Impact of Speech Emotions on Vowel Space

Daniel Cho, Catherine Watson, Jesin James

Department of Electrical, Computer, and Software Engineering,
The University of Auckland, New Zealand
dcho981@aucklanduni.ac.nz, c.watson@auckland.ac.nz, jesin.james@auckland.ac.nz

Abstract

This study explores how speech emotions affect vowel spaces, focusing on valence and arousal dimensions across two speech corpora: German Emotional Speech Corpus (EMO-DB) and New Zealand English Emotional Speech Corpus (JLCorpus). Formant extraction on the primary and secondary emotions from JLCorpus gives the F1 and F2 frequencies for vowels. Results show high arousal emotions correlate with higher mean F1 values, while low arousal emotions correspond to lower mean F1 values. Positive valence emotions tend to have higher mean F2 values. Statistical analyses confirm significant emotional influences on vowel formants, supporting existing research and extending insights into secondary emotions.

Index Terms: Formants, emotional speech, vowel space, primary emotions, secondary emotions

1. Introduction

People can recognise emotional states through facial expressions, body movements, gestures and speech [1]. The focus of this study is emotion expression via speech. A listener's ability to recognise affective states just by listening to vocal cues is based on the fact that human speech contains prosody, spectral and other suprasegmental features which convey emotional information [2]. The importance of pitch and loudness as acoustic features that differentiate emotions is well established [3, 4, 5]. Here, we contribute to the studies which look at the impact that emotions have on articulation by focusing on the resonances of the vocal tract, known as formant frequencies.

As the resonances of the vocal tract are not fixed, any change in the position of the tongue and the jaw will change the resonance of the vocal tract. Specifically, the first and second formants can be related to the jaw opening and tongue movement during articulation, respectively. As the tongue moves forward and backward, it changes the length and shape of the vocal tract. Generally, higher formant frequencies (e.g., F2 and F3) correspond to the position of the tongue body (front to back). A more open jaw creates a longer vocal tract, which tends to lower all formant frequencies (F1, F2, F3, etc.). This is because longer vocal tracts resonate at lower frequencies. Hence, visualising the vowels in a 2D space where first and second formants represent the y and x axes respectively (called the vowel space) is used as an approach to relate the formant values to the jaw opening and tongue movement during articulation. This makes the vowel space, and the formants a powerful tool to study factors that impact articulation. Among the other articulation strategies, lip rounding tends to lower F1 frequency of vowels. This is because when the lips are rounded, they create a constriction in the vocal tract that effectively increases its length acoustically. A longer vocal tract resonates at lower frequencies, so F1 decreases when the lips are rounded [6]. Formants, influenced by speech production physiology, can contribute to analyses of

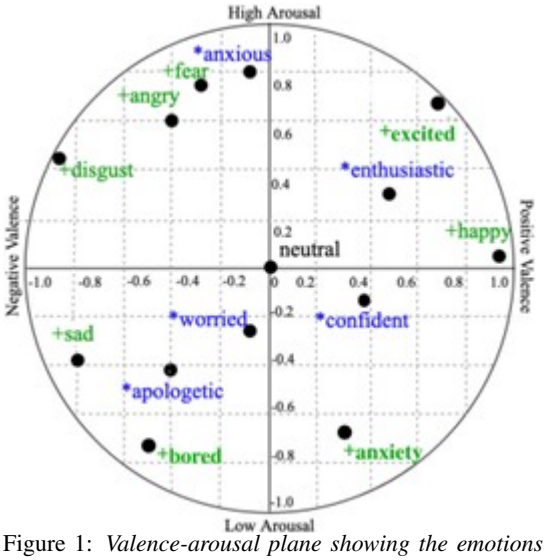

Figure 1: *Valence-arousal plane showing the emotions in the EMO-DB and JLCorpus, primary emotions are labelled with '+' and secondary emotions are labelled with '*' [12]*

mental health [7, 8, 9], speech disorders [10], and speaker characteristics [11], but cannot be viewed as standalone indicators.

