2006-1

Generation of High Quality Audio Natural Emotional Speech Corpus using Task Based Mood Induction

Charlie Cullen
Technological University Dublin, charlie.cullen@tudublin.ie

Brian Vaughan
Technological University Dublin, brian.vaughan@tudublin.ie

Spyros Kousidis
Technological University Dublin

See next page for additional authors

Funder: Salero Project

2006-01-01

Generation of High Quality Audio Natural Emotional Speech Corpus using Task Based Mood Induction

Charlie Cullen
Dublin Institute of Technology, charlie.cullen@dmc.dit.ie

Brian Vaughan
Dublin Institute of Technology

Spyros Kousidis
Dublin Institute of Technology

Yi Wang
Dublin Institute of Technology

Ciaran McDonnell
Dublin Institute of Technology

See next page for additional authors

Generation of High Quality Audio Natural Emotional Speech Corpus using Task Based Mood Induction

C. Cullen [a1], B. Vaughan [a], S. Kousidis [a], Wang Yi [a], C. McDonnell [a] and D. Campbell [a]
[a] *Digital Media Centre, Dublin Institute of Technology, Aungier Street, Dublin 2, IRELAND*

Detecting emotional dimensions [1] in speech is an area of great research interest, notably as a means of improving human computer interaction in areas such as speech synthesis [2]. In this paper, a method of obtaining high quality emotional audio speech assets is proposed. The methods of obtaining emotional content are subject to considerable debate, with distinctions between acted [3] and natural [4] speech being made based on the grounds of authenticity. Mood Induction Procedures (MIP's) [5] are often employed to stimulate emotional dimensions in a controlled environment. This paper details experimental procedures based around MIP 4, using performance related tasks to engender activation and evaluation responses from the participant. Tasks are specified involving two participants, who must co-operate in order to complete a given task [6] within the allotted time. Experiments designed in this manner also allow for the specification of high quality audio assets (notably 24bit/192Khz [7]), within an acoustically controlled environment [8], thus providing means of reducing unwanted acoustic factors within the recorded speech signal. Once suitable assets are obtained, they will be assessed for the purposes of segregation into differing emotional dimensions. The most statistically robust method of evaluation involves the use of listening tests to determine the perceived emotional dimensions within an audio clip. In this experiment, the FeelTrace [9] rating tool is employed within user listening tests to specify the categories of emotional dimensions for each audio clip.

Keywords: MIP, audio quality, emotion, corpus

1 INTRODUCTION

This paper considers the means by which an emotional speech corpus of high quality audio assets can be obtained for the purposes of analysis. The methods of obtaining emotional content are subject to considerable debate, with distinctions between acted [3] and natural [4] speech being made based on the grounds of authenticity. Having said this, it is difficult to obtain natural emotion speech, and so Mood Induction Procedures (MIP's) [5] are often employed to stimulate emotional dimensions in a controlled environment. This paper details experimental procedures based around MIP 4, using performance related tasks to engender activation and evaluation responses from the participant. Tasks are specified involving two participants, who must co-operate in order to complete a given task [6] within the allotted time. By varying factors such as the amount of materials available, the instructions provided or the time allowed to complete each task, the participants can be placed in situations where one or other is perceived to be resisting co-operation, thus inducing situations where emotional dimensions may be obtained.

2 EMOTIONAL DEFINITIONS

2.1 Comparison of Basic Emotional Categories

Schroeder [2] details a list of emotional categories compiled with regard to several leading commentators [10-15] that suggest many elements of cross-pollination (Figure 1).

Emotion	Lazarus	Ekman	Buck	Lewis & Haviland	Banse & Scherer	Cowie et al
Anger	X	X	X	X	X	X
Fear	X	X	X	X	X	X
Sadness	X	X	X	X	X	X
Happiness	X	X	X	X	X	X
Anxiety	X		X	X	X	X
Disgust	X	X	X	X	X	
Pride	X	X	X	X		
Shame	X	X	X	X	X	
Guilt	X	X	X	X		

Fig. 1: Compilation of corresponding emotional definitions for various leading commentators

From this table, it can be seen that some form of general agreement has been reached as regards the

[1] Charlie Cullen: Digital Media Centre, Dublin Institute of Technology, Aungier Street, Dublin 2, IRELAND. charlie.cullen@dit.ie

definition of at least four key emotions. Although discrepancies do exist (e.g. Banse & Scherer define 4 levels of anger) there can said to be some form of general consensus in several of the main categories. The definition of basic, full-blown emotions such as anger, fear, sadness and happiness conforms well to both the Darwinian [14] and Jamesian [16] emotional perspectives.

