has simply been an extension of automatic speech recognition (ASR), i.e. acquiring a corpus of acted speech, then annotating sequences containing affect within the corpus. In the case of automatic recognition of *episodic* emotions, this approach is plausible, based on the assumption that clear-cut bursts of episodic emotion will look and sound somewhat similar in most contexts (Cowie et al., 2005). However, recognition of *pervasive* emotions present a much greater challenge and, intuitively, one would think that awareness of personal and contextual information needs to be integrated into the recognition process. Richard Stibbard (Stibbard, 2001) who undertook the, somewhat difficult, Leeds Emotion in Speech Project reported,

> "The use of genuine spoken data has revealed that the type of data commonly used gives an oversimplified picture of emotional expression. It is recommended that future work cease looking for stable phonetic correlates of emotions and look instead at dynamic speech features, that the classification of the emotions be reconsidered, and that more account be taken of the complex relationship between eliciting event, emotion, and expression. "

5. Proposed solution

The proposed solution consists of 1) a generic model of affective communication; and 2) a set of *ontologies*. An ontology is a statement of concepts which facilitates the specification of an agreed vocabulary within a domain of interest.

The model and related ontologies are intended to be used in conjunction to describe 1) affective communication concepts; 2) affective computing research; and 3) affective computing resources.

5.1. A more comprehensive model

Figure 1 presents the base model. The first point to note is that it includes speaker and listener, more in keeping with the Brunswikian lens model as proposed by Scherer (Scherer, 2003). The reason for modelling attributes of both speaker and listener is that the listener's cultural and social presentation vis-à-vis the speaker may also influence judgement of emotional content. The second point to note is that it includes a number of factors that influence the expression of affect in speech. Each of these factors and the motivation for their inclusion is briefly discussed. More attention is given to context as this is seen as a much neglected factor in the study of automatic emotion recognition and synthesis.

5.1.1. Context

Context is linked to modality and emotion is strongly multi-modal (Cowie et al., 2005) in the way that certain emotions manifest themselves favouring one modality over the other. Physiological measurements change depending on whether a subject is sedentary or mobile. A stressful context such as an emergency hot-line, air-traffic control or a war zone is likely to yield more examples of affect than everyday conversation.

(Stibbard, 2001), recommended,

> "...the expansion of the data collected to include relevant non-phonetic factors including contextual and inter-personal information."

His findings underline the fact that most studies take place in an artificial environment, ignoring social, cultural, contextual and personality aspects which, in natural situations, are major factors modulating speech and affect presenta-

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Modulating factors			Production and detection factors		
Cultural	**Social**	**Context**	**Agent Characteristics**	**Physiological**	**Internal State**
Speaker's vis-à-vis listener's age and gender Language Customs Race	Education Familiarity/ rapport with listener Gender	Group situations Ambient conditions Dialogue turn Familiarity with system Sedentary/active Overt/covert Location	Extrovert/ Introvert Authoritarian/ control freak Child vs elderly Appearance, eg spectacles, facial hair, head and eye movement	Voice quality Child vs elderly Gender Illness/ Infirmity Impairment Vocal tract length Skin colour	Recent events, eg lottery wins, losses

Figure 2: Use of the model in practice

tion. The model depicted at Figure 1 takes into account the importance of context in the analysis of affect in speech.

A recent study by (Devillers et al., 2005) included context annotation as metadata to a corpus of medical emergency call centre dialogs. In that study, context information is treated as either task-specific global in nature and "Call origin" and "reason for call" are considered task-specific. The model proposed in this paper does not differentiate between task-specific and global context as the difference is seen merely as temporal, i.e. pre-determined or established at "run-time".

Other researchers have included "discourse context" such as speaker turns (Liscombe, Riccardi & Hakkani-Tür, 2005) and specific dialog acts of greeting, closing, acknowledging and disambiguation. Inclusion in a corpus of speaker turns would be useful but annotation of every specific type of dialog act would be extremely resource intensive.

(HUMAINE, 2006) included a proposal that at least the following issues be specified:

- Agent characteristics (age, gender, race)
- Recording context (intrusiveness, formality, etc.)
- Intended audience (kin, colleagues, public)
- Overall communicative goal (to claim, to sway, to share a feeling, etc.)
- Social setting (none, passive other, interactant, group)
- Spatial focus (physical focus, imagined focus, none)
- Physical constraint (unrestricted, posture constrained, hands constrained)
- Social constraint (pressure to expressiveness, neutral, pressure to formality)

but went on to say,

> "It is proposed to refine this scheme through work with the HUMAINE databases as they develop."

(Millar, Wagner & Göcke, 2004) developed a methodology for the design of audio-video data corpora of the speaking face in which the need to make corpora re-usable is discussed. The methodology, although aimed at corpus design, takes into account the need for *speaker* and *speaking environment* factors.