Emotions also impact articulation, and hence analysing the formant frequencies can provide insights into how the vocal tract is modified during emotion production. The authors in [1] were able to extract more valuable emotional content by segmenting vowels from utterances and analysing the formant frequencies of the vowels. Other studies [13, 14, 15] also report an improvement in emotion classification accuracy when formants are used as features. While these studies provide valuable insights into the impact of emotions on formants, detailed analysis on the effect of emotions on each of the vowels can provide insights into articulatory changes during emotion production. Hence, the research questions answered in this paper are:

1. How do formant frequencies of vowels change during the expression of emotions?
2. How do the articulatory modifications of the vocal tract necessary for emotive speech vary based on the levels of arousal and valence associated with emotions?

In this paper, the first two formants (F1 and F2) will be focused on as these two formants can be used to differentiate vowels, thereby teasing out the impact that emotions have on vowel production. Using a dimensional emotion model - Russel's circumplex model of emotions [16], a visual representation of the emotions on a valence arousal (V-A) two-dimensional plot is shown in Figure 1 [17]. The valence dimension indicates pleasantness (E.g., happy having positive valence, sad having negative valence) and the arousal dimension indicates the level

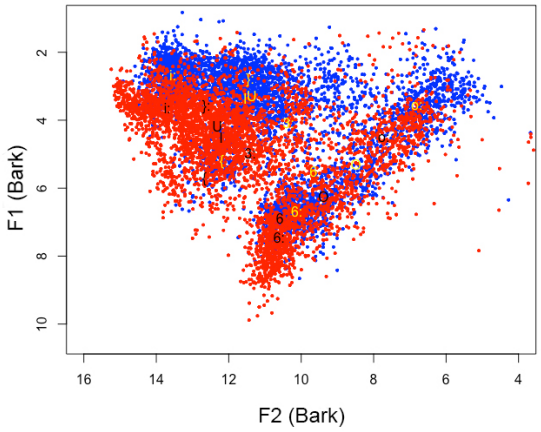

Figure 2: *Untransformed male (blue) and female (red) vowels from JLCorpus and the centroids, male (yellow), female (black)*

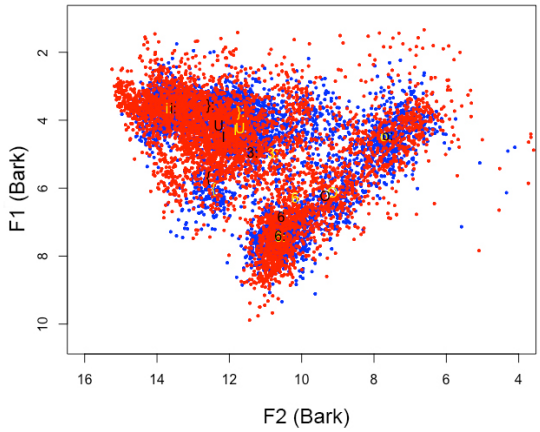

Figure 3: *Transformed male (blue) and female (red) vowels from JLCorpus and the centroids, male (yellow), female (black)*

of reaction to a stimulus (E.g., excited having high arousal, sad having low arousal. In the V-A dimension as seen in Figure 1, an emotion such as excited has positive valence, high arousal, while an emotion such as angry has negative valence, high arousal. This V-A two-dimensional representation will be used to interpret the impact that the valence and arousal levels of the emotions have on the formant frequencies, and thereby the articulation.