2.2 Circumplex Emotional Modelling

Regardless of the perspective upon which emotions are defined, there is still enough overlap of definition to consider basic models for their representation. Circular structures [17-19] of emotional definition have been suggested as a robust method of establishing contrast between basic (or prototype) emotional categories (Figure 2).

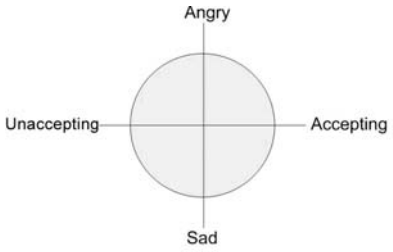

Fig.2: Example of circumplex model of basic emotional categorisation, adapted from Plutchik [19]

Circumplex models allow simple visualisations to be made of contrasting emotional categories, whilst also displaying underlying emotions that could conceivably be found between both. This method conforms well to Darwinian, Jamesian, Cognitive and Social Constructivist [20] perspectives on emotion. All of the perspectives contend that a certain level of activation takes place with emotion whether it is passive or active. Circumplex modelling provides a means by which stimuli may be rated visually in an easily understood manner.

2.3 Emotional Dimensions

A relatively recent consideration in emotional speech is the use of emotional dimensions as descriptors, although the principle is well defined within psychological literature [2]. Psychological experimentation over several decades [18, 21] led to the determination of contrasting basic factors such as pleasure and excitement. In 1984, Scherer suggested dimensions of positive/negative evaluation and activity [22]. These definitions were implemented by Cowie [23], notably within the FeelTrace system which provides a user interface for emotional ratings (Figure 3).

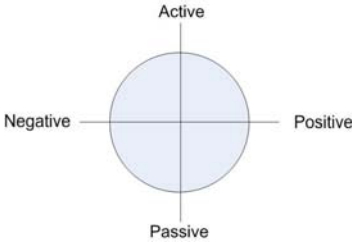

Fig.3: Example of the FeelTrace emotional dimension rating interface, adapted from Cowie [23]

The FeelTrace tool allows emotional content to rated quickly and intuitively by listeners, providing a simple and effective platform for emotional definition. It is argued that assessment of emotional content using such dimensions is an efficient method, and so should be considered as the basis for development.

2.4 Discussion

We argue that common ground is found by the use of basic emotional dimensions of positive/negative evaluation and activity [22]. It is essential that any model be based on criteria which can be effectively implemented and ascertained. Induced emotion experiments [5] have been performed to engender many differing responses from participants. The authors believe that a suitable emotional induction experiment should seek to produce assets that could be effectively classified using existing emotional dimensions, as defined by the FeelTrace tool; any suitable experiment must be designed according to basic dimensions such as activation or evaluation.

3 CURRENT EMOTIONAL SPEECH CORPORA

It is essential that all audio (and video) assets concerned are robustly defined in terms of their emotional content, in order that subsequent analysis of those assets is defendable. Several approaches to asset generation such as real speech [4, 24], acted speech [3, 25] and stimulated (emotive text) speech [26-28] have been undertaken by existing work, as outlined below.

3.1 Natural emotional databases

Few examples exist of natural emotional asset database [29-31], although the justification for such content is its 'purity' when compared with stimulated or simulated content. The method of obtaining such assets varies, but the premise of 'real' emotional content is the key factor. A corpus of real, non-induced emotional speech would provide the most accurate assets for analysis. Having said this, the definition of a neutral or contrasting emotional state is necessary when performing such analysis so that reasonable comparisons may be performed. For this reason, it may often be difficult to determine such a contrast in real emotional assets, although wider emotional definitions may allow more scope for categorisation.

Emotional speech as it is required for the specification of a rule based emotional definition model must ideally be as 'real' as possible. This term has come to mean many things in the field of emotional speech analysis, but fair compromise has often been sought using acted assets. Although easier to obtain and control, acted emotion is at best a good facsimile of real emotional content and so can reasonably be discounted for effective analysis. Similarly, the procurement of natural speech has come to include the use of assets obtained from broadcast sources. This type of asset can be considered in many ways as either an amateur or semi-professional performance by all concerned, and so is again open for great debate about its true validity for analysis purposes. Taking all these factors in to consideration the authors believe that a corpus must contain real emotional assets, not obtained from broadcast sources and not using actors. In this way we can be sure that the emotional content of the recorded audio really exists.

4 MOOD INDUCTION PROCEDURES

Mood Induction Procedures (MIP's) are a set of experiments that are designed to induce a specific emotional state in a test subject in a controlled situation. This emotional state is temporary and specific, as the goal of the MIP is to induce a single precise emotion.

