

doi: [10.5821/conference-9788419184849.52](https://doi.org/10.5821/conference-9788419184849.52)

MODELLING EMOTIONAL VALENCE AND AROUSAL OF NON-LINGUISTIC UTTERANCES FOR SOUND DESIGN SUPPORT

Ahmed KHOTA^{1a}, Eric COOPER^b, Yu YAN^c, Mate KOVACS^d

^a *Ritsumeikan University, Japan, gr0341xp@ed.ritsumei.ac.jp*

^b *Ritsumeikan University, Japan, cooper@is.ritsumei.ac.jp*

^c *Ritsumeikan University, Japan, yuyan@fc.ritsumei.ac.jp*

^d *Ritsumeikan University, Japan, kovacsm@fc.ritsumei.ac.jp*

ABSTRACT

Non-Linguistic Utterances (NLUs), produced for popular media, computers, robots, and public spaces, can quickly and wordlessly convey emotional characteristics of a message. They have been studied in terms of their ability to convey affect in robot communication. The objective of this research is to develop a model that correctly infers the emotional Valence and Arousal of an NLU. On a Likert scale, 17 subjects evaluated the relative Valence and Arousal of 560 sounds collected from popular movies, TV shows, and video games, including NLUs and other character utterances. Three audio feature sets were used to extract features including spectral energy, spectral spread, zero-crossing rate (ZCR), Mel Frequency Cepstral Coefficients (MFCCs), and audio chroma, as well as pitch, jitter, formant, shimmer, loudness, and Harmonics-to-Noise Ratio, among others. After feature reduction by Factor Analysis, the best-performing models inferred average Valence with a Mean Absolute Error (MAE) of 0.107 and Arousal with MAE of 0.097 on audio samples removed from the training stages. These results suggest the model infers Valence and Arousal of most NLUs to less than the difference between successive rating points on the 7-point Likert scale (0.14). This inference system is applicable to the development of novel NLUs to augment robot-human communication or to the design of sounds for other systems, machines, and settings.

¹ Corresponding author. Ahmed KHOTA, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577 JAPAN, gr0341xp@ed.ritsumei.ac.jp

Keywords: *Non-Linguistic-Utterances, Robots, Machine-Learning*

INTRODUCTION

1.1 Background and introduction to Non-Linguistic Utterances (NLUs)

Non-Linguistic Utterances (NLUs) for communication from machine to human have been popularized in fiction, for example in the Star Wars movies where the robot R2D2 communicates with his human counterparts using squeaks, beeps, and other robotic sounds. This has inspired scientific research to try to understand whether there is any real communicative capability behind these sounds (Bethel & Murphy, 2006). NLUs are also used in daily life, for example in train stations to indicate passing a turnstile or an approaching train. Although these are one-way signals. NLUs are sounds that contain no discernible words, aren't musical, and exclude laughing or onomatopoeia. They are used to convey information, affect, or to communicate. (Yilmazyildiz et al., 2016)

R2D2, Walle, cartoons and movies where robots use squeaks, beeps, and whirrs to communicate are enjoyed by audiences, who might not fully understand an actual message but can go along with the context and generally understand what is being conveyed. The character pose and design, and context of the situation also assist in conveying the intended message.

Since the 1970's, NLUs have been used in psychology, where researchers have explored how tones can be used to communicate affect (Yilmazyildiz et al., 2016). Mavridis (2015) also argues for the importance of affect in robot human communication. Affect is communicated in both the semantic content of an utterance as well as the prosodic content. Also, across different cultures, affect can be understood and communicated differently. Language-independent communicative capabilities would allow robots to communicate with a wider range of people, suitable for settings such as tourist attractions and multi-cultural societies.

1.2 Overview of the current state of related work

NLUs and other SFUs have been successfully interpreted in terms of affect and emotional expression. Previous research has investigated whether NLUs can successfully convey emotion or affect. Reasons for using NLUs include, natural language programming is costly and difficult, not all applications require natural language communication, programming for multiple languages adds additional complexity (R. Read & Belpaeme, 2012, 2016; Yilmazyildiz et al., 2016). NLUs provide two main benefits: they are not linked to any language, and they can communicate a message in a very short time (Luengo et al., 2017).

