Workshop on Speech, Music and Mind (SMM 2020)
23 October 2020, Nanjing, China

Role of emotion words in detecting emotional valence from speech

Venkata S Viraraghavan, Rahul D Gavas, Ramesh K Ramakrishnan

TCS Research and Innovation, Tata Consultancy Services
{venkatasubramanian.v,rahul.gavas,ramesh.kumar}@tcs.com

Abstract

An important task in several wellness applications is detection of emotional valence from speech. Two types of features of speech signals are used to detect valence: acoustic features and text features. Acoustic features are derived from short frames of speech, while text features are derived from the text transcription. In this paper, we investigate the effect of text on acoustic features. Some studies show that acoustic features of phones carry specific emotion information. We also observe that emotion words and the emotional valence of the spoken sentence need not always match (e.g. the usage of 'not happy'). We thus propose that acoustic features of speech segments carrying emotion words must be treated differently from other segments that do not carry such words. In this paper, we propose that all speech segments carrying emotion words are excluded from the training set. Standard emotion words from a language, words from Plutchik's wheel of emotion, and their synonyms are considered. We report performance results on the the Elderly Emotion Sub-Challenge corpus of the Computational Paralinguistics Challenge 2020. We show that exclusion of emotional words show significant improvements for both OpenSMILE ($p < 0.05$) and OpenXBoAW features ($p < 0.01$).

Index Terms: emotional valence, speech, emotion words

1. Introduction

According to the World Health Organization, "Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity". Emotional well-being, an important aspect of mental well-being, is the emotional aspect of everyday experience [1]. Recently, several organizations and enterprises have increased attention on emotional well-being at the workplace.

Emotions are expressed by humans through multiple modes, such as facial expressions, speech and body language, in general. Emotion is commonly measured using up to five dimensions, but nearly all models use at least two: valence and arousal [2]. An important task in understanding an individual's mental state is detecting their valence and arousal. Since audio recording is available on smartphones, detecting valence and arousal from speech [3] is an important approach.

Emotion in speech may be detected from samples of speech as short as exclamations or even long sentences. For training machine learning/deep learning algorithms, emotion may be annotated continuously (e.g. [4]) or on larger chunks depending on the context. For example, if a person is narrating a happy event, all sentences in the narration can be annotated as happy. In these situations, the typical approach is to split long signals into many shorter segments that share the ground truth of the source speech. This method has been followed in the Computational Paralinguistics Challenge (ComParE) 2020, for the Elderly Emotion Sub-Challenge (ESC) [5], where each segment is about 5 seconds long. In the ESC, each speaker narrates a story that is classified as having Low (**L**), Medium (**M**), or High (**H**) Arousal (**A**). Similarly, the Valence (**V**) is classified as low (**L**; negative or sad), medium (**M**; neutral), or high (**H**; positive or happy). The details may be found in [5].

In reality, segments of a story do not always carry the same emotion as the overall story [6]. For example, a happy or sad story may be interspersed with segments of neutral emotion. In this paper we focus on detection of valence from speech. In our previous work [7], we found that the agreement between sentences and its constituent 'utterances' is under 70%. Another consideration, in the consistency between the valence of a sentence and that of its utterances, is the text that each utterance carries. Intuitively, it is expected that this text has a role to play. Inspired by these two observations, we propose a technique that selects a subset of annotated speech samples to train a classifier that predicts the valence of speech.

1.1. Literature Survey

Machine-learning for detecting valence from speech follows the typical process of building classifiers on extracted features. The ground truth associated with speech samples is provided by human observers/experts, or is self-assessed. Some recently used examples of *acousitc* features are OpenSMILE [8, 9], OpenXBoAW [10], time-domain differences [11,12]. Deep networks are being used increasingly, but hand-crafted features are still relevant [13]. End-to-end approaches use time-domain audio samples or spectrograms as inputs for classification [14].

From the current state of the art, it appears that detecting valence from speech alone has limitations in classification-performance. Thus, multi-modal detection approaches have been proposed; a thorough review is done in [15]. Specifically, the combination of speech-features and text-features has shown promise in valence detection [5]. This combination was used in the winning entry [16] of two sub-challenges (atypical and self-assessed affect) in ComParE 2018 [17].

