An Open Source Emotional Speech Corpus for Human Robot Interaction Applications

Conference Paper · September 2018

DOI: 10.21437/Interspeech.2018-1349

3 authors, including:

Jesin James
University of Auckland
12 PUBLICATIONS **28** CITATIONS
SEE PROFILE

Catherine I Watson
University of Auckland
90 PUBLICATIONS **1,136** CITATIONS
SEE PROFILE

Interspeech 2018
2-6 September 2018, Hyderabad

An Open Source Emotional Speech Corpus for Human Robot Interaction Applications

Jesin James[1], *Li Tian*[1], *Catherine Inez Watson*[1]

Department of Electrical and Computer Engineering, University of Auckland, New Zealand

jjam194@aucklanduni.ac.nz , tli725@aucklanduni.ac.nz, c.watson@auckland.ac.nz

Abstract

For further understanding the wide array of emotions embedded in human speech, we are introducing a strictly-guided simulated emotional speech corpus. In contrast to existing speech corpora, this was constructed by maintaining an equal distribution of 4 long vowels in New Zealand English. This balance is to facilitate emotion related formant and glottal source feature comparison studies. Also, the corpus has 5 secondary emotions and 5 primary emotions. Secondary emotions are important in Human-Robot Interaction (HRI) to model natural conversations among humans and robots. But there are few existing speech resources to study these emotions, which has motivated the creation of this corpus. A large scale perception test with 120 participants showed that the corpus has approximately 70% and 40% accuracy in the correct classification of primary and secondary emotions respectively. The reasons behind the differences in perception accuracies of the two emotion types is further investigated. A preliminary prosodic analysis of corpus shows significant differences among the emotions. The corpus is made public at: github.com/tli725/JL-Corpus.

Index Terms: emotional speech corpus, primary and secondary emotions, perception test, prosody

1. Introduction

A simple one-channel speech producing system that transmits explicit verbal messages is not sufficient for modern Human-Robot Interaction (HRI) applications. The interacting robots at the receiving end should be able to detect the implicit intentions, motives, and physiological clues from emotions embedded in speech spoken by the transmitting end and react accordingly. Hence, with the rapid advancements of artificial intelligence in HRI, a more in-depth and systematic study of emotions becomes essential for their automatic recognition and synthesis. The design of an emotional speech corpus can deeply affect the quality of emotion recognition and synthesis studies. The attributes of emotional speech corpora are *naturalness, scope*, and *context* [1-2]. The *naturalness* of a corpus refers to whether the emotions were produced artificially or spontaneously. Emotional speech corpora with simulated, induced (semi-natural) and natural emotions have been developed [4]. Whilst corpora with naturally-produced emotions contain spontaneous emotional speech, the precise nature of the underlying emotion is not often captured due to the uncertainty in expressed emotion intensity and uncontrolled recording manner. This uncertainty makes it difficult to categorize some recordings to target emotional states. Also, retrieving natural speech data is complicated by privacy issues. So, for practical applications where prototypical emotional data are needed, induced or simulated corpora are preferred [1, 5]. The *scope* of a corpus is determined by the variations that are incorporated into it, like the emotion categories, gender, number of speakers and the modalities of emotion expressions. A wider scope is essential for a generalized corpus. The *context* of the corpus means the choice of speech material included. A common strategy for induced and simulated corpora is to use sentences without any emotion salient words [20] (e.g. interjections like "Alas!", "Heavens!"). These type of sentences are called emotionally neutral. These neutral sentences remove the semantic influence when speakers express the emotions. The naturalness, scope and context depend on the application in which the corpus is intended to be used. A summary of 5 commonly used open-source speech corpora with simulated emotions is summarized in Table 1. Most of these corpora contain only Primary emotions [1], with a few corpora (eg: IITKGP-SESC [12]) including some Secondary emotions. The sentences used in the corpora are all emotionally neutral. The research goal addressed in this paper is to develop robots that interact socially with people. This requires the robot to be able to recognize and speak a wider range of emotions than only the primary emotions (Secondary emotions are heavily used in social interactions by humans). An affective response (via speech - emotional speech synthesis) of the robot to human emotions that it can sense (emotion recognition) is the proposed research requirement. This can be extended to the modelling of empathetic behaviour expressed by social robots [27, 28]. Analysing the emotions in the existing databases were insufficient to cater to the wide range of emotions people portray in social conversations. Most of the existing corpora have a good coverage of all phonemes in the language. But for the study of vocal expression of emotions at the intra-segmental level, the balance of different vowel types whose features were identified as emotion-related [8] is also required. Such balance is absent in most existing corpora (e.g. the number of /i:/ tokens is two times more than /o:/ tokens in EmoDB [9]). To address these two concerns, this paper introduces a balanced open source emotional speech corpus which can be used for a systematic analysis of emotions used in HRI.