2. Methodology

2.1. Emotional Speech Corpora

Two speech databases are used to investigate the effects of speech emotion on formats - German Berlin Emotional Speech Database EMO-DB [18] and New Zealand English JLCorpus [12]. EMO-DB consists of seven emotions (anger, boredom, anxiety, happiness, sadness, disgust and neutral) conveyed by ten professional actors (five male and five female) through ten German utterances. In the JLCorpus, 10 emotions are divided into two sections: primary and secondary. The primary emotions are angry, excited, neutral, happy, and sad. The secondary emotions are anxious, apologetic, confident, enthusiastic and worried. EMO-DB and JLCorpus have a variety of high and low arousal emotions as well as positive and negative valence emotions which are labelled in Figure 1. To understand the placement of emotions along the arousal dimension, arrange them vertically from anxious (high arousal) at the top to bored (low arousal) at the bottom. For the valence dimension, arrange emotions horizontally from happy (positive valence) on the right to sad (negative valence) on the left. Although secondary emotions are nuanced in contrast to primary emotions [12], the significance of secondary emotions in social interactions is undeniably substantial and therefore must be studied [19, 17]. The emotions within the JLCorpus are well spread amongst the valence-arousal plan allowing for an extended scope compared to EMO-DB. EMO-DB has 12 German monophthongs distributed within a total of 535 utterances used for formant analysis in this paper. Whereas the JLCorpus has 11 New Zealand English monophthongs distributed within a total of 2400 utterances that are used for formant analysis. Sampling frequency rates are 16 kHz for EMO-DB and 44.1 kHz for JLCorpus.

2.2. Segmentation and Formant Extraction

EMO-DB speech recordings are labelled at the word and phonetic levels using forced alignment, enabled by the Munich Automatic Web Segmentation System, webMAUS [20] using the German language setting. The labelled database was then converted to an EMU-formatted database [21] for exact phonetic querying and segmentation. After segmentation, first two formants F1 and F2 are extracted for the vowels /aː/, /eː/, /ɛː/, /iː/, /ɪ/, /oː/, /uː/, /yː/ and /ʏ/ using the signal processing function forest within the R package wrassp [22]. The default forest parameters are used for the male speakers and the nominal F1 was set to 600 Hz for the female speakers. The formant frequencies for each vowel instance are time normalized and extracted from temporal midpoints.

JLCorpus recordings were segmented, and formants were extracted using the same method as EMO-DB, although the New Zealand English setting was used for webMAUS. Formant extraction was conducted for the vowels /iː/, /ɪ/, /ɛ/, /æ/, /ɒː/, /ɔ/, /oː/, /uː/, /ɜː/. There was no hand correction of phonetic boundaries and formant tracks for EMO-DB and JLCorpus data.

2.3. Data Transformation

Due to differences in vocal tract length between the speakers identifying as male and female, it can be challenging to empirically compare the vowel spaces of the speakers [23]. Therefore, for statistical analysis, a linear transformation is performed to make formant frequency observations comparable. This transformation allows for a conversion of male speaker values closer to female speaker values or vice versa; the choice is made by the researchers. The transformation method was first proposed in [24] and has also been used in other studies e.g. [25]. Both EMO-DB and JLCorpus have equal number of female and male speakers in their databases. A male-to-female transformation was conducted as reported in [25].

Figure 2 shows the male (blue) and female (red) vowel spaces respectively for all the vowel data in the JLCorpus before the transformation was performed. Figure 3 shows the transformed male vowel space and the female vowel space. The anchoring vowels for the JLCorpus were /iː/ /aː/ and /oː/. A similar transformation process was conducted on the male data of the EMO-DB with the anchoring vowels for this corpora being /iː/, /aː/ and /uː/. All formant plots in this paper show the female data and the transformed male data on the same plots.

3. Results

The F1 and F2 values are analysed separately for each vowel and presented in a centroid vowel space in bark frequency scale.

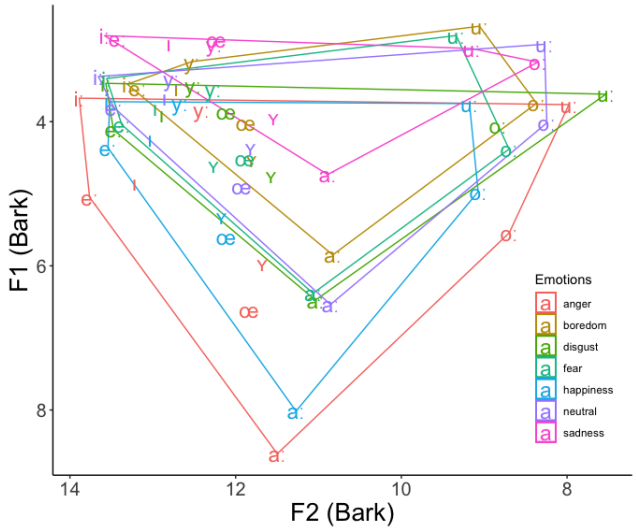

Figure 4: *Vowel space of all German speakers after transformation in different emotions from EMO-DB*