4.1 Mood Induction Procedures

Gerrards-Hesse et al [5] details five different Mood Induction Procedures (MIP) used to elicit real emotion from test subjects. Each MIP has a different method for accomplishing this, from using films and music to using hypnosis and gifts to elicit emotional responses from test subjects. We will briefly discusses the various different MIP's and then examine the MIP that is best suited to the experiment that we intend to carry out. The studies performed on the various methods of emotional induction [32], has led to the categorisation of induction within one of five MIP's:

1. MIP's where emotion is freely generated using mental techniques such as imagination or even hypnosis.
2. MIP's where guidance towards an emotional state is made using suggestions of positive or negative evaluation.
3. MIP's using emotion inducing material such as film, music or text without instruction.
4. MIP's using need-related emotional situations, such as the Success/Failure method where false positive or negative information is related about performance in a task.
5. MIP's employing artificial stimulants such as drugs to induce physiological states relevant to certain emotions.

4.2 Best MIP for Experiment

The authors decided that the MIP best suited to the experiment was MIP 4. Kehrein [6] carried out experiments using isolation booths and a cooperative task. In this case a Lego construction was the task, with one party giving instructions to what was to be built and the other party following the instructions. By manipulating the Lego available and the time allowed the subjects could easily be hindered or aided in the attainment of their goal. Similarly Johnstone [33] used computer games in order to induce real emotional

states in test subjects. Johnstone contends that computer games can be changed and manipulated in order to induce the desired emotions. A cooperative, task based MIP, along with isolation booths, provides the most controlled and natural situation possible in order to elicit real natural emotions.

5 Experimental Procedures

It is argued that the only truly effective method of obtaining quality audio assets is under inducement within laboratory conditions. This provides the essential advantage of full control over recording quality, as was demonstrated by Kehrein [6] using isolation booths. The argument for professional audio recording quality is also bolstered by the need for the creation of a comfortable experimental environment. Participants cannot fairly be expected to elicit natural reactions when faced with an unfamiliar environment, and thus the use of audio recording studios (as with music) would serve only to induce either performance (as with broadcast) or reticence on the part of the participant. For this reason, the best solution is the use of an acoustically controlled space (such as isolation booths) which can be specified as an environment for sensory control. Many experiments [26-28, 34] have been carried out using emotional induction, and it is argued that a controlled, isolated environment is the best method of relaxing a participant as required.

The basic set-up of the experiment has two participants, each in an isolation booth, engaged in a cooperative task. This task could be a form of isolated competition using a video game [33] as an impetus for participation. There would be two-way communication between the booths using microphones and headsets. The participants will be able to hear but not see each other and each participant will be observed and recorded (video and audio) by the researcher who is external to the two booths. Video and audio recordings will be made of the two participants. This approach has two main advantages: (1) though it is a dialogue we have each participants contribution as a separate audio stream, thus we achieve a segregated dialogue allowing us to analyse each speaker's audio signal without interruption from the other participant. (2) The audio from each speaker is free from any outside interference or external noise, a clean signal is achieved making it easier to analyse and listen to. While the participants are aware that they are being observed they will be told the experiment is to test a new piece of software or to assess reaction times (depending on the type of task they are asked to carry out)

The researcher involved will be able to hinder or aid the participants by changing various parameters of the test (such as time limits, equipment performance or materials provided); in this way emotional states can be induced and the resulting audio dialogue cleanly recorded. We must also consider variations on the experiment, which could conceivably be a reaction time assessment involving students performing tasks at a computer terminal in financial competition (best performance wins a predetermined sum). Various factors (such as poor software performance) could be coded into the test routine, thus engendering emotions such as frustration, anger and boredom. Conversely, the software could also be configured to indicate faster reaction times (or better performance) for certain participants to cultivate feelings of achievement, confidence, arrogance or happiness. We must also consider replicating Kehrein's experiment [6], using cooperative tasks based around the construction of Lego structures.

Although ideal experimental scenarios have been considered, it is also best practice to plan for contingencies that could arise from the use of non-ideal environments. Tasks could be performed in poor acoustic environments (such as computer laboratories) which would then subsequently be assessed in better surroundings. With such a compromise, all participants could also be asked questions about their performance post-test, with all responses being recorded. Although participants would be unaware that they were being emotionally observed, the responses would be genuine answers to questions related to performance or software usability. In this manner, ethical issues of surreptitious observation could be largely avoided while still maintaining a fairly unobtrusive monitoring of a subjects speech.

