2008-05-26

Emotional Speech Corpus Construction, Annotation and Distribution

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

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

Spyros Kousidis
Technological University Dublin, spyros.kousidis@tudublin.ie

See next page for additional authors

Recommended Citation

Vaughan, B. et. al. (2008) Emotional speech corpus construction, annotation and distribution. *Corpora for research on Emotion & Affect, LREC 2008 conference,* Marrakesh, Morocco. 28-29-30 May.

Funder: Salero Project

2008-05-26

Emotional speech corpus construction, annotation and distribution

Brian Vaughan
Dublin Institute of Technology, brian.vaughan@dit.ie

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

Spyros Kousidis
Dublin Institute of Technology, spyros.kousidis@dit.ie

John McAuley
Dublin Institute of Technology, john@dmc.ie

Emotional Speech Corpus Construction, Annotation and Distribution

Dr. Charlie Cullen, Brian Vaughan, Spyros Kousidis, John McAuley
Digital Media Center, Dublin Institute of Technology, Aungier Street, Dublin 2, Ireland
E-mail: charlie.cullen@dit.ie, brian.vaughan@dit.ie, spyros.kousidis@dit.ie, john@dmc.dit.ie

Abstract

This paper details a process of creating an emotional speech corpus by collecting natural emotional speech assets, analysisng and tagging them (for certain acoustic and linguistic features) and annotating them within an on-line database. The definition of specific metadata for use with an emotional speech corpus is crucial, in that poorly (or inaccurately) annotated assets are of little use in analysis. This problem is compounded by the lack of standardisation for speech corpora, particularly in relation to emotion content. The ISLE Metadata Initiative (IMDI) is the only cohesive attempt at corpus metadata standardisation performed thus far. Although not a comprehensive (or universally adopted) standard, IMDI represents the only current standard for speech corpus metadata available. The adoption of the IMDI standard allows the corpus to be re-used and expanded, in a clear and structured manner, ensuring its re-usability and usefulness as well as addressing issues of data-sparsitiy within the field of emotional speech research.

1. Introduction

Advances in both speech/emotion recognition and emotional speech synthesis largely depend on the availability of annotated, emotional speech corpora. Although it is common that corpora are purpose-built for specific applications or research purposes, it would be desirable to re-use existing corpora. However, there is a lack of widely accepted standards in such areas as audio quality, annotation with metadata in order to perform queries, as well as mutually agreed definitions, as in 'what is emotion?'(Cowie and Cornelius 2003). The work described here is a developing process of emotional asset acquisition, annotation and on-line publishing for emotional rating by end users, which attempts to address some of the above issues, while being flexible in practical issues such as re-usability, standardisation and access. The paper is divided into three parts: (1) A method for obtaining "genuine" emotional speech recordings, namely Mood Induction Procedures (MIP 4) (Gerrards-Hesse, Spies et al. 1994), while recording in a controlled environment; (2) the analysis and annotation of the recorded assets via a purpose-built audio analysis tool (Cullen 2008) and (3) an implementation of the IMDI corpus annotation schema.

2. Genuine Emotional Speech

There are three main forms of asset used in existing speech corpora: simulated assets, broadcast assets and induced assets. A few examples of claimed 'natural' emotional speech databases exist (Scherer and Ceschi 1997;2000; Chung 1999; Douglas-Cowie, Campbell et al. 2003) although the justification for such content is that it is more natural when compared with simulated content. In the majority of cases what is termed 'natural' emotional assets are obtained from a broadcast source (mainly television) (Douglas-Cowie, Campbell et al. 2003). However it can be argued that assets obtained from such sources may not be natural or contain genuine emotional content.

