Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech

Conference Paper · August 2021

DOI: 10.21437/Interspeech.2021-2217

3 authors, including:

Aaron Keesing
University of Auckland
3 PUBLICATIONS **34** CITATIONS

SEE PROFILE

INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech

Aaron Keesing, Yun Sing Koh, Michael Witbrock

School of Computer Science, University of Auckland, New Zealand
akee511@aucklanduni.ac.nz, y.koh@auckland.ac.nz, m.witbrock@auckland.ac.nz

Abstract

Many features have been proposed for use in speech emotion recognition, from signal processing features to bag-of-audio-words (BoAW) models to abstract neural representations. Some of these feature types have not been directly compared across a large number of speech corpora to determine performance differences. We propose a full factorial design and to compare speech processing features, BoAW and neural representations on 17 emotional speech datasets. We measure the performance of features in a categorical emotion classification problem for each dataset, using speaker-independent cross-validation with diverse classifiers. Results show statistically significant differences between features and between classifiers, with large effect sizes between features. In particular, standard acoustic feature sets still perform competitively to neural representations, while neural representations have a larger range of performance, and BoAW features lie in the middle. The best and worst neural representations were wav2veq and VGGish, respectively, with wav2vec performing best out of all tested features. These results indicate that standard acoustic feature sets are still very useful baselines for emotional classification, but high quality neural speech representations can be better.

Index Terms: speech emotion recognition, computational paralinguistics, affective computing

1. Introduction

Speech emotion recognition (SER) is the analysis of speech to predict the emotional state of the speaker. Previous research in this field uses various classifiers, features, datasets and methodologies that have been developed. Results are often reported on only one or two datasets, which may or may not be public. Additionally, research often uses different methodologies, such that direct comparability of results is reduced.

To aid in the comparison of different features, we perform thorough testing of several classifiers and features that have been used in previous literature, on 17 public datasets. In particular, we compare acoustic feature sets against bag-of-audio-words (BoAW) features and neural representations. Our methodology uses speaker-independent cross-validation in a full factorial design where each combination of classifier, features and dataset are tested. This is consistent with previous work, although more recently there has been a shift towards predefined train/validation/test splits in larger datasets. Since we test simpler classifiers and use datasets of between 400 and 8000 instances, we use cross-validation to better estimate a classifier's performance on unseen test data. Reproducibility is also important for validating research. As such, we aim to promote reproducible results by detailing our methodology, using public and licensed datasets and providing open-source code. Our code, along with supplementary results, is publicly hosted on GitHub[1] so that our results may be replicated.

There are two main benefits to our work. Firstly, the results from our factorial experiments allow us to compare the effect of classifiers, features and datasets, with a focus on ranking features that tend to be more predictive. Secondly, it serves as a reference for future research that uses the datasets present in this study, so that meaningful comparisons can be made.

The paper is structured as follows. In Section 2 previous comparative experiments are discussed. Section 3 lists the datasets used in this research. The tested classifiers and features are outlined in Section 4, and the corresponding results given in Section 5. These results are discussed in Section 6 and a conclusion and some future work is given in Section 7.

2. Related Work

There has been some previous work in comparing SER techniques on a number of datasets. Schuller et al. [1] compare HMM/GMM and SVM on nine datasets for three classification tasks. The HMM/GMM model used 12 MFCCs and log-frame-energy features, along with speed and acceleration values. For the SVM, 6552 features are extracted based on 39 statistical functionals (mean, min, max, IQR, etc.) of 56 low-level descriptors (LLDs). Testing was carried out in a leave-one-speaker-out (LOSO) or leave-one-speaker-group-out (LOSGO) cross-validation setup. The only three datasets in common with the present study are EMO-DB, eNTERFACE and SmartKom, for which recognition rates of 84.6%, 72.5%, and 23.5% UAR, respectively were achieved, with a polynomial-kernel SVM.

Schuller et al. expand upon this work using stacked restricted Boltzmann machines (RBMs) [2]. The RBMs are pretrained in an unsupervised manner and fine-tuned to maximise Fisher's criterion. They test using the same datasets as in [1] but results are only better on 5 out of 9 datasets. In particular, their model performs slightly better on SmartKom, but slightly worse on EMO-DB and eNTERFACE. In the present work, we adopt a similar testing methodology to these studies but use 17 datasets and 16 feature sets.