6 Audio Recording Quality

The specification of suitable equipment for the recording and analysis of audio assets is an often overlooked aspect of emotional speech analysis. Few examples in the literature make any mention of the audio equipment involved and thus an acoustic perspective towards speech recordings is a glaringly overlooked issue. It is suggested that high sample rates and bit depths are *essential* for the effective analysis of speech audio assets, as lack of resolution in either could potentially omit crucial acoustic features in the signal.

6.1 Audio Quality

Audio assets should be recorded at 24bit/192Khz [35] quality wherever possible. The accurate and rigorous recording of audio assets has been a glaring omission in much of the speech analysis field, notably when acoustic parameters are the focus of scrutiny. Assets obtained in poor quality environments or with sub-standard equipment cannot be claimed as anything other than working examples. It is strongly suggested that all audio assets obtained for the purposes of analysis should be of the highest possible recorded quality, which at present specifies a 24 bit/192Khz level file format. It is not sufficient to standardise on CD quality audio (16 bit/44.1Khz) purely on the grounds that it is a common benchmark.

6. 2 Recording Environment

The recording environment should be as acoustically isolated as possible, with the reduction of sound waves from other sources (and also as reverberations from the original source) being the primary goal. The transduction of a direct sound in isolation is practically near to impossible, though major steps can be taken towards this with careful consideration and planning. A suitable isolation booth can be obtained pre-constructed from many manufactures, offering typical reductions of 60dB at some frequencies. This method allows robust manufacturing processes to be utilised in the creation of suitable materials for an isolation booth, without recourse to manufacturing them within the course of this work. The use of isolation booths fulfils both the cognitive and acoustic requirements of this project, providing means by which robust and defendable emotion speech assets may be fairly obtained.

7 Evaluation by Listening Tests

In order to accurately rate the emotional content of the audio clips obtained from the experiment the authors will construct a web based, on-line interface, in order to carry out listening tests using the feeltrace tool [23] for evaluation purpose. Users will be able to log on and listen to the audio clips and use the feeltrace tool to define what they believe the emotional content of the clip to be, with the results being stored in a backend database. This allows many users to participate thus creating a robust statistical analysis of the audio clips. This statistical approach to evaluation allows a robust corpus of high quality emotional speech assets, graded by their emotional dimensions, to be created. Participants used in listening tests do not take part in any experimental task related to obtaining emotional speech assets, and the statistical compilation of all evaluations is used to specify emotional dimensions for each audio asset within the corpus. Once specified, an asset can thus be analysed for acoustic features which will potentially form the basis of a set of rules for basic emotional speech dimensions. While the original audio assets will be of a very high quality they will be stored in such a manner that they can be used effectively for web streaming.

8 Conclusions

In this paper we outline an experimental procedure for obtaining real, high quality, emotional audio assets. We also argue that existing speech corpuses are not of high enough quality and that little or no consideration has been given to the audio quality of emotional audio assets. By considering: (1) the main theoretical perspectives on emotion and the theory of full-blown and underlying emotion, (2) Mood Induction Procedures, (3) audio quality of existing corpuses and the *reality* of emotion in these corpuses and (4) existing experimental design using isolation booths [6] we have developed a robust, yet flexible experimental procedure that will ideally yield high quality emotional assets. These assets can then be assessed by listening tests using the feeltrace tool, to determine the emotional dimensions present within them. We can then proceed to analyse these clips to determine the acoustic parameters concerning emotional dimensions in speech.

Acknowledgment

The authors wish to thank Charlie Pritchard of the DMC and the Salero Project, a European funded research project into the use of ‘intelligent content’ in multimedia production which supports this research.

