World Wide Journal of Multidisciplinary Research and Development (September-2022)

WWJMRD 2022; 8(09): 96-102
www.wwjmrd.com
International Journal
Peer Reviewed Journal
Refereed Journal
Indexed Journal
Impact Factor SJIF 2017: 5.182 2018: 5.51, (ISI) 2020-2021: 1.361
E-ISSN: 2454-6615

Jaini Gandhi
JBCN International School,
Borivali Mumbai,
Maharashtra, India.

Aqsa Temrikar
JBCN International School,
Borivali Mumbai,
Maharashtra, India.

Correspondence:
Jaini Gandhi
JBCN International School,
Borivali Mumbai,
Maharashtra, India.

Identification of formant values in pre-recorded audio samples displaying various emotions using Praat software.

Jaini Gandhi (Student), Aqsa Temrikar (Mentor)

Abstract
Detection of emotions in voices can be crucial to understanding how these emotions can be displayed in synthesized voices. Emotion detection can be done by extracting standard formant frequencies from an existing dataset which displays 6 basic emotions according to Paul Ekman – happiness, sadness, fear, disgust, surprise, and anger – and using neutral voice audio to then compare the frequency values. Analysis is done using RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song). This led to finding average formant values F1, F2, F3 for aforementioned emotions. Displaying emotions in synthesized voices can open an entire new realm in terms of enhanced user experience and better AI-human understanding.

Keywords: speech, emotions, formant, synthesis, audio.

1. Introduction

Inside Out, a 2015 American animated film, portrayed the effect of emotions and memories on one's behaviour. It displayed how the same expression can have multiple meanings when said with different emotions. This is one of the prominent concerns encountered in the voice synthesis industry. Voice synthesis is a computer simulation of human speech; hence, they are devoid of emotions which cause dissimilarity in the replication of voice.

Being the most vital technology, speech synthesis has numerous significant and substantial applications. Speech synthesis aids in eradicating barriers for people with impairments, most frequently among visually impaired people and those with speech impairments. One of the most notable personalities to make use of a synthesised voice to communicate was Stephen Hawking. While it gave Professor Hawking the ability to communicate, many described his voice to also be robotic sounding.

In addition, speech synthesis techniques are also used in the entertainment and games industry. Animo, a speech synthesiser developer, and Celsys, an Anime and Manga creation tool software, collaborated in 2007 to create a software based on FreeSpeech which was used to create line narrations according to user specifications. Moreover, virtual assistant devices like Google Home, Alexa, Siri, and Cortana also make use of natural language processing interfaces to create valuable aid for interactions in systems with ARM architecture. In recent years, synthesised voices are also being applied in robots, specifically care bots, for interactions with patients. Introduction of emotions in these synthesised voices, therefore, will enhance user experience and allow better communication and care.

Furthermore, when Sundar Pichai the CEO of Google announced the launch of LaMDA 2 at Google I/O conference 2022, many problems in the field of synthetic voice and communication were resolved. However, despite the excellent conversation skills and high EQ the one thing that made it lack the nature of "human" communication was emotion. Being able to fine-tune on that aspect and develop a model which can detect and give an emotional speech will take this model to the next level.

There are various methods to synthesise voice, which are classified mainly into three categories: concatenative synthesis, articulatory synthesis, and formant synthesis. The latter was the most appropriate method for this research paper, considering the computational load, the preciseness of the speech, and its flexibility (can produce an infinite number of sounds as compared to the concatenation method).

Even though emotions are an integral part of human speech, emotional simulation in synthesised voices is not yet an evident feature. Emotional expression in speech can be influenced by multiple parts of the sound signal: spectral, prosodic, phonetic, idiolect, semantic. However, for practical purposes, this paper is only going to analyse the formants[1] produced by the phonetics to idealise the average F1, F2, F3 values which can be added in a neutral speech to produce a particular emotion.

This report begins with a detailed literature review which was used to develop the framework and methodology of the research as well as familiarise with research in voice synthesis. This is followed by a brief description of the methodology used. Later the influence of formants on emotions is explained with an elaborate detail of the primary research that took place. A comprehensive analysis of the data collected is detailed and a reasonable conclusion is drawn from the data.