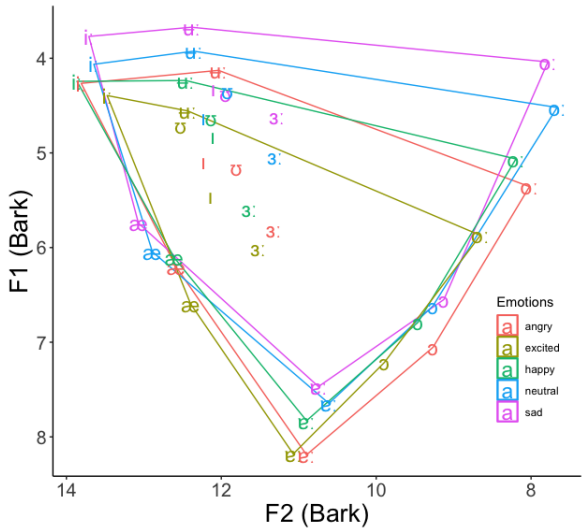

Figure 5: *Vowel space of the primary emotions after transformation of all New Zealand English speakers from JLCorpus*

3.1. Analysis of Emotions in EMO-DB Database

3.1.1. Vowel Space Analysis

Figure 4 depicts the German vowel space across different emotions. We will compare the vowel space with the relative valence-arousal values of the emotions shown in Figure 1. It is evident that the F1 axis effectively distinguishes between high and low arousal emotions, ranging from anger (high arousal) to sadness (low arousal). Emotions characterized by high arousal such as anger, disgust, fear, and happiness exhibit higher F1 values, resulting in an expanded vowel space. Conversely, emotions with low arousal like boredom and sadness display lower mean F1 values, leading to a more condensed vowel space. These findings align with previous studies on formant analysis of EMO-DB [14], which also noted a strong correlation between mean F1 values and speakers' emotional arousal.

3.1.2. Statistical Analysis

Statistical analysis of individual vowels revealed significant effects when comparing models that included both formant type and emotions against a null model with fixed effects of formant type alone. This comparison aimed to ascertain whether there were statistically significant differences in vowels across emotions compared to neutral. Notably, vowels /aː/, /eː/, /iː/, /ɪ/, /oː/, /ʊ/, /yː/, /ʏ/ showed significant emotion-related impacts, evidenced by p-values below 0.05. Subsequently, pairwise comparisons using t-tests (Table 1) were conducted to pinpoint where these significant differences occurred among the fixed effects. In this analysis, neutral served as the reference.

Specifically, emotions such as anger, sadness, and happiness exhibited statistically significant differences in both F1 and F2 formant frequencies compared to neutral. This underscores the influence of these emotions have on vowel production across the dataset.

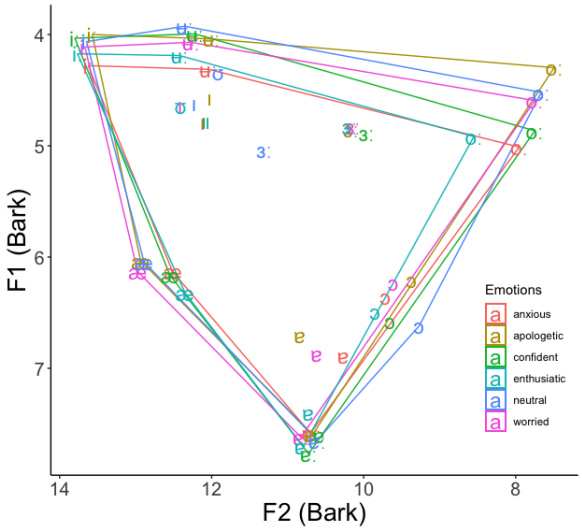

Figure 6: *Vowel space of the secondary emotions after transformation of all New Zealand English speakers from JLCorpus*