Komatsu (Komatsu, 2005) explored how to communicate positive, negative, agreement and disagreement using NLUs. He found that sounds with rising frequencies were interpreted to be positive while those with falling frequencies were found to be negative, as in 'earcons' (Blattner et al., 1989) for computers and mobile phones.

Read and Belpaeme (2012) found that children were able to readily interpret NLUs in terms of their affect (happy, sad, angry, scared). However, they were not always consistent amongst each other when considering a given utterance. They (R. Read & Belpaeme, 2016) also found that

adults interpreted NLUs in terms of their affect categorically, and that subtle changes between the NLUs did not result in subtly different interpretations.

The Keepon robot used NLUs to communicate with pre-verbal children with some success. Not using natural language makes designs easier to implement and suitable for the morphology of the robot (Yilmazyildiz et al., 2016). Read and Belpaeme (2014) recommend that NLUs be used alongside other more standard means of communication to augment functionality, as opposed to replacing them.

1.3 Machine Learning and speech emotion classification

Researchers have used machine learning methods to analyze mainly the linguistic parts of speech, but also non-linguistic components, to model the emotional meaning. In general, combinations of techniques for both feature extraction (Mel Frequency Cepstral Coefficients (MFCCs), spectrograms, pitch, intensity) and emotion classification using machine learning methods like Convolutional and LSTM neural networks, as well as Random Forest, have tended to outperform standard or non-machine learning oriented approaches. (Chen et al., 2020; Iliou & Anagnostopoulos, 2009; Issa et al., 2020). Previous work was done to model NLUs for communication in dialogue using their dialogue parts, and prosodic trends for dialogue part factors were established (Khota et al., 2019, 2020).

1.4 Objectives

The objective of the current research is to model the Valence and Arousal of NLUs, to evaluate candidate sounds for public facing social robots or similar agents that make use of such sounds.

This paper describes a novel inference model relating the features of NLUs to their affect.

2 MODEL DEVELOPMENT

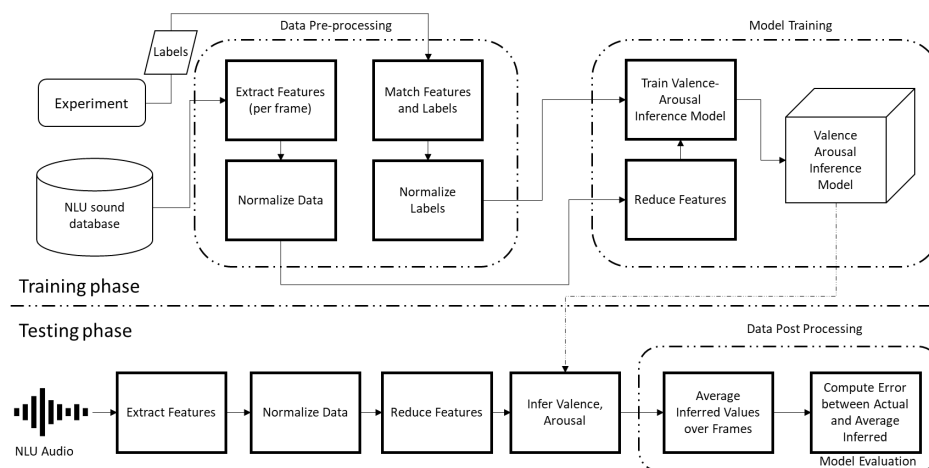


Figure 1: Proposed model

Figure 1 shows the method for developing the model. Firstly, audio files are labelled in experiments according to their Valence and Arousal. Next, important features are extracted from each audio file, and together these data are preprocessed, including being normalized, before

being split into training and test datasets and input into a machine learning model which is trained on the training data set to infer Valence and Arousal on the testing data set. Inferences are collected per frame of audio data (described in section 4) and averaged over the number of frames per audio file to compare predicted and actual Valence and Arousal per audio file.

