2011

Obtaining Speech Assets for Judgement Analysis on Low-pass Filtered Emotional Speech

John Snel
Dublin Institute of Technology, john.snel@student.dit.ie

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

Obtaining speech assets for judgement analysis on low-pass filtered emotional speech

John Snel and Charlie Cullen

Abstract— **Investigating the emotional content in speech from acoustic characteristics requires separating the semantic content from the acoustic channel. For natural emotional speech, a widely used method to separate the two channels is the use of cue masking. Our objective is to investigate the use of cue masking in non-acted emotional speech by analyzing the extent to which filtering impacts the perception of emotional content of the modified speech material. However, obtaining a corpus of emotional speech can be quite difficult whereby verifying the emotional content is an issue thoroughly discussed. Currently, speech research is showing a tendency toward constructing corpora of natural emotion expression. In this paper we outline the procedure used to obtain the corpus containing high audio quality and 'natural' emotional speech. We review the use of Mood Induction Procedures which provides a method to obtain spontaneous emotional speech in a controlled environment. Following this, we propose an experiment to investigate the effects of cue masking on natural emotional speech.**

I. INTRODUCTION

The acquisition and type of speech data obtained play an important role for emotional recognition, with implications that have several theoretical issues. Firstly, the type of speech samples used affect analysis accuracy of acoustical data (encoding studies), and secondly, it determines the quality of the labels that are provided by annotators (decoding studies). A correlation of acoustic analysis with listeners' ratings can provide information on which vocal cues determine the inference of emotion (cue measurement and statistical association).

It is difficult to determine the boundaries to which listeners infer emotion from acoustical cues or semantic content. Studies regarding decoding have shown that listeners are able to infer actual or simulated emotion from suprasegmentals and paralinguistic patterns (vocal expression that refers to qualities of speech rather than the verbal content) via vocal cues disregarding lexical content [1]. In most of these studies actors portray a set number of different emotions by producing speech utterances with standardized or nonsense content (e.g. [8]). This way the speech does not contain any linguistic information that could indicate the underlying emotion of the speaker and hence provides a way to see how well listeners decode the emotion from the speech variables (distal cues) alone. For non-acted emotions, however, one can remove semantic content and retain certain acoustic parameters through cue masking (inference studies). Cue masking is used to distort the speech sample to remove certain vocal cues. This method determines the influence the modified material has on the emotion inference. Several techniques exist for this such as, low-pass filtering, randomized splicing, playing backwards, pitch inversion, and tone-silence coding (e.g. [32], [46], [42], [27]).

In this paper we review the different types of assets available for emotion research. We argue for the use of natural speech assets and outline the methods used to obtain the current corpus composed of high audio quality spontaneous speech. We further propose an experiment that uses low-pass filtering, on natural speech data, in order to investigate the impact cue masking has on the listeners' perception of emotion in speech. We refer our research to several aspects of the Brunswik lens model, which will be the immediate following topic discussed in Section 2. This paper will be further presented as follows - Section 3 reviews the type of existing data available for emotional speech research. This is followed by Section 4, a brief summary of obtaining natural speech data through the use of Mood Induction Procedures. In Section 5, we propose an experiment to examine the users ability to infer emotion from content-free speech by the use of frequency filtering. Section 6 discusses the current rating tool. We conclude in Section 8 and closing the paper in Section 9, future work.

II. BRUNSWIK MODEL FOR SPEECH

The various facets of the expression and perception process, that are possible in studies of vocal emotion expression can be illustrated within a certain framework. The Brunswikian lens model [7], originally designed as a model of visual perception, has been adopted and modified for vocal communication (see Figure 1) [35]. Scherer [36] provides a review of the different research paradigms and suggests to base research on vocal communication of emotion on this model. The model considers the different aspects of communication, including *encoding* (expression), *transmission*, and *decoding* (impression) of vocal emotion communication. The emotional state of the speaker is encoded by certain voice and speech characteristics that can be objectively measured in the speech signal. Physiological changes are assumed to accompany the emotional arousal that produce emotion-specific patterns of acoustic patterns i.e. respiration, phonation, and articulation.