Table 1: *T Test of significant vowels impacted by emotion relative to neutral from EMO-DB*

Emotion	Vowel	F1/F2	Value	t-value	p-value
Anger	/iː/	F1	0.3	2.0	4e-2
Anger	/oː/	F1	1.5	7.5	3e-10
Anger	/oː/	F2	-1.1	-2.8	5e-3
Anger	/aː/	F1	2.01	4.5	3e-5
Anger	/aː/	F2	-1.4	-3.2	1e-3
Sadness	/iː/	F1	-0.6	-4.9	6e-6
Sadness	/iː/	F2	0.6	2.7	6e-3
Sadness	/oː/	F1	-0.8	-4.1	1e-4
Sadness	/oː/	F2	0.9	2.4	1e-2
Sadness	/aː/	F1	-1.9	-4.1	1e-4
Sadness	/aː/	F2	1.9	4.5	7e-6
Happiness	/iː/	F1	0.3	2.3	2e-2
Happiness	/iː/	F2	-0.5	-2.2	2e-2
Happiness	/oː/	F1	0.9	4.1	1e-4
Happiness	/aː/	F1	1.5	3.2	2e-3

3.2. Vowel Space of Emotions in JLCorpus Database

Figure 5 shows New Zealand English vowel space of the primary emotions from the JLCorpus. The relative location of each vowel within that space agrees with the general F1-F2 distribution of the New Zealand English vowels [26]. A substantial contrast between the vowel spaces of the primary emotions is seen which can be attributed to the different valence and arousal levels of each emotion. Higher aroused emotions such as angry, excited and happy have a similar vowel spread with a correspond-

ing higher mean F1. At the same time, lower arousal emotions such as sad and neutral can be seen with a vowel space with an overall lower mean F1. Positive valence emotions happy and excited have a moderately higher mean F2, which can also be seen from the vowel centroid (can be quite clearly seen for the /oː/ vowel) shifted to the left more than other emotions.

The resulting New Zealand English vowel space of the secondary emotions in Figure 6 shows a similar trend regarding mean F1-F2 values relating to valence and arousal of emotions. The vowel space of the secondary emotions is not as spread out, therefore, it is more difficult to distinguish clear boundaries between each emotion's vowel space. Regardless, higher arousal emotions - notably anxious and enthusiastic have a higher mean F1 overall whereas, low arousal emotions namely apologetic and worried have an apparent lower mean F1. It is also evident that the positive valence emotion enthusiastic has a higher mean F2 than other negative valence emotions. Formant positioning is also significantly affected by the valence dimension as well as vowel position.

3.2.1. Statistical Analysis

Statistical analysis of the JLCorpus was performed in the same method as EMO-DB, where a fixed effects formant type model was compared with a model based on formant and emotions. Results indicate significant emotional impact on vowels /ɒː/, /iː/, /oː/ and /uː/ with p-values less than 0.05. Afterwards, t-tests are conducted for these vowels with neutral used as the emotion reference level. The results showed that the F1 values for angry, excited, happy, sad, anxious, and enthusiastic emotions are each (statistically) significantly different from those for neutral. Significant differences in F2 were found for angry, excited, sad, and anxious compared to neutral.

3.3. Discussion

[14] performed formant analysis on EMO-DB and identified a direct correlation between the F1/F2 positions at the vowel level and the level of the speaker's emotional arousal. The study was a gender-dependent evaluation of the complete list of German phonemes. In contrast to Vlasenko et al's research [14], the formant analysis in our study took a phonetic approach, emphasizing the acoustic characteristics of vowel production. Our approach involved visualizing vowel space with F1 on the y-axis and F2 on the x-axis to highlight the phonetic relationship between vowel articulation and acoustic properties. Also, our analysis was solely emotion focused, therefore, a data transformation of the formants was conducted due to the impact of the gender differences in vocal tract length on the formants.

The phonetic approach allows linking the vowel space to possible articulation strategies. For example, for the the vowel /ɒː/ from the JLCorpus, there is a clear difference between sad and excited in F1, suggesting that the jaw is more open for excited than sad. In contrast, there is hardly any difference in F1 of vowel /ɒː/ for the secondary emotions, suggesting jaw opening is not contrasting those emotions.