2.1 Simulated Assets

Corpora consisting of simulated assets use acted emotional states, read texts and imagined/recalled emotional situations (Banse and Scherer 1996; Enberg 1997; Amir 2000; Kienast 2000; Pereira 2000). However very little is actually known about how simulated emotion compares to natural emotion (Douglas-Cowie, Campbell et al. 2003). Simulated emotion that involves reading from a text is not a spontaneous expression of emotion with read speech having distinct characteristics from spontaneous speech (Johns-Lewis 1986), with vowel substitution and reduction being more likely to occur in spoken as opposed to read speech (Van Bael 2004). Emotional states can be considered to be an important factor in maintaining and negotiating social interaction and relationships (Cornelius 2000), communicating information about our intentions and possible behaviour to those around us: they compel us to action and regulate social communication (Plutchik 2001). Simulated assets are often non-interactive (Banse and Schrerer 1996; Enberg 1997; Amir 2000; Kienast 2000; Pereira 2000), consisting of monologues with little or no interaction from other agents. The neglect of the social dimension of emotional speech means that obtained assets may contain only a limited range of emotions.

Numerous commentators have argued that there are fundamental biological and physiological aspects to emotion (Darwin 1872; James 1884; Bindra 1969; Frijda 1988; McGuire 1993) with Johnstone (Johnstone 1996) arguing that emotion can induce changes in speech that the speaker cannot control and that these changes reflect the underlying physiological changes taking place. It is debatable whether these uncontrollable changes are

present in simulated emotional speech, therefore, simulated emotions may well be nothing more than a resemblance of real emotional states (Pugmire 1994). Thus the voluntary and non-spontaneous nature of simulated emotion may undermine its authenticity and its suitability as a method of obtaining natural emotional speech assets.

2.2 Broadcast Assets

Some corpora use assets obtained from broadcast sources, mainly television (Chung 1999; Douglas-Cowie, Campbell et al. 2003), the justification being that they are 'natural' compared to simulated assets (Douglas-Cowie, Campbell et al. 2003). Some of the problems associated with the use of simulated assets are also of concern in using broadcast assets. Furthermore, it can be argued that any broadcast is a performance, as the speakers are usually very aware of the recording process taking place. It is recognised in anthropological research that the presence of a researcher and equipment may cause people to act differently or even feel constrained in what can be said and done (Geer 1957; Gottdiener 1979). It is possible that this distortion and constraint means that televised emotional displays, like simulated emotion, may only be a facsimile of real emotion. The only way to prevent this distortion is to conceal the equipment and covertly record subjects; however this is a highly questionable practice and ethically unsound. The distorting effect may lessen over time as subjects become used to being recorded (Erickson 1982). This would suggest that it would be more relevant to use clips taken from the middle or towards the end of a televised program as opposed to clips taken from the start. However, there is an inherent perceptual bias to the recording process (Bellman 1977). This perceptual bias is inherent in the subjective decisions of the cameraman, the director, the producers and the editor and it cannot be known how this affects the final outcome of a broadcast piece.

2.3 Audio Quality

Assets taken from broadcast sources can be of varying audio quality, as 'broadcast quality' is a term rather than a definition; one cannot assume that assets obtained from broadcast sources are of uniform quality. Audio quality will also vary depending on the nature of the program, whether it is recorded in a studio or outside in public spaces (as many reality television programs are). Various other factors will affect the audio quality: noise from studio audiences, people talking across each other and environmental noise from outside broadcasts. The equipment used will also affect the sound quality: different broadcast situations may use different recording apparatus (microphones, cameras etc) and methods. The greatest single advantage of simulated assets is the potential for control of the recording environment, such that most simulated assets are obtained using studio equipment and conditions. The huge variation in recording quality found in other types of corpora (such as those using broadcast assets) precludes the definition of cohesive standards, and thus simulated assets are often preferred for this reason.

