**Predicting Human Valence Ratings of Natural Spoken Narratives  
from Lexical and Acoustic Information**

Jess Alexander | CompMethods Final Project | Fall 2024

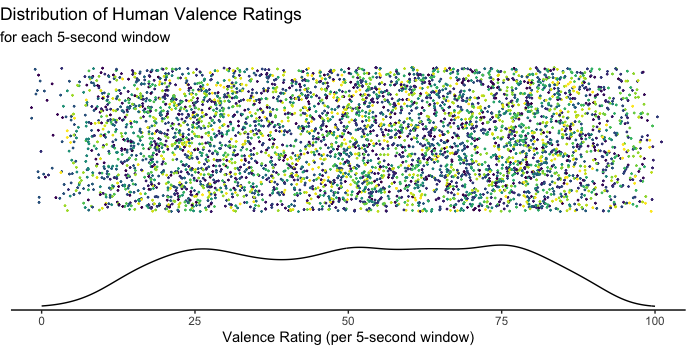
The cues used by human observers to predict the emotional state of an interlocutor are ambiguous, and this is particularly true for vocal emotion (Atias et al., 2019). Efforts to predict the negative-to-positive degree of “pleasantness” of emotional speech (termed “valence”) have been especially fraught (Busso & Rahman, 2012). According to prior work using time-series models (Ong et al., 2021), the perception of speaker valence by human listeners in an auditory-only modality is predominantly based on signal semantics, with acoustic features such as prosodic contour and voice quality demonstrating weaker explanatory power. Here, I investigate several linear regression models to compare the explanatory power of semantic and acoustic information, both alone and in combination. Furthermore, I explore the extent to which a self-supervised transformer model (Wav2Vec2.0) is able to simulate human behavior in valence ratings of natural, spoken narratives.

**Human Ratings: Our Gold Standard**

The Stanford Emotional Narratives Dataset (Ong et al., 2021) offers a naturalistic context in which to explore the cues used by human observers to assess a speaker’s emotional state. The corpus consists of 194 videos of unscripted narratives, provided by 49 unique speakers who were asked to share positive and negative personal experiences. Each video is approximately 2.25 minutes in length, and the entire corpus comprises 24 minutes of data. After the recording of these narratives, 700 raters provided independent valence ratings, with at least 20 raters viewing each video. Ratings were collected with a sliderbar, which the rater dynamically adjusted during the video; ratings were recorded as the sliderbar value every five seconds. A “gold standard rating” was calculated by weighting the recorded valence values of each rater by the extent to which that rater’s ratings correlated with the mean group rating, then averaging across all raters for a given video. This project reviews several models that predict this gold standard rating in order to shed light on the cues that human raters rely upon in their assessment of a speaker’s emotional state.

Given that human valence ratings are available for each five-second window of the SEND narratives, my models aim to predict this rating for each window, independently. However, I discard the first five seconds of each narrative on the premise that human ratings recorded in the first moments of each video will lag behind the ability of the rater to physically provide a rating with the sliderbar. I additionally discard any five-second window in which one of the predictors cannot be computed. For instance, in some windows the speaker may not produce any content words that could be used to extract lexical predictors. In order to compare models on an even footing, such windows are excluded from all models. In total, 277 windows were excluded, leaving 4860 five-second windows used as input to all models.

A quasi-normal distribution of ratings for these 4860 windows was confirmed visually.