Valence detection from speech and annotated text usually assumes that there is emotion information in the acoustic features, and in the text features, especially in words related to emotion, such as happy, sad, etc. In the Emo-DB [18], neutral sentences had to be spoken with different emotions so that text features were unlikely to carry their own emotion. In lexical compensation, acoustic features are compensated for the variability in words/sounds[1] by synthesizing the entire speech [19]. This compensation works does not consider emotion words separately. A recent work [20, 21] challenges the view that there is no inherent emotion information in phones. The authors show that phonetic (acoustic) features influence valence: e.g. words with hissing sibilants ('sss') feel more negative. In our approach, the training set consists of select speech segments without emotion words. This approach draws from an ongoing study on two datasets [22,23]. In this paper, we study the effect

[1] This technique also compensates for speaker variability.

10.21437/SMM.2020-6

of emotion words on valence detection using the ESC corpus.

2. Training-set based on emotion words

As explained in Section 1, all segments of a speech sample do not carry the same valence. We consider the effect of emotion words on the acoustic features of such segments. From the work of [20, 21], it appears that irrespective of the emotion of the overall narration, specific emotion words carry two types of emotion. The first is in their text-meaning and the second, in their acoustic features due to the phones in the emotion words. One possibility is that we use the acoustic features of only such emotion words for training a classifier. This approach leads to much less data than the original segments (see Table 1). In addition, the context may be such that the expressed emotion is opposite to that of the emotion word. An example from the ESC corpus[2] is:

> We didn't know how to deal with the situation, we didn't know how to help her, she was so withdrawn. And it only got ***better*** when she, uh, took help and started therapy.

In this case (and in other usage such as 'not happy'), the emotion of the story being narrated does not match the emotion of the word 'better' (comparative form of 'good'). Thus, we hypothesize that *excluding* emotion words would result in fewer contradictions such as the example above. The rest of this section details how the emotion words and speech segments for training are selected by excluding these words. The description is for valence, but it can be equally applied to arousal, where pronouns could also play a part.

2.1. Listing valence words

The choice of words that express high valence (positive) and low valence (negative) is based on emotion words in a language. For the ESC sub challenge, we use two websites that list standard German emotion words [24, 25]. Together, these result in a set of 68 emotion words. We then add the translations of 32 emotion words in Plutchik's wheel of emotions [26]. These are given in Figure 1. This results in 92 unique emotion words. Next, their synonyms and forms (predicative, adverbs, attributive, comparative in German) are found. The set chosen thus consists of 364 unique emotion words, and is denoted by $\mathcal{E}$.

2.2. Training-set selection

Let a long speech sample S marked as a training example, have a valence V_S, and let S be split into N segments, $s_n, n \in \{1, 2, \ldots, N\}$. Typically, but not always, $s_n \forall n$ are of the same length. By definition, the valence of s_n is $V_S \forall n$. A text-transcription for S can be obtained using manual or (semi-) automatic speech recognition. For the ESC corpus, both manual and automatic transcriptions of each story are available (a story corresponds to the speech sample S). Let there be K_n words in Segment s_n. We denote these words as $w_k^{(n)}, k \in \{1, 2, \ldots, K_n\}$. Algorithm 1 is used to obtain the selection of training segments $\mathcal{T}_S$ from S; $\mathcal{T}_S$ has no emotion words from $\mathcal{E}$. The complete training set, $\mathcal{T}$, consists of selections of training segments obtained from all training speech samples: $\mathcal{T} = \bigcup \mathcal{T}_S$.

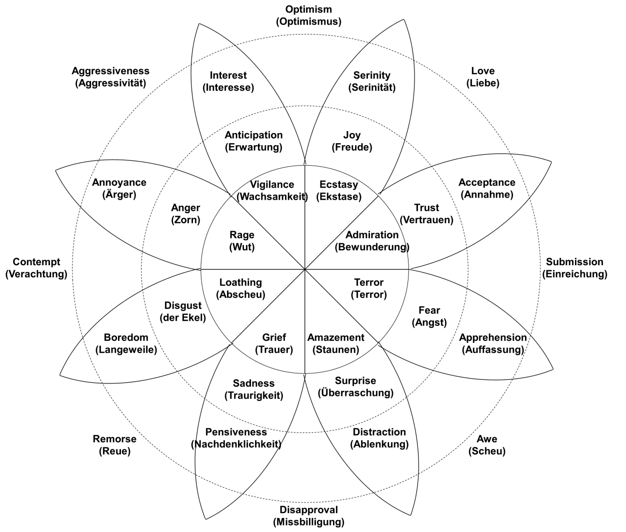


Figure 1: *English words on Plutchik's wheel of emotion and their German translations in brackets. Recommended magnification: $\geq 200\%$.*