2.4 Natural Assets and Mood Induction Procedures

In order for assets to be considered natural for the purpose of analysis, the authors argue that they should be derived from non-simulated and non-broadcast sources with audio quality being of paramount importance. The induction of natural emotional responses in a laboratory environment, thus ensuring audio quality can be maintained, is achieved through the use of Mood Induction Procedures. Mood Induction Procedures (MIPs) are procedures that are designed to induce specific emotional states in a test subject within a controlled situation. The Success/Failure MIP (Forgras 1990; Henkel 2004) uses false feedback (positive or negative) concerning a subject's performance in a test that they believe is testing their cognitive ability. By placing subjects in a situation where certain needs are activated, such as the need to succeed at a certain task, frustrating or aiding the subject in the attainment of their need can induce emotional states. While other MIPs have been found to be more successful in some cases (Gerrards-Hesse, Spies et al. 1994), their effectiveness may be overestimated due to demand effects (Westermann 1996). Demand effects pose a problem to the validity of MIPs due to the fact that participants may guess the purpose of the procedure (to elicit emotional responses) and so pretend to be experiencing the desired emotion. Any instruction given regarding required emotional states can cause a demand effect. The Success/Failure MIP avoids the creation of demand effects: the true nature of the experiment is not evident and can be further disguised if needed. Participants are engaged in a task and can be led to believe that the completion of the task is the purpose of the experiment. The use of false feedback, either positive or negative, further conceals the true purpose of the experiment. The use of a task based Success/Failure MIP may remove the subjective nature associated with some other MIPs, and allows the researcher to control and manipulate the experiment in greater detail. By frustrating or aiding the subjects in their task, without their knowledge, they can be guided towards natural negative or positive emotional states without being aware that a certain emotional state is required, thus avoiding the creation of demand effects.

2.4.1. MIP Audio Quality

The use of Mood Induction Procedures to stimulate emotion has the potential for the same recording conditions to be applied as with simulated assets. The difficulties associated with such conditions using MIPs are related to the concealment of recording equipment to avoid revealing the true purpose of the experiment prior to commencement. In Kehrein's experiment (Kehrein 2002), the fact that the participants were seated in separate sound proofed rooms, allowed the conversational interaction to be recorded as two separate high quality audio channels. This allowed both sides of the conversation to be analysed, including overlaps. Participants were aware of the presence of the recording equipment but believed it was used for them to communicate with each other.

A task-based MIP offers a high degree of control, either hindering or aiding participants, while the use of

separate sound proofed rooms enables high quality audio assets to be obtained. This approach ensures that obtained assets are natural, compared to simulated and broadcast assets, while the co-operative nature ensures the social aspect of emotional expression is not neglected. The resulting emotional assets can be claimed to be natural and spontaneous, arising out of the manipulation of the task and the interaction of the participants as opposed to voluntary or knowingly coerced attempts to generate emotional states.

2.5 MIP Experiment

Taking into consideration the arguments presented above, an MIP was devised using modern games consoles and games, in conjunction with sound isolation booths (Vaughan 2007).The main advantage of these console systems is that a large amount of the games are usually designed with extensive multiplayer options that are cooperative and/or competitive in nature. Computer games have been used before by Jonhnstone (Johnstone, Reekum et al. 2005) and as far back as 1978 by Isen et. Al (Isen 1978) to elicit emotional responses. Johnstone in particular noted that they are particularly suited to this task due to the fact that they can easily be changed and manipulated in order to suit the experiment. Little or no external manipulation is necessary as modern game design focuses very much on immersing the gamer in the gaming world while the majority of games have been designed to be competitive and challenging, usually with an emphasis on competitive goal achievement. The overall game play and style of most games is therefore conducive to inducing emotional states in participants. External manipulation can be achieved, where needed, through unplugging a participant's game controller, changing the time limit, giving false information regarding the amount of time left or through offering a cash or material reward (for elated states).

Participants in the experiment are aware they are being recorded, all must sign a consent form giving permission for the recordings to be used for research purposes. However the true nature of the experiment is not revealed; participants are led to believe that competitive gaming or in game communication is being studied, thus minimizing the chance of a demand effect manifesting. The audio is recorded at 24bit/9Khz and once enough recordings have been attained, it is then analysed and annotated.

3. Analysis and Annotation

The audio recordings obtained from the gaming MIP are segmented into short phrases/clips. At present this is done by hand using a digital audio editor. However it is envisaged, and work is being undertaken in this area, that this will eventually be a semi-automatic or automatic process. These audio clips are then processed using the LinguaTag application in preparation for inclusion in the speech corpus.