2. Emotional Corpus Design

The design of the Emotional speech corpus was done to address the requirements for following attributes:

1) *Naturalness*- A speech corpus with strictly-guided simulated emotions has been developed. To ensure that the emotions expressed by the speakers were an adequate reflection of reality, the emotional state was aroused into the speakers using two strategies. Firstly, a variation of the Stanislavski method [23] was used to induce the emotion into the speakers. Before each emotion recording, a situation was explained to the speaker along with some images conveying the emotion. This made the speakers' mental state aligned to that emotion before speaking.

[1]Primary Emotions are emotions that are innate to support fast and reactive response behavior. Eg: angry, happy, sad, fear. Secondary emotions are assumed to arise from higher cognitive processes, based on an ability to evaluate preferences over outcomes and expectations. Eg: relief and hope [14, 15]

10.21437/Interspeech.2018-1349

Table 1: *Common Emotional speech corpora*

Corpora	Emotion Types	Speakers	Language	Context
German emotional corpus [9]	Angry, boredom, disgust, fear, happy, neutral,sad.	5 male speakers, 5 female speakers	German	Neutral sentences
PAVOQUE corpus [10]	Angry, fear, happy, neutral, sad, poker	1 male speaker	German	Neutral sentences
SAVEE emotion corpus [11]	Angry, disgust, fear, happy, sad, surprise, neutral	4 male speakers	English	Neutral sentences
IITKGP-SESC corpus [12]	Angry, compassion, disgust, fear, happy, neutral, sarcastic, surprise	5 male speakers, 5 female speakers	Telugu	Neutral sentences
Danish emotional corpus [13]	Angry, happy, neutral, surprise, sad	2 male speakers, 2 female speakers	Danish	Neutral sentences, words, passages

These images were displayed throughout the recording process so that the emotional state can be maintained. Secondly, for secondary emotions, the speakers were asked to speak leading phrases before the emotionally neutral sentences. For example, before sentences with encouraging emotion, the speakers said "Well done", and then continued with a neutral sentence like "Linda asks for more darts" (These leading phrases were not included in the actual corpus.). These 2 methods of emotion induction makes this a strictly-guided simulated corpus. Preparing different speakers with the same leading phrases and images helped them reach the same level of emotion intensity and they were required to maintain it throughout the recording.

2) *Scope*- This corpus provides the resources for emotion synthesis and recognition using speech as the modality, thus only the audio recordings were included. This corpus is divided into 2 sections based on the emotions included, which are the 5 primary and 5 secondary emotions. The primary emotions are happy, angry, neutral, sad, excited and the secondary emotions are enthusiastic, apologetic, pensive, worried, and anxious. The choice of the primary emotions has been made based on current research works focusing on these emotions [9-13,17], making it possible to compare the developed corpus with existing ones. The choice of the secondary emotions is based on the conclusions from [31], that these subtle emotions are relevant for the speech synthesis of social robots in HRI. A visual representation of the emotions on a Valence Arousal (V-A) 2D [2] plot is shown in Figure 1. In the plot, the position of primary emotions has been marked in upper-case blue letters and secondary emotions in lower-case red letters. It can be seen that the corpus has been designed to span over a large part of V-A plane. The addition of secondary emotions has extended the scope of the corpus to regions that are variations and combinations of the primary emotions. To increase the scope, the speech was recorded from 4 speakers. All the speakers (two male and two female) were trained voice actors (two current broadcasters, one broadcasting tutor and one broadcasting trainee) of New Zealand English.