3 EXPERIMENT

3.1 Sound data sources

Sounds from popular media draw from many different sound designers who may have used various methods to generate and record sounds. This selection is intended to make the dataset agnostic in terms of generation method, potentially removing some influence from specific methods, recording sources, software, tools, or instruments.

The sounds used in experiments were gathered from popular movies, TV shows, and videogames. These sources were intended to explore the sample space and involve a wide range of sounds in terms of their source character, timbre, context, mode, as well as pitch, amplitude, and other prosodic properties. Table 1 shows the movies, TV shows, and videogame sources used, as well as the number of sounds from each, many of which are children's shows, or popular science fiction. A total of 560 NLUs were used. The sounds were extracted from lossy MP4 sources. The bitrate of all audio files is 768kbps except for the sounds synthesized (Khota et al., 2019, 2020) from PureData (*Pure Data — Pd Community Site*, n.d.), which are 705kbps.

Table 1. Sound data sources

Source	No. of Sounds	Character Type	Source	No. of Sounds	Character Type
Aladdin	21	Animal	Tangled	9	Animal
Castle in the Sky	5	Robot	The Emperor's new Groove	15	Animal
Dark Star	7	Robot	The Iron Giant	17	Robot
District 9	8	Alien	Toy Story	6	Other
E.T.	40	Alien	Treasure Planet	11	Other
Gremlins	4	Other	Wallace and Gromit	1	Robot
Guardians of the Galaxy	4	Alien	Walle	31	Robot
How to Train your Dragon	6	Animal	Synthesized Sounds from Pure Data	29	Other
Ice Age	2	Animal	Blues Clues and You: TV Series	10	Other
Monsters Inc	19	Other	Bucket's Quest Star Wars Resistance: TV Series	5	Robot
My Neighbor Totoro	6	Animal	The Muppets Dr Bunsen and Beaker: TV Series	13	Other
Shrek	6	Animal	Curious George: TV Series	21	Animal
Silent Running	8	Robot	In the Night Garden: TV Series	12	Other
Spirited Away 2001	12	Other	Pokémon Journeys: TV Series	13	Other
Star Wars Episode 4: A New Hope	75	Robot	Scooby Doo: TV Series	7	Animal

Star Wars Episode VII: The Force Awakens	37	Robot	Star Wars Rebels: Video Game	66	Robot
Surfs Up	3	Animal	Star Wars Jedi Fallen Order: Video Game	31	Robot

About half the sounds come from characters that are robots. 17% were from animals, 9% by alien characters, and others from various other imaginary characters.

3.2 Experiment process

The experiment proceeds as follows, after starting the experiment program: The sounds are played in random order for the subject. Each time a sound is played, the subject must rate it in terms of Valence and Arousal. Each of these is rated on a seven-point scale where a Valence rating of -3 is the lowest, representing a negative emotional state. -2 is a negative valence, -1 is a slightly negative valence. 0 is a neutral valence, 1 is a slightly positive valence, 2 is a positive valence, and 3 is the most positive valence, representing a positive emotional state. Arousal rating of -3 is the lowest arousal rating, representing a very low level of excitedness or energy. -2 is a low level of arousal, and -1 is slightly low arousal. 0 represent a neutral level of arousal. 1 is slightly high arousal, 2 is high arousal, and 3 is the highest level, representing an extremely energetic or excited state. The subject can repeat each sound as many times as necessary and, once satisfied with the ratings, proceeds to the next sound. Once all sounds have been rated, the experiment ends, and the rating data is saved into a text file. The sounds were labelled according to Russell's Circumplex model of affect (Posner et al., 2005)

3.3 Statistical analysis

A total of 17 subjects participated in the experiment, with each subject rating batches of 140 files each, 5 unique batches totaling 560 sounds. Results from the labelling experiment showed that most NLUs were rated to be slightly negative Valence and slightly high Arousal, as shown in figure 2. The ratings appear to be normally distributed for both Valence and Arousal.