3.1 LinguaTag

LinguaTag (Cullen 2008) is a purpose-built application, has been developed for the acoustic analysis and first stage of annotation (tagging) of the corpus. LinguaTag is written in Eiffel (Software 2008) and makes use of the PRAAT (Boersma and Weenink 2006) engine to obtain low-level acoustic data from the recorded signal, while providing a user-friendly interface for transcription, segmentation, labeling and emotional rating of the sound clips . The low-level analysis includes automatic vowel identification, with pitch, intensity and formant contours, as well as voice quality measures (Johnstone and Scherer 1999; Gobl, Bennett et al. 2002) calculated for each vowel. Separate tiers in the annotation schema allow for acoustic analyses of larger clips, as well as the other linguistic annotations mentioned above. In addition, three sliders are available (pitch, intensity, duration) for setting thresholds for detection of stressed vowels. Emotional rating is performed using a circumplex model, comprised of two axes (activation and evaluation), adapted from Scherer (Scherer 1984; Vaughan 2007). LinguaTag outputs this data in a separate XML file, following the SMIL format (Consortium 2008; XML 2008). This XML file is then uploaded with the original WAV file and an MP3 file for use in future online listening tests.

3.2 IMDI Corpora

Consideration has been given to the annotation schema itself, as the existence of metadata is arguably as crucial as the content of the corpus: metadata can be used to query data in a corpus, thus expanding its usability and re-usability. Developing powerful emotional speech technology applications and in-depth analysis in emotional speech research require ever larger amounts of data, both to overcome problems such as data-sparsity (Xiao, Dellandrea et al. 2005) and to enable use of the most appropriate data available. Therefore, corpora need to be sufficiently annotated for such queries to be possible, and there has to be a standardisation of the annotation form, to allow for easy universal access.

Unfortunately, there is a lack of standardisation for annotating speech corpora, particularly in relation to emotive content. The only cohesive attempt at corpus metadata standardization performed thus far has been by the EAGLE/ISLE consortium (ISLE 2003), which has led to the development of the ISLE Metadata Initiative (IMDI). Although not a comprehensive (or universally adopted) standard, IMDI represents the only current standard for speech corpus metadata available. For this reason, it was decided to implement the IMDI standard within the speech corpus detailed in this paper in order to maintain as cohesive a standard as possible within current developments. The IMDI schema is extensive and so it was decided that initially only the four higher tiers of the schema (Project, Session, Actor and Content) would be implemented. It was felt that these were the most relevant elements of the schema to the corpus. This does not preclude the inclusion of other elements of the schema from being implemented at a later date should it be deemed necessary.

3.3.1. Implementation

The implementation of the IMDI annotation schema is structured as follows: A project groups together different bundles of sessions. A session is defined as the common bundle for linguistic events within IMDI metadata, and thus all speech assets are defined relative to a specific

session. This allows an audio clip to be taken from a longer recording for specific analysis, while still retaining the same overall metadata as all other files in that session bundle. Within each session, actors, i.e. participants in the recordings, are documented (with anonymity preserved at all times for ethical reasons) so that database queries involving geographical information or age can be performed. The content metadata relates to specific activities for a given session, such as the type of emotional content (induced, acted, etc) or other types of content categorization (the vocabularies are open for some of the tags). Finally, the asset metadata relates to the low-level acoustic information and the linguistic and emotional annotation that is performed by LinguaTag in SMIL format.

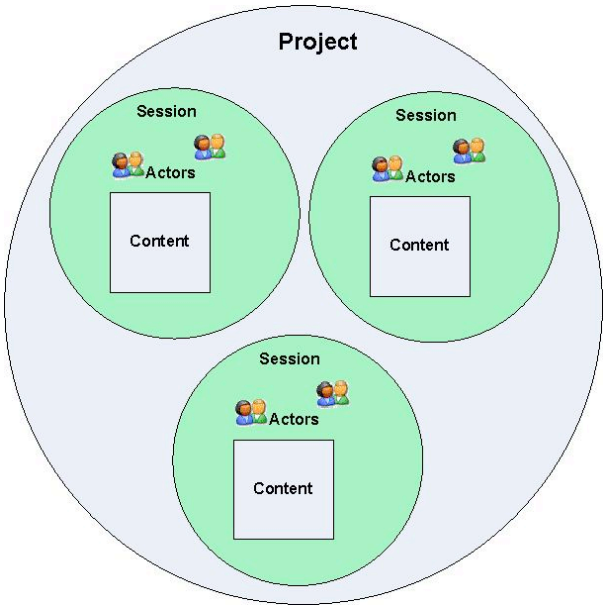

Figure 1: Example block diagram of the IMDI schema organisation. In this example, 3 separate session bundles are grouped logically under a single project.

