Naturalistic Audio-Visual Emotion Database

Sudarsana Reddy Kadiri[1], **P. Gangamohan** [2], **V.K. Mittal** [3] **and B. Yegnanarayana** [4]
Speech and Vision Laboratory,
Language Technologies Research Center,
International Institute of Information Technology-Hyderabad, India.
[1]sudarsanareddy.kadiri@research.iiit.ac.in, [2]gangamohan.p@students.iiit.ac.in,
[3]vinay.mittal@iiit.ac.in,[4]yegna@iiit.ac.in

Abstract

The progress in the areas of research like emotion recognition, identification, synthesis, etc., relies heavily on the development and structure of the database. This paper addresses some of the key issues in development of the emotion databases. A new audio-visual emotion (AVE) database is developed. The database consists of audio, video and audio-visual clips sourced from TV broadcast like movies and soap-operas in English language. The data clips are manually segregated in an emotion and speaker specific way. This database is developed to address the emotion recognition in actual human interaction. The database is structured in such a way that it might be useful in a variety of applications like emotion analysis based on speaker or gender, emotion identification in multiple emotive dialogue scenarios etc.

Keywords: Emotion analysis, Emotion recognition, Expressive synthesis, Simulated parallel database, Semi-natural database, Audio-visual data.

1 Introduction

Emotion databases provide an important experimental foundation for analysis when researchers aim at building emotion-aware speech systems (M. Gnjatovic et al., 2010). The basic requirement of database for studies in emotion analysis, identification, classification and synthesis is guided primarily by its suitability to the application chosen. Several emotion databases developed by different research groups can be categorized as simulated, semi-natural and natural databases (E. Douglas-Cowie et al., 2003; B. Schuller et al., 2011; D. Ververidis et al., 2003; S. G.Koolagudi, et al., 2012).

Simulated parallel emotion corpus are recorded from speakers (artists) by prompting them to enact emotions through specified text in a given language. The simulated parallel emotion corpus reported in (Zhihong Zeng et al., 2009; F. Burkhardt et al., 2005; I. S. Engberg et al., 1997; B. Schuller et al., 2010; S. G. Koolagudi et al., 2009), were collected from speakers by asking them to emote same text in different emotions. Their main disadvantage is that the deliberately enacted emotions are quite at variance from the natural 'spontaneous' emotions, and also at times they are out of context (D. Ververidis et al., 2003; D. Erickson et al., 2006).

Semi-natural is a kind of enacted corpus where the context is given to the speakers. The semi-natural emotion database in German language was developed by asking speakers to enact the scripted scenarios, eliciting each emotion (R. Banse et al., 1996; I. Sneddon et al., 2012). Similar semi-natural databases in English and Russian languages were reported in (E. Douglas-Cowie et al., 2000; N. Amir et al., 2000), respectively.

The third kind of emotion database is *natural database*, where recordings do not involve any prompting or the obvious eliciting of emotional responses. Sources for such natural situations could be like talk shows, interviews, panel discussions, and group interactions, etc., in TV broadcast. The Belfast natural database in English language was developed by segmenting 10-60 seconds long audio-visual clips from TV-talk shows (E. Douglas-Cowie et al., 2000). Similar kind of databases were developed in Korean (Zhihong Zeng et al., 2009), German and English languages such as, FAU Aibo (Steidl, S et al., 2009), USC-IEMOCAP (C. Busso et al., 2008), (S. Chung et al.,) etc. Geneva airport lost-luggage study database was developed by videotaping the interviews of passengers at lost-luggage counters (K. Scherer et al., 1997). The "Vera am

D S Sharma, R Sangal and J D Pawar. Proc. of the 11th Intl. Conference on Natural Language Processing, pages 206–213,
Goa, India. December 2014.

Mittag" German audio-visual emotion database (M. Grimm et al., 2008), was developed by segmenting the audio-visual clips from the talk-show "Vera am Mittag". More details of the various types of databases, issues and important aspects of databases were given in (E. Douglas-Cowie et al., 2003; B. Schuller et al., 2011; D. Ververidis et al., 2003; Koelstra, S et al., 2012).