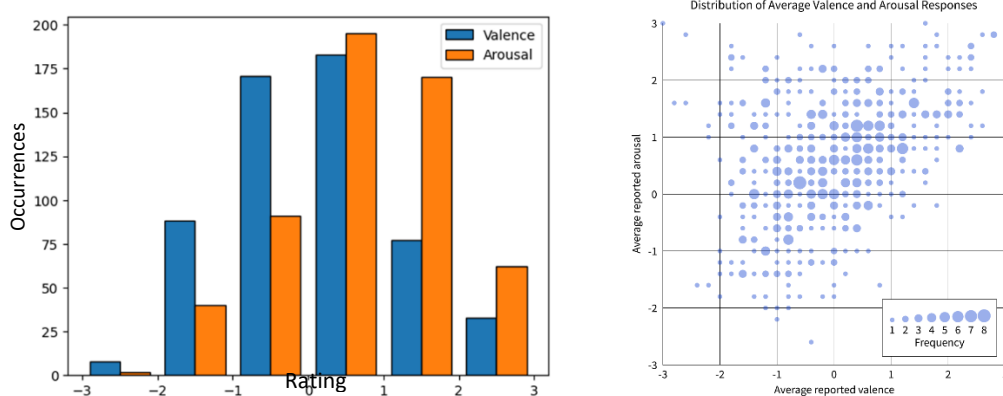


Figure 2. Valence and Arousal ratings data distribution and bubble scatter plot

To assess the reliability of the rating data, each batch of 5 ratings per file was analyzed for inter-rater agreement. Two measures were used to gauge inter-rater agreement. First, the standard deviation of ratings between all 5 raters was computed, for both Valence and Arousal, and the average of these values was calculated to be 1.18 for Valence and 1.12 for Arousal, indicating that

subjects were generally within just over one rating point for a given evaluation. Next, Krippendorff's Alpha was calculated for the 5 raters and for Valence and Arousal was found to be 0.36 each. Since Krippendorff's Alpha does not consider the interval nature of the rating scale, this result shows that there is reasonable agreement between the raters.

4 MODEL EXECUTION

4.1 Audio feature extraction

Three audio analysis packages extracted various audio feature sets from the audio files: pyAudioAnalysis, openSMILE Low Level Descriptors and openSMILE Functionals.

The pyAudioAnalysis audio feature library (Giannakopoulos, 2015) contains 136 frame-based audio features. Each sound file is separated into frames by defining frame length in seconds and frame step (difference between starting point of successive frames) in seconds. The 136 features are calculated based on these window sizes such that each audio file has 136 features per frame. Features include: zero crossing rate, energy, energy entropy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral rolloff, Mel Frequency Cepstral Coefficients (MFCCs), and chroma.

The openSMILE (open-Speech and Music Interpretation by Large-space Extraction) toolkit (Eyben et al., 2015, 2016) was developed to create a standardized freely available feature extractor for problems within the audio analysis domain. It is made up of two main feature sets, The first being Low Level Descriptors (LLDs), which are frame-based in that they are calculated over discrete frames or sections of the audio file. These include zero-crossings, signal energy, loudness, cepstral features such as MFCC and PLP-CC, as well as fundamental frequency, jitter, shimmer, chroma etc. The second feature set are functionals, which are calculated over the duration of the audio file. These include extreme values, means, moments, peaks, segments, coefficients of Discrete Cosine Transform (DCT), rise and fall times, etc.

4.2 Factor analysis

Factor analysis was performed on the normalized openSMILE Functionals and the pyAudioAnalysis features to reduce the dimensions before applying the machine learning models described in the following sections. 35 factors were used for the pyAudioAnalysis feature set and 12 factors for the openSMILE Functionals features set, based on the scree plot and eigenvalues (of greater than 1). As for the openSMILE Low-Level Descriptors features (22 features) the best performance of the model was achieved without using factor-transformed data.

