Emotional Speech Classifier Systems: For Sensitive Assistance to support Disabled Individuals

Vishnu Vidyadhara Raju V, Priyam Jain, Krishna Gurugubelli, and Anil Kumar Vuppala

Speech Processing Laboratory, LTRC, KCIS,
International Institute of Information Technology, Hyderabad, India
{vishnu.raju, priyam.jain, krishna.gurugubelli}@research.iiit.ac.in,
anil.vuppala@iiit.ac.in

Abstract

This paper provides the classification of emotionally annotated speech of mentally impaired people. The main problem encountered in the classification task is the class-imbalance. This imbalance is due to the availability of large number of speech samples for the neutral speech compared to other emotional speech. Different sampling methodologies are explored at the back-end to handle this class-imbalance problem. Mel-frequency cepstral coefficients (MFCCs) features are considered at the front-end, deep neural networks (DNNs) and gradient boosted decision trees (GBDT) are investigated at the back-end as classifiers. The experimental results obtained from the EmotAsS dataset have shown higher classification accuracy and Unweighted Average Recall (UAR) scores over the baseline system.

Index Terms: MFCCs, GBDT, DNNs.

1. Introduction

Emotions present in the human speech reflect the perception of various things related to the psychology of the human body and the mental health. Many studies [1, 2] have provided proper and strong evidences showing the inter-relation between the factors of emotion and health [3, 4]. Researchers have identified these emotion cognitions have an effect on the human health and there is a need to handle these emotional disorders [5]. Hence there is a need to perform the analysis and classify these emotions.

Emotion recognition from the input speech signal involves three stages, namely signal processing, feature selection and the classification [6]. The main difficulty of emotion classification lies in the identification of the features which work well in different emotional conditions. The selection of the features is purely dependent on the datasets. MFCC [7, 8] features evoked a big impact as a representation for the frame-level analysis of short-term features for the better classification task at the front-end when compared to linear predictive cepstral coefficients (LPCC) [9] and perceptual linear predictive (PLP) [10] features. With the advancement of DNNs in the recent years, they have been vastly used to model the emotions for the emotion recognition task [11, 12] at the back-end. Dynamic models such as hidden markov models (HMMs) are used for the frame-level dynamic spectral features of MFCC at the back end [13]. For the supra-segmental prosodic features [14] which are estimated for the entire utterance, global models such as gaussian mixture models (GMMs), DNNs, support vector machine (SVMs) are used [15]. DNNs capture and model the data at the linear or nonlinear manifold effectively. Training of these DNNs is initialized by a pre-training algorithm.

The main aim of the Atypical affect sub-challenge dataset is to propose a methodology for the classification of the four basic emotions and to develop a possible emotional speech-driven application to support disabled individuals. Spectral subtraction is performed prior to the extraction of MFCC features, which are used at the front-end and DNN models are used for the emotion classification at the back-end. The performance of DNN classifiers are compared with GBDT [16].

This paper is organized as follows. The description of the emotion speech dataset and the proposed approach which includes front-end features and back-end classifiers is explained in detail in Section 2 and 3 respectively. The experimental setup with detailed feature representation and classifier parameters are presented in Section 4. Section 5 discusses about the experimental results. Finally, conclusion of study is presented in Section 6.

2. Description of the Emotion Speech Dataset

In this study, the experiments are carried out on EmotAsS dataset which is provided as a part of Interspeech 2018 Computational Paralinguistics: The Atypical Affect Sub-Challenge. The EmotAsS dataset [17], has the recording from 15 different speakers having physical, mental and neurological disabilities. These speech samples were recorded in a familiar room at their workplace while speaking about their personal and health issues. These speakers comprises of about 8 female and 7 male speakers.

Table 1: *Number of instances per class in the train/development/test cases; Test case distributions are blinded.*

Emotion-sensitive Assistance Systems (EmotAsS)			
	Train	Dev	Test
Angry	125	50	blinded
Happy	743	965	blinded
Neutral	2287	2842	blinded
Sad	187	329	blinded
Total	3342	4186	3099

The age group of these speakers were between 20 to 58 years.From these recordings of 15 speakers, 12 speakers were mentally and 2 speakers were neurologically disabled. 1 speaker had multiple disability. These recordings were done using a Zoom-H6 voice recorder and a Jabra speak 510 microphone at 44.1 kHz 24-bit mono mode. The speech collected was for a duration of 9.2 hours with 10,627 segmented chunks. The annotations were performed using 12 volunteers using gamified

10.21437/SMM.2018-2

crowdsourcing platform iHEARu-PLAY [18]. Four basic emotions of anger, happiness, sadness, along with neutral speech were considered. The details of the train, development and test distributions for the four basic emotions is shown in Table 1.

