Proceedings of Meetings on Acoustics

DECEMBER 08 2020

Automatic detection of instances of focused crowd involvement at recreational events FREE

Eric Todd; Mylan R. Cook; Katrina Pedersen; David S. Woolworth; Brooks A. Butler; Xin Zhao; Colt Liu; Kent L. Gee ; Mark K. Transtrum; Sean Warnick

Check for updates

Proc. Mtgs. Acoust. 39, 040003 (2019)
https://doi.org/10.1121/2.0001327

View Online Export Citation

Articles You May Be Interested In

Noise dynamics in city nightlife: Assessing impact and potential solutions for residential proximity to pubs and bars

Proc. Mtgs. Acoust. (April 2024)

Data-driven decomposition of crowd noise from indoor sporting events

J. Acoust. Soc. Am. (February 2024)

Characterizing natural soundscapes and understanding human response to human-caused noise in a Hong Kong country park

Proc. Mtgs. Acoust. (December 2007)

15 October 2024 14:10:43

ASA ACOUSTICAL SOCIETY OF AMERICA

Volume 39 http://acousticalsociety.org/

178th Meeting of the Acoustical Society of America

San Diego, CA

2-6 December 2019

Noise: Paper 2aNSb7

Automatic detection of instances of focused crowd involvement at recreational events

Eric Todd
Brigham Young University, Provo, UT, 84602, USA; eric.w.todd@gmail.com

Mylan R. Cook
Department of Physics and Astronomy, Brigham Young University, Provo, Utah, 84602, USA; mylan.cook@gmail.com

Katrina Pedersen
Brigham Young University, Provo, UT, USA; katrina.pedersen@gmail.com

David S. Woolworth
Roland, Woolworth, and Associates, Oxford, MS, USA; dwoolworth@rwaconsultants.net

Brooks A. Butler, Xin Zhao and Colt Liu
Brigham Young University, Provo, UT, USA; brooks.butler93@gmail.com; xinzhao0822@gmail.com; colt.liu7@gmail.com

Kent L. Gee and Mark K. Transtrum
Department of Physics and Astronomy, Provo, UT, 84602, USA; kentgee@byu.edu; mkt24@byu.edu

Sean Warnick
Brigham Young University, Provo, UT, USA; seanwarnick@gmail.com

This paper describes the development of an automated classification algorithm for detecting instances of focused crowd involvement present in crowd cheering. The purpose of this classification system is for situations where crowds are to be rewarded for not just the loudness of cheering, but for a concentrated effort, such as in Mardi Gras parades to attract bead throws or during critical moments in sports matches. It is therefore essential to separate non-crowd noise, general crowd noise, and focused crowd cheering efforts from one another. The importance of various features—both spectral and low-level audio processing features—are investigated. Data from both parades and sporting events are used for comparison of noise from different venues. This research builds upon previous clustering analyses of crowd noise from collegiate basketball games, using hierarchical clustering as an unsupervised machine learning approach to identify low-level features related to focused crowd involvement. For Mardi Gras crowd data we use a continuous thresholding approach based on these key low-level features as a method of identifying instances where the crowd is particularly active and engaged.

Published by the Acoustical Society of America

© 2020 Acoustical Society of America. https://doi.org/10.1121/2.0001327

15 October 2024 14:10:43

1. INTRODUCTION

Audio processing of acoustic data is a rich research field. With the advent of machine learning (ML), audio datasets can be processed much more rapidly, and machines can find connections that are not always intuitive to human observers, but which can be useful in various ways, from classifying music to generating sports highlights.[1-5] Unsupervised learning is particularly useful because it requires minimal user input, making it a rapid, low-cost method of identifying patterns in data.[6-8] While these methods can be applied to various settings, we focus our attention on crowd noise audio data.

Processing acoustic crowd noise is a rather challenging endeavor, as there can be any combination of noise produced by any number of individuals, who are not always in agreement with one another. While a human listener can distinguish between a cheering crowd and a booing crowd with relative ease, developing a ML algorithm to separate these types of events is challenging.