Algorithm 1 Training selection from speech with N segments.

```
$\mathcal{T}_S \leftarrow \{\}$
for $n$ in $\{1, \ldots, N\}$ (each signal in $S$) do
    flagSelect ← True
    for $k$ in $\{0, 1, \ldots, K_n\}$ (each word in $s_n$) do
        if $w_k^{(n)} \in \mathcal{E}$ (speech signal has an emotion word) then
            flagSelect ← False (do not select it)
            break (for-loop)
        end if
    end for
    if flagSelect is True (no emotion words in $s_n$) then
        $\mathcal{T}_S \leftarrow \mathcal{T}_S \cup s_n$
    end if
end for
```

2.3. Features used

In this paper, we consider the two most successful acoustic features for the baseline valence prediction of the ESC [5]: OpenSMILE (6373 dimensional) and OpenXBoAW (dimension: 250 to 4000). To address class imbalance, we use the Synthetic Minority Over-sampling TEchnique (SMOTE) [27].

2.4. Classification

We use the Linear Support Vector Classifier (Linear SVC) following the methodology of the challenge baseline [5]. We also explore the use of other classifiers from the `sklearn` python package[3] in order to study the effect of excluding emotion words. The various classifiers use: Decision Trees (DT), Logistic Regression (LR), Support Vector Classifier (SVC), Gaussian Naive Bayes (GNB) modeling, Random Forest (RF). In addition, we use the One-vs-Rest (OvR) paradigm with the Multi-layer perceptron (OvR-MLP) and the eXtreme Gradient Boosted (XGB) classifiers. Further, for the Linear SVC, we optimize over the complexity values: $\{10^{-4}, 10^{-3}, 10^{-2}\}$.

[2] Story st2, Speaker ID ALBI110. For this paper, German text was translated to English at `https://translate.google.co.in/`

[3] Classifier parameters are left as default values if not specified. Our implementation uses open-source python libraries.

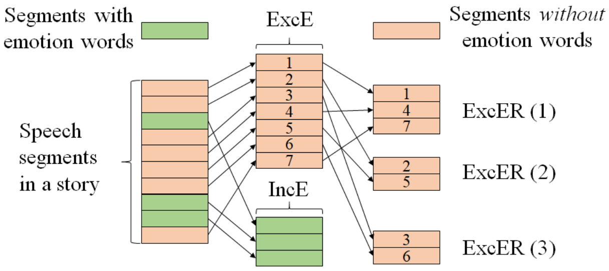


Figure 2: *Illustration of training set construction with an example of 1 in 3 segments selected for ExcER sets.*

3. Evaluation and Results

3.1. Database

We use the ESC database to evaluate the effect of presence and absence of emotion words in training. This database, formally called the Ulm State-of-Mind in Speech-elderly (USoMS-e) corpus, is described in [5]. We retain the suggested partitions of 'Train' (training set) and 'Dev' (development set or validation set) in order to be able to compare with baseline results on Dev. The Train, Dev, and Test partitions have 87 stories each, but they differ in distribution. There are 33 **L** stories and 24 **H** stories in Train, while there are 40 **L** stories in Dev and only 19 **H** stories. The evaluation criterion is the unweighted average recall (UAR); i.e. the average of the recall for each of the three classes **L**, **M**, and **H**. UAR penalizes predictions that favor any class, and is thus stricter than accuracy in unbalanced test sets. The lengths of the stories vary from $\sim$ 30 seconds to $\sim$275 seconds. The non-overlapping, individual speech segments split from a story are about 5 seconds long. Thus, the number of speech segments per story (N) varies.