Some ensemble methods are tested in [3, 4], including bagged C4.5, boosted C4.5, and a stack of SVM, naïve Bayes, C4.5 and kNN. These are compared against standalone classifiers naïve Bayes, C4.5, kNN and SVM. Multi-layer perceptrons are also compared in [4]. Their results showed that SVM performed better than the ensembles on EMO-DB when fusing acoustic and linguistic features. [5] found that stacking all four base classifiers performed best on a private dataset. Similar findings were reported in [4, 3]. In contrast to these studies we opt to test random forests instead of bagging, boosting or stacking.

Our work complements previous research. We compare

[1] https://github.com/Broad-AI-Lab/emotion

http://dx.doi.org/10.21437/Interspeech.2021-2217

more recent approaches to feature generation, namely BoAW and neural representations, to standard acoustic feature sets. Our methodology is consistent for all tests, and is similar to previous comparative testing methodologies.

3. Datasets

Seventeen open or academic-licensed datasets are used in this study, all of which have a set of categorical labels for primary emotions. For MSP-IMPROV and IEMOCAP, the majority label assigned by annotators is used, consistent with previous work. For JL, we use only primary emotions. For CREMA-D and VENEC we opted to use the acted emotion as ground truth. For VENEC, the few neutral clips are excluded and only speakers that represent all emotions are included. Information about each dataset is in the relevant citation.

Open datasets are under a free and permissive license. The open datasets used in this study are: *CaFE* [6], *CREMA-D* [7], *EMO-DB* [8], *eNTERFACE* dataset [9], *JL corpus* [10], *Portuguese* dataset [11], *RAVDESS* [12], *ShEMO* [13], *TESS* [14], *URDU* [15], and *VENEC* [16].

Academic datasets require signing an EULA in order to gain access. The licensed datasets used in this study are: *DEMoS* [17], *EmoFilm* [18], *IEMOCAP* [19], *MSP-IMPROV* [20], *SAVEE* [21], and *SmartKom-Public* [22].

4. Methodology

We use a full factorial design which tests each combination of classifier, features and dataset using cross-validation, for a total of $9 \times 17 \times 17 = 2601$ results. Figure 1 shows our experimental process. Since our focus is comparing the predictive quality of various feature vectors, we only test a handful of simpler classifiers, namely support vector machine (SVM), small dense neural networks, random forest (RF) and a small convolutional network, all described below. The features we compare are acoustic feature sets, BoAW models, neural network embeddings.

4.1. Classifiers

We test three different classes of classifiers: SVM, RF and neural networks.

For the SVM classifiers we use linear, polynomial and radial basis function (RBF) kernels, which have the following equations, respectively:

$$k_{linear}(\mathbf{x}, \mathbf{y}) = \frac{1}{p}\mathbf{x}^\top \mathbf{y} \tag{1}$$

$$k_{poly}(\mathbf{x}, \mathbf{y}) = (\frac{1}{p}\mathbf{x}^\top \mathbf{y} + 1)^d \tag{2}$$

$$k_{rbf}(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2) \tag{3}$$

where p is the number of features, $\mathbf{x}$ and $\mathbf{y}$ are two instance vectors. We use $d = 2$ and $d = 3$ for the polynomial kernels, and optimise γ using cross-validation.

We implement 1, 2, and 3 hidden layer fully connected feed-forward networks, called multi-layer perceptrons (MLPs), where each hidden layer has 512 nodes with ReLU activation and 50% output dropout. We also implement a small convolutional network proposed by de Pinto et al. [23]. The neural networks were implemented with TensorFlow[2]. SVM and random forests were implemented with scikit-learn[3].

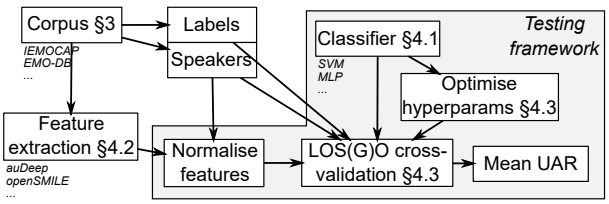

Figure 1: *Our proposed methodology. The gray box constitutes the processes for each experiment.*

4.2. Features

Acoustic feature sets define acoustic low-level descriptors (LLDs) and statistical functionals to apply to an utterance to yield a feature vector. We use the openSMILE toolkit[4] to extract features for the IS09 [24], IS13 [25], GeMAPS and eGeMAPS [26] standard feature sets. We also use a feature set containing 64 temporally averaged mel-frequency cepstral coefficients (MFCCs).