The simulated emotion parallel speech corpus are mainly used in the area of emotion conversion (M. Schroder et al., 2001; I. Murray et al., 1993; H. Kawanami et al., 2003). In these cases Gaussian Mixture Models (GMMs) and Artificial Neural Networks (ANNs) are used for developing these systems. The output speech from these systems is unnatural and has the constraints of parallel speech corpus. Also, the emotion recognition systems based on these parallel speech emotion corpus are not reliable in real life scenarios. The semi-natural emotion databases also have promptness, but they are useful for developing emotion recognition systems because the recorded utterances have context. The requirement is to have the sentences with context, and the artists to have good performance skills. For the collection of natural databases from TV talk shows and interactive sessions, the main difficulty is to label the emotion or expressive state of the dialogue. Also, it is possible that emotion states like extreme anger, sad, fear, etc., may not occur in such TV broadcasts. Therefore, in "Vera am Mittag" emotion corpus (M. Grimm et al., 2008), the annotation of the utterances was described by three basic primitives: Valence (positive or negative), Activation (calm or excited) and Dominance (weak or strong).

Speakers involved in TV broadcast like talk shows (M. Grimm et al., 2008), interviews, panel discussions, and group interactions, etc., control their emotions/expressive states, i.e., they cannot express the feelings that occur in natural communication among humans. There is always a tradeoff between the controllability and naturalness of the interaction (M. Grimm et al., 2008).

In this paper, we describe an audio-visual emotion database named as *IIIT-H AVE*, developed at Speech and Vision Laboratory, IIIT Hyderabad. We have decided to use TV broadcast such as movies and soap-operas for data collection, because the emotions produced are more generic towards the natural communication, even though they are enacted.

The remaining part of the paper is organized as follows. Section II describes the challenges involved in the collection of emotion data for different applications. Section III describes data collection, recording parameters and various stages involved. In Section IV, the structure of the database and and in Section V, issues encountered, proposed solutions and limitations are reported. In Section VI, possible applications of the proposed database are discussed briefly. Finally, Section VII gives a summary and scope of future work.

2 Challenges involved in the collection of emotion data for different applications

In order to develop high quality text to emotion speech synthesis systems, large sized natural databases of each target emotion are required (M. Schroder et al., 2001). But it is impractical to develop a large sized natural emotion database with spontaneity (naturalness). Hence emotion conversion systems are adopted as a post-processing block for speech synthesis from neutral text. In this, a large database of neutral speech is used by text-to-speech (TTS) system to generate a neutral speech first, which is then fed to emotion conversion system where input neutral speech is converted to desired emotional speech. Since emotional speech is produced from emotion conversion systems, it is reasonable to use enacted parallel corpus (D. Erro et al., 2010).

Although it is practically reliable to use simulated parallel corpus for emotion synthesis systems, it does not serve the purpose of developing the emotion classification system because it consists of enacted speech. The original state of the speaker might be different as well. Most of the time, semi-natural and close to natural databases are used for developing emotion recognition systems.

The problem with semi-natural emotion type of databases is, whether the produced emotion is real or it is produced for the purpose of emotion data collection because the speakers know that they are being recorded.

Ideally, natural databases with multiple number of speakers, styles and contextual information are required to design emotion recognition systems for realistic applications. The collection of natural databases mostly from talk shows and interactive sessions in TV broadcast, call centers, interaction with robots, conversations in public places

etc. The main difficulty is to identify and label the emotion or expressive state of the dialogue. The emotive states like extreme anger, sad, fear, etc., may not occur some times in such TV broadcasts because the expression of emotion is continuum in nature. Therefore, for natural emotion corpus the annotation of the utterances was described mostly by three basic primitives or dimensions: valence, activation and dominance because the labelling of the naturalistic emotions as highly subjective and categorization of emotions is always debatable (M. Grimm et al., 2008; K. P. Truong et al., 2012). The difficulties involved in natural databases are overlapping multiple speakers data in audio or video or both, background noise or music etc. The good ground truth for natural emotion databases is a difficult task as there are inconsistencies in the annotation. Databases with good emotion labels/annotation would be helpful for emotion recognition tasks.

There are some challenges involved in collection of audio-visual data of naturally occurring emotions. Different people annotate different emotions/expressive states for the same data (audio visual clips). Also, there is a possibility of inconsistency in annotation done by the same person. It is impossible to define strict boundaries for the occurrence of emotion, as presence of emotion is a continuum in speech. Also, emotion depends on the semantic and contextual information.

