HLT-NUS DiCOVA 2021 Challenge System Report

Rohan Kumar Das, Maulik Madhavi and Haizhou Li

Department of Electrical and Computer Engineering
National University of Singapore
{rohankd, maulik.madhavi, haizhou.li}@nus.edu.sg

Abstract

This report presents the submission made for DiCOVA 2021 challenge from Human Language Technology (HLT) Laboratory, National University of Singapore (NUS). The challenge focuses on two tracks that aims to detect COVID-19 using voice. We participate the Track 1 of the challenge, which deals with detection using cough sounds from individuals. In this challenge, we use a few novel acoustic cues based on long-term transform, gammatone filterbank and equivalent rectangular bandwidth spectrum. We evaluate these representations using logistic regression, random forest and multilayer perceptron classifiers for detection of COVID-19. On the blinded test set, we obtain an area under the ROC curve (AUC) of 83.49% for the best system submitted to the challenge.

Index Terms: COVID-19, acoustics, cough, Gammatone, Constant Q, equivalent rectangular bandwidth

1. System Description

In this work, we focus on novel acoustic front-ends for detection of COVID-19 given the fact that the data for the challenge is very limited. We consider long-term transform, gammatone filterbank and equivalent rectangular bandwidth spectrum based acoustic cues for capturing discriminative signal characteristics for detection of COVID-19.

1.1. Methodology Overview

We follow the baseline recipe given by the organizers for training the model for 5-fold cross validation. For testing, we use the entire training data to build the model and test scores are computed using this model.

1.2. Feature Description

In our system, we use the following acoustic features:

1. **Mel frequency cepstral coefficients (MFCC):**
   We consider the most widely used mel frequency cepstral coefficient (MFCC) features [1], which is also used as for challenge baseline given from the organizers. The MFCC extraction process remains the same as that of the challenge baseline, to obtain 39-dimensional features after energy based voice activity detection (VAD).
2. **Constant-Q transform (CQT) spectrum**:

   The constant-Q transform (CQT) [2] is a long term window transform, which has been provided to be effective for various classification tasks previously [3, 4]. We consider the log power spectrum of CQT as one of the front-ends in these research studies. The extraction process follows our previous work given in [5]. We use `LibROSA` toolkit in python to extract CQT features from the speech signal.
3. **Gammatone cepstral coefficients (GTCC)**:

   Literature shows that the impulse response associated with basilar membrane vibrations are highly correlated with Gammatone signals [6, 7]. In addition, Gammatone filterbanks have been widely used for many sound classification research problems [8]. The GTCC features are extracted along with dynamic features delta and delta-delta using `audioFeatureExtractor` of MATLAB audio toolbox.
4. **Equivalent rectangle bandwidth (ERB) spectrum:**

   The equivalent rectangular bandwidth (ERB) frequency scale is a psychoacoustic measure of the auditory filter width on different location of cochlea [9]. The ERB frequency scale is used in the Gammatone filterbank design. The ERB spectrum (erbSpec) is extracted with 43 number of bands using `audioFeatureExtractor` of MATLAB audio toolbox. To expand the dynamic range of feature vectors, we applied logarithm to spectral features. Furhter for feature computation, the frames containing very smaller values are expected to be the part of silence and hence removed.

1.3. Classifier Description

In this work, we used the three classifiers, namely, logistic regression (LR), Random Forest (RF) and Multilayer Perceptron (MLP). They are also considered by the organizers for the challenge baseline [10]. Therefore, we keep the parameters related to each of them same following that of the baseline.