The perception process consists of cues that are distal and proximal. For an observer, the acoustic changes that serve as cues to the speaker's emotional state, are called distal cues (cues distant from the observer). These cues, as part

This work was supported by the Science Foundation Ireland under Grant No. 09-RFP-CMS253

J. Snel and C. Cullen are with Digital Media Centre, Dublin Institute of Technology, Aungier St., Dublin 2, Ireland john.snel@student.dit.ie, charlie.cullen@dit.ie

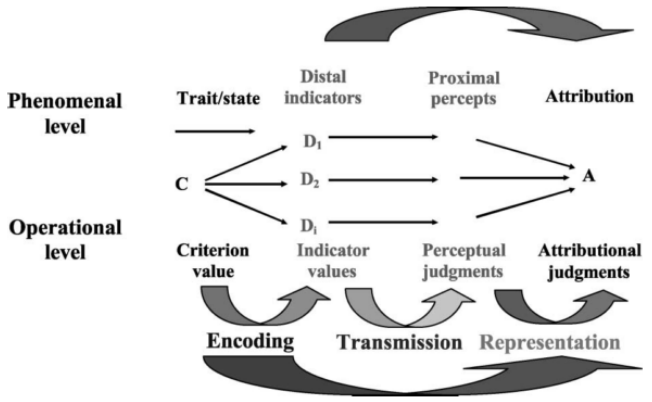

Fig. 1. Scherer's [36] adopted Brunswikian lens model adopted for vocal communication of emotion.

of the speech signal, are transmitted to the listener who then perceives them via the auditory perceptual system. The perceived cues, or subjective dimensions, are referred to as proximal cues (cues close to the observer) and are used for attribution by the listener to infer the speaker's emotional state. The degree to which the speaker's underlying emotional state correlates with the externalizations in the form of specific distal cues, is termed by Brunswik as *ecological validity*. An example of a proximal cue for fundamental frequency would be pitch, while the attribution would be the perceived emotion of the speaker [37]. According to Scherer, a vital aspect to the model is the awareness that objectively measured distal cues are not necessarily identical to the proximal cues produced by the observer. The distal cues may be misrepresented due to distortion of the acoustic signal in the transmission channel (e.g. noise, distance or medium) or the physical characteristics of the auditory perceptual system which will impact the transduction and coding process (e.g. some enhancement of certain frequency bands). Finally, if the proximal cues reliably map the valid distal cues, it is still possible for the listener to incorrectly infer the speaker's emotional state from the cognitive process when judging the encoded emotion from all the cues. Studies can focus on one particular aspect or use a combination of the perceptual inference process, for example, encoding and decoding. When acoustic properties are measured, it is referred to the encoding process, and perception tests of emotional speech (e.g. the rating of speech assets) are considered as decoding aspects [37]. A review of the different research paradigms as described by Scherer are outlined as follows:

- *Encoding studies*: in encoding studies, one tries to establish an association between the emotional state and the measurable acoustic parameters of the speech
- *Decoding studies*: research in this area examines to what extent listeners are able to infer natural or simulated emotions from acoustic information regardless of the semantic content of speech.
- *Inference studies*: inference studies are concerned with the distal cues that produce certain types of emotion inferences in listeners. This can involve manipulating or varying certain acoustic cues to determine the relative parameters that are associated with emotion judgement.
- *Transmission studies*: work in this area examines the transmission of the distal signal of the sender to the receiver, at proximal level, which may be distorted causing inference errors.
- *Representation studies*: Representation studies focus on the subjective impression of the vocal qualities and are regarded as the most difficult aspect of the research guided by the Brunswikian lens model [36].