3 Data collection

The objective of this audio-visual emotion data collection is to have an emotion annotated database with adequate context and large number of speakers. We have chosen English movies and soap operas in TV broadcast as source for data collection.

3.1 Selection of sources

We began by watching a range of source videos over a period of time, and eventually identified a few sources that are potentially useful. For example, if the story of a source had some drama revolving around a group of characters then it was considered as useful source to yield good clips of emotional content. This collection of source videos is named as raw data.

3.2 Emotive and Expressive states

The emotive and expressive states are chosen based on the examples derived from the selected sources. It is also observed that the communication among people always exhibits expressions. The extreme cases of these expressions leads to different emotions. We have identified 7 basic emotions (anger, disgust, fear, happy, neutral, sad and surprise) and 6 expressive states (confusion, excited, interested, relaxed, sarcastic and worried) (K. Scherer et al., 2003; R. Cowie et al., 2003). The list of emotive and expressive states considered is shown in Table 1.

Table 1: *List of emotive and expressive states.*

Emotive states	Expressive states
1. Anger	1. Confusion
2. Disgust	2. Excited
3. Fear	3. Interested
4. Happy	4. Relaxed
5. Neutral	5. Sarcastic
6. Sad	6. Worried
7. Surprise	

3.3 Segregation of raw data

Segregation of audio-visual clip segments (or specific scene-selection) from the chosen source video is carried out on the basis of perceived significance of emotion/expressive state. The duration of such audio-visual clips ranges from 0.5-30 seconds, with average being around 5 seconds. The criteria adopted for selecting 'good source clip' are the following:

- *The audio-visual clips with no background music or noise*
- *Clips with only one actor speaking at a time*

There were 6 subjects, each a research scholar, involved in the segregation of source videos. The basic challenge was to annotate the segregated clips. Subjects were asked to label the clip with one of the emotion/expressive state, and also to specify the confidence level.

If a particular soap-opera has many episodes then during segregation of the clips, the prominent characters of that particular soap-opera are also labelled with speaker numbers.

3.4 Recording quality

From the segregated audio-visual clips, the audio and video streams are extracted. The video files are MPEG4 coded image sequences of frame sizes mostlt 1280×720 pixels, with frame rate of either 24 fps. Files are in *x.avi* and *x.wmv* formats. All the extracted audio wave files have sampling rates of 44.1/48 kHz and are in stereo/mono mode. The data is downsampled to 16 kHz.

4 Structure of database

This database consists of segregated emotion clips in three formats namely, audio, video, and audio-visual. It has 1176 clips in each format. For ease of usage, a consistent structure is maintained for labelling the database, which is explained as follows.

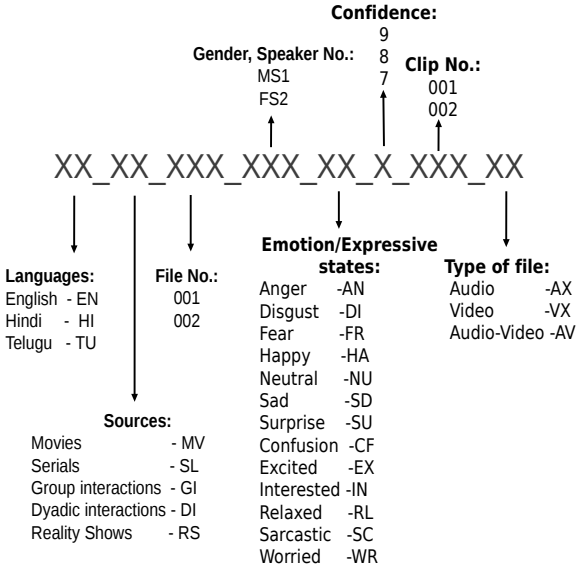

Figure 1: *Labelling structure of the segregated clips*

4.1 Labelling the raw data

The raw videos selected from the chosen source are labelled with a string of 9 characters as follows:
"XX_XX_XXX", where

- Characters 1-2 refer to *language code* [for example, the database collected in English language is coded as 'EN'].
- Characters 4-5 refer to the *kind of source* [for example code 'MV' specifies the source video as movie].
- Characters 7-9 refer to *source video number*.