2. Literature Review

2.1 Ekman, P. (1992). Are there basic emotions? Psychological Review, 99(3), 550–553

Paul Ekman discussed in this study that as shown in the book 'Emotion in the Human Face: Guidelines for Research and an Integration of Findings' (Ekman, P et.al 1971), there has been evidence of only 6 basic emotions - happiness, sadness, fear, surprise, anger, disgust. He shares that while these are not the only types of emotions, almost all the emotions can be placed into these basic emotion families.[2] Each of these emotion families differ from each other through certain characteristics and each emotion in a family has similar characteristics to others in the same family. Ekman's theory also postulates that (except the emotion families mentioned above) no other emotion shows cross-cultural evidence of facial expression. Using this evidence, the author has reduced the wide range of emotions to the 6 basic emotions suggested by Paul Ekman.

2.2 BANSE, Rainer, SCHERER, Klaus R. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 1996, vol. 70, no. 3, p. 614-636

BANSE, Rainer, SCHERER, Klaus R. (1996) noted that listeners can, with an accuracy of 50%, infer the speaker's affective state from their voice. Different emotional states of the speaker can be determined through their vocalisations. They strongly suggest that several variables of the sound signal are directed by the changes in articulation, respiration and phonation (which are heavily influenced by input of emotions in voice). These variables include range and contour of fundamental frequency, amplitude, distribution of energy in the spectrum and the location of the formants. Taking inspiration from this paper, the author has decided to take one of the variables into account (the location of the formants) for the 6 basic emotions. Rainer and Klaus also predict that unpleasant emotions lead to a rise in the F1 mean and a fall in the F2 mean, which is anatomically related to the faucal and pharyngeal constriction and tensing of the vocal tract walls. The conclusion shown in this paper was that 12 out of 14 emotions were identified with high accuracy levels, ensuring that emotions can be recognised through full utterances.

2.3 Carlson, R., Granström, B., Nord, L., 1992. Experiments with emotive speech—acted utterances and synthesised replicas. In: Internat. Conf. on Spoken Language Processing, Banff, Canada, pp. 671–674.

This paper used pre-recorded speech material and analysed the spectrograms produced for basic emotions - happy, sad, angry - and neutral. The differences in the spectrograms were highlighted mainly between happy and sad: 1] duration and fundamental frequency (F0) 2] change in tempo and pitch movement 3] phonetic differences like angry and happy sounds showed more energetic articulations. These differences were carefully noted and then applied to synthetic voices and put to listening tests, which showed that sad was successfully conveyed when originated from happy and angry utterances, angry was also strongly signalled when originated from happy and sad utterances. This model gave a good understanding of how the phonetic differences make a difference in spectrograms hence, making it evident that formants play an important part in displaying emotions.

3. Methodology

This research demands high technology, fully functioning sound labs to facilitate complete primary research. Due to unavailability of those, the author made a conscious decision to source the audio database over the internet.

Thus, the research was conducted in two parts:

1. Secondary sourcing: This paper makes use of the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) (Livingstone and Russo, 2018) database where 24 actors, with an average age of 26, vocalise lexically matched statements in a neutral North American accent. This dataset was rated 10 times, by 247 untrained individuals, for emotional intensity and validity. The time constraint and the academic capabilities have restricted the author from analysing all 24 actors in this paper; therefore, this paper makes use of only actor 7 vocal channel speech, statement 1 ('kids are talking by the door') audios at strong intensity for all emotions: happiness, sadness, fear, anger, disgust, surprise and neutral.
2. Primary research: The analysis of the sound was done using an open-source software Praat (designed and developed by Paul Boersma and David Weenink). The sounds were first carefully edited to remove any blank spaces in speech to ensure that false values were discarded and then the formants were extracted from the formant listing feature of Praat. An average value of F1, F2 and F3 was calculated for each emotion. These estimated values could be used as variables in the source-filter model of voice synthesis and create sounds with emotions.

4. Emotions, speech and formants

As this research's primary emphasis is the relation between emotion and speech, these two domains need to be

[1] each of several prominent bands of frequency that determine the phonetic quality of a vowel

[2] a group of things related by common characteristics

appropriately described and defined because of the notoriously fluid use of both in everyday language.