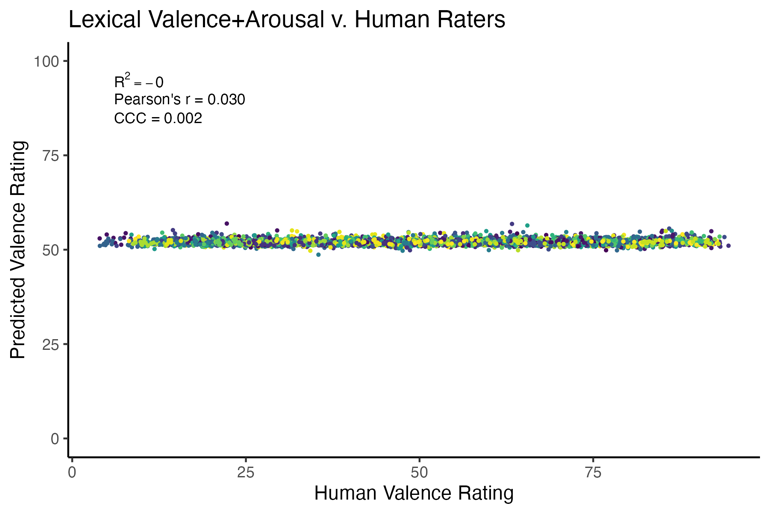


*Fig 1.* Jittered distribution and density of the gold standard human valence rating for each five-second window of the final dataset. Each dot represents one five-second window (4860 in total). Each unique speaker (N = 49) is represented by a different color.

**Lexical Models**

Individual words differ in lexical characteristics associated with emotionality, including valence and arousal (Larsen et al., 2008). One possibility, therefore, is that human observers simply average over each word of an utterance to compute a “valence score,” which they then use to assess the speaker’s emotional state. I examined this possibility (Model 1) by constructing a mixed effects model to predict human valence ratings, with fixed effects of average lexical valence and average lexical arousal, and a random intercept by speaker. For each five-second window, average lexical valence and arousal were computed on the basis of norms compiled by Warriner et al. (2013), which provide a single value between 1 and 9 for 13,915 English lemmas. SpaCy (Honnibal & Montani, 2017) was used to lemmatize the transcriptions included in the SEND dataset, after which valence and arousal values were averaged across all lemmas available in the Warriner norms. The resulting average values were median-centered by subtracting the median valence (5.2) and arousal (4.11) of all lemmas in the Warriner norms; in this way, a median value was set equal to 0.

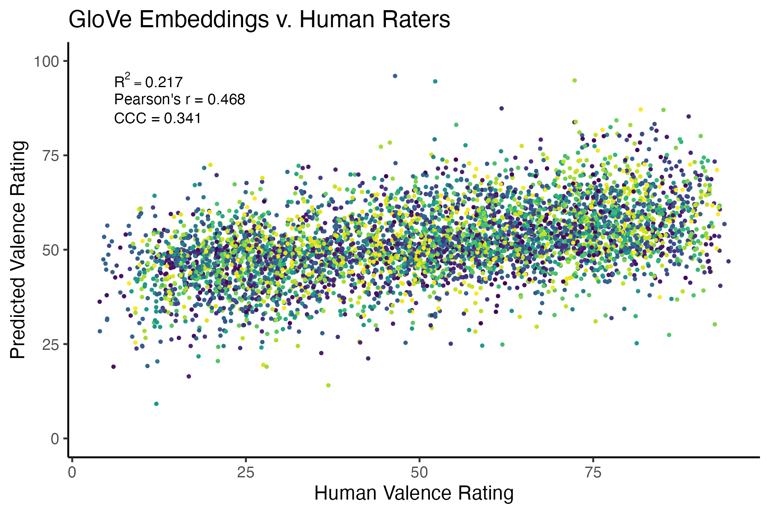
This model was highly unsuccessful at predicting human valence ratings. As the average valence and arousal values for each five-second window hover near the median of the norms, the model fails to explain the wider variance in the human gold standard.



*Fig 2.* Results of Model 1, which predicts human valence ratings by average lexical valence and average lexical arousal in each five-second window on the basis of the Warriner et al. (2013) norms.

Clearly, human observers do not simply compute average lexical values to assess a speaker’s emotional state. However, humans do rely predominantly on lexical information (Ong et al., 2021) to assess emotional valence, so perhaps a model based on a larger semantic space would better account for human ratings. I therefore constructed my second model using GloVe embeddings via gensim (Pennington et al., 2014; Rehurek & Sojka, 2011). For each five-second window, I computed the centroid of all GloVe embeddings for lemmas in the window. Stop words, punctuation, and lemmas not available in ‘glove-wiki-gigaword-300’ were excluded from the centroid computation. The 300 embedding dimensions were submitted as fixed effects in a mixed effects model to predict the gold standard human valence rating, with a random intercept by speaker.

