2009

Emotional Speech Corpus Creation, Structure, Distribution and Re-Use

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

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

Emotional Speech Corpus Creation, Structure, Distribution and Re-Use.

Brian Vaughan [1], *Charlie Cullen* [2]

[1] Digital Media Center, Dublin Institute of Technology, Aungier Street, Dublin 2.

brian.Vaughan@dit.ie charlie.cullen@dmc.dit.ie

Abstract

This paper details the on-going creation of a natural emotional speech corpus, its structure, distribution, and re-use. Using Mood Induction Procedures (MIPs), high quality emotional speech assets are obtained, analysed, tagged (for acoustic features), annotated and uploaded to an online speech corpus. This method structures the corpus in a logical and coherent manner, allowing it to be utilized for more than one purpose, ensuring distribution via a URL and ease of access through a web browser. This is vital to ensuring the reusability of the corpus by third party's and third party applications.

Index Terms: speech corpora, annotation, distribution, corpus interface, re-usability.

1. Introduction

Creating applications that make use of speech corpora is dependent on their availability and the overall structure and annotation strategy used to create them. While corpora are usually purpose built, the amount of time and effort required to create said corpora makes their re-use highly desirable. This is hampered by a lack of widely accepted standards regarding audio quality, metadata annotation, emotional definitions and overall corpus structure [1]. This paper details on-going PHD work in the creation of a natural emotional speech corpus using a task based Mood Induction Procedure (MIP) and discusses how the speech assets within the corpus are analysed, tagged and annotated. Work has been carried out on the design of an MIP, a tagging and annotation strategy, software to carry out the strategy and the creation of a database backend and front-end. Current work is focusing on refining and improving this process, as well as analysing the data to determine an emotional rule-set for the detection of emotion in speech.

2. Corpus Creation

A number of key areas were considered when creating the speech corpus:

- The type of speech asset within the corpus (acted, simulated or natural).
- The audio quality of the assets.
- The method of acoustic or linguistic analysis (depending on the type of corpus being created)
- The structure of the corpus: the method of uploading the assets and distribution/access to the corpus.

2.1. Type of Asset

Cowie et.al [2] have compiled a list of key databases for emotion research. The majority of the corpora were created using acted emotions, with the rest using induced and 'natural' emotion. This presents a number of problems to the field. Acted emotion is a form of simulated emotion, which also includes emotional recall, imagined events and read texts (Velten MIPs etc) [1]. The use of acted emotion is potentially problematic. Little is known about how acted emotion compares to natural spontaneous emotion[3]. It has further been argued that there are physiological aspects to emotion [4-7] inducing uncontrollable changes in speech [8] that reflect the underlying emotion. The presence of these uncontrollable changes may not be present in acted speech. Natural emotional speech usually has a degree of spontaneity, while acted speech is a thought out, voluntary expression of emotion. This may well render acted speech an unrealistic facsimile of natural emotional speech [1].

What is usually termed natural emotion is usually obtained from TV broadcast sources, comprising of clips from reality TV shows, talk shows, interviews etc: any televised program where 'natural' emotion is judged to be present. As with the use of acted emotion this has a number of potential flaws. Arguably, any broadcast is a performance (similar to acting), as the speakers and participants are constantly aware of the presence of recording equipment. Anthropological research recognises that the presence of the researcher and equipment distorts the reality of the situation, causing people to feel constrained and act differently [9, 10]. This renders the televised displays of emotion highly dubious. Furthermore there is an inherent perceptual bias to the whole recording process on the part of the director, cameraman, producers and editors (to name just a few involved in the overall process). It is almost impossible to know or account for this subjective bias, and it is therefore almost impossible to judge the veracity of televised emotional displays. Additionally, the audio quality of broadcast material is not of uniform quality with a lack of consistency across broadcasts.