3) *Context*- Most existing corpora have only addressed the overall phonetic balance without equal distributions for specific vowel types. The waveforms of long monophthongs (eg: /a:/, /o:/) can be easily decomposed into vocal tract and vocal source parts due to their superior quasi-periodicity. Previous studies suggest these vowels can be used to effectively predict emotions [19, 22]. A balanced design can facilitate emotion related formant and glottal source feature comparison across vowel types, as different vowel types do not perform equally

[2]V-A 2D plot is based on Russel's circumplex model of emotions [7]. Valence level indicates the pleasantness of the voice ranging from unpleasant (Eg: sad, fear) to pleasant (Eg: happy, calm). The arousal level specifies the level of reaction to stimuli and range from inactive (Eg: sleepy, sad) to active (Eg: anger, surprise).

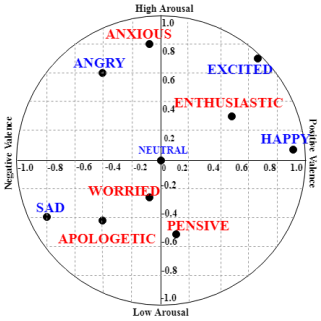

Figure 1: *V-A plane showing the emotions in the corpus*

in emotion recognition [21]. So, the speech material of this corpus has been chosen carefully to allow for equal number of 4 English long vowels - /a:/, /o:/, /i:/ and /u:/, each with 10 occurrences in 15 different sentences. The semantic context of all the sentences in the corpus were kept the same for all primary emotions without any emotional bias [16]. For the secondary emotions, a different strategy was adopted. During the pilot recording experiment, emotionally neutral sentences were used for Secondary emotions as well. Secondary emotions have only subtle characteristics compared to primary emotions. Without any emotion salient words the speaker found it difficult to reach the emotion state. For example, it was challenging to portray apology without words like "sorry". Due to this difficulty, first two sentences were chosen to have emotion salient words. Eg: for apologetic emotion, a sentence included was "I am sorry for your loss". These were followed by 13 emotionally neutral sentences. With the addition of these sentences the gap between speakers' normal mental state and the targeted emotional state could be narrowed and they also found it easier to maintain the required emotion intensity. Occasionally in a long sentence, the expected emotion may deviate to a different one, especially for some pairs, like "excited" and "happy". To ensure that only one emotion was maintained throughout the whole utterance, short sentences with 4-7 syllables were chosen. An example of a sentence used was "Jack vi**ew**s an **ar**t p**ie**ce". The monophthongs have been marked with the bold letters.

These speakers were asked to portray the emotions naturally, without any exaggeration. The recording was split into two sessions, conducted on different days to account for speaker's possible psychophysiological abnormality at the time of recording due to unexpected events beforehand. Each session was further split into primary emotion and secondary emotion sections. There was a 30 minutes rest interval in between the two sections. Each recording session took 120 minutes to complete. To conduct the recording in a natural and reverberation free condition, the recording took place in a soundproofed

room with an acoustic separation between the speaker and the recording computer. The level of ambient noise in the room was 18 dB(A). An AKG C460B microphone and a Roland Octa-Capture pre-amplifier (set to 25 dB) were used to collect the speech data. The speech signal was sampled at 44.1kHz and stored as 16-bit numbers. While recording, the speakers were required to sit upright, with their lips at a distance of 15 cm from the microphone. The sentence prompt was changed to the next one only when the researcher hearing the recorded sentences in real-time was satisfied with the quality and intensity of the emotion. In total, there are 4 (speakers) $\times$ 5 (primary emotions) $\times$ 2 (repetitions) $\times$ 15 (sentences) $\times$ 2 (sessions) = 1200 primary emotion sentences and 4 (speakers) $\times$ 5 (secondary emotions) $\times$ 2 (repetitions) $\times$ (13 (emotion neutral sentences) + 2 (emotion salient sentences)) $\times$ 2 (sessions) = 1200 secondary emotion sentences, making a total of 2400 sentences, with a footprint of 520 MB. (The full list of sentences is included in the online repository). Overall, this corpus was designed to cater to requirements of researchers currently working in the areas of emotion classification and synthesis studies for HRI. Human perception tests (to validate the quality of the corpus) and pilot prosody analysis were conducted. This is explained in detail in the following sections.