III. Existing emotional speech data

Research in the recognition of emotion from speech is dependent on speech data but there exists practically no standardized corpora and test-conditions, hence low comparability of results [39]. A large amount of available data does not contain realistic data but acted speech. Acoustics from acted material cannot simply be compared to realistic data [39]. A study by Cullen et al. [11] addresses issues such as audio quality (transmission studies), annotation (decoding studies) and obtaining genuine emotional speech recordings (encoding studies). Obtaining a corpus of emotional speech can be quite difficult whereby verifying the emotional content is an issue thoroughly discussed [2]. Ververidis [45] provides an overview of the emotional speech data collections available, identifying 64. As mentioned above, in the context of the Brunswikian model, encoding studies involve identifying distinctive acoustic patterns (distal cues) associated with the speakers emotional state. These studies differ in the emotion elicitation type contained in the speech data investigated. In this section the main types of speech data are discussed, divided into simulated, natural and induced vocal expressions.

A. Simulated (portrayed) vocal expressions

Vocal affect expressions, which are based on discrete categories, can be produced by professional or lay actors (see [36] for a review). The great advantage of using acted speech is the ability to produce speech and maintain the same verbal content. This allows for direct comparisons of acoustic content and its paralinguistic patterns associated with the different emotional states expressed. This method makes it easy to obtain speech assets and the environment in which it is recorded is easy to control, allowing for high quality recordings. However, little is known about how simulated and natural speech assets compare [16]. It is likely for simulated vocal expression to represent culturally shaped prototypical expressions [23] that are more intense than induced states or natural emotional states [36]. Generally, acted material is more reliable for classification performance in comparison to realistic and spontaneous speech because the emotions portrayed through acting produce a higher arousal level [45], [5]. Scherer [35] mentions that actors may over-emphasize certain emotional cues whereby they may disregard more subtle cues. He recognizes that actors are likely to portray conventionalized stereotypical expressions of emotion but notes that all public displays of affect-expressions are to

some extent acted [36], [4]. One can anticipate a particular outcome when dealing with simulated emotions. The recognition ratings that provide a measurement of success are compared to the probability of it being by chance [37]. Both correct answers and wrong answers (confusions) are of interest. A confusion matrix can be used to contain information about actual and predicted classifications and has been used more regularly in recent studies [36], [19], [5], [37].

Although the use of simulated assets has its advantages, there are certain drawbacks when choosing acted material. Some are listed as follows:

- Simulated assets are often non-interactive [4], [3] and hence may only provide for a limited range of emotions [11].
- Emotion portrayals reflect conventionalized stereotypes [36] and may not include involuntary physiological responses to stimuli normally associated with emotion [22].
- Actors can over emphasize certain vocal cues while missing more subtle ones [36].
- Actors may have different subjective interpretations of the emotions they are instructed to simulate [14].
- Actors encoding abilities, when simulating an emotion, differ considerably between actors [46].
- Reading from text is not spontaneous and differs in acoustic characteristics [21].

Recent efforts address some of these issues that include interactive dialogues and avoidance of stereotypical portrayals by improving elicitation techniques modeled on theatrical performance strategies [9]. Other recent studies in the area of speech and emotion are showing a tendency towards developing natural, real-life speech corpora [38], [29].