4.3 Random forest machine learning model

A Random Forest Machine Learning model was trained and used to infer valence and arousal on a test data set. Firstly, the factor-transformed (except for openSMILE LLDs) dataset was split into training and testing data, and the Random Forest classifier was trained on the training data. The trained Random Forest model was then run on the test dataset to infer the Valence and Arousal values for each frame of audio data. The results were averaged over the number of frames per audio file to relate the predictions back to the original labels obtained from

experiments, in which the subjects had assigned Valence and Arousal ratings per audio file. The model was trained and tested with various combinations of the input parameters, namely, frame length, train-test split, and number of factors, and also with all 136 features.

The key metric used to evaluate a given model was Mean Absolute Error (MAE):

$$MAE = \frac{\sum_i^n |\hat{y}_i - \bar{y}_i|}{n}$$

Where MAE is the Mean Absolute Error, \hat{y}_i are the predicted values for Valence or Arousal, \bar{y}_i are the actual values for Valence or Arousal, n is the number of items in the test data set.

4.4 Transfer learning Mel spectrogram neural network model

A transfer learning Neural Network Mel Spectrogram based model was also used and results compared to the main model described in this work. Such techniques have been used in speech emotion recognition (Shor et al., 2020). In the current work, Mel Spectrograms were generated for each of the 560 sounds. The InceptionV3 Neural Network (Szegedy et al., 2015) was used to classify the sounds. The InceptionV3 Neural Network is a freely available Neural Network trained to recognize images trained on the ImageNet database, which contains around 14 million images in 20,000 categories. The network is pre-trained and its weights frozen, as per the transfer learning method. The results of the Neural Network model are included alongside the Random Forest model results in the current work.

5 RESULTS

Results of the audio feature random forest model, as well as the Mel spectrogram neural network model, are shown in Table 6. In the table, the transfer learning neural network model is referred to as TLNeuralNet and the simple sequential neural network model as NeuralNet. The openSMILE features are listed as follows: openSMILE Functionals after factor analysis are openSMILEFuncFT, openSMILE Low Level Descriptors are openSMILELLD. The pyAudioAnalysis features after factor analysis are called pyAudioFT.

Table 1. Model execution results: Mean Squared Error (MAE) – results with lowest MAE are in bold

Model Type	Feature Set	MAE Valence	MAE Arousal	Factors/Features	Train/Test Split
TLNeuralNet	Mel Spectrograms	0.147	0.161	N/A	0.9
NeuralNet	openSMILEFuncFT	0.136	0.106	12	0.9
Random Forest	pyAudioFT	0.121	0.105	35	0.9
	openSMILEFuncFT	0.116	0.101	12	0.9
	openSMILELLD	0.107	0.097	22	0.9

Figure 3 shows scatter plots of actual vs predicted values produced by the model for Valence and Arousal. The Valence, the correlation is 0.63 and for Arousal 0.75.