Our previous work focused on training a ML algorithm to distinguish spectral data taken at sporting events—primarily college basketball—into instances of active crowd noise vs. non-crowd noise, e.g., noise produced by the crowd versus noise produced by the public announcement (PA) system or band.[9] We used the K-means clustering algorithm in conjunction with a jump analysis to identify six main clusters for the basketball data when clustering on spectral data calculated over short time increments.[9] We assigned each cluster a distinct color for ease of visualization. By listening to data from each cluster, we also assigned labels to each cluster that represented a type of crowd engagement or other sounds heard, as well as identified types of events present in the cluster. Table 1 presents a summary of these results with cluster colors, labels, and descriptions.

Table 1. Assigned basketball cluster colors, descriptions, and examples of typical events.

Cluster Color	Descriptive Cluster Label	Examples of events
Green	Null	Background noise, national anthem, silence
Blue	Murmur	Crowd neither loud nor quiet
Red	Moderate Involvement	Applause, cheering
Magenta	High Involvement	Loud cheering, screaming, booing
Cyan	Softer Music	Quieter music (band/PA), minimal crowd noise
Yellow	Louder Music	Louder music (band/PA), minimal crowd noise

Using the cluster centroids, we show a representative spectrum for each cluster.[9] Figure 1 shows these spectra labeled with identifying colors and descriptive labels. As can be seen, we identify various levels of both crowd involvement and music. Notice that the crowd involvement clusters have a similar shape, with differing spectral slopes at lower frequencies, while the music clusters have higher levels of low-frequency energy.

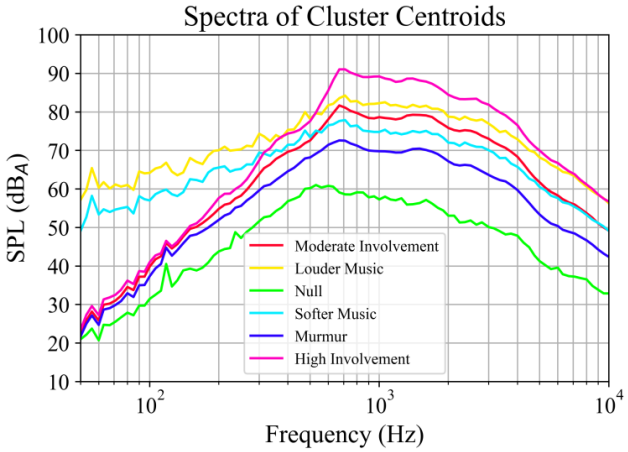

Figure 1. A-weighted spectral levels for each cluster centroid.

15 October 2024 14:10:43

In this report, we extend this approach to a different type of crowd noise that was collected during a 2019 Mardi Gras parade. Applying the methodology to these different datasets allows us to see how robust our process is, to identify differences between disparate crowd noise environments, and to see whether we can label other datasets through the same unsupervised approach. Particularly, we seek to identify instances of focused crowd involvement in the Mardi Gras crowd noise dataset. We expect that these instances will contain concentrated and active crowd engagement as well as directed or unified cheering from members of the crowd. One application would be to detect and reward a section of the crowd that is more engaged in their cheering and desire to attract beads thrown from a float, as opposed to more passive cheering of other spectators in the crowd.

We find that due to the differences between the different types of crowds, the previous clustering approach cannot be used on the Mardi Gras data. This is because clustering techniques rely on an assumption of discrete differences in the data. While the basketball data does indeed have discrete events, so clustering makes sense, the Mardi Gras data is too homogeneous for simple clustering to be effective, in part due to the difference in crowd dynamic. Instead, we apply a combination of hierarchical spectral clustering in conjunction with low-level features common to audio processing.[10-13] Though the process is not identical to that used with the basketball data, the lessons learned from processing the basketball data are useful to develop similar methods for Mardi Gras data. Ultimately, we generalize the discrete partitioning of basketball clusters to a continuous thresholding approach based on key low-level features we identify as related to focused crowd involvement, which we then apply to the Mardi Gras data. The paper is organized as follows. Section 2 contains an outline of data collection and processing techniques. Section 3 then discusses the application of the basketball data processing approach applied to the Mardi Gras data, along with the results. Conclusions are given in Section 4.