B. Natural vocal expressions

Speech data containing natural vocal expression has been obtained from various sources such as, dangerous flights situations, journalists reporting emotion-eliciting events, therapy sessions, or talk and game shows on TV (see [36] for a list of studies). Devillers and Vidrascu [15] have worked on several real life data corpora from financial call centers, EmoTV clips and a corpus with real agent-client spoken dialogs from a medical emergency call center. Assets from broadcast sources can be said to be more natural compared to simulated assets [16]. Grimm et al. [19], in conjunction with the EMA corpus, worked with the VAM corpus that contains spontaneous speech with authentic emotions taken from a German TV talk-show. The two different production styles allowed them to compare between natural and acted speech that was explored over two languages; English and German. Most full-blown, prototypical emotions seem to be absent in most realistic databases, which consist more of pervasive emotions like boredom, interest, etc. [6]. This would influence the decision for discrete categories or a dimensional model. Otherwise, specific affective states can also be chosen that would be more fitted for the particular data collected. For example, studies on affective states such as interest, approval, attention and prohibition, and motherese and empathic [28]. Natural vocal changes have very high ecological validity and therefore seem to be the ideal research paradigm [36]. However, the detection of reliable characteristics of emotion in natural speech is considered more complex than in acted speech [5] and natural speech is generally more difficult to obtain, due to ethical issues and problems with audio quality. For a list of natural databases refer to [45].

C. Induced vocal expressions

Ideally, experiments should make use of the controlled environment that simulated assets provide but broadcast assets lack and still maintain the natural element in emotions expressed. A method, that is not as widely used, is to actually induce emotions in a laboratory environment. Mood Induction Procedures (MIP's) are designed to induce emotional responses from a test subject, in an environment that favors high audio quality, and has shown to be successfully used by numerous researchers [26], [24]. Studies have used a variety of techniques that can be categorized into five different MIP groups [17]:

1) MIP's that involve emotion-inducing techniques such as mental imagery [18] or hypnosis [47].
2) MIP's that use emotion inducing material that include the Velten MIP [44], the Film MIP [33] and Music MIPs [13]. Participants are guided towards the suggested emotional state.
3) MIP's that use external material similar to the previous, with the additional Gift MIP [34], but without explicit instructions of how to feel about the suggested material (it is assumed the material automatically induces the emotion).
4) MIP's that include tasks with the need for achievement, such as the Success/Failure method where subjects are given false-positive or false-negative feedback to manipulate appraisal dimensions [30], [40].
5) MIP's that induce physiological states associated with emotion with the use of drugs (Drug MIP) [20] or facial feedback (Facial MIP) [41].

D. MIP Choice

Advantages for the use of MIP's include authenticity of natural assets, and experimental control for both audio quality and manipulation of cognitive appraisal. Juslin and Scherer [25] note that this approach allows the researcher to investigate the complete scope of the Bruswik lens model. For example, manipulating the dimensions of cognitive appraisal, indexing the physiological changes, measuring the voice alterations, which can be subsequently used in judgement studies. This MIP approach may, however, have some drawbacks. It has been noted by Scherer [36] that it can't be assumed that similar emotion states are produced by each participant. To overcome this our assets will be

rated (i.e. via emotion scales, see section VI) to establish consistency decoded by listeners (judgement studies). The authenticity of the assets may differ between the different MIP groups. Several MIPs are ethically questionable. The authors of this paper believe that for their specific goal the best suited experiment is the Success/Failure MIP (MIP 4). The Success/Failure MIP is advantageous when concerned with demand effects[1] [48] because the true nature of the experiment can be disguised (see also [40]). The controlled environment allows for recording conditions achieved by those compiling simulated assets.

IV. An MIP based corpus

In this section we review how the assets were obtained. A study by Vaughan et al. [43] investigated 3 different experiments incorporating the MIP 4 group (Success/Failure and social interaction MIP) and the MIP 3 group (Gift MIP). The build of this corpora [11], [43] took several factors into consideration. Amongst these, authenticity of emotional content, demand effects, ethical issues, annotation, and audio quality. The assets were obtained in a controlled environment providing high quality audio [12], thus preventing unwanted noise that may disrupt effective analysis.