3. Proposed Approach

In this section, the components used in the proposed system specifically front-end features and back-end classifiers are explained in detail. The proposed approach is shown in detail in Figure 1 and the procedure involved in extracting the front-end features and back-end classifiers is explained in detail in the next subsections.

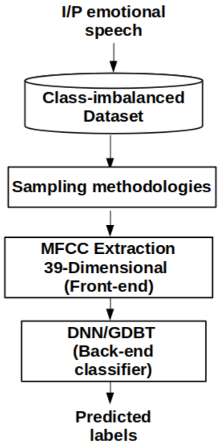

Figure 1: *Proposed Approach*

3.1. Sampling Methodologies

Imbalanced data refers to a classification problem where the classes are not represented equally. In the machine learning context this problem arises where the total number of a class of data (positive) is far less than the total number of another class of data (negative). Majority of the machine learning algorithms works better when the number of instances of each classes are roughly equal. Whenever the number of instances of one class far exceeds the other, this class-imbalance problem arises.

Re-sampling techniques such as random under-sampling and over-sampling techniques are the data level approaches used to handle the class-imbalance problem. Dealing with imbalanced datasets entails strategies such as improving classification algorithms or balancing classes in the training data (data preprocessing) before providing the data as input to the machine learning algorithm. The main objective of balancing classes is to either increasing the frequency of the minority class or decreasing the frequency of the majority class. This is done in order to obtain approximately the same number of instances for both the classes. The main aim of random under-sampling is to balance class distribution by randomly eliminating majority class examples. This is done until the majority and minority class instances are balanced out whereas over-Sampling increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority class in the sample. In our approach under-sampling is preferred to handle the class-imbalance for the EmotAsS dataset.

3.2. Front-end Features: MFCC

39-Dimensional MFCC features are extracted from the input speech. Frame size of 20 ms with 10 ms shift and 24 triangular filter-banks are considered to compute mel-frequency coefficients. For each frame the first 13 coefficients are computed first and later the second order derivatives is calculated, therefore finally ending up with 39 dimension feature vector for each 20 ms frame. By taking derivates, the dynamic nature of speech is incorporated into the feature vector.

3.3. Back-end Classifiers: DNN

DNNs are a powerful tool when it comes to modeling complex and high dimensional non linear relationships because of their structure that involves multiple hidden layers and non linear activations. Deep Neural Network architecture is shown in Figure 2. It is seen that every layer is fully connected to the layer before and after it. Each circle in that figure represents a neuron. A Neuron in l^{th} layer will be computed as,

$$\eta_j^l = f(y_j^l) = max(0, y_j^l), \tag{1}$$

$$and y_j^l = \Sigma_{i=1}^{n_{l-1}} W_{ij}^{l-1} \eta_i^{l-1}, \tag{2}$$

where f is a Relu activation function,η_j^l is the j^{th} neuron in the l^{th} neuron andW_{ij}^l is weight corresponding to j^{th} neuron in $l+1^{th}$ layer and i^{th} neuron in l^{th} layer.

DNNs are trained for a particular task using back-propagation algorithm, which involves updating weights of the network in order to reduce a pre-defined cost function. Weights are updated by the backward flow of derivative of the cost function, which in turn can be done in various ways. This is termed as optimizing the network. Adam, an algorithm for first order gradient based optimization is preferred in this paper. The choice of cost function, number of hidden layers, number of neurons in each hidden layer, activation function and the optimizer depends upon the classification task and the dataset.

3.4. Back-end Classifiers: GBDT

Gradient boosting is used to convert weak learners to strong learners by following an iterative mechanism. In this technique, at each iterative step, the classifier stresses more on the misclassified samples from the previous step. Hence the convergence is fast in gradient boosted classifiers. In this paper Gradient boosting with decision trees is used for the classification task. GBDTs like other supervised learning algorithms, employs use of a loss function and minimize it. Any loss function can be preferred until it is differentiable. The logistic regression function is used for classification task.

$$J(\theta) = \frac{-1}{m} \Sigma_{i=1}^{m} [y^i log(h_\theta(x^i))] + (1 - y^i) log(1 - h_\theta(x^i))] \tag{3}$$

Where y^i is the class label, $h_\theta x^i$ is the output probability(or score) of the classifier. This loss function is a binary loss function. In each iteration step, we construct number of trees equal to the number of classes. Where each tree is corresponding to a class, it is like the one-vs-rest approach at each iteration step. The number of iteration steps is a hyper-parameter which needs to be set.