The key belief about the meaning of emotion is that they are episodes which are brief and highly distinctive. However, there are many proposed terms to describe these emotions. Like (Pultchik, 1984) describes this as primary emotions, while (Stein and Oatley, 1992) and (Ekman, 1999) term it as basic emotions. Emotions are referred to as an aspect of most mental states instead of a special kind of state. These emotion categories were first assumed (Anscombe and Geach, 1970) to be categorically designed as a colour wheel, where there were few key categories - primary, and the other emotions were acquired from mixing these primary emotions. This paper moves ahead with the widely accepted and supported referral to the term "basic emotions" proposed by (Ekman and Davidson, 1994) that suggests that qualitatively prominent states are considered universal and linked with the evolution of the brain system to be able to cope with different situations. This was also backed with (Ekman, 1992) which suggested the evidence of only 6 basic emotions - happiness, sadness, anger, fear, surprise, disgust.

Long throughout the past, emotional expression in speech has held immense interest. Speech is a vocal communication method through languages used by humans. Each language has distinct features and uses different phonetic and semantic features of words, in addition to the paralinguistic aspects of vocalisation, to express the meaning of a phrase. Emotions are used to "colour" the language, ensuring a meaningful and direct conversation between two humans. However, only in recent years were the paralinguistic and pragmatic features of speech considered – with special attention to speaker's attitudes and emotions – while building speech technology applications. Over the years, it has been believed that prosodic features of speech are primary indicators of the emotional state of the speaker; several research have indicated that energy and formant frequencies are one of the potentially effectual ways to measure the emotional state (Khanna and Sasikumar, 2011).

Formants are spectral peaks in a spectrum having high degrees of energy. They are also referred to as acoustic resonance of the human vocal tract; these are sculptured using filter models whose frequencies fall near the formants. When a speaker talks, for example, there is a change in the shape of their vocal tract which results in a variable acoustic filter (Kamiloğlu, Fischer and Sauter, 2020). The formants are numbered as F1, F2 and so on. Each formant corresponds to a resonance produced in the vocal tract (Bageshree and Ashish, 2012).

F1: pharynx
F2: oral cavity
F3: nasal cavity

While formants are mainly used to distinguish between vowels (as they are most notable for vowel frequencies), recent research has used formants to note the prosodic features of the speech. Formant F1 and F2 affect the vocal sound being produced the highest. As the frequencies get higher, we can further differentiate between phonetic sounds and different speakers. This makes it ideal to make use of formants F1, F2 and F3 to extract relevant conclusions about the speech.

RAVDESS has a database of actor produced audios; therefore, this research is clearly established on the premise that these audios are equivalent to "natural" vocal emotion expressions. This assumption can be questioned with the case that these actors' performances are not commensurable to real-life expressions of the emotions. (Arndt and Janney, 1991) used linguistic pragmatics to address this problem and made a convincing argument that as emotive communication is a widespread interactional phenomenon, actors producing convincing emotive vocal expressions has reasonable ecological validity and will not differ strongly from instinctive emotional expressions. This is also backed by (Banse and Scherer, 1996).

5. Analysis

To analyse the sound wave, Praat software was used. This software gave the ability to edit the sound wave facilitating deletion of blank speech, extraction of formant values, and drawing visible formant contours. Average formant values (F1, F2 and F3) are reported and trends are discussed.

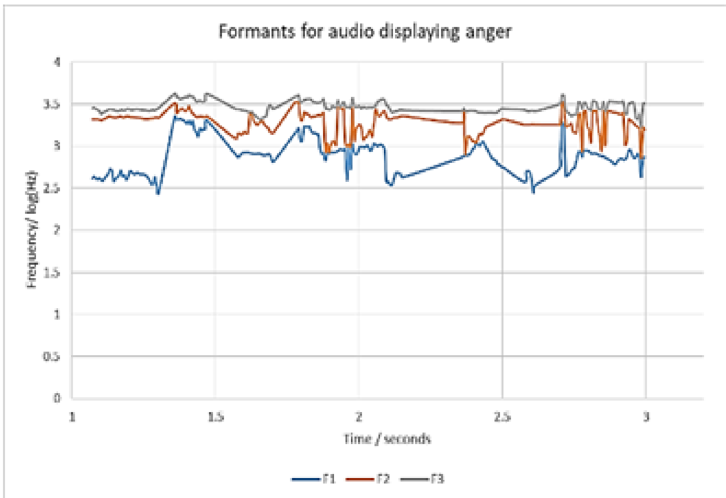

Fig. 1: Extracted values of formants F1, F2 and F3 for the audio displaying anger.