References

1. Scherer, K.R., *Emotion as a multicomponent process: A model and some crosscultural*

data. Review of Personality and Social Psychology, 1984. **5**: p. 37-63.
2. Schroeder, M., *Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*, in *Faculty of Philosophy*. 2004, Universit¨at des Saarlandes. p. 288.
3. Roach, P., et al., *Transcription of Prosodic and Paralinguistic Features of Emotional Speech.* Journal of the International Phonetic Association, 1998(28): p. 83-94.
4. Campbell, N. *Databases of emotional speech*. in *ISCA Workshop on Speech and Emotion*. 2000. Northern Ireland.
5. Gerrards-Hesse, A., K. Spies, and F.W. Hesse, *Experimental inductions of emotional states and their effectiveness: A review.* British Journal of Psychology, 1994. **85**(1): p. 55-78.
6. Kehrein, R. *The prosody of authentic emotions*. in *Speech Prosody*. 2002. Aix-en-Provence, France.
7. Watkinson, J., *Art of Digital Audio, Third Edition*. 2000, Burlington, MA: Focal Press.
8. Howard, D. and J. Angus, *Acoustic and Psychoacoustics*. 1999: Focal Press.
9. Cowie, R., et al. *'FEELTRACE': An instrument for recording perceived emotion in real time*. in *ISCA Workshop on Speech and Emotion*. 2000. Northern Ireland.
10. Cowie, R., et al., *What a neural net needs to know about emotion words*, in *Computational Intelligence and Applications*, N. Mastorakis, Editor. 1999, World Scientific & Engineering Society Press. p. 109-114.
11. Banse, R. and K.R. Scherer, *Acoustic profiles in vocal emotion expression.* Journal of Personality and Social Psychology, 1996. **70**(3): p. 614-636.
12. Lewis, M. and J.M. Haviland, *Handbook of Emotions*, ed. M. Lewis and J.M. Haviland. 1993, New York: Guilford Press.
13. Lazarus, R.S., *Stress and Emotion: A new synthesis*. 1999, New York: Springer.
14. Ekman, P., *Basic emotions*, in *Handbook of Cognition & Emotion*, T. Dalgleish and M.J. Power, Editors. 1999, John Wiley: New York. p. 301-320.
15. Buck, R., *Biological affects: A typology.* Psychological Review, 1999(106): p. 301-336.
16. Cornelius, R.R., *Theoretical approaches to emotion.* SpeechEmotion-2000, 2000. **1**: p. 3-10.
17. Russell, J.A., *A circumplex model of affect.* Journal of Personality and Social Psychology, 1980(39): p. 1161-1178.
18. Schlosberg, H., *A scale for the judgement of facial expressions.* Journal of Experimental Psychology, 1941. **29**: p. 497-510.
19. Plutchik, R., *Emotion: A psychoevolutionary synthesis*. 1980, New York: Harper and Row.
20. Cornelius, R.R., *The Science of Emotion*, in *Research and Tradition in the Psychology of Emotion*. 1996, Prentice-Hall: Upper Saddle River, NJ.
21. Osgood, C.E., G.J. Suci, and P.H. Tannenbaum, *The measurement of meaning*. 1957, Urbana, USA: University of Illinois Press.
22. Scherer, K.R., *Emotion as a multicomponent process: A model and some crosscultural data.* Review of Personality and Social Psychology, 1984. **5**: p. 37-63.
23. Cowie, R., et al., *Emotion recognition in human-computer interaction.* IEEE Signal Processing Magazine, 2001. **18**(1): p. 32-80.
24. Douglas-Cowie, E., R. Cowie, and M. Schröder. *A new emotion database: considerations, sources and scope*. in *ISCA Workshop on Speech and Emotion*. 2000. Northern Ireland.
25. Higuchi, N., T. Hirai, and Y. Sagisaka, *Effect of speaking style on parameters of fundamental frequency contour*, in *Progress in speech synthesis*, J.v. Santen, et al., Editors. 1997, Springer-Verslag: New York. p. 417-428.
26. Picard, R.W., E. Vyzas, and J. Healey, *Toward Machine Emotional Intelligence: Analysis of Affective Physiological State*. IEEE Trans Pattern Analysis & Machine Intelligence, 2001(23): p. 1175-1191.
27. Gross, J.J. and R.W. Levenson, *Emotion elicitation using films.* Cognition and Emotion, 1995(9): p. 87-108.
28. Iida, A., N. Campbell, and M. Yasumura, *Design and Evaluation of Synthesised Speech with Emotion.* Journal of Information Processing Society of Japan, 1998. **40**(2): p. 479-486.
29. Chung, S. *Vocal expression and perception of emotion in Korean.* in *14th International Conference of Phonetic Sciences*. 1999. San Fransisco, USA.
30. Scherer, K.R. and G. Ceschi, *Lost luggage emotion: A field study of emotion-antecedent appraisal.* Motivation and Emotion, 1997(21): p. 211-235.
31. Douglas-Cowie, E., et al., *Emotional speech: towards a new generation of databases.* Speech Communication Special Issue Speech and Emotion, 2003. **40**(1-2): p. 33–60.
32. Gerrards-Hesse, A., K. Spies, and F.W. Hesse, *Experimental inductions of emotional states and their effectiveness: A review.* British Journal of Psychology, 1994(85): p. 55-78.
33. Johnstone, T., et al., *Affective speech elicited with a computer game.* Emotion, 2005(5): p. 513-518.
34. Fernandez, R. and R. Picard. *Modelling drivers' speech under stress*. in *ISCA Workshop on Speech and Emotion*. 2000. Northern Ireland.
35. Katz, B., *Mastering Audio: The Art and the Science*. 2002, Burlington, MA: Focal Press.