This model (Model 2) explained 21.7% of the variance in human ratings. It also demonstrated a concordance correlation coefficient–a measure of agreement between the gold standard ratings and the ratings predicted by the model–of 0.341. While this is substantially better than Model 1, it still leaves most of the variance in human assessments unexplained.

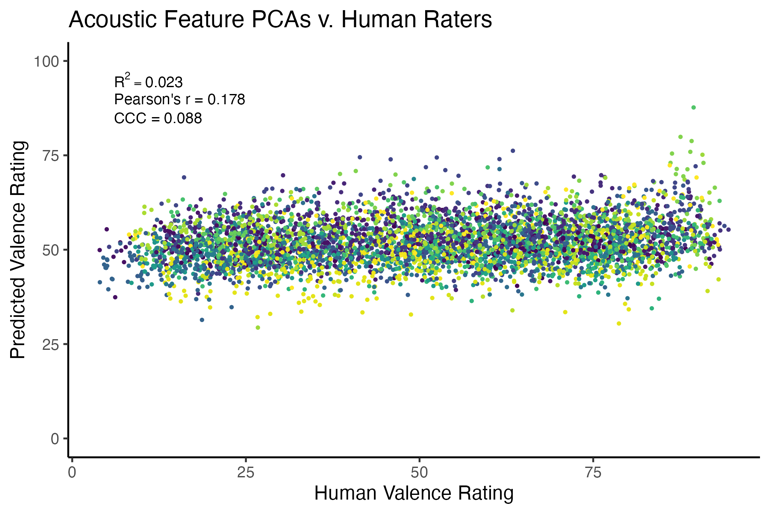


*Fig 3.* Results of Model 2, which predicts human valence ratings by the 300-dimensional GloVe embeddings of the semantic centroid for each five-second window.

**Acoustic Model**

Acoustic features of speech, such as fundamental frequency and intensity, have been associated with different vocal emotions (Banse & Scherer, 1996; Grichkovtsova et al., 2012; Wang et al., 2014). Given that listeners are reliably able to identify a speaker’s intended emotion even when the signal is comprised of neutral semantics, such as the innocuous phrase “It’s eleven o’clock” (Cao et al., 2014), human valence ratings may be predictable from signal acoustics. I probed this by extracting 88 acoustic features using the extended Geneva Minimal Acoustic Parameter Set (eGeMAPS: Eyben et al., 2016) to serve as model predictors in Model 3. However, acoustic features demonstrate high levels of collinearity. I therefore derived 88 principal components from the acoustic features in order to submit 88 orthogonal predictors to the model. As before, these predictors served as fixed effects in a mixed effects model to predict the gold standard valence rating, with a random intercept by speaker.

Only a small amount of variance is explained by acoustic features alone (2.3%).

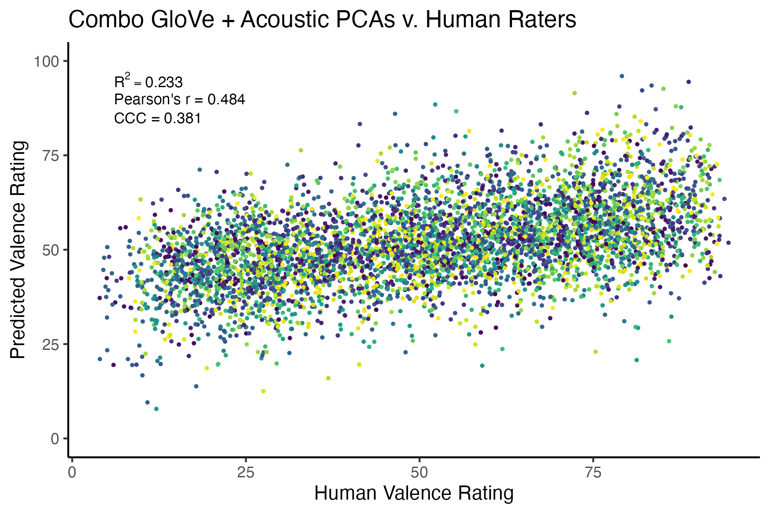


*Fig 4.* Results of Model 3, which predicts human valence ratings on the basis of 88 principal components derived from eGeMAPS acoustic features extracted for each five-second window.