3. Emotional Corpus Evaluation

Once the corpus was prepared, a large-scale human perception test was conducted to evaluate it. Such an evaluation is needed to standardize the corpus by removing recordings that may not satisfy the required emotion levels. A testing environment was set up, where each participant had to listen to 60 recordings randomly chosen from the 2400 recordings. As the corpus consists of 2 parts - Primary and Secondary emotions, the perceptual validation was also conducted on each part separately. Moreover, based on the feedback of a pilot perception experiment, participants found it very difficult to discriminate a mixture of 5 primary and 5 secondary emotions via speech. Each participant listened to a sentence set and judged the emotion from a pre-set list of emotions, including an "unsure" option. The perception test was conducted via an online survey platform. So, the users could finish the test at their preferred location with a headset and a computer. An online survey platform was chosen to maximize the number of participants, which helps to generalize the findings. The participants were not required to have any specific knowledge about the role of emotions in speech. This generalized participation was chosen as the corpus will be used in applications where the users may also not have any such knowledge. The perception test was completed by 120 participants aged 16-45, with 60 people each evaluating the primary and secondary emotions. Based on their self-reporting, all participants had above average hearing ability, with 50 participants being first language New Zealand English speakers and the remaining 70 were bilingual speakers who spoke English on a daily basis. All participants completed the perception test, with 20% and 80% of them using loudspeakers and headphones respectively. Each participant took approximately 25 minutes for the test. The design of the perception test was done such that each recording was evaluated by three participants.

The emotion perceptual recognition performance is shown in the form of confusion matrices (Figure 2 (a) and (b) for primary and secondary emotions respectively). The "unsure" choices (6.7% of total responses) were neglected for the analysis. The last column of the confusion matrix summarizes the hit rate for each emotion (in grey cell) and the overall hit rate for all emotions (in blue cell). The bold black numbers in each cell

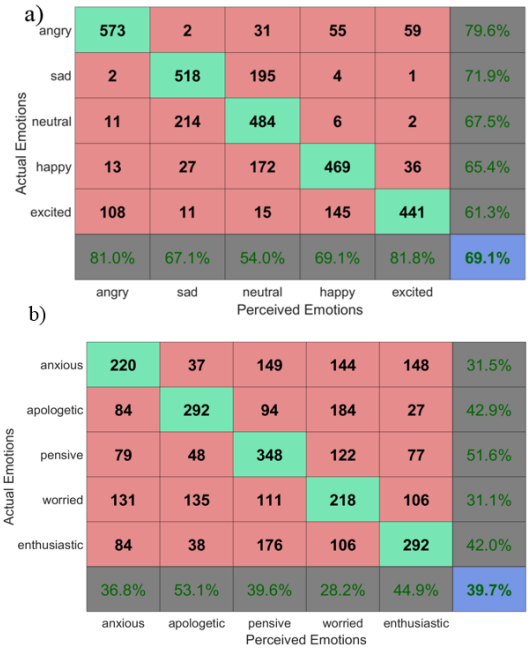

Figure 2: *Primary emotions (a) and Secondary emotions (b) perception confusion matrices*