2.2. Mood Induction Procedures as a Source of Natural Emotion

The use of induced emotion is a useful approach, holding out the promise of authentic natural speech being obtained. Researchers have had great success utilizing Mood Induction Procedures (MIPs) to elicit emotional responses from participants [11-16]. While some MIPs suffer the same problems as assets obtained using actors and from broadcast sources, MIPs can be an ideal method for obtaining natural emotional speech assets. In particular, the success/failure MIP has distinct advantages: the true nature of the experiment can easily be concealed, avoiding demand effects by focusing the participants on the task they have been set. False-positive or false-negative feedback can be used to manipulate the experiment along with subtle external manipulation. Kehrin successfully used sound-booths and a Lego based success/failure MIP to elicit emotional responses from

participants [14]. This method allows for natural emotional speech to be obtained at a high quality level, something which is an oft overlooked consideration.

2.3. MIP Experimental Design

Building upon Kehrins basic design, a number of experimental designs were initially tested as part of a wider case study [17]:

1. A Tetris based design
2. A replication of the Lego based task
3. The use of gaming consoles (Xbox 360's)

All the experiments used two separate sound booths and high quality audio equipment, recording the two participants on separate independent audio tracks at a rate of 192Khz/24Bits.

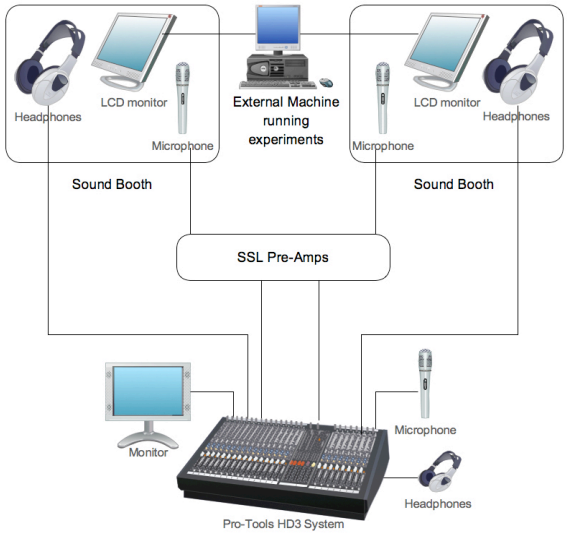

Figure 1: Diagram of the recording and MIP experimental setup.

While these designs proved successful in eliciting emotional responses, the Tetris task was prone to demand effects, the Lego task had only a minimum level of manipulation and the gaming console experiment yielded a very wide a range of emotional responses.

Therefore, another task based MIP has been designed to elicit a more focused and specific range of emotional responses, to avoid demand effects and to be easily manipulated. This MIP consists of two participants being given an imaginary shipwreck survival scenario task to complete. 15 items are shown on a screen in each sound booth, and participants are told that they have to correctly rank them to achieve a total score of 15. They receive points for correctly ranking items and lose points for incorrect rankings. A target score of 15 is displayed on screen throughout the experiment along with a 'Your Score' that changes as items are ranked. A ten-minute countdown timer is constantly displayed with the scores and a reward of twenty Euro is offered if the 15 items are correctly ranked before the timer reaches zero.

As the participants rank the items, the value in the 'Your Score' box changes. However the scores displayed are part of an overall scoring pattern: no matter what choices are made in each experiment, the same pattern of scores is used. The score changes every time a ranking choice is made leading the participants to believe it is in direct response to their choices. The use of a scoring pattern allows the experimental manipulation to be standardised across all the experiments. This ensures that differences in responses are attributable to the participants and not the method of manipulation. A simple application was created using Adobe Flex [18] to run the experiment and with external manipulation.

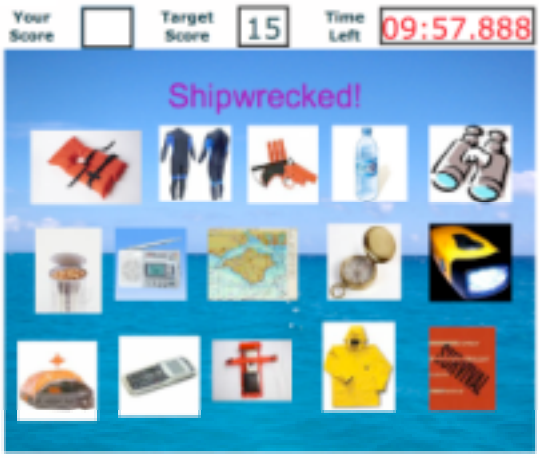

Figure 2: Screenshot of the MIP experiment

The 15 items that are ranked remain static on the screen throughout the experiment; participants are told to remember the ranking choices between them. This is done for two reasons:

1. It puts them under pressure as they have to keep track of what order they have the items in, constantly talking to each other and thus keeping a continuous dialogue going. One of the problems with the console gaming MIP [17] was the lack of conversation and dead-air: at times subjects lapsed into quiet concentration. Keeping the participants talking is important to ensure that as much of the ten-minute audio recording contains dialogue that can be used for emotional speech assets.

2. It makes it harder to perceive the external manipulation. Since the scores are unrelated to the choices made it might have been possible to discern this. If the 15 items changed order on screen accordingly, it would have been relatively easy to spot discrepancies in the scores given when choices were made. By making the participants remember their choices, they attribute the different score to not remembering the order properly or to the complex scoring system. At the very least they have to constantly communicate to verify the ranking choices and make changes in an attempt to get a higher score.

3. NASA TLX Tests

Participants complete a brief modified NASA TLX test after each experiment:

"NASA-TLX is a subjective workload assessment tool. NASA-TLX allows users to perform subjective workload

assessments on operator(s) working with various human-machine systems" [19].

The TLX tests are designed to test various aspects of a task: mental demand, physical demand, temporal demand, performance, effort and frustration. For the purposes of this research the physical demand section has been omitted. A control group will complete the same MIP, without the monetary reward, without the timer and with no scoring at all. This control group will also complete TLX tests. This will provide a comparison to determine to what degree the MIP frustrated the original participants, thus eliciting dimensionally rated negative/active emotions.

4. Tagging and Annotation

Once collected, the assets are then processed using the Linguatag application [20]. Linguatag uses the PRAAT engine [21] to obtain low-level acoustic data from each asset. This data is outputted in an XML file, using the SMIL format [22, 23].The assets (as a WAV and MP3 file for listening tests) and corresponding XML files are uploaded to a database backend and annotated using the EAGLE/ISLE Metadata Initiative (IMDI) [24], as part of the upload process. The IMDI is the only standard for speech corpus metadata currently available. The annotation method allows for a highly structured and easily accessible speech corpus. The lack of standardization of speech corpora is a major obstacle to their usability and re-use. Applications that make use of speech corpora need easily accessible and properly structured corpora. While the IMDI schema is extensive, only parts of the schema need be implemented. In this case the Project, Session, Actor and Content tiers have been utilised. However further tiers can easily be added without effecting the overall structure of the corpus. The diagram below describes the current structure:

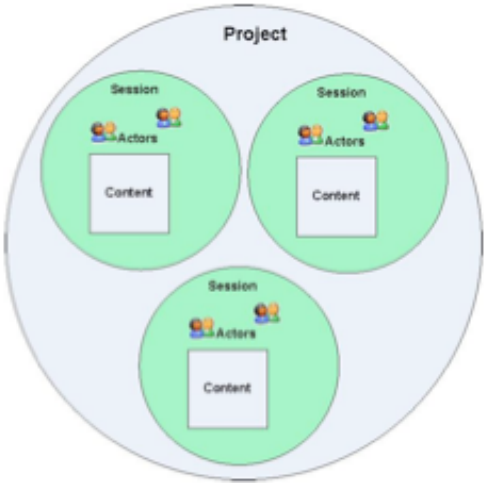

Figure 3: Representation of the IMDI schema organization, showing the current DIT Speech Corpus structure. (Reprinted from: [1])

The main advantage, aside from a consistent and logical hierarchical structure, is that assets can be queried for a number of different properties (prosodic values, emotional dimensions and speaker characteristics etc).