This approach is advantageous in that the definition of a particular project allows various sessions to be grouped in a logical form. Thus, in the case of the emotional speech corpus described in this paper, all sessions are organised relative to the project. Grouping sessions logically, allows for future expansion of the speech database to include other corpora developed for different purposes. The session definition provides a convenient way to group assets for analysis, allowing assets taken from different experiments to be assessed either in isolation or within a wider common context. The definition of an actor(s) within a session is a very useful aspect of the IMDI standard, as it allows the various participants in a speech recording to be documented for later consideration. In many instances, actor details may be vague and non-specific to ensure that ethical standards are adhered to (this is given as an option for each testing participant). However, as mentioned more detailed actor information would be of use for certain types of queries. Future work may consider the multi-lingual definition of assets within a corpus for analysis, and so actor information would be crucial for this.

The annotation schema provides the flexibility of querying assets for different properties, such as speaker characteristics, emotional dimensions (Cullen 2006; Vaughan 2007) (e.g. 'only negative' or 'extremely active'), or certain audio quality. In addition, the overhead associated with tagging the audio files is greatly reduced by the use of automation functionality (auto-suggest) during the tagging process. Metadata previously entered can be reused, e.g. metadata that is the same for the whole session need only be entered once. Similarly, any metadata shared between any number of assets in any combination, need not be re-entered, as it is available through the autosuggest functionality.

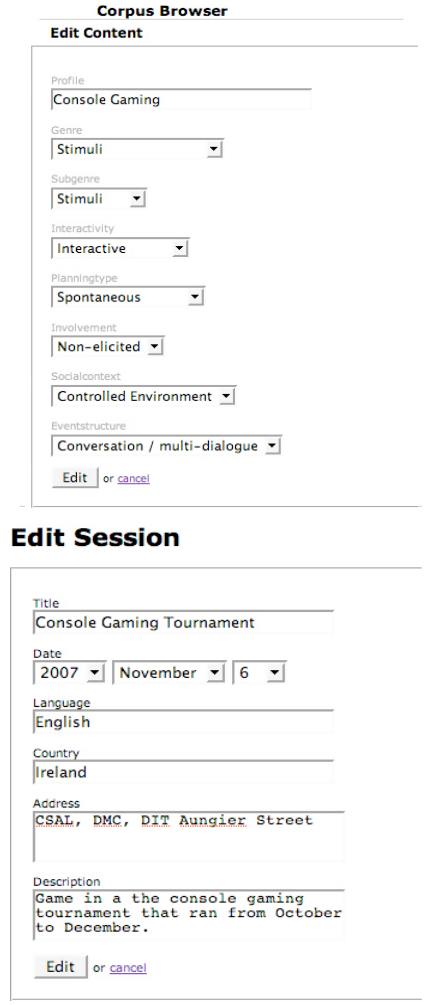

Figure 2: Screenshots of the Session and Content screens

There were, from the outset, several considerations that helped to define the technical architecture of the corpus. Firstly, the prototype must provide editors with the ability to insert assets, in the form of WAV files, and related Linguatag data, in the form of SMIL files. The prototype must parse the SMIL file and populate the corresponding database tables. The corpus, therefore, necessitates a storage layer or database as a persistent back-end. Secondly, editors require remote access to corpus assets. This allows for the addition, deletion and alteration of corpus assets and related metadata. At first, each asset was to be uploaded and annotated individually. However, following initial trials, it was decided to provide the ability for batch uploads, thereby

allowing an editor to upload several assets at the any one time. In this case, each asset is annotated with the same metadata.