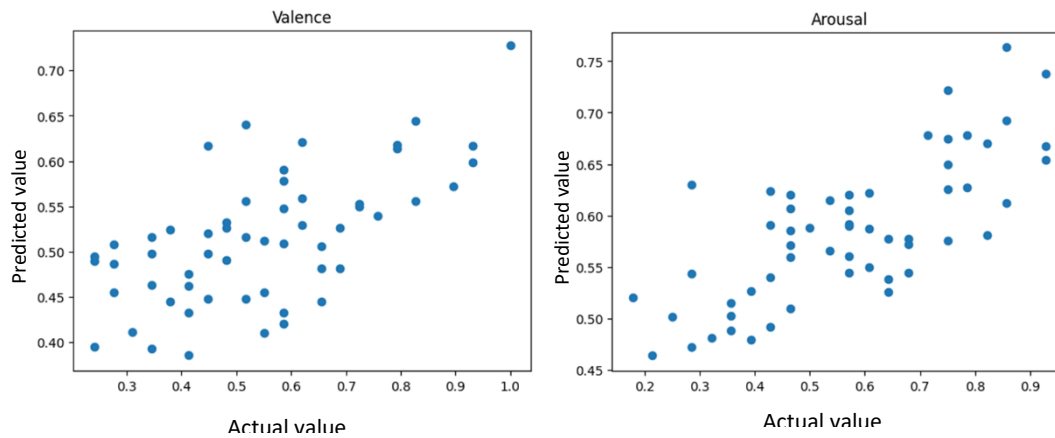


Figure 3: Actual vs predicted values of Valence and Arousal: scatter plots

6 DISCUSSION

The lowest average error achieved for Valence was 0.107, and for Arousal 0.097, both from the Random Forest model and using the openSMILE LLD Feature Set with 22 Features and a 90/10 Training Test split. The Random Forest model performed best. However, a simple sequential Neural Network using the openSMILE Functionals produced comparable results. The Mel-Spectrogram Transfer Learning InceptionV3 model performed worse, possibly due to the spectrograms containing too much information that was not useful for the neural network, and there being only 560 spectrograms. The model might yield better results with more sounds/spectrograms. Considering the original 7-point rating scale, a difference of one rating point would be equal to $100/7 = 0.14$. Therefore, it can be deduced that the model is able to predict the Valence and Arousal of most sounds tested with an accuracy within one rating point difference along that scale.

The difference between the three feature sets was almost negligible. Combining features from each feature set might improve results. Regarding the openSMILE LLD feature set and pyAudioAnalysis feature set models, a frame-based method was used whereby results were averaged over the number of frames per sound as a post processing step to establish the predicted values of Valence and Arousal per sound and compare them to the actual rating data. Accounting for the contribution of each frame to the sound might also yield better results.

The model is slightly better at inferring the Arousal of a sound than its Valence. Arousal results also had higher correlations between predictions and actuals, as seen in figure 3. This result is not surprising since Arousal is mostly related to the amplitude and rhythm of a sound, whereas more factors make up Valence. Some experiment participants also commented that they found it easier to assess the Arousal of each sound than its Valence. The model was run multiple times and the average MAE for each sound each time it occurred in the test data set was computed using the absolute value of the difference between predicted and actual Valence and Arousal. This value was averaged over the number of model runs and used to indicate whether a given sound was able to be successfully classified by the model or not. Sounds with higher average MAE were analyzed and the most common characteristics were; sounds too long and/or containing too many utterances, containing too much background music, or containing speech.

7 CONCLUSIONS AND FUTURE WORK

A model was developed to infer Valence and Arousal of NLUs from popular media sources. The model predicted the Valence and Arousal of an NLU within one rating point on the 7-point scale used in the experiment. This system can be used to evaluate candidate sounds for sound design contexts such as social public-facing robots or interactive agents, for commercial and industrial settings, public spaces such as train stations, or in media and entertainment. Possible applications also include machine sounds such as those used in smart devices or even hospital machines.

In future, the modification of the model such that results are weighted in terms of each frame's contribution to the audio file may improve results. Weighting by frame for features extracted at a frame level might better relate the model back to the original experiment context in which sounds were evaluated by audio file. Future work will also focus on expanding and developing the model in terms of exploring wider ranges of sounds. Increasing the number of sounds used could lead to better model results and more suitability to deep learning methods. For example, Generative Adversarial Networks could be used to create sounds to augment the dataset, and to validate the model by creating novel unseen sounds to test it. Sound generation with support from this type of model would lead to useful applications and systems to generate sounds for specific interaction scenarios. This model is the first of its kind to use machine learning to accurately relate the features of Non-Linguistic Sounds to their affective dimensions. In the future, the aim is to use the current model to develop a sound design support system for social robots, entertainment, and other media that use such sounds.