4.1. Database Backend

The corpus backend is constructed using Ruby-on-Rails [25] and MySQL [26] while the front end is browser based, using a combination of HTML and Adobe Flex [27]. This method of construction allows the corpus to be easily accessed by third party applications utilising readily available Web 2.0 technologies/APIs.

5. Determining the Emotional Content of the Speech Assets

Once the assets have been collected, tagged, annotated and uploaded, they are rated for their emotional dimensions. This is achieved through the use of an online listening tool using a modified circumplex model [28]. The tool pulls 10 random assets from the corpus (in the form of Mp3s) and plays them one at a time. The user has two separate sliders to rate the clip: a Negative-Positive slider and a Passive-Active slider. Once they have been adjusted accordingly a RATE button is pressed to store the rating and play the next clip. There is also the option to not rate a clip by pressing a DO NOT RATE button. The rating values are written back to the corpus for future analysis. The rating tool is accessed using a URL, opening in the users browser and was created using Adobe Flex [18]. This method allows a large amount of people to rate the assets in the corpus. In this way a robust statistical definition of assets can be built up over time; clips with a high amount of similar ratings can then be analysed to determine common acoustic elements and values. The ratings are stored in the database backend alongside the acoustic and IMDI data and can be searched and utilised as easily as the other data.

6. Asset Analysis

Once enough assets have received a large amount of ratings, the acoustic properties (measured using Linguatag/PRAAT) will be analysed to determine if there are acoustic similarities between assets with the same/similar rating. A small subset of highly rated assets will initially be analysed. This will be achieved manually by running a number of queries on the database. These queries can be used to search similarly rated assets for certain prosodic properties (intensity, pitch, pitch range, intensity range, number of vowels, average vowel length etc), indicating a possible a correlation between the ratings and the acoustic properties. Following on from this, WEKA machine learning algorithms [29] will be used to process assets to determine if this correlation exists for a wider data set, as well as discerning possible patterns within the prosodic content of similarly rated assets. This is the final focus of the PHD research.

7. Distribution and Re-use

One of the biggest problems with corpora is that they are usually designed and created for a specific purpose, thus limiting their usefulness and re-usability. Advances in the field of emotion and speech synthesis increasingly depend on large, accessible datasets. The overall corpus creation strategy is just as important as the assets it contains. While the asset collection strategy is important to obtain authentic emotional speech, the corpus construction strategy is important to ensure that the data within is easily distributable and re-usable in various different ways.

The overall web based methodology of the discussed corpus allows it to be accessed through a web browser as well as allowing it to be utilised by various APIs. The Adobe Flex listening tool is one such example, plugging

into the corpus using a few lines of code and utilising the data for a specific purpose. Another example is a prototype corpus visualisation interface that is currently being implemented, allowing the complex data to be visualised in an easily understood manner. Numerous other Web 2.0 APIs can connect in a similar manner, while the Ruby-on-Rails/MySql backend can be scaled and modified as needed.

8. Conclusions

This paper discussed and considered a number of different elements related to the creation of a natural emotional speech corpus:

- The type of speech asset,
- The tagging and annotation of the collected assets,
- The creation and structure of a database to contain the assets,
- The rating and analysis of the emotional dimensions of the assets
- The distribution and the re-use of the corpus.

MIPs provide an ideal method for obtaining authentic emotional speech assets. The use of high quality audio equipment allows the assets to be captured at a very high sampling and bit-rate (192Khz/24 Bit) ensuring clarity and a low noise level. Creating and structuring the speech corpus using the IMDI schema allows for ease of distribution and use by third parties. The information and assets contained in the corpus can be easily accessed through the existing online interface or using readily available Web technologies and APIs. This allows the corpus to be used in a number of different ways and by a number of different applications. The emotional rating tool and corpus visualisation prototype are two such examples. The emotional rating tool is used to gather a large amount of ratings to determine the most statistically robust emotional assets. These can then be analysed to determine a speech based emotional rule set. The new MIPs are still ongoing, but so far results have been promising. The next stage of the research will focus on analysing the assets manually and using machine learning algorithms.

9. Acknowledgements

This work was funded by the SALERO European project. Special thanks to Evin McCarthy.