4. Conclusion

This paper considered a method for obtaining natural emotional assets and annotating them as part of a speech corpus. MIPs were determined as the best method for obtaining natural emotional speech assets. A gaming based MIP experiment was developed to elicit natural emotional responses from participants. In order to analyse these assets an application, LinguaTag, was developed (Cullen 2008), providing a SMIL file with detailed acoustic information that can be parsed by a relational database. The IMDI corpus standardisation was adopted and implemented in order to annotate the assets and provide a clear and concise method by which the data could be structured in a 3-tiered system. This approach goes someway to avoiding data-sparsity and improving the inter-operability of the corpus.

At time of writing, the corpus contains over 650 fully annotated and tagged assets, and this figure is intended to grow. There is no defined headroom for the size of the corpus, but the experimental criteria, recording conditions and annotation metadata will be upheld in all future work. An on-line listening tool is also being developed and tested and will be the method by which on-line listening tests are carried out to rate the emotional dimensions of the assets. The intention of the online rating system is to obtain a statistical definition of emotional dimensions for each clip in the corpus, and rate each clip both in terms of its dimensional values and also the confidence rating for that clip. Thus, a clip which has been rated by more listeners will be defined as having a higher confidence level relative to its emotional dimension values, allowing statistical analysis to be performed on groups of assets in as robust a manner as possible.

5. Acknowledgements

This work was funded by the SALERO project. Special thanks to Charlie Pritchard and Evin McCarthy.

6. References

Amir, N., Ron, S., Laor, N. (2000). Analysis of an emotional speech corpus in Hebrew based on objective criteria. ISCA ITRW on Speech and Emotion, Belfast.

Banse, R. and K. R. Scherer (1996). "Acoustic profiles in vocal emotion expression." Journal of Personality and Social Psychology **70**(3): 614-636.

Banse, R. and K. R. Schrerer (1996). "Acoustic profiles in vocal emotion expression." Journal of Personality and Social Psychology **70**(3): 614-636.

Bellman, B. L., & Bennetta Jules-Rosette. (1977). A Paradigm for looking. Norwood, Ablex Publishing.

Bindra, D. (1969). "A unified interpretation of emotion and motivation." Annals of the New York Academy of Science. **159**: 1071-1083.

Boersma, P. and D. Weenink (2006). Praat: doing phonetics by computer.

Chung, S. (1999). Vocal expression and perception of emotion in Korean. 14th International Conference of Phonetic Sciences, San Fransisco, USA.

Consortium, W. W. W. (2008). "Synchronized Multimedia." from http://www.w3.org/AudioVideo/.

Cornelius, R. R. (2000). "Theoretical approaches to emotion." Speech Emotion **1**: 3-10.

Cowie, R. and R. R. Cornelius (2003). "Describing the emotional states that are expressed in speech." Speech Communication Special Issue on Speech and Emotion **40**(1-2): 5-32.

Cullen, C., Vaughan, B. ,Kousidis, S., Wang, Yi ., McDonnell, C. and Campbell, D. (2006). Generation of High Quality Audio Natural Emotional Speech Corpus using Task Based Mood Induction International Conference on Multidisciplinary Information Sciences and Technologies Extremadura, Merida.

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

Cullen, C., Vaughan, B., Spyros,. K. (2008). LinguaTag: an emotional speech analysis application. 12th World Multiconference on Systemics, Cybernetics and Informatics (WMSCI 2008). Orlando, Florida. USA.**:** 7.

Darwin, C. R. (1872). The Expression of the Emotions in Man and Animals. London., Albermarle.

Douglas-Cowie, E., N. Campbell, et al. (2003). "Emotional speech: towards a new generation of databases." Speech Communication Special Issue Speech and Emotion **40**(1-2): 33–60.

Enberg, I. S., Hansen, A.V., Anderson, O., Dalsgaard, P. (1997). Design, recording and verification of a Danish Emotional Speech Database. Eurospeech '97, Rhodes, Greece.

Erickson, F., and Schultz, J. (1982). The Counsellor as Gatekeeper: Social Interaction in Interviews. Language, Thought and Culture: Advances in the Study of Cognition. E. Hammel. New York, Academic Press.