2. METHODS

A. MARDI GRAS PARADE DATA

Crowd noise was recorded from a float in the Muses parade for the 2019 Mardi Gras in New Orleans, Louisiana. As shown in Figure 2, a portable recorder (Zoom Hn1) was placed inside of a cup and suspended by a nylon stocking on the side of a float as it traveled down the parade route. In contrast to the basketball data that consists of about two hours of recordings per game of a mostly stationary crowd composed of the same individuals, the Mardi Gras crowd is fundamentally different. As the float moved through the crowd, an individual is in range of the microphone for only about 15 seconds before being drowned out by the crowd, and so while there is constant crowd noise for four hours and the distance to the crowd is mostly constant, the individuals that comprise the nearby crowd are continually changing; effectively the peak of the crowd sound follows the float like a wave. The depth of the crowd varies from 5’ to 30’ along the route flanking the float on either side, and the distance from the nearest person (mouth as sound source) in the crowd to the microphone is as small as 10’, emphasizing the closest voices in the crowd, and devaluing sound pressure created by the more distant crowd voices.

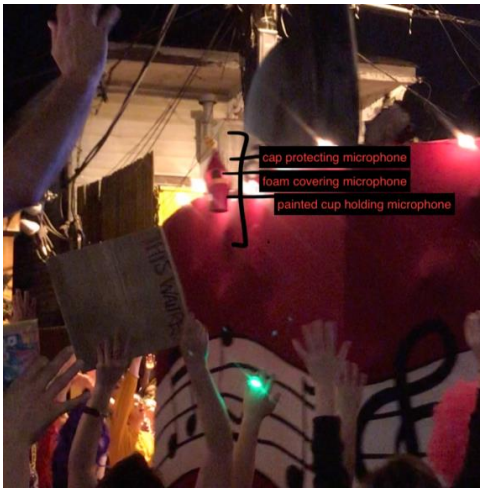

Figure 2. Microphone setup used for taking data at Mardi Gras.

15 October 2024 14:10:43

B. DATA PROCESSING

The data were processed in the same manner as the basketball data,[9] extracting spectral levels for each half-second interval of the parade data. However, the data were not manually labeled with crowd responses as was done for the basketball data since the constant involvement of the crowd throughout the recording made it challenging to identify the beginning and endings of isolated crowd responses. While this limits the interpretability of the data somewhat, we seek to apply the same unsupervised ML approach to see if we can extract meaningful instances of focused crowd involvement. The motivation for this is to see if we can automatically detect focused involvement without the need for manual review so that during future events crowds can be rewarded for their efforts.

Spectral levels alone yield some information, but much more can be learned by also calculating what are known as low-level signal parameters commonly used in audio processing.[10,12] For each half-second interval of the parade data, we compute 20 different feature values measuring various statistical metrics, such as the spectral centroid, root-mean-square (rms), and spectral roll-off. While different low-level features can be of varying importance for different applications, for our purposes we identified eight main low-level features, namely: (1) spectral flux, (2) rms, (3) tonal power ratio, (4) spectral decrease, (5) spectral flatness, (6) spectral roll-off, (7) kurtosis, and (8) spectral slope. While we do not discuss them all in detail here, a full list of features and their descriptions can be found in a prior paper.[9]

Two of these low-level features which are of particular importance are the slope and the flux. The spectral slope is a measure of the linear change of the spectral shape. Larger slope values indicate that the audio spectrum tails off towards higher frequencies more quickly. The spectral flux is the root-mean-square of the change in spectral slope over a given interval.[10] Higher values of flux indicate greater changes in the amount of sound energy present. How these relate to crowd noise is discussed subsequently.

3. RESULTS

A. CLUSTERING THE MARDI GRAS SPECTRA

While clustering can be performed on the Mardi Gras spectral data alone, it is more difficult to interpret different clusters without the context that comes from manually labeling the data. In addition, when interpreting the basketball clusters, we were able to review videos of the game for further confirmation of the cluster descriptions. This approach has limited utility without manual labels or accompanying video of the crowd to guide interpretation. Indeed, the clusters found after conducting a jump analysis were not as clearly interpretable as those identified by the basketball crowd data.