indicate the number of participants' judgments which classified each of the actual emotions listed on the left side as the perceived emotions listed at the bottom of the table. It can be observed from Figure 2(a) that angry and sad have relatively high perception hit rates, reaching accuracies of 79.6% and 71.9% respectively. Excited is the emotion that was most difficult to identify (hit rate = 61.3%). It was misclassified as either happy or angry. The emotions highly confused with angry were happy and excited. Also, it can be noted that sad sentences were highly misclassified as neutral. A majority of the wrongly classified neutral sentences were marked as sad and happy. It can be seen that angry or excited are rarely perceived as sad. Also, sad and neutral were rarely confused with excited. These findings can be explained by the Valence-Arousal emotion model [7] (refer Figure 1). The emotions well apart on the arousal dimension are more easily differentiated than those on the valence dimension. This implies a complete emotion separation cannot be perceptually achieved and more emphasis should be put on the valence level when complementing with other modalities like video for a multi-modal emotion recognition. In summary, the overall hit rate combining all of the primary emotions is 69.1%. This result is comparable with some existing corpora [12, 13]. The secondary emotion perception test confusion matrix is shown in Figure 2(b). The best perception hit rate was obtained for pensive (51.6%), with some tokens being confused with worried. This can be related to their closeness in the V-A plot (Figure 1). Enthusiastic and apologetic have classification rates around 42%, and worried was the most wrongly identified emotion. The few misclassifications between apologetic vs. anxious, apologetic vs. enthusiastic shows that these emotions are less confused. Some confusions like worried vs. apologetic, anxious vs. enthusiastic can be expected as people often relate to certain words to perceive them. In order to understand whether the emotionally salient words had an impact on the perception of emotions, a comparison of the results of two separate sets of sentences in the secondary emotion corpus- emotionally neutral sentences and emotion salient sentences was done. The

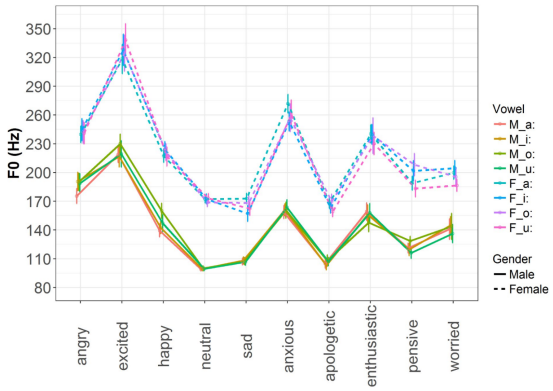

Figure 3: *Mean F_0 trend for Primary and Secondary emotions*

average hit rate of the emotion salient sentences alone was 60%, while the neutral sentences alone had a much lower hit rate of 33%. This indicates that emotion salient words enhanced the perception rate of the participants, while the context neutrality of the sentences made it difficult to perceive the secondary emotions. This proves that the emotion salient words in the sentences play a significant role in influencing peoples perception of secondary emotions whose variations are subtle compared to primary emotions. Overall, the reasons for the misclassifications for both primary and secondary can be associated to the participants familiarity with understanding emotions in speech (the test was conducted with a broader population, and not with people trained in speech studies) as well as the context-neutrality of the sentences used.

The corpus was screened by selecting sentences that were correctly classified by at least 2/3 people, thus retaining 60% of the total corpus (comparable to existing corpus retention percentages after perception test [9]). After removing the poorly expressed recordings, the remaining corpus achieved 86% and 73.1% emotion perception accuracy for primary and secondary emotions respectively. Post-processing was done on the perceptually evaluated corpus to make it suitable for feature extraction and prosody analysis. The corpus were labeled at the word and phonetic levels by the Munich Automatic Web Segmentation System, webMAUS [24]. MAUS has a New Zealand English option. However, hand checking of vowel boundaries is still required. This need has been noted by other studies [26]. In this case, we found around 15% of the webMAUS marked boundaries in the corpus were not accurate and they were manually corrected afterwards. We converted the labeled data into an EMU formatted corpus [25], followed by the EMU-supported extraction of each voiced segments pitch contour. Phonetic labelled recordings, their F_0 contours and all of the preparation material of this corpus (JL corpus) are made public at: github.com/tli725/JL-Corpus.