4. Experimental Setup

4.1. Sampling the dataset

The dataset provided can be treated as a case of class imbalance problem. Class imbalance is when one(or few) of the classes have a lot more samples than rest of the classes. In our case, neutral and happy are the emotions which have most number of samples as shown in Table 1. So, in order for our classifiers to not be biased towards these large sample classes, we have tested sampling techniques like random over-sampling, random under-sampling, Cluster-Centroid under-sampling. Python library imbalanced-learn [19] was used to implement various sampling techniques. Random under-sampling technique is preferred as it has yielded better UAR and accuracy scores compared to other sampling techniques.

4.2. Classifiers: DNN

Keras [20] library was used to implement the DNN architecture. In the DNN architecture 7 hidden layers were used with number of neurons being in the power of 2 at each layer and Relu activation was followed by Dropout of 0.2. Mini-batch training approach is preferred with batch size of 4000 for under sampled dataset and 10000 for over-sampled dataset. Each one of those 4000/10000 was a 39 dimensional vector. Number of epochs was set at 50. Adam was chosen as the optimizer with the following parameters with a learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Mean squared error was chosen as the loss function to be optimized.

4.3. Classifiers: GBDT

In GBDT, the logistic regression loss function which is also called deviance is used. The learning rate is set at 0.1 and the number of estimators for performing the number of boosting stages was set to 100. So, the total number of trees constructed is equal to 400 i.e.(number of estimators*number of classes). The Maximum Depth of each individual estimator tree was kept at 3. Unlimited number of leaf nodes were considered to grown in a tree. The minimum number of samples required at a node for splitting is set at 2 and at last this split has been allowed for resulting in the improvement of the performance. This classifier was implemented using scikit-learn python library [21].

5. Results and Discussion

As per the Atypical Affect Sub-Challenge, the base-line results were reported in terms of accuracy and UAR scores. These accuracy and UAR scores are computed using scikit-learn python library [21] . The experimental results on development and evaluation datasets are discussed in detail in the following section.

5.1. Results on Development Data

Table 2: *UAR and accuracy(%) for the DNN and GBDT classifiers on the development data for MFCC features*

Random Under-Sampling		
	UAR(%)	Accuracy (%)
DNN	36.6	37.7
GBDT	36.7	23.2

In order to handle the class-imbalance problem, experimentation is done on four different sampling techniques. From these four sampling techniques random under-sampling is preferred as it randomly eliminates the majority class examples until balance of the classes is observed for the large dataset. The UAR and accuracy scores for the proposed system which uses MFCCs as the front-end features, DNNs and GBDT as the back-end classifiers for the under-sampled dataset is shown in Table 2. It is shown that the better UAR score is observed for the case of GBDTs and accuracy score is better for the case of DNNs.

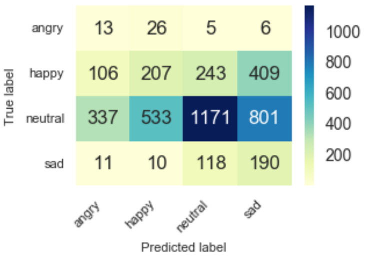

Figure 2: *Heat map(confusion matrix) on development set for DNN classifier*

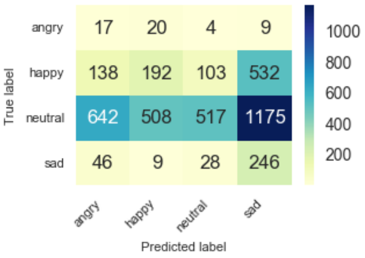

Figure 3: *Heat map(confusion matrix) on development set for GBDT classifier*

The scores of DNN and GBDT were considered in separate and the heat maps or the confusion matrix of both the classifiers is shown in Figure 3 and 4 respectively. From the confusion matrix or the heat maps of the GBDT classifier shown in Figure 4. It is evident that misclassification is more for the case of GBDT. Misclassification for the case of neutral as sad emotion was the highest in Figure 4. When the comparison of the GBDT with DNN heat maps or confusion matrices is done, it has been observed that the misclassification is less for DNN classifiers.

5.2. Results on Test Data

The experimentation is done on the test data and their results are reported in Table 3. From the table it is evident that random under-sampling outperformed random over-sampling for the provided data regardless of the classifier used. The performance of DNNs and GBDT are somewhat similar in terms of UAR but if we bring accuracy in consideration, then DNN performs better.