**Combination Model**

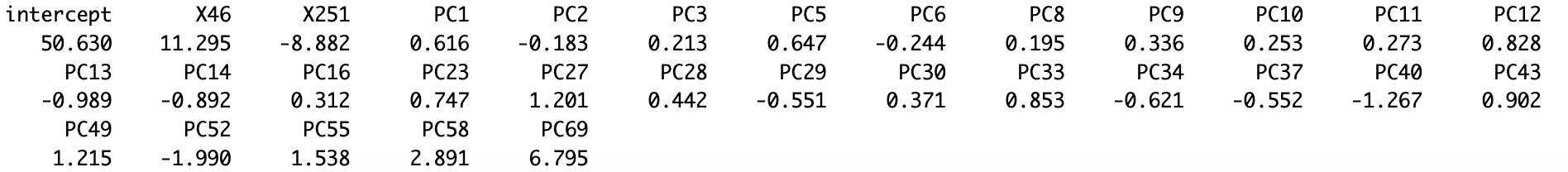
Human observers plainly use both semantic and acoustic information in judging the emotional state of an interlocutor. It therefore seems reasonable to combine our lexical and acoustic models in order to determine to what extent lexical and prosodic information is simply additive in terms of explaining human valence ratings. The resulting mixed effects Model 4 thus predicts the gold standard rating with 389 fixed effects: intercept, a semantic centroid described by 300 GloVe dimensions, and 88 orthogonal features of acoustic information. As before, a random intercept by speaker was also included in the model.

The inclusion of acoustic information improves performance over the lexical-only Model 2. However, the effects of combining semantic and acoustic information is sub-additive: if human listeners integrated semantic and acoustic information in a purely additive manner, we would expect Model 4 to explain 24% of the variance in human valence ratings, but it only explains 23.3%. This suggests that signal semantics and speaker voice quality exhibit some (albeit minor) redundancy in terms of conveying the speaker’s emotional state. Based on the models I have investigated thus far, it would seem that the brunt of this labor is performed by signal semantics.



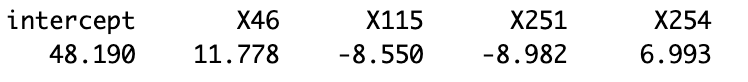
*Fig 5.* Results of Model 4, which combines all predictors from Models 2 and 3.

Interestingly, an exploration of the weights in Model 4 raises new questions. Of the 389 predictors in the model, 31 are statistically different (p < 0.05) from 0. Only two of these are GloVe dimensions; conversely, 28 are principal components of the acoustic feature space.



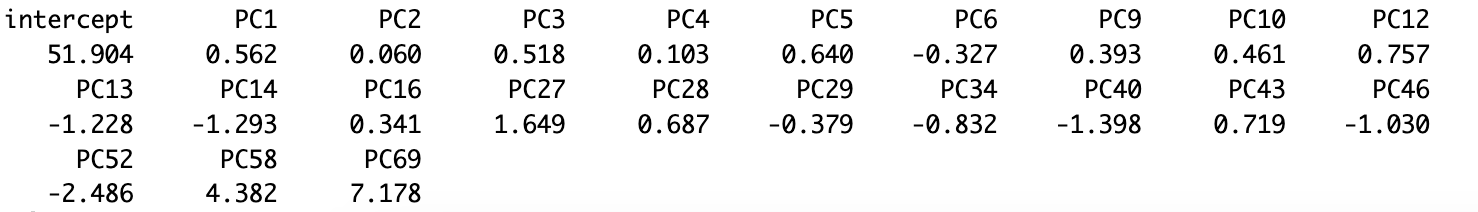
*Fig 6.* Significant model weights for Model 4. “X--” represents a GloVe dimension; “PC-” represents an acoustic principal component.

This mystery is resolved by drilling down into the weights for Models 2 and 3. In Model 2, only four of the GloVe dimensions actually contributed to model performance.



*Fig 7.* Significant model weights for Model 2. “X--” represents a unique GloVe dimension

In contrast, 22 of the principal components derived from acoustic features contributed to performance in Model 3:



*Fig 8.* Significant model weights for Model 3. “PC-” represents an acoustic principal component.