Sixty five low-level descriptors (LLDs) are obtained from frames of individual speech segments. The frame-size is 1 second, and the hop-size is 40 ms. The 6373 OpenSMILE feature-points per segment are obtained as functionals over its LLDs. The same LLDs and their deltas (difference from one hop to the next) are used to obtain C codebooks. For a speech segment, the logarithmic term-frequency weighted histogram of its LLDs and of their deltas are concatenated into the OpenXBoAW features ($2C$ dimensions). We observe the baseline results to choose $C = 1000$ for this paper. Further details of the features and their computation can be found in [5, 9, 10].

3.2. Analysis of selected training sets

The numbers of speech segments with and without emotion words, as well as the total number of segments, are given per class in Table 1. The table shows that about 1 in 8 speech segments contain emotion words. Four words (gut, glücklich, traurig and freude, and their related forms) account for about half the occurrences of emotion words. Forty four such segments contain more than one emotion word. In our initial analysis (i.e. does not account for stemming and root-words), we found that the most frequently occurring words are not emotion words. Further, the emotion words are distributed across **L**, **M**, and **H** classes.

We measure the UAR for classifiers trained on the set of speech segments with emotion words, the set without emotion words, and on all training samples. Table 1 suggests that a classifier trained on the set without all emotion words is expected to perform better than one trained on the set with emotion words due to the size of the training set. To eliminate this effect, we measure the performance on the following four sets, which are illustrated in Figure 2.

Table 1: *Number of speech segments in the Train set*

Class	IncE: With Emotion Words	ExcE: Without Emotion Words	All: Total
L	109	912	1021
M	73	690	763
H	110	602	712
Total	292	2204	2496

1. IncE: 292 training segments with emotion words
2. ExcE-reduced (ExcER): $\sim$ 292 training segments by choosing 1 in every 8 segments *without* emotion words, while ensuring that (nearly) all stories are represented[4]. Each choice starts at a different offset (which ranges from 0 to 7).
3. ExcE: 2204 training segments *without* emotion words
4. All: 2496 training segments

For the case of ExcER, each set chosen as above is hand-pruned to ensure that the total number of segments is 330 after applying SMOTE (which matches 3 × 110 for IncE in Table 1). We compare the results among two pairs; the first pair is IncE and ExcE, and the second pair is ExcER and All.

3.3. Classification performance

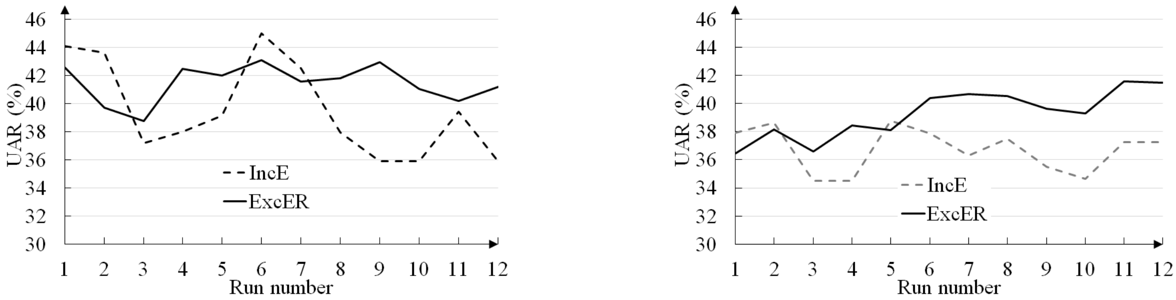
Table 2 shows the UAR obtained on the Dev set for various classifiers trained with OpenSMILE features. The training set is selected according to the technique in Section 2. For this measurement, the training sets considered are IncE, ExcE, and All (Section 3.2). The UARs given in the table are averaged over four runs.