In contrast to the HUMAINE report mentioned previously, the model presented in this paper treats agent characteristics and social constraints separate to context information. This is because their effect on discourse is seen as separate topics for research. It is evident:

1. That context is extremely important in the display rules of affect; and
2. Defining context annotation is in its infancy.

5.1.2. Agent characteristics

As Scherer points out (Scherer, 2003) most studies are either speaker oriented or listener oriented with most being the former. This is significant when you consider that the emotion of someone labelling affective content in a corpus could impact the label that is ascribed to a speaker's message.

Not much attention in the literature has been given to the role that agent characteristics such as personality type play in affective presentation which is surprising when one considers the obvious difference in expression between extroverted and introverted types. Intuitively, one would expect a marked difference in signals between speakers. Again intuitively, one would think that a baseline of the person's personality type would be of great benefit in applications that monitor an individual's emotions.

At a more physical level, agent characteristics such as facial hair, whether they wear spectacles, and their head and eye movements all affect the ability to visually detect and interpret emotions.

5.1.3. Cultural

Culture-specific display rules (Cowie et al., 2005) influence the display of affect. Gender and age are established as important factors in shaping conversation style and content in many societies.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

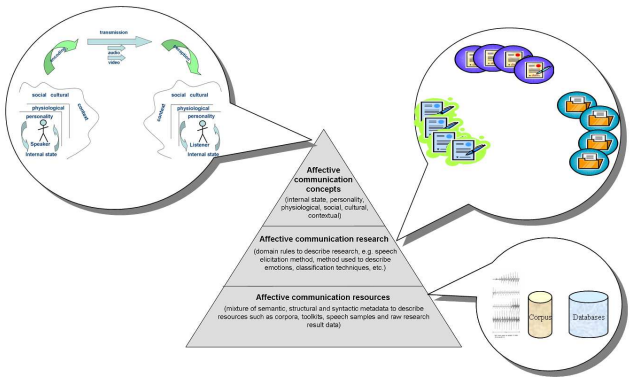

Figure 3: A set of ontologies for affective computing

Studies by (Koike, Suzuki & Saito, 1998) have shown that it is difficult to identify the emotion of a speaker from a different culture and that people will predominantly use visual information to identify emotion (Shigeno, 1998). Putting it in the perspective of the proposed model, cognisance of the speaker and listener's cultural backgrounds, the context, and whether visual cues are available, obviously influences the effectiveness of affect recognition

5.1.4. Physiological

It might be stating the obvious but there are marked differences in speech signals and facial expressions between people of different age, gender and health. The habitual settings of facial features and vocal organs determine the speaker's range of image and sound possibilities. The configuration of chin, lips, nose and eyes provide the visual cues, whereas the vocal tract length and internal muscle tone guide the interpretation of acoustic output (Millar, Wagner & Göcke, 2004).

5.1.5. Social

Social factors temper speech to the demands of civil discourse (Cowie et al., 2005). Affective bursts are likely to be constrained in the case of a minor relating to an adult, yet totally unconstrained in a scenario of sibling rivalry. A social setting in a library is less likely to yield loud and extroverted displays of affect than a family setting.

5.1.6. Internal state

Internal state has been included in the model for completeness. At the core of emotions is the person and their experiences. Recent events such as winning the lottery or losing a job are likely to influence emotions.

5.1.7. Examples

To help explain the differences between the factors that influence the expression of affect, Figure 2 lists some examples. The factors are divided into two groups. On the left, is a list of factors that modulate or influence the speaker's display of affect, i.e. cultural, social and contextual. On the right, are the factors that influence production or detection in the speaker or listener respectively, i.e. personality type, physiological make-up and internal state.

5.2. A set of ontologies

The three ontologies described in this paper are the means by which the model is implemented and are currently in prototype form. Figure 3 depicts the relationships between them and gives examples of each of the ontologies. Formality and rigour increase towards the apex of the diagram. The types of users are not confined solely to researchers. There could be many types of users such as librarians, decision support systems, application developers and teachers.

5.2.1. Ontology 1 - Affective communication concepts

The top level ontology correlates to the model discussed in section 5.1 and is a formal description of the domain of affective communication. It contains internal state, personality, physiological, social, cultural, contextual. It can be linked to external ontologies in related fields such as medicine, anatomy and biology.