A. MIP Experiments

Three experiments were performed to compile elicited emotional assets. This consisted of two participants, placed in two isolation booths, that were instructed to perform a cooperative-based task while the researcher monitored, manipulated, and recorded the procedure. The true time left for the completion of tasks was open to alteration. The MIPs devised consisted of computer games and a physical game. The first case study used the puzzle game, Tetris. It was constructed in such a way that both participants co-operatively played the game. Manipulations included, cutting the audio feed, altering the time, and remotely controlling the placing of the blocks. The second experiment used console gaming, allowing multiplayer options. The competitive nature of this game, determines the use for little or no manipulation during game play [11]. In saying this, external manipulations may consist of unplugging participant's game controllers, altering the time limit, or offering a reward (Gift MIP - group 3). In the third experiment two participants had to build a Lego construct where one had access to the pieces and the other gave instructions. Aspects of manipulation can include missing pieces of Lego, cutting the audio link, and altering the time to complete the task. All three experiments can make use of the MIP 4 group (Success/Failure MIP and social interaction MIP) combined with the MIP 3 group (Gift MIP). The MIP experiments consisted of emotion that was naturally inherent through the co-operative (social interaction) and competitive (Success/Failure) nature of the procedure that can combine the Gift MIP to further elate a participant and use manipulation to elicit frustration or satisfaction. Those aspects of the MIP that are controlled in order to elicit emotion can be noted against the time of the recordings in order to analyze the validity of those specific procedures (when later rated by listeners). Great consideration was taken in order to provide a high standard of audio quality. This will be outlined next.

B. Experimental conditions

As mentioned above, broadcasts assets may vary in audio quality. Such assets may be obtained from a studio, out in open spaces and telephone networks. Audio quality has been largely considered for the providing of the current corpus [12].

The above mentioned experiments consisted of two participants in two soundproof isolation booths. Currently, there is a possibility to use up to four booths. Two of the booths are of a larger size that can accommodate for tasks where space may be of importance. Soundproof isolation booths reduce unwanted acoustic factors within the speech signal, such as waves from other sources, and room reverberation. Additionally, the isolation booths prevent any external distractions (outside of experiment) effecting the participant's involvement.

Other equipment used comprise of professional Neumann U87 microphones, Beyerdynamic DT 150 headphones, and Pro Tools HD3 rig. The audio is recorded at its highest quality, 192Khz/24-bit. The consistency of the quality makes it easier for analysis. If one wants to compare results with that of lower sample/bit rate from other assets, it is possible to down-sample. Visa-versa, however, it is not possible to increase the sample-rate to try and regain the acoustic information from the original recording. Each booth is further fitted with monitors with an option for other required control devices (e.g. keyboard, or console controller).

V. Cue masking experiment

In this section, we propose an experiment to investigate the effects of filtering as a method for *cue masking*, but still retaining lower frequencies and tonal quality of speech. Other cue masking techniques, such as randomized splicing, playing backwards and pitch inversion, do not retain as much information about prosody and the intonation contour as filtering would. A similar study by Knoll et al. [27] investigated the use of filtering for vocal affect directed at infants, adults and foreigners. Their study investigated the affective salience of speech altered by different levels of low-pass filtering. They claim that the degree to which different levels of low-pass filtering impacts the inferring perception of emotion (or as they refer to as 'affect') in non-acted, natural speech remains unknown. They further argue that low-pass filtering is a useful tool in future affective speech research (amongst other speech related research).

A. Aims

When one listens to a conversation in a room next door (filtering out certain frequencies) the lexical content may not be audible, however, the 'tone'[2] of the conversation can

[1]Demand effects are those possibilities of the subject guessing the purpose of the procedure and hence act the desired emotion.

[2]here we refer 'tone' of voice to the non-lexical information transmitted with verbal messages.