Anger: Like expected, anger shows a high rise in all 3 formant values. It has a relatively higher change in F1 value compared to F2 and F3 which suggest a wider opening of the pharyngeal region. This is consistent with the conclusion of (Lee et al, 2006). It also has the highest F1 and F3 values compared to all the other emotions.

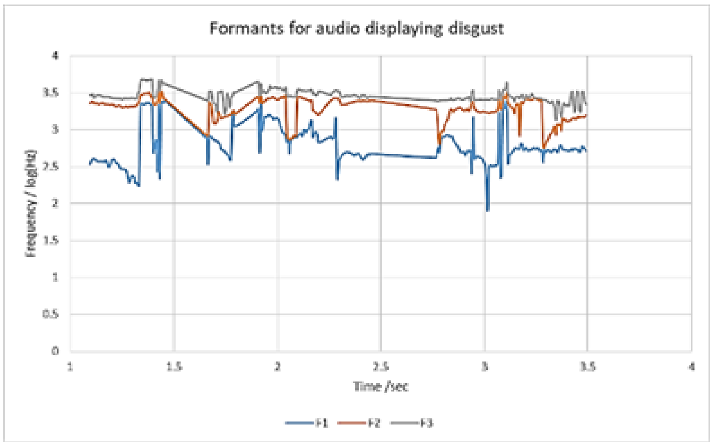

Fig. 2: Extracted values of formant 1, 2 and 3 for the audio displaying disgust.

Disgust: Like anger, disgust also has high values of formants, which is probably caused by the stress in the organs of the vocal tract. Disgust has the highest F2 value adhering to its significance that the tongue is placed toward the front of the oral cavity. (F2 and F3 values are very close to anger)

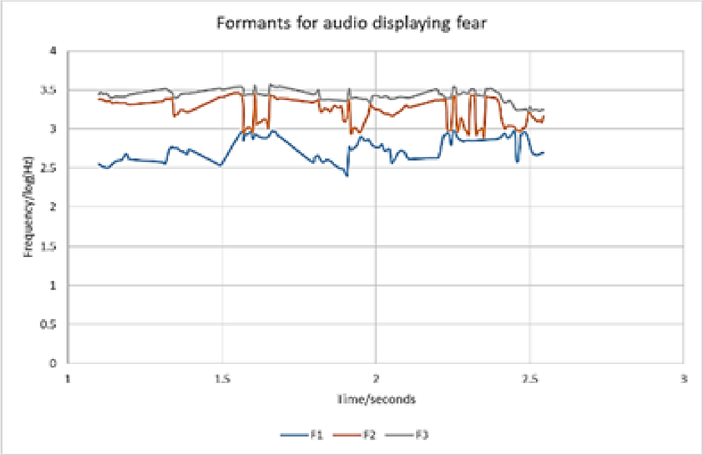

Fig. 3: Extracted values of formant 1, 2 and 3 for the audio displaying fear.

Fear: The formant F2 and F3 frequencies are very close to each other compared to the F1 frequency. Overall formants produced by fear lie somewhat in between of the 6 basic emotions. This might be due to the fact that often fear is expressed with underlying surprise and/or disgust.