BoAW features represent the distribution of quantised feature vectors from an instance [27]. The openXBOW toolkit[5] is used to generate BoAW features from the first 13 MFCCs and log frame energy. To smooth the frequency distribution, the counts for the a nearest clusters to a vector are incremented in addition to its own cluster. The counts in the BoAW vector are modified using $\log(1 + x)$ and normalised by the number of frames. We chose parameters $(a, n) \in \{(20, 500), (50, 1000), (100, 5000)\}$ that performed well on the EMO-DB dataset, where n is the number of clusters.

Neural representations come from intermediate layer outputs in a neural network. We test features from auDeep [28], DeepSpectrum[6], VGGish and YAMNet [29], and wav2vec [30]. We generate auDeep features from an autoencoder trained on the combined mel-spectrograms from all datasets. We tested DeepSpectrum features and found Densenet features yielded higher accuracy so we report only those. Finally, we test temporally averaged embeddings from pretrained wav2vec, vq-wav2vec, wav2vec 2.0, YAMNet and VGGish models.

4.3. Experiments

Unweighted average recall (UAR) is measured for all-class emotion classification, except in IEMOCAP and MSP-IMPROV where only four classes are used, and in ShEMO where fear is removed due to having few instances. All experiments use LOSO or LOSGO cross-validation and UAR is averaged over folds. If the number of speakers in the corpus is 12 or less, LOSO is used, otherwise the speakers are grouped into 6 groups and LOSGO is used. For IEMOCAP and MSP-IMPROV, each session is used as speaker group, as in previous work. Features are normalised per-speaker, except for URDU where we normalise over the whole corpus due to the extreme speaker imbalance. This normalisation method assumes speaker identities are known in test data and thus might overestimate UAR. "Online" normalisation using only training data may be more realistic in some situations, however we already generate auDeep and BoAW features in an offline manner, and preliminary testing using online normalisation yields similar trends but lower values.

[2] https://www.tensorflow.org/
[3] https://scikit-learn.org/
[4] https://github.com/audeering/opensmile/
[5] https://github.com/openXBOW/openXBOW
[6] https://github.com/DeepSpectrum/DeepSpectrum

Table 1: *Features used in our experiments, with type and number of features. SP = signal processing, NN = neural network.*

Name	Type	#Features
IS09	SP	386
IS13	SP	6373
GeMAPS	SP	62
eGeMAPS	SP	88
Mean MFCC	SP	64
BoAW (20, 500)	SP + BoAW	500
BoAW (50, 1000)	SP + BoAW	1000
BoAW (100, 5000)	SP + BoAW	5000
auDeep	NN	1024
Densenet-201	NN	1920
Densenet-169	NN	1664
Densenet-121	NN	1024
VGGish	NN	128
YAMNet	NN	1024
wav2vec	NN	512
vq-wav2vec	NN	512
wav2vec2	NN	1024

SVM and RF parameters are optimised by selecting the best cross-validated parameters. We optimise the number of RF trees in $\{100, 250, 500\}$ and the maximum depth in $\{10, 20, 50, \infty\}$. For SVMs we optimise the cost parameter C logarithmically over $[2^{-6}, 2^{6}]$ in multiples of 4. For polynomial kernels we optimise r over $\{-1, 0, 1\}$. For RBF kernels we optimise γ logarithmically over $[2^{-12}, 2^{-1}]$ in multiples of 4. We train all neural networks for 50 epochs using Adam optimisation and a learning rate of 10^{-4}.

5. Results

In this section we present summary data to compare corpora, classifiers and features. The mean UAR is reported in all cases; it tends to be preferred in the emotion recognition literature [31] because it ignores class imbalance. We perform Friedman tests with Nemenyi post-hoc tests to determine differences between classifiers and between features. Corpora were used as subjects and UAR was averaged over the omitted variable to satisfy the independent subjects assumption of the Friedman test. The ranks from the Friedman tests are used to order classifiers and features in Figures 3 and 4.

Figure 4 contains the maximum achieved UAR for each corpus-features pair over all classifiers. Taking the maximum instead of the mean makes the values more "spiky" although similar trends were observed using the mean. From the differences in UAR we see that there are different classification 'difficulties' between corpora. VENEC and SmartKom are the most difficult with the lowest UAR while EMO-DB, JL and eNTERFACE are the easiest corpora. Table 2 shows the best classifier-features combination for each dataset. The most common best feature set is wav2veq, although IS09 and GeMAPS are each best for one dataset. SVMs with various kernels tend to be the best classifiers, although random forests are the best for TESS using wav2veq features. Figure 3 shows the mean UAR for each classifier-feature pair, averaged over corpora.