1.4. Results

Table 1 shows the performance with different combinations of acoustic features and classifier that we explored in this study. The performance of the system is evaluated using area under the ROC curve (AUC), which is the official performance metric of the challenge. Our experimental results show that RF and MLP classifiers perform relatively better than LR for most of the front-ends. It is noted that as the test set results are available for only a limited number of submissions on the leaderboard for each team, we could not report results on the test set for a few systems. Therefore, we submitted the systems using RF and MLP classifiers for our investigated features as they performed better than LR classifier in most of the cases on the validation set. The erbSpec feature is relatively better acoustic feature among other features on both validation and test sets. In particular, erbSpec with RF performs the best among the different single systems as shown in Table 1. We find GTCC feature

Table 1: *Performance of various systems submitted to DiCOVA 2021 challenge.*

System (Feature-Classifier)	Validation AUC (%)	Test AUC (%)
MFCC-LR	64.04	60.83
MFCC-RF	67.71	66.77
MFCC-MLP	69.36	66.09
CQT-LR	65.48	-
CQT-RF	71.95	68.85
CQT-MLP	68.71	71.79
GTCC-LR	64.84	-
GTCC-RF	68.65	72.68
GTCC-MLP	68.61	78.61
erbSpec-LR	67.54	-
erbSpec-RF	**73.41**	**81.89**
erbSpec-MLP	68.17	65.38

Table 2: *Performance for score-level fusion of erbSpec-RF and GTCC-MLP systems submitted to DiCOVA 2021 challenge.*

Weight (α)	Validation AUC (%)	Test AUC (%)
0.5	69.98	82.42
0.9	71.95	**83.49**

with MLP classifier performs as the second best system. It is worth to be noted that both these single systems outperforms the challenge baselines with MFCC features by a large margin as can be viewed from Table 1.

We also perform some analysis for score-level fusion using the two best single systems described above. A weighted score level fusion is adopted for this study, where we fuse the scores obtained from erbSpec-RF and GTCC-MLP as follows:

$$S_{\text{fusion}} = \alpha S_{\text{erbSpec-RF}} + (1-\alpha) S_{\text{GTCC-MLP}} \quad (1)$$

where α is the weighted ratio and $S_{\text{erbSpec-RF}}$, $S_{\text{GTCC-MLP}}$ and S_{fusion} represent the scores from erbSpec-RF, GTCC-MLP systems and the final fused score.

Table 2 shows the analysis for the fusion studies carried out with the two best single systems. We consider two weighted combinations for this study. For the first one, we consider equal weights for the both systems and for the second one, we put a higher weight for erbSpec-RF system as it performed the best for both validation as well as test set among all our single systems. From Table 2 we observe that fusion of the two systems improve the performance for both cases, which is more significant when a higher weightage is given for erbSpec-RF system. Thus, our best system submitted to the challenge achieves an AUC of 83.49% on the test set, which is comparable to the top systems submitted to the challenge.

2. References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[2] J. C. Brown, "Calculation of a constant Q spectral transform," *Journal of Acoustical Society of America*, vol. 89, pp. 425–434, 1991.

[3] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[4] R. K. Das and H. Li, "Instantaneous phase and long-term acoustic cues for orca activity detection," in *Interspeech 2019*, 2019, pp. 2418–2422.

[5] R. K. Das, J. Yang, and H. Li, "Data augmentation with signal companding for detection of logical access attacks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2021*, 2021.

[6] X. Valero and F. Alías, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Trans. Multim.*, vol. 14, no. 6, pp. 1684–1689, 2012.

[7] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2009*, 2009, pp. 4625–4628.

[8] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1d convolutional neural network," *Expert Systems with Applications*, vol. 136, pp. 252–263, 2019.

[9] R. D. Patterson and J. Holdsworth, "A functional model of neural activity patterns and auditory images," *Advances in speech, hearing and language processing*, vol. 3, no. Part B, pp. 547–563, 1996.

[10] A. Muguli, L. Pinto, N. R., N. Sharma, P. Krishnan, P. K. Ghosh, R. Kumar, S. Ramoji, S. Bhat, S. R. Chetupalli, S. Ganapathy, and V. Nanda, "DiCOVA Challenge: Dataset, task, and baseline system for covid-19 diagnosis using acoustics," 2021. [Online]. Available: https://arxiv.org/abs/2103.09148