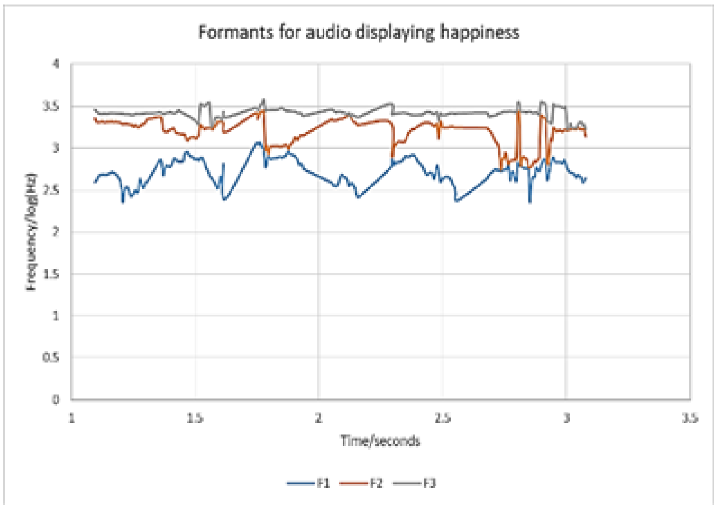

Fig. 4: Extracted values of formant 1, 2 and 3 for the audio displaying happiness.

Happiness: This also converges towards the centre of the graph of the 6 emotions, where the F1 mean averages to 571Hz and F2 mean to 1614 Hz. These values are supported by the conclusion of (Whiteside, 1999) and (Kamiloğlu et al. 2020) where there are consistent results found where the F1 and F2 mean are greater than the neutral parameters.

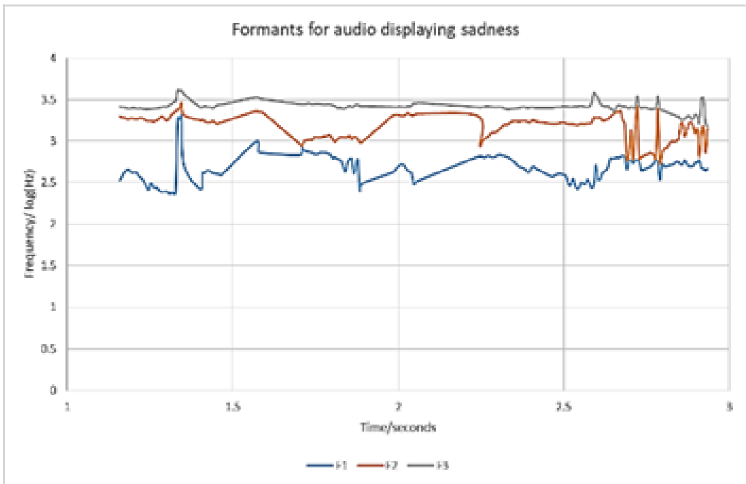

Fig. 5: Extracted values of formant 1, 2 and 3 for the audio displaying sadness.

Sadness: The formant F1 and F2 average values are almost equal to the F1 and F2 mean of neutral; however, there is an increased F3 mean value. This suggests that it uses less intensity but higher vocal energy.

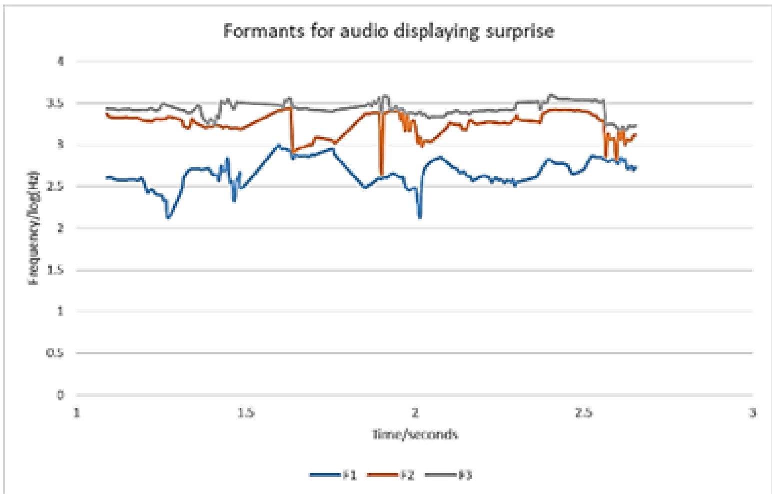

Fig. 6: Extracted values of formant 1, 2 and 3 for the audio displaying surprise.