4. Emotional Corpus Analysis

Preliminary analysis of the corpus was conducted based on 2 prosody parameters - Fundamental frequency, F_0 and Speech rate, with changes in the emotions. Figure 3 shows the F_0 analysis for the primary and secondary emotions. The plots depict how the mean F_0 of 4 vowels vary along with the emotions for male and female speakers. The error bars represent the F_0's 95% confidence interval (CI). The equal distribution of 4 vowels allows direct comparison across vowel types. There are substantial changes in F_0 across different emotions. These trends are steady and comparative for both genders, with female F_0 values higher than male F_0 values. The F_0 of different vowels behave similarly across different emotions and the between-vowel within-emotion differences are less pronounced than the between-emotion within-vowel ones. For low arousal emotions (eg: sad, apologetic), the F_0 of different vowels are lower and more converged, especially for male speakers, as expected [29]. An analysis of the speech rate in syllables/second [30] across different emotions was done (Figure 4 - significantly different emotions pairs obtained from a pairwise T-test is marked by arrows). The emotion with the lowest average speech rate is sad, and is significantly different from all other emotions except apologetic. Excited has the highest average speech rate, followed by anxious and enthusiastic. Pensive showed significant difference in speech rate from all other emotions, and another significantly different pair is enthusiastic and neutral. The difference in the arousal levels between worried vs. anxious, excited vs. happy (having similar valence levels) is evident from their speech rate variation. The speech rate variations are not as pronounced as the F_0 due to the short duration of the sentences in the corpus (Noted in other studies [13]). The preliminary results with Mean F_0 and speech rate do provide reasonable separation of the emotions, but a deeper prosodic analysis is still required. Even though the secondary emotions did not perform well in the perception test, their prosodic features still suggest significant differences. This shows that the corpus can be used as a reliable tool to study the secondary emotions.

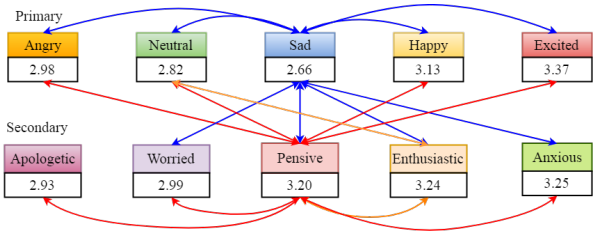

Figure 4: *Mean speech rate for all emotions. Arrows indicate significantly different emotion pairs ($p<0.005$**, Paired T-Test)*

5. Conclusion

This paper discusses the development of a strictly-guided simulated emotional speech corpus in New Zealand English. The investigation of emotion-related features in phonetic units like vowel segments can be supported by this corpus. Due to the equal distribution of the 4 vowels in the corpus, the intra-segmental feature comparison extracted from them can yield more reliable findings without any quantity bias due to more tokens from any vowel type. Another novelty of this corpus is the addition of secondary emotions. Secondary emotions are of significance in social interactions between humans and also need to be studied in the context of HRI. Synthesizing such subtle emotions poses challenges as they are not easily differentiated as the primary emotions from the valence and arousal levels. This difficulty in differentiation has been noted in the Perception test conducted. Even so, such a corpus is required to enable researchers to study these emotions and synthesize them. Preliminary acoustic analysis of the corpus has been conducted, which shows variations in the basic prosody parameters across vowels and emotions. The variations are being used for emotion classification using the long vowels and emotional speech synthesis of the secondary emotions for HRI applications.

6. Acknowledgements

This work is funded by Centre for Automation & Robotic Engineering Science (CARES), University of Auckland. We thank the participants of the perception test for their time and effort.

7. References

[1] Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication Journal, EURASIP*, 2015.

[2] Murray, Iain R and Arnott, John L, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *The Journal of the Acoustical Society of America, 93 (2)*, pp. 1097–1108, 1993.

[3] Cowie, Roddy and Cornelius, Randolph R, "Describing the emotional states that are expressed in speech," *Speech Communication, 40 (1-2)*, pp. 5–32, 2003.

[4] Scherer, Klaus R and Banse, Rainer and Wallbott, Harald G, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-cultural psychology, 32 (1)*, pp. 76–92, 2001.

[5] Scherer, Klaus R, "Vocal communication of emotion: A review of research paradigms," *Speech Communication, 40 (1-2)*, pp. 227–256, 2003.

[6] Koolagudi, Shashidhar G and Rao, K Sreenivasa, "Emotion recognition from speech: a review," *International journal of speech technology, 15 (2)*, pp. 99–117, 2012.

[7] J. A. Russel, "A circumplex model of affect," *J.Personality and Social Psychology*, pp. 1161–78, 1980.

[8] Airas, Matti and Alku, Paavo, "Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient," *Phonetica, 63 (1)*, pp. 26–46, 2006.