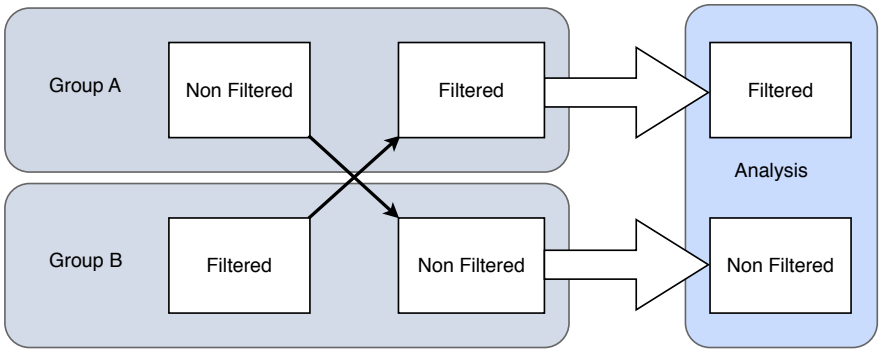

Fig. 2. Listeners randomly assigned to one of two groups are asked to rate non-filtered and filtered speech assets.

often be clearly heard. Following from this, the experimental method seeks to evaluate the following hypothesis:

if certain acoustic cues are masked to disguise the linguistic content but retaining certain aspects of the acoustic cues, in this case through low-pass filtering, then listeners are still able to infer vocal affect from natural speech.

B. Method

If the participant rates a speech utterance and subsequently listens to the same speech utterance but filtered, he/she may match the previous rating due to recognition and familiarity from the original stimuli — either from implicit or explicit memory. In order to minimize any carryover and order effects, a between-subject design will be adopted. Subjects will be randomly assigned to one of two groups, presenting the filtered and original stimuli in different orders. Group A will be asked to rate the non-filtered clips before the filtered clips, whereas Group B will be, inversely, presented with the filtered clips and then the non-filtered clips (see Figure 2). The speech stimuli will be presented to the listener via a web-based tool. A repeated measure design will be adopted to reduce the subjects retention of the tone of voice (or F0) from the first condition (perceptual priming). Participants will be asked to rate the conditions in two separate sessions two weeks apart; in each session each participant will be asked to rate 6 speech clips from one of the conditions. The number of speech clips chosen are kept to a minimum, rather than a large number of clips, chosen on the basis of a preliminary/pilot study, to minimize fatigue effects. Finally, the ratings will be compared for both the original audio and the filtered audio to determine the hypothesis. The dependent variables will be vocal affect ratings; activation and evaluation.

VI. Rating Strategy

Listening tests for rating the emotional speech corpus will be delivered to the user via an online webpage. Our ratings strategy is based on the dimensional model [31]. To use a dimensional approach for rating emotional speech, one can retain information about relationships between emotions. Bipolar scales within a circumplex model provide an indication to the similarity and contrast of certain categories. Using dimensions further allows for visualizing the intensity of an emotion. It is best suited when considering the physiological component of emotion, for example breathing rates, which may effect the vocal chord activity. Further, the possibility for temporal assessment for emotion as it varies over time would seem more feasible with a dimensional model (for example the 'feeltrace' tool [10]). Yu et al. [49] adopted a two-dimensional emotion space using three and five different levels for valence and arousal. Their results showed that increasing the number of classes can decrease classification accuracy. The authors of this position paper propose the use of two affective scales, namely evaluation and activity, both rated using a five point Likert-scale.

VII. Conclusion

In this position paper we refer the different aspects of our research to the Brunswik lens model. We summarized the different types of assets available for emotional research and outline the methods for obtaining high quality, natural assets through the use of Mood Induction Procedures. We also propose an experiment to isolate semantic content from acoustic information within the speech signal by masking acoustic cues via low-pass filtering. To our knowledge, few recent studies exist that investigate decoding accuracy on content-free natural emotional speech through the use of filtering. Both speech asset sets, original and filtered, will be rated and compared according to evaluation and activity.

VIII. Future work

In order to proceed to analysis of our speech assets, a sufficient amount of ratings need to be collected. A web-based tool will be developed to collect the ratings for annotation. In order to investigate lexical and acoustic aspects, other means for cue masking and manipulation will be considered. Text-only assets can be rated and compared against the original speech sample and the filtered speech signal.