REFERENCES

- Bethel, C. L., & Murphy, R. R. (2006). Auditory and Other Non-Verbal Expressions of Affect for Robots. *Fall Symposium Series, Aurally Informed Performance: Integrating Machine Listening and Auditory Presentation in Robotic Systems*.
- Blattner, M. M., Sumikawa, D. A., & Greenberg, R. M. (1989). Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction*, 4(1), 11–44. https://doi.org/10.1207/s15327051hci0401_1
- Chen, L., Su, W., Feng, Y., Wu, M., She, J., & Hirota, K. (2020). Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Information Sciences*, 509, 150–163. <https://doi.org/10.1016/j.ins.2019.09.005>
- Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2), 190–202. <https://doi.org/10.1109/TAFFC.2015.2457417>
- Eyben, F., & Schuller, B. (2015). openSMILE:). *ACM SIGMultimedia Records*, 6(4), 4–13. <https://doi.org/10.1145/2729095.2729097>
- Fernandez De Gorostiza Luengo, J., Alonso Martin, F., Castro-Gonzalez, A., & Salichs, M. A. (2017). Sound synthesis for communicating nonverbal expressive cues. *IEEE Access*, 5, 1941–1957. <https://doi.org/10.1109/ACCESS.2017.2658726>

- Giannakopoulos, T. (2015). PyAudioAnalysis: An open-source python library for audio signal analysis. In *PLoS ONE* (Vol. 10, Issue 12). <https://doi.org/10.1371/journal.pone.0144610>
- Iliou, T., & Anagnostopoulos, C. N. (2009). Comparison of different classifiers for emotion recognition. *PCI 2009 - 13th Panhellenic Conference on Informatics*, 102–106. <https://doi.org/10.1109/PCI.2009.7>
- Issa, D., Fatih Demirci, M., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59, 101894. <https://doi.org/10.1016/j.bspc.2020.101894>
- Khota, A., Kimura, A., & Cooper, E. (2019). Modelling of Non-Linguistic Utterances for Machine to Human Communication in Dialogue. *International Symposium on Affective Science and Engineering, ISASE2019*, 1–4. <https://doi.org/10.5057/isase.2019-C000037>
- Khota, A., Kimura, A., & Cooper, E. (2020). Modelling Synthetic Non-Linguistic Utterances for Communication in Dialogue. *International Journal of Affective Engineering*, 19(2), 93–99. <https://doi.org/10.5057/ijae.ijae-d-19-00011>
- Komatsu, T. (2005). Toward making humans empathize with artificial agents by means of subtle expressions. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3784 LNCS(2004), 458–465. https://doi.org/10.1007/11573548_59
- Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective. *October*, 141(4), 520–529.
- Pure Data — Pd Community Site. (n.d.). Retrieved July 8, 2019, from <https://puredata.info/>
- Read, R., & Belpaeme, T. (2012). How to use non-linguistic utterances to convey emotion in child-robot interaction. *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '12*, 219. <https://doi.org/10.1145/2157689.2157764>
- Read, R., & Belpaeme, T. (2014). Non-linguistic utterances should be used alongside language, rather than on their own or as a replacement. *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction - HRI '14*, 276–277. <https://doi.org/10.1145/2559636.2559836>
- Read, R., & Belpaeme, T. (2016). People Interpret Robotic Non-linguistic Utterances Categorically. *Intl Journal of Social Robotics*, 8(1), 31–50. <https://doi.org/10.1007/s12369-015-0304-0>
- Shor, J., Jansen, A., Maor, R., Lang, O., Tuval, O., de Chaumont Quiry, F., Tagliasacchi, M., Shavitt, I., Emanuel, D., & Haviv, Y. (2020). Towards learning a universal non-semantic representation of speech. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2020-Octob*, 140–144. <https://doi.org/10.21437/Interspeech.2020-1242>
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015). Going Deeper with Convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.
- Yilmazyildiz, S., Read, R., Belpaeme, T., & Verhelst, W. (2016). Review of Semantic-Free Utterances in Social Human–Robot Interaction. *International Journal of Human-Computer Interaction*, 32(1), 63–85. <https://doi.org/10.1080/10447318.2015.1093856>