Surprise: It has an almost equal F1 value to neutral but a significant increase in the F2 and F3 mean values. This could be mainly since surprise is a very brief emotion expressed due to sudden and unexpected movement/thought. Therefore, the most difference in the resonance is caused by the oral and nasal cavity.

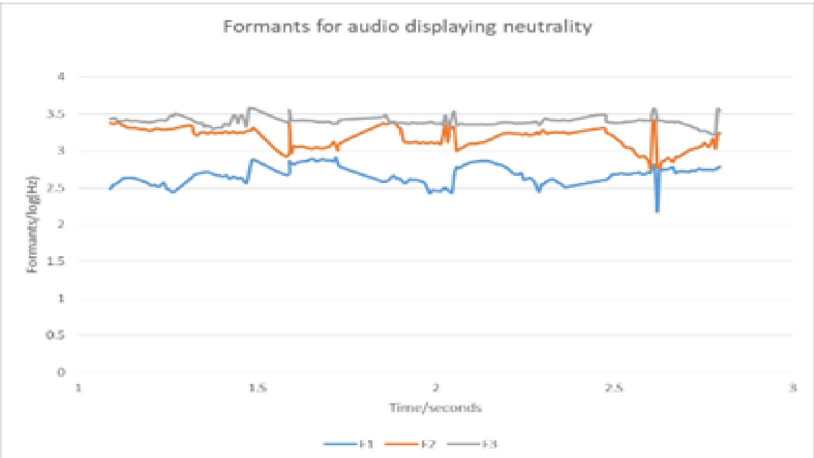

Fig. 7: Extracted values of formant 1, 2 and 3 for the audio displaying neutrality.

The neutral graph shows an almost constant frequency throughout the timeframe of the audio with occasional rises in the formant values suggesting that it can be an accepted base value to compare the rest of our frequencies with.

6. Conclusion

Table 1: Average values of formants F1, F2 and F3 for each of the 6 basic emotions: anger, disgust, fear, happiness, sadness and surprise. Neutrality is also included as a base value for comparison.

Emotions	Average value/ Hz			Average value/ log (Hz)		
	F1	F2	F3	F1	F2	F3
Anger	850.84	1987.45	2936.76	2.880	3.279	3.463
Disgust	730.28	2021.54	2917.69	2.779	3.280	3.456
Fear	601.10	1724.03	2642.67	2.755	3.210	3.416
Happiness	570.74	1613.96	2613.27	2.734	3.182	3.413
Sadness	512.15	1532.43	2591.54	2.679	3.161	3.409
Surprise	494.85	1859.87	2726.19	2.676	3.250	3.427
Neutrality	492.41	1533.42	2485.52	2.675	3.165	3.392

This research had started with the thought that unpleasant emotions like anger and disgust will have high formants while positive emotions like happiness will have a comparatively lower formant average and sadness will have its value close to the neutral audio. The above table showcases similar results; therefore, the author believes that these values are accurate and can be applied to the neutral sounds to achieve the required emotion. Moreover, based on past research also, these formant values match the assumptions made by several researchers, hence proving its validity.

7. Future Scope

The author is a high school student and so has access to limited resources which restricted the scope of the research. Based on the conclusions drawn, there could be further development done where these formant values F1, F2 and F3 are inputted into the source-filter model as peak frequencies and tested whether the audio produced reflects the desired emotion or not. Moreover, this research focuses on only 6 emotions, so the range of emotions can be widened, the intensities as well can be changed. For example, the anger used in this research is hot anger and there can be further research into cold anger and different strengths of anger as well. Due to the time constraint, this research only analyses the audio produced by a single male. We can investigate more male audios and test whether the average formant frequencies remain the same or how the formant frequencies change based on the vocal model of a person. Furthermore, the female and kids' audios can also be taken into consideration as the speech frequencies of females and kids are again different than males. Additional development can be made where there is an addition of nonverbal vocalisations into the audio based on the emotion required. For instance, filler words are most associated with nervousness and sighs signify sadness or distress. Incorporation of these vocalisations into the emotion-based audio can produce human-like sounds and accomplish the much-needed development in the field of voice synthesis.