References

[1] N. Amir, B. Almogi, and R. Gal. Perceiving Prominence and Emotion in Speech – a Cross Lingual Study. *Communications*, 2004.

[2] N. Amir, Ori Kerret, and Dimitry Karlinski. Classifying emotions in speech : a comparison of methods. *Science*, 2001.

[3] N. Amir, Samuel Ron, and Nathaniel Laor. Analysis of an emotional speech corpus in Hebrew based on objective criteria. *and Research Workshop (ITRW) on Speech and*, 2000.

[4] R. Banse and K. R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614–36, March 1996.

[5] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and Others. Combining efforts for improving automatic classification of emotional user states. *Proc. IS-LTC*, pages 240–245, 2006.

[6] A. Batliner, S. Steidl, B. Schuller, Dino Seppi, Thurid Vogt, Johannes Wagner, L. Devillers, L. Vidrascu, Vered Aharonson, and Loic Kessous. Whodunnit – Searching for the most important feature types signalling emotion-related user states in speech. *Computer Speech & Language*, 2010.

[7] E. Brunswik. *Perception and the Representative Design of Psychological Experiments*. University of California Press, Berkeley, 1956.

[8] F Burkhardt, A Paeschke, M Rolfes, W Sendlmeier, and B Weiss. A Database of German Emotional Speech. *in Interspeech Lissabon, Portugal*, 2005.

[9] C. Busso and S. S. Narayanan. Scripted dialogs versus improvisation : Lessons learned about emotional elicitation techniques from the IEMOCAP database. *Interspeech*, 2008.

[10] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Citeseer, 2000.

[11] C Cullen, S. Kousidis, and J. McAuley. Emotional speech corpus construction, annotation and distribution. *Arrow.dit.ie*, 2008.

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

[13] SK De l'Etoile. The effect of a musical mood induction procedure on mood state-dependent word retrieval. *Journal of music therapy*, 39(2):145, 2002.

[14] Salero Deliverable. SALERO-D6.2.1 -Limited Emotional Speech Corpus for Analysis. *Analysis*, 2008.

[15] L. Devillers, L. Vidrascu, and Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, 2005.

[16] E. Douglas-Cowie. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):33–60, April 2003.

[17] A. GERRARDS-HESSE, K. SPIES, and F. W. HESSE. Experimental inductions of emotional states and their effectiveness: a review. *British journal of psychology*, 85(1):55–78, 1994.

[18] Anja S. Göritz. The Induction of Mood via the WWW. *Motivation and Emotion*, 31(1):35–47, November 2006.

[19] M. Grimm, K Kroschel, E Mower, and S. S. Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10-11):787–800, 2007.

[20] Hede Helfrich, Reiner Standke, and Klaus R. Scherer. Vocal indicators of psychoactive drug effects. *Speech Communication*, 3(3):245 – 252, 1984.

[21] C Johns-Lewis. Prosodic differentiation of discourse modes. In *In: Johns-Lewis, C. (Ed.), Intonation in Discourse. College-Hill Press*, pages 199–220, 1986.

[22] T. Johnstone. Emotional speech elicited using computer games. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, 1996.

[23] T. Johnstone and K. R. Scherer. *Vocal Communication of Emotion*, volume 2, chapter 14, pages 220–235. New York: Guilford Press, 2000.

[24] T. Johnstone, C.M. van Reekum, Kathryn Hird, Kim Kirsner, and K. R. Scherer. Affective Speech Elicited With a Computer Game. *Emotion*, 5(4):513–518, 2005.

[25] Patrick Juslin and K. R. Scherer. *The new handbook of methods in nonverbal behavior research*, chapter 3, pages 65–135. Oxford University Press, Oxford, series in edition, 2005.

[26] Roland Kehrein. The prosody of authentic emotions. *Speech Prosody, Aix-en-Provence, France*, 2002.