5.3. Comparison with Baseline

This emotion classification system results are compared with that of Atypical affect sub-challenge baseline which is achieved

Table 3: *UAR and accuracy(%) for the DNN and GBDT classifiers on the development data for MFCC features*

Random Under-Sampling		
	UAR(%)	Accuracy (%)
DNN	36.42	51.27
GBDT	31.04	25.08

Table 4: *Comparison with baseline system*

UAR(%)	Dev	Test
Confidence based Fusion of END2YOU, OPENSMILE, OPENXBOW and AUDEEP features	37.8	43.4
MFCC+DNN	36.6	36.42
MFCC+GBDT	36.7	31.04

by confidence-based fusion of four different classifiers whose results are reported in Table 4 below. It is observed that the classification system has performed better than the baseline system. The feature set for the base-line is formed from the fusion of END2YOU, OPENSMILE,OPENXBOW and AUDEEP features. The size of the resultant fused static feature set is 6,373 dimensional one formed from the computaton of various functionals over low-level descriptor contours. The baseline system uses a RNN-CTC classifier at the back-end. From the results obtained from the proposed approach uses a 39-dimensional MFCC features extracted at the front-end with a DNN classifier, which is a less-dimensional feature set. The proposed approach shows a performance near to the baseline with the smaller dimension feature set which is simpler to compute rather than the fusion of the complex baseline system. DNN classifiers have provided slightly better performance over the GBDTs. The performance of the two classifiers on the development set is equivalent and some degradation is observed on the test set for the case of GBDTs, which can be due to slight overfitting.

6. Summary and conclusions

This study presents the classification of emotionally annotated speech of mentally impaired people. The class-imbalance present in the dataset is handled by random under-sampling approach. 39-dimensional feature set is extracted at the front-end by considering DNN and GBDT classifiers at the back-end where the DNN classifier has yielded slightly better performance over GBDTs. Though DNN with 39-dimensional feature set could not beat the performance of fused scores of higher dimensional feature set, it has given an performance equivalent to the baseline. This proposed approach is simple to implement when compared to the complex fusion approach of the base-line system. As the results on development set are relatively convincing than the results on evaluation set, further investigations are needed for generalization of provided methodologies. The scope of GBDTs can be explored to further extents in speech applications.

7. Acknowledgements

The first author would like to thank Tata Consultancy Services (TCS), India for supporting the PhD program.

8. References

[1] L. E. Francis, "Emotions and health," in *Handbook of the sociology of emotions*. Springer, 2006, pp. 591–610.

[2] R. Pandey and A. K. Choubey, "Emotion and health: An overview." *SIS Journal of Projective Psychology & Mental Health*, vol. 17, no. 2, 2010.

[3] S. D. Pressman and S. Cohen, "Does positive affect influence health?" *Psychological bulletin*, vol. 131, no. 6, p. 925, 2005.

[4] J. F. Finch, L. E. Baranik, Y. Liu, and S. G. West, "Physical health, positive and negative affect, and personality: A longitudinal analysis," *Journal of Research in Personality*, vol. 46, no. 5, pp. 537–545, 2012.

[5] J. Huang, X. Xu, and T. Zhang, "Emotion classification using deep neural networks and emotional patches," in *International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017, pp. 958–962.

[6] S. Majuran and A. Ramanan, "A feature-driven hierarchical classification approach to emotions in speeches using svms," in *International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2017, pp. 1–5.

[7] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient mfcc extraction method in speech recognition," in *Proc. of International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2006, pp. 4–pp.

[8] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[9] Y. Yujin, Z. Peihua, and Z. Qun, "Research of speaker recognition based on combination of lpcc and mfcc," in *International Conference on Intelligent Computing and Intelligent Systems (ICIS)*, vol. 3. IEEE, 2010, pp. 765–767.

[10] F. Grezl and P. Fousek, "Optimizing bottle-neck features for lvcsr," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4729–4732.

[11] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 3687–3691.

[12] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[13] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[14] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. of Tenth Annual Conference of the International Speech Communication Association (ISCA)*, 2009.

[15] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[16] Y. Xia, C. Liu, Y. Li, and N. Liu, "A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring," *Expert Systems with Applications*, vol. 78, pp. 225–241, 2017.

[17] S. Hantke, H. Sagha, N. Cummins, and B. Schuller, "Emotional speech of mentally and physically disabled individuals: Introducing the emotass database and first findings," *Proc. of Interspeech*, pp. 3137–3141, 2017.

[18] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "ihearu-play: Introducing a game for crowdsourced data collection for affective computing," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 891–897.

[19] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017. [Online]. Available: http://jmlr.org/papers/v18/16-365.html

[20] F. Chollet *et al.*, "Keras," https://github.com/keras-team/keras, 2015.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.