Example: EN_MV_123

4.2 Labelling of segregated data

Each segregated clip of source video includes the raw label of source video along with labels specifying the gender, speaker, emotion category, confidence score and the type of file. The labelling scheme of segregated clips is as follows:
"XX_XX_XXX_XXX_XX_X_XXX_XX",
where

- Initial characters 1-9 are same as the *label of the raw source video*.
- Character 11 refers to the *gender* (M/F).
- Characters 12-13 refer to *speaker number*. Speaker numbering is kept consistent for all episodes of a particular soap-opera or movie [for example, code 'FS2' represents female speaker number 2].
- Characters 15-16 refer to *emotion category* [for example code 'AN' refers the particular clip to anger state].
- Character 18 refers to the *confidence score* [Range 2 to 9, 9 being highest].
- Characters 20-22 refer to the *clip number for a particular source video*.
- Characters 24-25 refer to the *type of clip* [for example codes 'AX', 'VX' and 'AV' specify the clip in audio, video and audio-visual formats, respectively].

Example: EN_MV_123_FS2_AN_9_106_AV

More details of labelling structure are given in Fig. 1.

The data can be sub-structured as per emotion, gender and speaker. It also has further levels of sub-structuring as per speaker-emotion and gender-emotion categories. The database consists of 1176 labelled clips, of which 741 clips are of male and 435 clips are of female speakers. The statistics of data as per emotion and per speaker is given in Tables II and III respectively.

Database also contains multiple emotions (one emotion followed by another) that occurred in a sentence continuously.

For example Anger followed by Frustration, Excitement followed by Anger or Happy etc.

Table 2: *Number of the clips per emotion/expressive state, with confidence score in each column (CX),(2 to 9, 9 being highest).*

Emotion	C9	C8	C7	C6	C5	C4	C3	C2	Total
1. Anger	5	24	60	44	27	7	2	2	171
2. Disgust	-	6	35	18	8	8	-	-	75
3. Fear	-	-	4	5	6	1	-	-	16
4. Happy	3	27	50	19	21	9	-	-	130
5. Sad	6	17	41	19	17	16	8	-	117
6. Surprise	5	11	28	27	8	12	2	-	93
7. Neutral	4	34	90	31	14	1	-	-	174
8. Confusion	-	1	2	4	4	-	-	-	11
9. Excited	5	19	77	28	11	8	1	-	149
10. Interested	-	4	46	11	5	4	1	-	71
11. Relaxed	-	-	5	2	4	3	-	-	14
12. Sarcastic	-	3	9	8	8	8	2	1	39
13. Worried	-	19	33	7	11	4	1	0	75

Table 3: *Number of the clips per speaker (SX).*

Speakers	S1	S2	S3	S4	S5	S6	S7	S8	S9
No. of clips	11	86	16	24	33	45	24	12	3
Speakers	S10	S11	S12	S13	S14	S15	S16	S17	S18
No. of clips	125	99	30	30	78	132	18	54	60

These are named as Multiple emotion Files (M-files). There are 41 such files obtained in this database. The labelling scheme of M_files is as follows:

"XX_XX_XXX_XXX_XXX_X_XXX_XX", where

- Initial characters 1-13 are same as the *label of the segregated data*.
- Characters 15-17 refers to the *starting emotion clip number*.
- Character 19 as M, refer to *M-file*.
- Characters 21-23 refers to the *ending emotion clip number*.
- Characters 25-26 refer to the *type of clip* [for example codes 'AX', 'VX' and 'AV' specify the clip in audio, video and audio-visual formats, respectively].

Example: EN_MV_123_FS2_106_M_107_AV

To analyze the inter-evaluator agreement, Fleiss Kappa statistic was computed (Fleiss J et al., 1981). The result for the entire database is 0.31. Since the emotional content of the database mainly span the target emotions (see Table. 1), the Kappa statistic was calculated for the emotional states alone and it turns that 0.41. These levels of agreement, which are considered as fair agreement, are expected since people have different perception and interpretation of the emotions and these values are consistent with the agreement levels reported in previous work (Steidl, S et al., 2009; C. Busso et al., 2008; M. Grimm et al., 2008).