Table 2: *Dev set performance with OpenSMILE features: UAR (%) for* 10 *classifiers and three types of training sets. The complexity parameter is shown in brackets for Linear SVC and SVC.*

	Training set		
Classifier	**IncE**	**ExcE**	**All**
DT	32.67	35.60	34.08
GNB	39.51	44.49	45.51
LR	39.00	42.45	42.33
Linear SVC (10^{-4})	40.72	47.83	46.03
Linear SVC (10^{-3})	41.15	44.92	49.02
Linear SVC (10^{-2})	36.79	43.33	45.47
OvR-MLP	38.06	42.49	43.23
OvR-XGB	36.67	47.08	44.93
RF	34.76	41.38	40.28
SVC (10^{-4})	42.69	40.81	42.93

Table 2 cannot be used to choose between ExcE and IncE because the sizes of these two training sets are disparate. Instead, we use Table 2, to choose the classifier that performs best: Linear SVC. For this classifier, we obtain results for the four training sets listed in Section 3.2. An average performance

[4] This is not possible for stories with less than 8 speech segments.



(a) OpenSMILE features

(b) OpenXBoAW features

Figure 3: *Dev set performance using Linear SVC: UAR (%) for two types of training sets and two feature sets.*

over multiple random sets is used for ExcER. Further, this measurement is repeated for the OpenXBoAW features also. These results are presented in Tables 3 and 4, in which the pairs {IncE,ExcER} and {ExcE,All} are demarcated. All results in the table are averaged over four runs. These results show that the UAR for ExcER is mostly higher than for the IncE. Since these two training sets are of identical size, it suggests that the exclusion of emotion words has an effect on the UAR. To verify this conclusion, the UAR using the Linear SVC for 12 runs (= 4 runs for 3 complexity values) per feature set are shown in Figure 3. The figure shows that the UAR is indeed mostly higher for the ExcER training set than for the IncE training set. A t-test reveals that the differences are significant for OpenSMILE ($p < 0.05$) and OpenXBoAW features ($p < 0.01$) and the combined results ($p < 0.002$).

Table 3: *Dev set performance with OpenSMILE features: UAR (%) for four types of training sets.*

Classifier	Training set IncE	ExcER	ExcE	All
Linear SVC (10^{-4})	40.72	**40.88**	**47.83**	46.03
Linear SVC (10^{-3})	41.15	**42.11**	43.33	**49.02**
Linear SVC (10^{-2})	36.79	**41.35**	44.92	**45.47**

Table 4: *Equivalent of Table 3 for OpenXBoAW features. Dev set performance: UAR (%) for four types of training sets*

Classifier	Training set IncE	ExcER	ExcE	All
Linear SVC (10^{-4})	36.38	**39.16**	41.28	**41.99**
Linear SVC (10^{-3})	37.62	**38.95**	**46.87**	40.24
Linear SVC (10^{-2})	36.16	**39.01**	**42.28**	39.93

In another experiment with OpenSMiLE features and the Linear SVC (complexity $= 10^{-2}$), the Dev set is split into four, approximately-equal-size subsets of stories, and one of them is left out for testing. The other three are used for training (only those segments that do not have emotion words). These four splits are named ExcD1 to ExcD4 depending on the subset that is left out for testing. The UARs for these splits are given in Table 5. Finally, a model is trained on the combined Train and Dev sets, with speech segments that do not have emotion words. The ESC baseline is 49.0%, which was obtained using only text features. The UAR for this model's predictions of valence on the Test set, as returned by the challenge website, is 36.3%. The value being below the baseline is not unexpected as text features have not been used yet. Our experiments with other features (e.g. Dynamic Mode Decomposition [28]) suggest that text features show a very large improvement (57% on Dev Set) even though the features themselves were not performing well. Thus, we choose to measure the impact of acoustic features without fusing text features. Our result of 36.3% is comparable to the best UAR among the baseline results that use only acoustic features (36.9%). Since our approach locates emotion words, and text features also consider emotion words, fusing the two sets requires further analysis, which is in progress.

Table 5: *Dev set and test performance with OpenSMILE and Text features fused: UAR (%) for Linear SVC (complexity = 10^{-2}). Training segments with emotion words are excluded.*

Dev ExcE	ExcD1	Subsets of Dev Set ExcD2	ExcD3	ExcD4	Test set
44.92	36.67	25.00	51.15	45.44	36.30

4. Conclusion

In this paper, we examined the effect of emotion words on the unweighted average recall of valence for several classifiers operating on acoustic features. We proposed that the training set should be selected based on the presence of emotion words. For equal sizes of training sets, excluding speech segments with emotion words results in improved performance compared to including them. In practice, the number of speech segments with emotion words is a small fraction of the total number of segments. Consequently, the size of the training set that excludes emotion words is only slightly smaller than the complete training set. Thus, only a minor difference in performance is expected between the two sets. Even so, the best UAR with OpenXBoAW features was obtained when the training set excluded emotion words. In summary, emotion words in speech segments affect acoustic features, and their influence on the performance of valence-classification needs to be accounted for.