5.2.2. Ontology 2 - Affective communication research

This ontology is more loosely defined and includes the concepts and semantics used to define research in the field. It has been left generic and can be further subdivided into an affective computing domain at a later stage, if needed. It is used to specify the rules by which accredited research reports are catalogued. It includes metadata to describe, for example, classification techniques used; the method of eliciting speech, e.g. acted or natural; and manner in which corpora or results have been annotated, e.g. categorical or dimensional. Creating an ontology this way introduces a common way of laying down the knowledge and facilitates intelligent searching and reuse of knowledge within the domain. For instance, an ontology just based on the models described in this paper could be used to find all research reports where:

SPEAKER(internalState='happy',
physiology='any',
agentCharacteristics='extrovert',
social='friendly',context='public',
elicitation='dimension')

Again, there are opportunities to link to other resources. As an example, one resource that will be linked is the Emotion Annotation and Representation Language (EARL) which

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

is currently under design by Schröder et al. within the HUMAINE project (Schröder, 2006). EARL is an XML-based language for representing and annotating emotions in technological contexts. Using EARL, emotional speech can be described either using a set of forty-eight categories, dimensions or even appraisal theory. Examples of annotation elements include "Emotion descriptor" - which could be a category or a dimension, "Intensity" - expressed in terms of numeric values or discrete labels, "Start" and "End".

5.2.3. Ontology 3 - Affective communication resources

This ontology is more correctly a repository that contains both formal and informal rules, and data. It is a mixture of semantic, structural and syntactic metadata. It contains information about resources such as corpora, toolkits, speech samples and raw research result data.

6. Conclusions

Affective computing is an important domain for the next generation of computer user interfaces and cannot be ignored. To move from being seen as something that produces things that critics only describe as "cute", it is essential that research into affective computing embraces notions such as *systematic*, *verifiable means* and *acquiring knowledge through empiricism*. However, the study of emotions in speech is complex and there are some difficult problems to address. It is suggested that providing a complete model of affective communication and a set of ontologies will encourage pluridisciplinary research and lay a solid foundation for future research.

Further information will be available at

http://users.rsise.anu.edu.au/~gmcintyr/index.html

as the ontology is developed.

References

Cornelius, R. (1996). *The science of emotion*. New Jersey: Prentice Hall.

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

Cowie, R., Douglas-Cowie, E., & Cox, C. (2005). Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural Networks*, *18*, 371–388.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schroder, M. (2000). FEELTRACE: an instrument for recording perceived emotion in real time. In *Proceedings of the International Speech Communication Association Workshop on Speech and Emotion*.

Daviss, B. (2005). Tell Laura I love her; She's attractive, charming and always there for you. *New Scientist*, *188*(2528), 42–46.

Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, *18*, 407–422.

Ekman, P. (1999). *The Handbook of Cognition and Emotion*, Facial Expressions, (pp. 301–320). Sussex, U.K: John Wiley and Sons, Ltd.

Ekman, P. & Oster, H. (1982). *Emotion in the human face* (2nd ed.). New York: Cambridge University Press.

HUMAINE (2006). Retrieved 26 October, 2006 from. http://emotion-research.net/.

Koike, K., Suzuki, H., & Saito, H. (1998). Prosodic parameters in emotional speech. (pp. 679–682). International Conference on Spoken Language Processing.

Liscombe, J., Riccardi, G., & Hakkani-Tür, D. (2005). Using context to improve emotion detection in spoken dialog systems. (pp. 1845–1848). EUROSPEECH'05, 9th European Conference on Speech Communication and Technology.

Millar, J. B., Wagner, M., & Göcke, R. (2004). Aspects of speaking-face data corpus design methodology. In *International Conference on Spoken Language Processing 2004*, volume II, (pp. 1157–1160)., Jeju, Korea.

Scherer, K. R. (1999). *Handbook of Cognition and Emotion*, Appraisal theory. New York: John Wiley.

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

Scherer, K. R. (2004). HUMAINE Deliverable D3c: Preliminary plans for exemplars: theory. Retrieved 26 October, 2006 from. http://emotion-research.net/publicnews/d3c/.

Schröder, M. (2006). HUMAINE project D6e: Report on Representation Languages. Retrieved 26 October, 2006 from. http://emotion-research.net/deliverables/D6efinal.

Schröder, M. & Cowie, R. (2005). HUMAINE project: Developing a Consistent View on Emotion-oriented Computing. Retrieved 26 October, 2006 from. http://emotion-research.net/aboutHUMAINE.

Shigeno, S. (1998). Cultural similarities and differences in the recognition of audio-visual speech stimuli. (pp. 281–284). International Conference on Spoken Language Processing -1998, paper 1057.

Stibbard, R. (2001). *Vocal expression of emotions in non-laboratory speech: An investigation of the Reading/Leeds Emotion in Speech Project annotation data.* PhD thesis, University of Reading, UK.

Velten, E. (1968). A laboratory task for induction of mood states. *Behaviour Research and Therapy*, *6*, 473–482.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review