The vowel spaces pf EmoDB and JLCorpus datasets were obtained to be ordered based on the arousal levels of emotions, as shown in Figures 4, 5, and 6. While some effects of valence levels were observed, they did not reach statistical significance.

High arousal emotion angry can be seen with a larger vowel space and higher centre of gravity (centre of vowel space) in comparison to low arousal emotion sadness when the red centroid is compared to the magenta centroid in Figure 4. The overall vowel space and centre of gravity of low arousal emotions can be seen lifted (lower F1) compared to high arousal emotions and neutral.

In the German EmoDB dataset, all corner vowels (/i/, /a/, and /u/) demonstrate sensitivity to arousal levels. In New Zealand English, it was found that the back vowel /o/ is particularly effective in distinguishing between emotions, as depicted in Figures 5 and 6. Although differences were noted for other corner vowels like /i/ and /ɒ/, these distinctions were less pronounced compared to /o/. This suggests that the tongue position required to produce the back vowel is a crucial articulatory strategy for conveying varied emotional states. Additionally, the presence of lip rounding may serve as an articulatory strategy for low-arousal emotions (such as sadness), given that lip rounding typically lowers the F1 value [27]. It is important to note that the first formant of the /o/ vowel, due to its proximity to the fundamental frequency, often faces challenges in accurate formant estimation. Therefore, significant deviations in the /o/ vowel cannot be conclusively verified without manual verification of formant estimates.

The vowels /iː/, /eː/, /aː/, /oː/, /ʏ/, /yː/ and /ɪ/ from EMO-DB were found to be better at conveying emotional differences than others. Five of the seven vowels listed are long vowels, therefore, vowel length plays an important role in distinguishing between the different emotions. The NZE vowels /iː/ , /ɪ/ , /ɒː/ , /oː/ and /uː/ from the JLCorpus conveyed emotional differences better than other vowels. The vowels listed were all long vowels with the exception of /ɪ/, further verifying the findings with EmoDB that vowel length is an articulation strategy to differentiate emotions.

The addition of statistical analysis was necessary to make the results more robust. Due to formant analysis on emotions not being vastly popular, secondary emotion's impact on the vowel space has rarely been studied. Although secondary emotions are more subtle and nuanced than secondary emotions [17], the JLCorpus emotional speech database provided and introduced secondary emotions which were analysed in this paper. Results showed that the vowel space of subtle secondary emotions such as anxious and enthusiastic still displayed significant differences compared to the neutral vowel space.

4. Conclusion

Many acoustic features have been used for emotion analysis. Among the most important of these are features that describe the pattern of resonances within the vocal tract, also known as formants. Formant position is a spectral property of the speech signal that reflects voice quality as well as linguistic vowel identity. Most past research carried out in the field has tended to focus on a smaller set of rather strong emotions, such as anger, joy, sadness, and fear. To bridge this research gap, secondary emotions from the JLCorpus were used in this study. This study explored the the first and second formants in emotional speech and linked it with articulation, thereby answering the first research question on how formants are impacted during emotion production. We also related the valence and arousal values of the emotions to the formant values via vowel spaces, thereby answering the second research question linking the valence-arousal dimension to emotion expression.

In the future, research relating higher and lower mean F1 values to the vocal tract resonance such as past research generating vocal tract shapes from formant frequencies [28] would allow insight into how different emotions impact the jaw and tongue movements. While formant analysis can aid in classifying emotions, it should be combined with other methods due to the lack of a direct relationship between formants and emotions.

5. References

[1] M. Goudbeek, J.-P. Goldman, and K. R. Scherer, "Emotion dimensions and formant position." in *Interspeech*, vol. 2009, 2009, pp. 1575–1578.

[2] M. Meyer, M. Keller, and N. Giroud, "Suprasegmental speech prosody and the human brain," *The Oxford handbook of voice perception*, pp. 143–165, 2018.

[3] J. James, L. Tian, and C. Watson, "An open source emotional speech corpus for human robot interaction applications," *Interspeech 2018*, 2018.