10. References

[1] Cullen, C., Vaughan, B., Kosidis, S., *Emotional Speech Corpus Construction, Annotation and Distribution*, in *The 6th edition of the Language Resources and Evaluation Conference.* 2008: Marrakech (Morocco).

[2] Roddy Cowie, E.D.-C., Cate Cox, *Beyond emotion archetypes: Databases for emotion modelling using neural networks.* Neural Networks, 2005(18): p. 371-388.

[3] 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.

[4] Frijda, N.H., *The Laws of Emotion.* American Psychologist, 1988. **43**(5).

[5] Bindra, D., *A unified interpretation of emotion and motivation.* Annals of the New York Academy of Science., 1969. **159**: p. 1071-1083.

[6] James, W., *What is an emotion?* Mind, 1884. **9**: p. 188-205.

[7] Darwin, C.R., *The Expression of the Emotions in Man and Animals*. 1872, London.: Albermarle.

[8] Johnstone, T. *Emotional Speech Elicited Using Computer Games*. in *4th International Conference on Speech and Language Processing*. 1996. Philadelphia, PA, USA.

[9] Gottdiener, M., *Field Research and Video Tape.* Sociological Inquiry, 1979. **4**(49): p. 59-66.

[10] Geer, B.a.H.S.B., *Participant Observation and Interviewing: A Comparison.* Human Organization, 1957. **16**(3): p. 28-32.

[11] 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.

[12] Gross, J.J. and R.W. Levenson, *Emotion elicitation using films.* Cognition and Emotion, 1995(9): p. 87-108.

[13] 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.

[14] Kehrein, R. *The prosody of authentic emotions*. in *Speech Prosody*. 2002. Aix-en-Provence, France.

[15] Fernandez, R. and R. Picard. *Modelling drivers' speech under stress*. in *ISCA Workshop on Speech and Emotion*. 2000. Northern Ireland.

[16] Johnstone, T., et al., *Affective speech elicited with a computer game.* Emotion, 2005(5): p. 513-518.

[17] Vaughan, B., Kosidis, S., Cullen, C., Wang, Yi. *Task-Based Mood Induction Procedures for the Elicitation of Natural Emotional Responses*. in *The 4th International Conference on Cybernetics and Information Technologies, Systems andApplications: CITSA 2007* 2007. Orlando, Florida.

[18] Adobe. *Adobe Flex Home Page*. 2009 [cited; Available from: http://www.adobe.com/products/flex/.

[19] (NASA), N.A.a.S.A. *NASA TLX: Task Load Index*. 2009 [cited; Available from: http://humansystems.arc.nasa.gov/groups/TLX/.

[20] Cullen, C., Vaughan, B., Kosidis, S., *LinguaTag: an emotional speech analysis application*, in *Accepted paper at: The 12th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2008*. 2008: Orlando, Florida, USA.

[21] Boersma, P. and D. Weenink, *Praat: doing phonetics by computer*. 2006.

[22] SMIL, W.W.W.C. *Synchronized Multimedia*. 2008 [cited; Available from: http://www.w3.org/AudioVideo/.

[23] XML, W.W.W.C. *Extensible Markup Language*. 2008 [cited; Available from: http://www.w3.org/XML/.

[24] ISLE. *Title: IMDI (ISLE Metadata Initiative), Metadata Elements for Session Descriptions*. 2003 [cited 2008].

[25] Hansson, D.H. 2009 [cited; Available from: http://rubyonrails.org/.

[26] Microsystems, S. *Java Speech API Programmer's Guide*. 1998 [cited; Available from: http://java.sun.com/products/java-media/speech/forDevelopers/jsapi-guide/index.html.

[27] Adobe. *Adobe Labs*. 2008 [cited; Available from: http://labs.adobe.com/technologies/flex/.

[28] Plutchik, R., *Emotion: A psychoevolutionary synthesis*. 1980, New York: Harper and Row.

[29] Wakiato, T.U.o. *Weka Machine Learning Project*. 2009 [cited January 2009]; Available from: http://www.cs.waikato.ac.nz/ml/weka/.