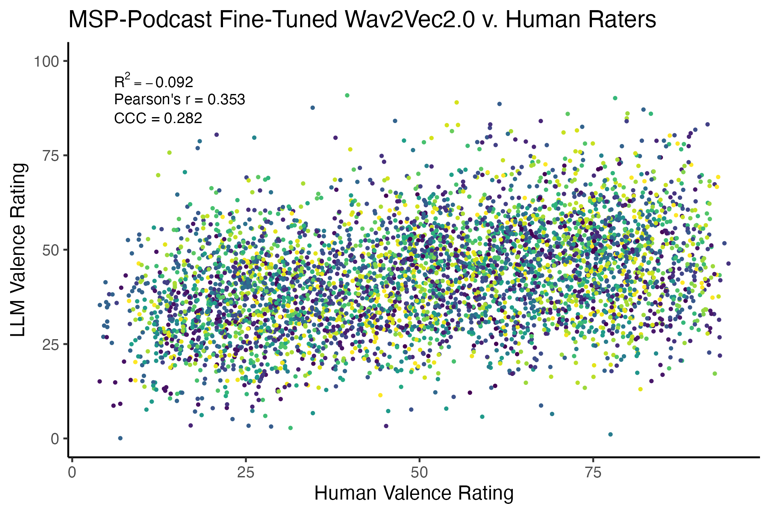
It thus appears that several semantic dimensions are doing all the heavy lifting in this combo model, and receive secondary support from multiple aspects of the acoustic signal.

**Accounting for Nonlinearity**

Thus far, I have explored multiple linear regression models that use either lexical or acoustic features, or a combination of these features, to predict listener valence ratings for five-second windows of natural spoken narratives. This methodology assumes a linear relation between lexical and/or acoustic features and human listener evaluations of speaker valence. However, human neural responses to auditory input, including emotional speech, are non-linear (Biesmans et al., 2017; Horikawa et al., 2020), and human rating behavior may thus be better explained by models that account for such non-linearity. In fact, Ong and colleagues (2021) used their dataset of natural, spoken narratives to analyze the accuracy with which deep learning models could predict human valence ratings in similar five-second windows. In their case, they explored long short-term memory (LSTM) and variational recurrent neural network (VRNN) models; in their final test set, these models demonstrated a concordance correlation coefficient of 0.09 and 0.35, respectively.

Given recent, substantive advancements in the performance of transformer-based models, I tested whether one such model, Wav2Vec2.0 (Baevski, et al., 2020), could outperform my best-performing Model 4. Wagner and colleagues (2023) fine-tuned Wav2Vec2.0 (large, robust) on the MSP-Podcast corpus of emotional speech (Loftian & Busso, 2019) with the multitask of predicting three different dimensional aspects of emotion–arousal, dominance, and valence–each in a range of 0-1.[[1]](#footnote-1) The classification accuracy of this fine-tuned model was then tested (Derington et al., 2023) on several existing corpora of emotional speech, achieving a concordance correlation coefficient of 0.50 for valence across all corpora. Notably, the more naturalistic narratives collected by Ong and colleagues (2021) were not among the corpora tested.

I used the pubicly-available model by Wagner and colleagues (2023) to make out-of-sample valence predictions (Model 5), based on the raw waveform to which human raters were exposed, for all 4860 five-second windows of data provided to the linear regression models described earlier. This achieved a concordance correlation coefficient of 0.282, underperforming the multi-corpus testing by Derington and colleagues (2023). It seems likely that the more ecologically valid SEND narratives are less “extreme” in their emotions than those used either for fine-tuning the model or in the testing corpora. The results here also fall short of those achieved by the VRNN model in the work of Ong and colleagues (2021). This is likely explained by the fact that Model 5 treats each five-second window as independent whereas the VRNN constructed by Ong and colleagues (2021) models both inputs and outputs as generated from a lower-dimensional latent state to capture sources of variation across speakers, narrative styles, etc., and thus makes it possible for their model to learn which sources of variation are relevant for valence predictions.



*Fig 9.* Results of Model 5, comprised of out-of-sample predictions extracted from Wav2Vec2.0 fine-tuned on the MSP-Podcast dataset.