8 Acknowledgments

This paper and the research behind it would not have been possible without the exceptional support of my supervisor, Ms Aqsa Temrikar. Her enthusiasm, knowledge and exacting attention to detail have been an inspiration and kept my work on track from my first encounter with the topic to the final draft of this paper.

I would also like to extend my gratitude to all the teachers and co-ordinators at JBCN International School, Borivali for providing extensive personal and academic guidance.

Finally, I wish to thank my family for their support and encouragement throughout the process.

References

Journals/Articles/Conference Papers

1. Anscombe E. and Geach P. (1970). Descartes Philosophical Writing, Nelson: The Open University
2. Arndt, H. and Janney, R.W., 1991. Verbal, prosodic, and kinesic emotive contrasts in speech. Journal of pragmatics, 15(6), pp.521-549.
3. Bageshree V. Sathe-Pathak, Ashish R. Panat, ―Extraction of Pitch and Formants and its Analysis to identify 3 different emotional states of a person‖, International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012
4. Banse, R. and Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. Journal of personality and social psychology, 70(3), p.614.
5. Burkhardt, F., & Campbell, N. (2015). Emotional Speech Synthesis. In: The Oxford Handbook of Affective Computing. Oxford University Press New York; 2014. p. 286.
6. Burkhardt, F., 2009. Rule-based voice quality variation with formant synthesis. In Tenth Annual Conference of the International Speech Communication Association.
7. Carlson, R., Granström, B. and Nord, L., 1992. Experiments with emotive speech-acted utterances and synthesized replicas. In the Second International Conference on Spoken Language Processing.
8. Descartes, René; Anscombe, G. E. M. & Geach, P. T. (1954). Philosophical Writings Nelson.
9. Ekman, P. (1992). Are there basic emotions? Psychological Review, 99(3), 550–553
10. Ekman, P., 1992. An argument for basic emotions. Cognition & emotion, 6(3-4), pp.169-200.
11. Ekman, P., 1999. Basic emotions. Handbook of cognition and emotion, 98(45-60), p.16.
12. Ekman, P and Davidson, R.J., 1994. The nature of emotion: Fundamental questions. Oxford University Press.
13. Goudbeek, M., Goldman, J.P. and Scherer, K.R., 2009. Emotion dimensions and formant position. In the Tenth Annual Conference of the International Speech Communication Association.
14. Homma, T., Sun, Q., Fujioka, T., Takawaki, R., Ankyu, E., Nagamatsu, K., Sugawara, D. and Harada, E.T., 2021. Emotional Speech Synthesis for Companion Robot to Imitate Professional Caregiver Speech. ArXiv preprint arXiv: 2109.12787.
15. R. Murray and J. L. Arnott, (1996) "Synthesizing emotions in speech: is it time to get excited?" Proceedings of Fourth International Conference on Spoken Language Processing. ICSLP '96, 1996, pp. 1816-1819 vol.3