Instead, we partitioned each half-second block of the Mardi Gras data into the six spectral basketball clusters according to the nearest cluster centroid. A portion of this analysis is shown in Figure 3. In this 8-minute segment, we see that the two music clusters (yellow and cyan) dominate near the beginning of the clip; however, the crowd's involvement increases and then alternates between the moderate- and high-involvement categories (red and pink labels, respectively) for a majority of the time. Finally, around the 7-minute mark, we see the crowd involvement drop and the labels shift to the murmur and null categories (blue and green, respectively), which occurs when the float nears the end of the parade route.

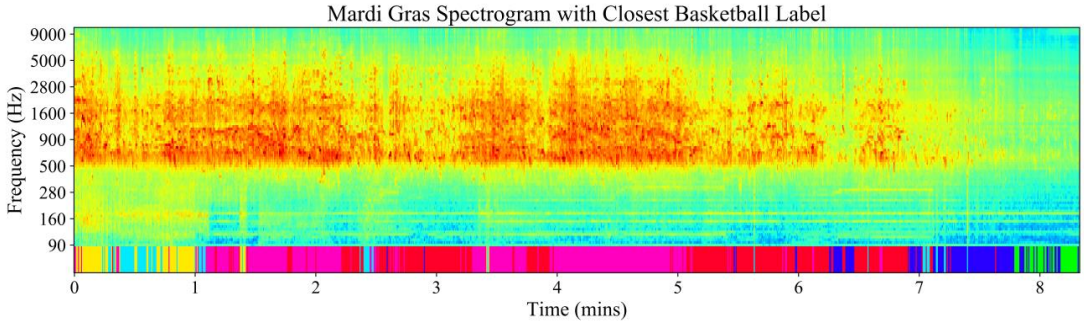

Figure 3. A spectrogram showing the closest basketball cluster for each half-second segment of an 8-minute clip of the Mardi Gras audio. The color at the bottom of the spectrogram corresponds to the colors of the six basketball clusters listed in Table 1.

15 October 2024 14:10:43

When analyzing the Mardi Gras data as a whole, more than 65% of the data is classified as moderate to high crowd involvement. This confirms the constant involvement of the crowd throughout the recording, as noted previously, and suggests that the nature of the Mardi Gras crowd noise is very different from the basketball crowd noise data. Figure 4 shows percentages for each of the categories. We can still use this approach to separate crowd noise vs. non-crowd noise; however, since such a high percentage of the audio contains either moderate or high levels of crowd engagement, using only the spectral data does not allow us to isolate instances of focused crowd involvement particularly well. Since using only the spectral data has had limited success, we next augment the clustering analysis with low-level features to further separate audio dominated by high levels of crowd involvement.

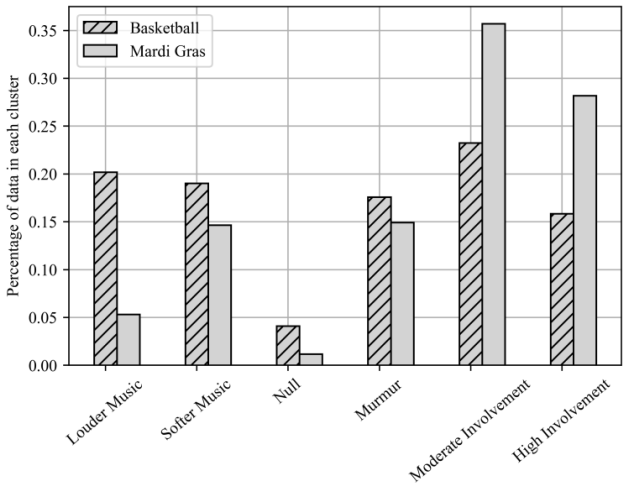

Figure 4. Cluster distribution of crowd noise for basketball and Mardi Gras data. Mardi Gras audio has a much higher percentage of moderate and high crowd involvement than does the basketball audio.

B. HIERARCHICAL CLUSTERING

The different classes of clusters found using the spectral data—namely noise created by the crowd and noise created by the PA system or band—naturally suggest using a hierarchical-clustering-type approach. Divisive hierarchical clustering starts with all the data in one cluster and partitions it into many clusters in a top-down fashion.[14,15] We have explored various clustering approaches using various data subsets and different feature sets. One such method uses the two most-involved crowd noise clusters (moderate and high involvement) found when clustering on spectral data—effectively eliminating all samples dominated by music and minimal crowd noise—which we then re-cluster using the previously processed low-level features to further partition the relevant data.