The database is also labelled in dimensional approach in two dimensions (primitives) namely arousal and valence. The labelling structure is same as described in Fig. 1. except the characters 15-16 refer to two dimensions. The codes using two primitives, arousal (active-A, passive-P) and valence (positive-P, negative-N), forms 4 combinations namely AP (active-positive), AN (active-negative), PP (passive-positive) and PN (passive-negative). The neutral samples are labelled as NN (neutral-neutral).

The statistics of the of data as per dimensions is shown in Table IV.

5 Issues in data collection

The ambiguity in annotating the emotion is indicated by specifying the confidence scores. There

Table 4: *Number of clips as per two dimensions - arousal (active/passive) and valence (positive/negative).*

	Positive	Negative
Active	309	245
Passive	198	209

are two reasons for ambiguity of annotating the emotions. One of them is occurrence of mixed emotions in a sentence. For example, there is a possibility of combinations like, surprise-happy, frustration-anger, anger-sad, etc., occurring in the dialogue at the same time. For these cases, the subjects were asked to annotate the clip with multiple emotions along with confidence score for each. If there exist two sub-dialogues in a dialogue, each corresponding to different emotions, then they are segregated separately as M-files. If there is only one dialogue which has mixed emotions, the emotion with maximum confidence is selected. These kind of clips with entire dialogue are considered as *special cases*.

The second reason for ambiguity is unsustainability of emotion throughout the dialogue. In the case of natural communication among human beings, emotion being non-normal (emotional) speech, may not be sustainable for the duration of entire dialogue. The emotion is mostly expressed in some segments of dialogue, like at the end or at the beginning of a dialogue, with the rest of the dialogue being neutral. Hence, the corresponding emotion is given in the annotation.

We have also given confidence score for each audio-visual clip. It indicates the degree of confidence in the labelled emotion actually being present in the clip. Since the confidence score is given by only one person, the clips with less confidence scores and ambiguities can be used better after performing the subjective evaluation. Some of the clips also have abrupt cut-off due to interruption made by other actors before completion of the dialogue. Although this database is more generic and is closer to the natural spontaneous communication, it is still from the enacted source.

6 Possible applications

Due to variety in this database, applications like emotion recognition based on speaker dependent and independent, gender dependent and independent cases can be studied in audio alone, video alone and audio-visual modes. Identification of non-sustainable regions in an entire dialogue will be an interesting research problem. The clips with multiple emotions can also be used to study how an individual can vary his/her emotive state in a dialogue. The perceptual evaluation of these clips with only audio, only video and audio-visual analysis can also be performed. The subjective scores with only audio can be used as ground truth for evaluation of emotion recognition system based on audio.

7 Summary

In this paper, we have described the audio-visual emotion data collection, segregation and labelling of audio-video clips from movies and soap-operas in TV broadcast. It is assumed that the generic and natural communication among the humans can be reflected closely in these sources. The data is collected in three modes: audio, video and audio-visual. The labelling of gender, speaker and emotion is described. Issues in special cases like multiple emotions and non-sustainability of emotions in a dialogue are addressed. The database is still limited in number of clips. Data with sufficient number of clips covering many other cases need to be developed. In order to standardize the data and to know the perception of emotions by human beings, subjective evaluation need to be carried out in all three modes (audio, video and audio-visual).

Acknowledgement

The authors would like to thank all the members of Speech and Vision Lab, especially to B. Rambabu, M. Vasudha, K. Anusha, Karthik, Sivanand and Ch. Nivedita, for spending their valuable time in collection of the IIIT-H AVE database.

References

M. Gnjatovic, D. Rosner, "Inducing Genuine Emotions in Simulated Speech-Based Human-Machine Interaction: The NIMITEK Corpus," *IEEE Transactions on Affective Computing*, vol.1, no.2, pp.132-144, July-Dec. 2010.

E. Douglas-Cowie, N. Campbell, R. Cowie, and P.Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, pp. 33-60, 2003.

B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realisitic emotions and affect in speech: State of the art and lessions learnt from the

first challenge," *Speech Communication*, vol. 53, pp. 1062-1087, 2011.

D. Ververidis, C. Kotropoulos, "A review of emotional speech database,". *In 9th Panhellenic Conf. on Informatics*, November 123, 2003, Thessaloniki, Greece, pp. 560-574.

S. G.Koolagudi, K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, Volume 15, Issue 2, pp 99117. 2012.

Zhihong Zeng, M. Pantic, G.I. Roisman, T.S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.31, no.1, pp.39-58, Jan. 2009.