The Friedman test for differences between classifiers detected statistically significant differences ($p < 10^{-14}$) and the Nemenyi critical difference was 2.9 ranks. SVM-RBF was ranked highest (mean rank 1.7), but not with statistical significance. The random forest ranked lower than all other classifiers (mean rank 9.0), but not with statistical significance. Effect sizes with reference to SVM-RBF were small (< 0.4), except for random forest which had a medium effect size of 0.53.

The Friedman test for features was also significant ($p < 10^{-26}$) and the critical difference was 6.0 ranks. The wav2vec features (mean rank 2.1) were ranked highest but not statistically significantly higher than other wav2veq variants or acoustic features. VGGish embeddings were ranked lowest (mean rank 16.1) but not with statistical significance. Figure 2 shows the feature ranking with horizontal bars indicating lack of statistical significance. The four standard acoustic feature sets are interspersed with wav2vec variants at the high end, with BoAW and Densenet features grouped in the middle, and mean MFCCs, VGGish, YAMNet near the bottom. Effect sizes relative to wav2veq are mostly large (> 0.8), very large (> 1.2) or huge (> 2).

Table 2: *Best classifier and feature set combination by UAR per dataset.*

Corpus (abbr.)	Classifier	Features	UAR
CaFE (CA)	SVM-R	wav2vec	76.3
CREMA-D (CR)	MLP-2	wav2vec	75.3
DEMoS (DE)	SVM-C	wav2vec	68.7
EMO-DB (ED)	SVM-L	wav2vec	93.0
EmoFilm (EF)	SVM-R	wav2vec	65.5
eNTERFACE (EN)	MLP-1	wav2vec	84.5
IEMOCAP (IE)	MLP-2	wav2vec	65.6
JL	SVM-R	wav2vec	82.6
MSP-IMPROV (MS)	MLP-3	wav2vec	61.6
Portuguese (PO)	SVM-R	wav2vec	72.6
RAVDESS (RA)	SVM-Q	wav2vec	76.9
SAVEE (SA)	SVM-C	wav2vec	78.5
ShEMO (SH)	DEP	wav2vec	64.0
SmartKom (SM)	SVM-L	IS09	29.3
TESS (TE)	RF	wav2vec	73.0
URDU (UR)	MLP-1	GeMAPS	65.4
VENEC (VE)	SVM-Q	wav2vec	34.4

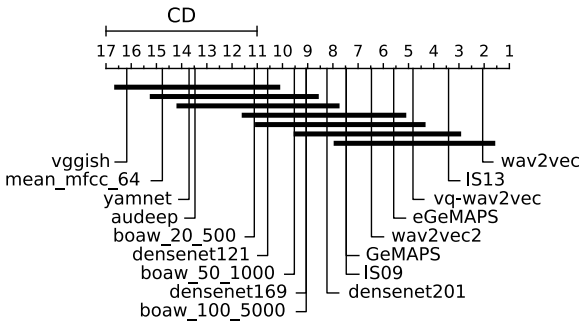

Figure 2: *Critical difference diagram for feature ranks.*

6. Discussion

Our results highlight interesting differences between tested features, but much smaller differences between classifiers. The fact that most of the classifiers perform similarly is expected because the SVM and MLP classifiers are generally useful for

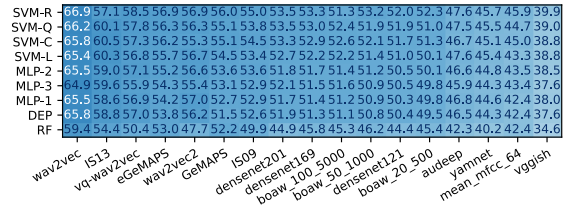

Figure 3: *Mean UAR over corpora for each classifier-feature pair.*

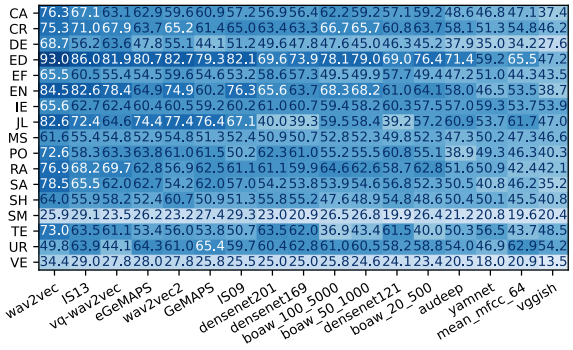

Figure 4: *Best UAR over classifiers for each feature set.*

classification in high-dimensional continuous feature spaces. It is interesting that random forests tended to perform worst on most datasets, perhaps due to the high dimensionality. The de Pinto convolutional network performed well, likely due to having a single convolutional layer which would have little effect, and thus perform similarly to a dense feed-forward network. The training for all of the neural networks was fixed in terms of number of training epochs and learning rate; hyper-parameter tuning, early stopping, learning rate decay may improve performance of the trained models.