Forgras, J. P. (1990). "Affective influences on individual and group judgements ." European Journal of Social Psychology **20**: 441-453.

Frijda, N. H. (1988). "The Laws of Emotio." American Psychologist **43**(5).

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

Gerrards-Hesse, A., K. Spies, et al. (1994). "Experimental inductions of emotional states and their effectiveness: A review." British Journal of Psychology **85**: 55-78.

Gerrards-Hesse, A., K. Spies, et al. (1994). "Experimental inductions of emotional states and their effectiveness: A review." British Journal of Psychology **85**(1): 55-78.

Gobl, C., E. Bennett, et al. (2002). Expressive Synthesis: How Crucial is Voice Quality? IEEE Workshop on Speech Synthesis, Santa Monica, CA (USA).

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

Henkel, M., J., Hinsz, V. (2004). "Success and failure in goal attainment as a mood induction procedure." Social Behavior and Personality **32**(8): 715-722.

Isen, A., Shalker, T., Clark, M., Karp., L. (1978). "Affect, acessibility of Material in Memory, and Behavior: A Cognitive Loop?" Journal of Personality and Social Psychology **36**(1): 1-12.

ISLE. (2003). "IMDI (ISLE Metadata Initiative), Metadata Elements for Session Descriptions." Draft Proposal Version 3.0.3. from http://www.mpi.nl/IMDI/Schema/IMDI.

James, W. (1884). "What is an emotion?" Mind **9**: 188-205.

Johns-Lewis, C. (1986). Prosodic differrentiation of discourse modes. Intonation in Discourse. C. Johns-Lewis. San-Diego, College Hill Press**:** 199-220.

Johnstone, T. (1996). Emotional Speech Elicited using computer games. Spoken Language, ICSLP 96. Proceedings., Fourth International Conference on, Philadelphia, PA, USA.

Johnstone, T., C. M. v. Reekum, et al. (2005). "Affective speech elicited with a computer game." Emotion(5): 513-518.

Johnstone, T. and K. R. Scherer (1999). The Effects of Emotions on Voice Quality. XIV Int. Congress of Phonetic Sciences, San Francisco.

Kehrein, R. (2002). The prosody of authentic emotions. Speech Prosody, Aix-en-Provence, France.

Kienast, M., Sendlmeier, W.F., (2000). Acoustical analysis of spectral and temporal changes in emotional speech. ISCA ITRW on Speech and Emotion, Newcastle, Belfast, Textflow.

McGuire, T. T. (1993). Emotion and behaviour genetics in vertebrates and invertabrates Handbook of Emotions. M. Lewis, Haviland, J.M. New York, Guilford Press.

Pereira, C. (2000). Dimensions of emotional meaning in speech. ISCA ITRW on Speech and Emotion, Newcastle, Belfast, Textflow.

Plutchik, R. (2001). "The Nature of Emotions." American Scientist **89**(4): 344-350.

Pugmire, D. (1994). "Real Emotion." Philosophy and Phenomenological research **54**(1): 105-122.

Scherer and Ceschi (1997;2000). "Geneva Airport Lost Luggage Study." Motivation and Emotion **21**: 211-235.

Scherer, K. R. (1984). On the nature and function of emotion: A component process

approach. Approaches to emotion. K. R. Scherer and P. Ekman. Hillsdale, NJ, Erlbaum**:** 293-317.

Software, E. (2008). "Eiffel Software Home Page." Retrieved February, 2008.

Van Bael, C., van den Heuvel, H., Strik, H. (2004). Investigating Speech Style Specific Pronunciation Variation in Large Spoken Language Corpora Large Spoken Language Corpora. Proceedings of Interspeech (ICSLP) Jeju, Korea.

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

Westermann, R., Spies, K., Stahl, G., & Hesse, F. W. (1996). "Relative effectiveness and validity of mood induction procedures: a meta analysis." European Journal of Social Psychology **2 6**: 557-580.

Xiao, Z., E. Dellandrea, et al. (2005). Features extraction and selection for emotional speech classification. IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), Como, Italy.

XML, W. W. W. C. (2008). "Extensible Markup Language." from http://www.w3.org/XML/.