Importantly, however, Model 5 demonstrates that the concordance correlation coefficient does not always tell the whole story. For a direct comparison with the linear regression models, I calculated the variance explained by the model’s predictions as a simple R2. Despite the decent concordance correlation coefficient, the predictions from Wav2Vec2.0 are actually slightly less reliable than simply using the mean of human valence ratings to predict any single five-second window, yielding a slightly negative coefficient of determination.

**Up Next**

I am currently in the process of further fine-tuning the model released by Wagner and colleagues (2023) on a subset of the SEND narratives. I have randomly assigned each of the 49 speakers into the training or test set, with a 50/50 split. Following further fine-tuning of the model, I plan to again perform an out-of-sample prediction on the test set to determine whether the additional fine-tuning on natural, spoken narratives is sufficient to increase the model’s ability to predict the gold star human valence ratings. I am especially curious to see how additional fine-tuning affects both the amount of variance that the model can explain, as well as its concordance correlation coefficient.

**References**

Atias, D., Todorov, A., Liraz, S., Eidinger, A., Dror, I., Maymon, Y., & Aviezer, H. (2019). Loud and unclear: Intense real-life vocalizations during affective situations are perceptually ambiguous and contextually malleable. *Journal of Experimental Psychology: General*, *148*(10), 1842–1848. <https://doi.org/10.1037/xge0000535>

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460. <https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html>

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636. <https://doi.org/10.1037/0022-3514.70.3.614>

Biesmans, W., Das, N., Francart, T., & Bertrand, A. (2017). Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(5), 402–412. <https://doi.org/10.1109/TNSRE.2016.2571900>

Busso, C., & Rahman, T. (2012). Unveiling the acoustic properties that describe the valence dimension. *Interspeech 2012*, 1179–1182. <https://doi.org/10.21437/Interspeech.2012-124>

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, *5*(4), 377–390. <https://doi.org/10.1109/TAFFC.2014.2336244>

Derington, A., Wierstorf, H., Özkil, A., Eyben, F., Burkhardt, F., & Schuller, B. W. (2023). *Testing Speech Emotion Recognition Machine Learning Models* (arXiv:2312.06270). arXiv. <http://arxiv.org/abs/2312.06270>

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>

Grichkovtsova, I., Morel, M., & Lacheret, A. (2012). The role of voice quality and prosodic contour in affective speech perception. *Speech Communication*, 54(3), 414–429. <https://doi.org/10.1016/j.specom.2011.10.005>

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Horikawa, T., Cowen, A. S., Keltner, D., & Kamitani, Y. (2020). The Neural Representation of Visually Evoked Emotion Is High-Dimensional, Categorical, and Distributed across Transmodal Brain Regions. *iScience*, *23*(5), 101060. <https://doi.org/10.1016/j.isci.2020.101060>

Larsen, R. J., Mercer, K. A., Balota, D. A., & Strube, M. J. (2008). Not all negative words slow down lexical decision and naming speed: Importance of word arousal. *Emotion*, *8*(4), 445–452. <https://doi.org/10.1037/1528-3542.8.4.445>

Lotfian, R., & Busso, C. (2019). Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings. *IEEE Transactions on Affective Computing*, *10*(4), 471–483. <https://doi.org/10.1109/TAFFC.2017.2736999>

Ong, D. C., Wu, Z., Zhi-Xuan, T., Reddan, M., Kahhale, I., Mattek, A., & Zaki, J. (2021). Modeling Emotion in Complex Stories: The Stanford Emotional Narratives Dataset. *IEEE Transactions on Affective Computing*, *12*(3), 579–594. <https://doi.org/10.1109/TAFFC.2019.2955949>

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1162>

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10745–10759. <https://doi.org/10.1109/TPAMI.2023.3263585>

Wang, T., Ding, H., Kuang, J., & Ma, Q. (2014). Mapping emotions into acoustic space: The role of voice quality. *Interspeech 2014*, 1978–1982. <https://doi.org/10.21437/Interspeech.2014-451>

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. <https://doi.org/10.3758/s13428-012-0314-x>

1. https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim [↑](#footnote-ref-1)