The critical difference diagram in Figure 2 suggests that acoustic feature sets still tend to perform competitively to the best neural representations such as wav2vec. However, neural representations are more varied in performance, with seemingly no direct correspondence to number of features. Useful acoustic features tend to comprise a diverse range of LLDs and functionals, as evidenced by the low predictive performance of mean MFCCs. IS13 outperforms all but one tested feature set most likely due to its size; the $\sim$ 6400 features derived from brute forced combinations of LLDs and functionals are simply too numerous not to be useful. However, this comes at the cost of feature redundancy and computation time. Indeed, eGeMAPS only has 88 features and performs only slightly worse than IS13 on most datasets.

The BoAW vectors are not statistically significantly different from one another, although it seems that the more features (clusters) are used the better the predictive performance. The BoAW features are generated by vector quantisation of only 13 MFCCs and log frame energy and yet rank only slightly below acoustic feature sets in general. We additionally tested using BoAW with eGeMAPS frame-wise LLDs but the UAR on all datasets were reduced between 10 to 30% and were not included in this study. An avenue for future research is determining the most useful frame-level features for a BoAW model.

Wav2vec variants seem to be best out of the tested neural representations, with plain wav2vec features best out of all features. The wav2vec models were trained unsupervised on the vast Librispeech 960h training set, and thus learned the most useful and specific speech representations compared to the other models. The Densenet neural representations perform similarly to the BoAW features, indicating the network designed for image recognition generates a useful but non-specific representation of the audio from a colour-mapped spectrogram. The more features Densenet generates, the higher the performance, with Densenet-201 performing best, and similarly to IS09 and GeMAPS features. On the other hand, the auDeep model is trained on spectrograms from all the datasets and yet performs worse than wav2vec and Densenet features. It is possible that other model structures, hyperparameters or spectrogram parameters may improve the usefulness of the auDeep features; we tested using larger spectrogram windows without improvement. VGGish embeddings perform worst in general, possibly because there are only 128 features, although YAMNet features are more numerous but yield second lowest accuracies from the neural representations. Another reason YAMNet and VGGish embeddings perform slightly worse than auDeep is that they were trained on all types of audio data, not just speech data, hence some of the features from the embedding may be suited to other types of audio and be less useful for representing speech. In general, the features from the neural embeddings are less speech-specific than the acoustic feature sets as the mappings are trained from backpropagation though a bottleneck layer deeper in the network.

Corpus-specific differences occur for a number of reasons. Some are due to different number of labels for classification, as in VENEC's low scores, or the annotation method (acted vs. majority vote). Another reason is naturalness of emotional expression, where natural expression is more subtle and harder to discriminate. An avenue for future research is quantifying the effect of naturalness, speaker variability, annotation methods and even emotion categories, on classification accuracy. While our metric was UAR for all-class classification, we expect similar trends for binary arousal/valence tasks and even regression tasks. There also appear to be corpus-specific oddities with some features: wav2vec performs sub-average on the URDU dataset, Densenet features perform sub-average on JL, BoAW features perform sub-average on TESS. Acoustic feature sets do not seem to have this issue, and are more consistent in this regard.

7. Conclusion

This research tests 9 classifiers and 17 features in a full factorial design, using speaker-independent cross-validation to measure categorical emotion classification performance on 17 datasets. We compared classical acoustic feature sets, BoAW models and neural embeddings, but also tested a diverse range of classifiers. The results indicate that the non-quantised wav2vec representations are most predictive, and are thus a useful baseline for future research. However, standard acoustic feature sets still have competitive performance and are more consistent across datasets. Further research may investigate the causes of large performance differences between representations.

8. Acknowledgements

The authors would like to thank the University of Auckland for providing computational resources to complete this research. This research was funded through a University of Auckland doctoral scholarship.

9. References

[1] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, Nov. 2009, pp. 552–557.