F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," *in Proc. Interspeech*, Lisbon, Portugal, pp. 1517-1520, 2005.

I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a Danish emotional speech database," *in Proc. Eurospeech*, Vol. 4, pp. 1695-1698, 1997.

B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, G. Rigoll , "Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies," *IEEE Transactions on Affective Computing*, vol.1, no.2, pp.119-131, July-Dec. 2010.

S. G. Koolagudi, S. Maity, V. A. Kumar, S.Chakrabarti, K.S. Rao, "IITKGP-SESC: speech database for emotion analysis," *In LNCS Communications in computer and information science*, Berlin: Springer, August 2009.

D. Erickson, K. Yoshida, C. Menezes, A. Fujino, T. Mochida, and Y. Shibuya, "Exploratory study of some acoustic and articulatory characteristics of sad speech," *phonetica*, Volume 63, p. 1-5, 2006.

R. Banse, and K. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, Vol. 70,no. 3, pp.614-636, 1996.

I. Sneddon, M. McRorie, G. McKeown, J. Hanratty, "The Belfast Induced Natural Emotion Database," *IEEE Transactions on Affective Computing*, vol.3, no.1, pp.32-41, Jan.-March 2012.

E. Douglas-Cowie, R. Cowie, and M. Schroeder, "A new emotion database: Considerations, sources and scope," *in proc. ISCA ITRW on Speech and Emotion*, Newcastle, pp. 39-44, Sep. 2000.

N. Amir, S. Ron, and N. Laor, "Analysis of an emotional speech corpus in Hebrew based on objective criteria," *in proc. ISCA ITRW on Speech and Emotion*, Newcastle, pp. 29-33, sep. 2000.

Steidl, S. Automatic classification of emotion-related user states in spontaneous childrens speech. Studien zur Mustererkennung, Bd. 28, ISBN 978-3-8325-2145-5, 1260 (January), 2009.

C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, *J. Language Resour. Eval.*, vol. 42,pp. 335-359, 2008.

S. Chung, "Expression and perception of emotion extracted from the spontaneous speech in Korean and English," Ph.D. dissertation, Sorbonne Nouvelle University, Paris, France.

K. Scherer and G. Ceschi, "Lost luggage emotion: A field study of emotion-antecedent appraisal," *Motivation and Emotion*, Vol. 21, pp. 211-235, 1997.

M. Grimm, K. Kroschel, and S. Narayana, "The Vera am Mittag German audio-visual emotional speech database," *in proc. IEEE int. Conf. Multimedia and Expo (ICME)*, Hannover, Germany, pp. 865-868, Jun. 2008.

Koelstra, S.; Muhl, C.; Soleymani, M.; Jong-Seok Lee; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I., "DEAP: A Database for Emotion Analysis ;Using Physiological Signals," *IEEE Tran. on Affective Computing*, vol.3, no.1, pp.18-31, Jan-Mar. 2012.

M. Schroder, "Emotional speech synthesis-a review," *in Proc. Eurospeech*, vol. 1, pp. 561-564, Aalborg, Denmark, 2001.

I. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," J. Acoust. Soc. Amer., pp. 1097-1108, 1993.

H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," *in Proc. Eurospeech*, pp. 2401-2404, Geneva, Switzerland, 2003.

D. Erro, E. Navas, I. Hernaez, and I. saratxaga, "Emotion conversion based on prosodic unit selection," *IEEE Trans. Audio. Speech, Lang. Process.*, vol. 18, no. 5, pp. 974-983, Jul. 2010.

C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 13, no. 2, pp. 293-303, Mar. 2005.

K. P. Truong, D. a. van Leeuwen, and F. M. G. de Jong, Speech-based recognition of self-reported and observed emotion in a dimensional space, *in Speech Communcation*, vol. 54, no. 9, pp. 1049-1063, Nov. 2012.

K. Scherer, "Vocal Communication of emotion: A review of research paradigms," *in Speech Communcation*, Vol. 40, pp. 227-256, 2003.

R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *in Speech Communcation*, Vol. 40, pp. 2-32, 2003.

Fleiss J, "Statistical methods for rates and proportions". New York, NY, USA: John Wiley & Sons, 1981.