[4] H. Nordström, "Emotional communication in the human voice," Ph.D. dissertation, Department of Psychology, Stockholm University, 2019.

[5] P. Laukka, H. A. Elfenbein, N. S. Thingujam, T. Rockstuhl, F. K. Iraki, W. Chui, and J. Althoff, "The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features." *Journal of personality and social psychology*, vol. 111, no. 5, p. 686, 2016.

[6] R. D. Kent, J. Dembowski, and N. J. Lass, "The acoustic characteristics of american english," in *Principles of Experimental Phonetics*, N. J. Lass, Ed. St. Louis, Missouri: Mosby, 1996, pp. 185 – 225.

[7] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Comparing objective feature statistics of speech for classifying clinical depression," in *The 26th annual international conference of the IEEE engineering in medicine and biology society*, vol. 1. IEEE, 2004, pp. 17–20.

[8] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[9] S. Scherer, L.-P. Morency, J. Gratch, and J. Pestian, "Reduced vowel space is a robust indicator of psychological distress: A cross-corpus analysis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4789–4793.

[10] W. R. Rodríguez and E. Lleida, "Formant estimation in children's speech and its application for a spanish speech therapy tool." in *SLaTE*, 2009, pp. 81–84.

[11] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[12] J. James, L. Tian, and C. Watson, "An open source emotional speech corpus for human robot interaction applications," *Interspeech 2018*, 2018.

[13] J. C. Kim and M. A. Clements, "Formant-based feature extraction for emotion classification from speech," in *2015 38th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2015, pp. 477–481.

[14] B. Vlasenko, D. Philippou-Hübner, D. Prylipko, R. Böck, I. Siegert, and A. Wendemuth, "Vowels formants analysis allows straightforward detection of high arousal emotions," in *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 2011, pp. 1–6.

[15] Z.-T. Liu, A. Rehman, M. Wu, W.-H. Cao, and M. Hao, "Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence," *Information Sciences*, vol. 563, pp. 309–325, 2021.

[16] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[17] J. James, B. Balamurali, C. I. Watson, and B. MacDonald, "Empathetic speech synthesis and testing for healthcare robots," *International Journal of Social Robotics*, vol. 13, no. 8, pp. 2119–2137, 2021.

[18] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[19] J. James, C. I. Watson, and B. MacDonald, "Artificial empathy in social robots: An analysis of emotions in speech," in *2018 27th IEEE International symposium on robot and human interactive communication (RO-MAN)*. IEEE, 2018, pp. 632–637.

[20] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[21] S. Cassidy and J. Harrington, "Multi-level annotation in the emu speech database management system," *Speech communication*, vol. 33, no. 1-2, pp. 61–77, 2001.

[22] R. Winkelmann, J. Harrington, and K. Jänsch, "Emu-sdms: Advanced speech database management and analysis in r," *Computer Speech & Language*, vol. 45, pp. 392–410, 2017.

[23] J. Harrington, "Acoustic phonetics," *The handbook of phonetic sciences*, pp. 81–129, 2010.

[24] C. Watson, B. Ross, E. Ballard, H. Charters, R. Arnold, and M. Meyerhoff, "Preliminary investigation into sound change in auckland," in *Proceedings of the 17th Australasian International Conference on Speech Science and Technology*, 2018, pp. 17–20.

[25] B. Ross, "An acoustic analysis of new zealand english vowels in auckland," Ph.D. dissertation, Open Access Te Herenga Waka-Victoria University of Wellington, 2018.

[26] C. I. Watson and A. Marchi, "Resources created for building new zealand english voices," in *Proc. 15th Australas. Int. Conf. Speech Science and Technology*, 2014, pp. 92–95.

[27] N. Lass, *Principles of Experimental Phonetics*. Mosby, 1996. [Online]. Available: https://books.google.co.nz/books?id=jUliAAAAMAAJ

[28] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, "Generating vocal tract shapes from formant frequencies," *The Journal of the Acoustical Society of America*, vol. 64, no. 4, pp. 1027–1035, 1978.