16. Kamiloğlu, R. G., Fischer, A. H., & Sauter, D. A. (2020). Good vibrations: A review of vocal expressions of positive emotions. Psychonomic Bulletin & Review, 27(2), 237–265.
17. Khanna, P. and Sasikumar, M., 2011. Recognizing emotions from human speech. In Thinkquest~ 2010 (pp. 219-223). Springer, New Delhi.
18. KR. Thakuria, L. and Hari Talukdar, P., 2022. Analysis and Detection of Emotional states of a person by extracting Pitch and Formants with respect to Assamese Language. International Journal of Scientific & Engineering Research, Volume 5(Issue 2).
19. Livingstone, S.R. and Russo, F.A., 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), p.e0196391.
20. Lemmetty, S (1999), "Review of Speech Synthesis Technology", Master's Thesis, Espoo: Helsinki University of Technology
21. Montero, J.M., Gutierrez-Arriola, J.M., Palazuelos, S.E., Enriquez, E., Aguilera, S. and Pardo, J.M., 1998, November. Emotional speech synthesis: from speech database to TTS. In ICSLP (Vol. 98, pp. 923-926).
22. Ortony, Andrew & Turner, Terence. (1990). What's Basic About Basic Emotions? Psychological review. 97. 315-31.
23. Plutchik, R., 1984. Emotions: A general psychoevolutionary theory. Approaches to emotion, 1984(197-219), pp.2-4.
24. R. Cowie and R. R. Cornelius, “Describing the emotional states that are expressed in speech,” Speech communication, vol. 40, no. 1-2, pp. 5–32, 2003
25. Sathe-Pathak, B.V. and Panat, A.R., 2012. Extraction of pitch and formants and its analysis to identify 3 different emotional states of a person. International Journal of Computer Science Issues (IJCSI), 9(4), p.296.
26. Sezgin, M.C., Gunsel, B. and Kurt, G.K., 2012. Perceptual audio features for emotion detection. EURASIP Journal on Audio, Speech, and Music Processing, 2012(1), pp.1-21.
27. Somoza M, Brave S, Foehr U, Nass C (2001) "The Effects of Emotion of Voice in Synthesized and Recorded Speech" AAAI Technical Report FS-01-02
28. Stein, N.L. and Oatley, K., 1992. Basic emotions: Theory and measurement. Cognition & Emotion, 6(3-4), pp.161-168.
29. Suthar, N., 2020. Role of Second Formant Frequency (F2) in Forensic Speaker Identification. Formal Approaches to South Asian Languages, 3.
30. Whiteside, S.P., 1999. Acoustic characteristics of vocal emotions simulated by actors. Perceptual and motor skills, 89(3_suppl), pp.1195-1208.
31. Yoshimura, Takayoshi, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi and Tadashi Kitamura. (1999) “Simultaneous modelling of spectrum, pitch and duration in HMM-based speech synthesis.” 6th European Conference on Speech Communication and Technology (Eurospeech 1999)

Websites

1. Baram, T., 2022. Classifying emotions using audio recordings and Python. [online] Medium. Available at: <https://towardsdatascience.com/classifying-emotions-using-audio-recordings-and-python-434e748a95eb> [Accessed 14 September 2022].
2. Boersma, P. and Weenink, D., 2021. [online] Praat.org. Available at: <http://www.praat.org> [Accessed 13 June 2022].
3. Dupuis, K. and Pichora-Fuller, M., 2022. Toronto emotional speech set (TESS) | TSpace Repository. [online] Tspace.library.utoronto.ca. Available at: <https://tspace.library.utoronto.ca/handle/1807/24487> [Accessed 14 September 2022].
4. Law of Attraction. 2022. A Thorough Break Down of the Emotion Chart by Frequency. [online] Available at: <https://lawofattractionblueprint.com/a-thorough-break-down-of-the-emotion-chart-by-frequency/> [Accessed 14 September 2022].
5. Miller, E., 2022. Speech Synthesis Software for Anime Announced. [online] Anime News Network. Available at: <https://www.animenewsnetwork.com/news/2007-05-02/speech-synthesis-software> [Accessed 14 September 2022].
6. Monster, B., 2022. What Is Speech Synthesis or Telephony Text to Speech? [online] Plum Voice. Available at: <https://www.plumvoice.com/resources/blog/speech-synthesis-text-speech/> [Accessed 14 September 2022].
7. O’Connor, K., 2022. Vowels, Vowel Formants and Vowel Modification — SingWise. [online] SingWise. Available at: <https://www.singwise.com/articles/vowels-formants-modifications> [Accessed 14 September 2022].
8. Stanley, J. and Lipani, L., 2022. Automatic Formant Extraction in Praat. [online] Joeystanley.com. Available at: <https://joeystanley.com/downloads/191002-formant_extraction#3_extracting_the_names_of_praat_objects> [Accessed 14 September 2022].
9. Thomas Martin, S., 2018. [online] Formants Explained and Demonstrated. Available at: <https://youtu.be/jpbFnsusfp0> [Accessed 21 April 2022].

Books

1. Ekman, P and Friesen, W. V (1975) Unmasking the face: A guide to recognizing emotions from facial clues. Los Altos: Maylor Books. reprinted 2003
2. Paget, R (1999) Human Speech, Taylor and Francis reprinted 2013
3. Ekman, P., Friesen, W.V., and Ellsworth, P. (1971) Emotion in the Human Face: Guidelines for Research and an Integration of Findings. Pergamon Press. 1971