[27] M Knoll, M Uther, and a Costall. Effects of low-pass filtering on the judgment of vocal affect in speech directed to infants, adults and foreigners. *Speech Communication*, 51(3):210–216, March 2009.

[28] Petri Laukka, Daniel Neiberg, Mimmi Forsell, Inger Karlsson, and Kjell Elenius. Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. *Computer Speech & Language*, April 2010.

[29] Hiroki Mori, Tomoyuki Satake, Makoto Nakamura, and Hideki Kasuya. Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics, August 2010.

[30] Lauri Nummenmaa and Pekka Niemi. Inducing affective states with success–failure manipulations: A meta-analysis. *Emotion*, 4(2):207 – 214, 2004.

[31] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. The measurement of meaning. *University of Illinois Press, Urbana, USA*, 1957.

[32] R. Rosenthal, J.A. Hall, M.R. DiMatteo, P.L. Rogers, and D. Archer. *Sensitivity to Nonverbal Communication: The PONS Test*. Johns Hopkins University Press, 1979.

[33] J. Rottenberg, R. D. Ray, and J. J. Gross. *Emotion elicitation using films*, chapter 1, pages 9–28. Oxford University Press., London, j. a. coan edition, 2007.

[34] B Ruffle. Gift giving with emotions. *Journal of Economic Behavior & Organization*, 39(4):399–420, July 1999.

[35] K. R. Scherer. Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99(2):143–165, March 1986.

[36] K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256, 2003.

[37] M. Schröder. *Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis.* PhD thesis, Universitat des Saarlandes: 288, 2004.

[38] B. Schuller, A. Batliner, Dino Seppi, S. Steidl, Thurid Vogt, Johannes Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and Others. The Relevance of Feature Type for the Automatic Classification of Emotional User States : Low Level Descriptors and Functionals. *Proc. INTERSPEECH 2007*, 101:2253–2256, 2007.

[39] B. Schuller, S. Steidl, and A. Batliner. The INTERSPEECH 2009 Emotion Challenge . *Interpretation A Journal Of Bible And Theology*, 2009.

[40] S. Steidl. *Automatic classification of emotion-related user states in spontaneous children's speech*. PhD thesis, Der Technischen Fakultät der Universität Erlangen-Nürnberg, 2009.

[41] F. Strack, L. L. Martin, and S. Stepper. Inhibiting and facilitating conditions of the human smile: A non-obtrusive test of the facial feedback hypothesisial expressions: A non-obtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54:768–777, 1988.

[42] R van Bezooijen and L Boves. The effects of low-pass filtering and random splicing on the perception of speech. *Journal of psycholinguistic research*, 15(5):403–17, September 1986.

[43] B. Vaughan, S. Kosidis, C Cullen, and Y. Wang. Task-based mood induction procedures for the elicitation of natural emotional responses. In *The 4th International Conference on Cybernetics and Information Technologies, Systems and Applications: CITSA 2007*, Orlando, Florida., 2007.

[44] E. Velten. A laboratory task for induction of mood states. *Behaviour Research and Therapy*, 6(4):473–482, 1968.

[45] D. Ververidis and C Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, September 2006.

[46] H.G. Wallbott and K. R. Scherer. Cues and channels in emotion recognition. *Journal of Personality and Social Psychology*, 51(4):690–699, 1986.

[47] Friedbert Weiss, Gerald S. Blum, and Lisa Gleberman. Anatomically based measurement of facial expressions in simulated versus hypnotically induced affect. *Motivation and Emotion*, 11(1):67–81, March 1987.

[48] R. Westermann, K. Spies, G. Stahl, and F. W. Hesse. Relative effectiveness and validity of mood induction procedures: a meta-analysis. *Eur. J. Soc. Psychol.*, 26(4):557–580, 1996.

[49] Chen Yu, PM Aoki, and Allison Woodruff. Detecting user engagement in everyday conversations. *Proceedings of Interspeech 2004 — ICSLP, Jeju, Korea*, 2004.