[2] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 5688–5691.

[3] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker Independent Speech Emotion Recognition by Ensemble Classification," in *2005 IEEE International Conference on Multimedia and Expo*. Amsterdam, The Netherlands: IEEE, 2005, pp. 864–867.

[4] B. Schuller, R. Villar, G. Rigoll, and M. Lang, "Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1, Mar. 2005, pp. I/325–I/328 Vol. 1.

[5] B. Schuller, M. Lang, and G. Rigoll, "Robust Acoustic Speech Emotion Recognition by Ensembles of Classifiers," in *Proceedings of the DAGA'05, 31, Deutsche Jahrestagung Für Akustik, DEGA*, München, Jan. 2005, pp. 329–330.

[6] P. Gournay, O. Lahaie, and R. Lefebvre, "A canadian french emotional speech dataset," in *Proceedings of the 9th ACM Multimedia Systems Conference*. ACM, 2018, pp. 399–402.

[7] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, Oct. 2014.

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech 2005 - Eurospeech*, Lisbon, Portugal, Sep. 2005, pp. 1517–1520.

[9] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE' 05 Audio-Visual Emotion Database," in *22nd International Conference on Data Engineering Workshops*, Apr. 2006, pp. 8–8.

[10] J. James, L. Tian, and C. I. Watson, "An Open Source Emotional Speech Corpus for Human Robot Interaction Applications," in *Interspeech*, 2018, pp. 2768–2772.

[11] S. L. Castro and C. F. Lima, "Recognizing emotions in spoken language: A validated set of Portuguese sentences and pseudosentences for research on emotional prosody," *Behavior Research Methods*, vol. 42, no. 1, pp. 74–81, Feb. 2010.

[12] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, p. e0196391, May 2018.

[13] O. Mohamad Nezami, P. Jamshid Lou, and M. Karami, "ShEMO: A large-scale validated database for Persian speech emotion detection," *Language Resources and Evaluation*, vol. 53, no. 1, pp. 1–16, Mar. 2019.

[14] K. Dupuis and M. K. Pichora-Fuller, "Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set," *Canadian Acoustics*, vol. 39, no. 3, pp. 182–183, Sep. 2011.

[15] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross Lingual Speech Emotion Recognition: Urdu vs. Western Languages," in *2018 International Conference on Frontiers of Information Technology (FIT)*, Dec. 2018, pp. 88–93.

[16] A. S. Cowen, P. Laukka, H. A. Elfenbein, R. Liu, and D. Keltner, "The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures," *Nature Human Behaviour*, vol. 3, no. 4, pp. 369–382, Apr. 2019.

[17] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. W. Schuller, "DEMoS: An Italian emotional speech corpus," *Language Resources and Evaluation*, Feb. 2019.

[18] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird, and B. Schuller, "Categorical vs Dimensional Perception of Italian Emotional Speech," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 3638–3642.

[19] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, Nov. 2008.

[20] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, Jan. 2017.

[21] S. Haq, P. J. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification," in *International Conference on Auditory-Visual Speech Processing 2008*, Tangalooma Wild Dolphin Resort, Moreton Island, Queensland, Australia, Sep. 2008, pp. 185–190.

[22] F. Schiel, S. Steininger, and U. Türk, "The SmartKom Multimodal Corpus at BAS," in *Third International Conference on Language Resources and Evaluation*. Las Palmas, Canary Islands, Spain: Citeseer, 29, pp. 200–206.

[23] M. G. de Pinto, M. Polignano, P. Lops, and G. Semeraro, "Emotions Understanding Model from Spoken Language using Deep Neural Networks and Mel-Frequency Cepstral Coefficients," in *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, May 2020, pp. 1–5.

[24] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *10th Annual Conference of the International Speech Communication Association*, Brighton, United Kingdom, Sep. 2009, pp. 312–315.

[25] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, and E. Marchi, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *14th Annual Conference of the International Speech Communication Association*, Lyon, France, Aug. 2013, pp. 148–152.

[26] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, Apr. 2016.

[27] S. Pancoast and M. Akbacak, "Softening quantization in bag-of-audio-words," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 1370–1374.

[28] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6340–6344, 2017.

[29] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN Architectures for Large-Scale Audio Classification," *arXiv:1609.09430 [cs, stat]*, Jan. 2017.

[30] A. Baevski, S. Schneider, and M. Auli, "Vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations," *arXiv:1910.05453 [cs]*, Feb. 2020.

[31] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, Mar. 2011.