[9] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B, "A database of German emotional speech," *Interspeech, Lisbon, Portugal,*, pp. 1517–1520, 2005.

[10] Steiner I, Schrder M, Klepp A. "The PAVOQUE corpus as a resource for analysis and synthesis of expressive speech," *In: Proc. Phonetik and Phonologie*, 2013.

[11] S. Haq and P.J.B. Jackson, "Multimodal Emotion Recognition," *In W. Wang (ed), Machine Audition: Principles, Algorithms and Systems, IGI Global Press, ISBN 978-1615209194*, pp. 398–423, 2010.

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

[13] Engberg, Inger S and Hansen, Anya Varnich and Andersen, Ove and Dalsgaard, Paul, "Design, recording and verification of a Danish emotional speech database," *Fifth European Conference on Speech Communication and Technology*, 1997.

[14] Damasio, A., "Descartes error, emotion reason and the human brain," *Grosset/Putnam*, 1994.

[15] Christian Becker-Asano, Ipke Wachsmuth, "Affective computing with primary and secondary emotions in a virtual human," *Autonomous Agents and Multi-Agent Systems*, 2010

[16] Cowie, R., Douglas-Cowie, E., Cox, C, "Beyond emotion archetypes: databases for emotion modelling using neural networks," *Neural Networks 18*, pp. 33-88,2005.

[17] Ververidis, D., Kotropoulos, "A review of emotional speech databases," *In: PCI 2003, 9th Panhellenic Conf. on Informatics,Greece*, pp. 560-574, 2003.

[18] Tamagawa, Rie and Watson, Catherine I and Kuo, I Han and MacDonald, Bruce A and Broadbent, Elizabeth, "The effects of synthesized voice accents on user perceptions of robots," *International Journal of Social Robotics, 3 (3)*, pp. 253-262, 2011.

[19] Ringeval, Fabien and Chetouani, Mohamed "A vowel based approach for acted emotion recognition," *In Ninth Annual Conference of the International Speech Communication Association* , 2008.

[20] Krothapalli, Sreenivasa Rao and Koolagudi, Shashidhar G "Characterization and Recognition of Emotions From Speech Using Excitation Source Information," *International Journal of Speech Technology, 16 (2)*, pp. 181-201, 2013.

[21] Tian, Li and Watson, Catherine "Continuous spoken emotion recognition based on time-frequency features of the glottal pulse signal within stressed vowels," *In: Proc. of of the 16th Australasian International Conference on Speech Science and Technology SST* , pp. 285-288,2016.

[22] Airas, Matti and Alku, Paavo "Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient," *Phonetica, 63 (1)*, pp. 26-46, 2006.

[23] Benedetti, Jean, "Stanislavski: An Introduction." *London: Methuen* , 1982.

[24] Kisler, T. and Schiel, F. and Sloetjes, H., "Signal processing via web services: the use case WebMAUS," *Proceedings Digital Humanities, Hamburg, Germany*, pp. 30-34,2012.

[25] Cassidy, S. and J. Harrington, "Multi-level annotation in the Emu speech corpus management system," *Speech Communication, 33*, pp. 61-77,2001.

[26] Kisler, T. and Reichel U. D. and Schiel, F., "Multilingual processing of speech via web services," *Computer Speech and Language*, pp. 326-347,2017.

[27] Achim Stephan, "Empathy for Artificial Agents," International Journal of Social Robotics, 2015.

[28] Minoru Asada, "Towards Artificial Empathy: How Can Artificial Empathy Follow the Developmental Pathway of Natural Empathy?," International Journal of Social Robotics, 2015.

[29] Vogt, Thurid and Andr, "Improving automatic emotion recognition from speech via gender differentiation," *In Proc. Language Resources and Evaluation Conference , Genoa*, 2015.

[30] Chung Ting Justine Hui, Teh June Chin, Catherine Watson, "Automatic detection of speech truncation and speech rate," *In Proc. Australasian International Speech Science and Technology Conference* , 2014

[31] Jesin James, Catherine Watson, Bruce MacDonald, "Artificial Empathy in Social Robots: An analysis of Emotions in Speech," *In Proc. IEEE International Conference on Robot and Human Interactive Communication* , 2018