5. References

[1] D. Kahneman and A. Deaton, "High income improves evaluation of life but not emotional well-being," *Proceedings of the national academy of sciences*, vol. 107, no. 38, pp. 16 489–16 493, 2010.

[2] J. A. Russell and J. M. Carroll, "On the bipolarity of positive and negative affect." *Psychological Bulletin*, vol. 125, no. 1, p. 3, 1999.

[3] S. S. Narayanan, "12 speech in affective computing," *The Oxford Handbook of Affective Computing*, p. 170, 2015.

[4] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 3–9.

[5] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, H. Antonia, S. Amiriparian, A. Baird, R. Georgios, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Proceedings of Interspeech*, 2020.

[6] P. Koval and P. Kuppens, "Changing emotion dynamics: individual differences in the effect of anticipatory social stress on emotional inertia." *Emotion*, vol. 12, no. 2, p. 256, 2012.

[7] G. Deshpande, V. S. Viraraghavan, M. Duggirala, V. R. Reddy, and S. Patel, "Comparing manual and machine annotations of emotions in non-acted speech," *Engineering in Medicine and Biology*, 2018.

[8] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[9] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[10] M. Schmitt, F. Ringeval, and B. W. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech." in *Interspeech*, 2016, pp. 495–499.

[11] G. Deshpande, V. S. Viraraghavan, M. Duggirala, and S. Patel, "Detecting emotional valence using time-domain analysis of speech signals," *Engineering in Medicine and Biology*, 2019.

[12] G. Deshpande, V. S. Viraraghavan, and R. Gavas, "A successive difference feature for detecting emotional valence from speech," in *Proc. SMM19, Workshop on Speech, Music and Mind 2019*, 2019, pp. 36–40.

[13] J. Wagner, D. Schiller, A. Seiderer, and E. André, "Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?" *Proc. Interspeech 2018*, pp. 147–151, 2018.

[14] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[15] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.

[16] C. Montacié and M.-J. Caraty, "Vocalic, lexical and prosodic cues for the interspeech 2018 self-assessed affect challenge," *Proc. Interspeech 2018*, pp. 541–545, 2018.

[17] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny *et al.*, "The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats." in *Interspeech*, 2018, pp. 122–126.

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

[19] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Communication*, vol. 57, pp. 1–12, 2014.

[20] A. Aryani, M. Conrad, D. Schmidtke, and A. Jacobs, "Why'piss' is ruder than'pee'? the role of sound in affective meaning making," *PloS one*, vol. 13, no. 6, 2018.

[21] A. Aryani, C.-T. Hsu, and A. M. Jacobs, "The sound of words evokes affective brain responses," *Brain sciences*, vol. 8, no. 6, p. 94, 2018.

[22] G. Deshpande, V. S. Viraraghavan, M. Duggirala, V. R. Reddy, and S. Patel, "Empirical evaluation of emotion classification accuracy for non-acted speech," in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2017, pp. 1–6.

[23] R. D. Gavas, D. Das, T. Bhattacharjee, M. B. Sheshachala, L. K. Hissaria, R. R. Vempada, V. S. Viraraghavan, A. D. Choudhury, K. Muralidharan, R. K. Ramakrishnan *et al.*, "A sensor-enabled digital trier social stress test in an enterprise context," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 1321–1325.

[24] "Feelings in German," https://www.rocketlanguages.com/german/lessons/feelings-in-german, accessed: 2020-05-15.

[25] "German Emotion Vocabulary," https://study.com/academy/lesson/german-emotion-vocabulary.html, accessed: 2020-05-15.

[26] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[28] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *Journal of fluid mechanics*, vol. 656, pp. 5–28, 2010.