I. BASKETBALL CROWD NOISE

Because our goal is to identify focused crowd involvement, we first used our labeled basketball data to help identify low-level features that may help distinguish between different crowd responses. For each of the nine labeled responses (“cheer”, “applause”, “positive chant”, “negative chant”, “singing”, “angry noise”, “disappointment”, and “silence”) and each of the nine basketball games, data were standardized to have mean of zero and standard deviation of one. Note that data were standardized for individual games and responses to prevent games or low-level features with higher levels overall to dominate. The mean of the data for each response was then calculated from all data for a given response. Low-level features for which the difference in minimum and maximum level was greater than 1.5 were selected for further use in clustering. The threshold of 1.5 was chosen because it provided a natural break for reducing the low-level feature set; it gives us a total of eight low-level features that can be used to distinguish between different crowd responses, as described previously. We found conducting a jump analysis using these features suggested there were three subclusters within the subset of basketball data we had selected.

By clustering this data subset using these eight low-level features, we were able to produce 3 new subclusters that help further differentiate noise dominated by crowd involvement. A plot of the low-level feature centroid

15 October 2024 14:10:43

values is shown in Figure 5. As is evident, subcluster 2 in the plot (purple) shows high values for the flux and the slope. By manually listening to data where these subclusters are active, it is clear that this subcluster is characterized by instances of focused crowd involvement—where the crowd is very actively engaged in high-level screaming after a noteworthy game event. The other subclusters, however, are more passive cheering such as applause and shouting after a regular basket. This distinction between the subclusters fits the description of spectral flux and slope that was given above. The high values of flux in subcluster 2 suggest there is a large increase in the total amount of sound energy, and the high values of slope mean that most of the energy is being pushed to higher frequencies, where cheering occurs, resulting in a sharper decrease of levels at the highest frequencies.

These subclusters are helpful in that they not only identify differences in levels of crowd involvement but separate different crowd behaviors. Thus, we can use this hierarchical approach with low-level features to effectively identify instances of focused crowd involvement in basketball crowd noise. We now investigate the effectiveness of this approach for identifying focused crowd involvement using the Mardi Gras parade data.

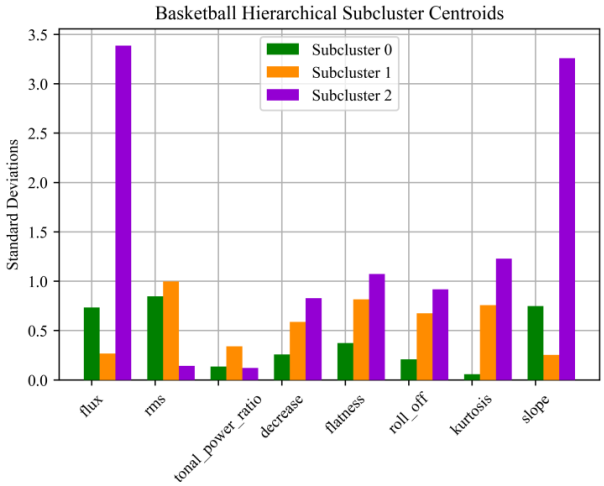

Figure 5. Hierarchical subclusters found using normalized low-level features from the basketball data.

II. MARDI GRAS CROWD NOISE

Hierarchical clustering for the Mardi Gras data was performed similarly to the basketball data. However, due to the preponderance of high crowd involvement, only the data subset classified as being moderate or high crowd involvement was clustered using low-level features. The subclusters found using this approach are very different from those found in the basketball data and are shown in Figure 6.

In Figure 6, no subcluster was found that had such extreme values for the slope and flux, which for the basketball crowd corresponded to instances of focused crowd involvement. This is likely due to the fundamental differences between crowd types. Flux and slope are both related to the change in crowd levels: in the basketball data, we saw a greater fluctuation between types of crowd involvement because we had one crowd that was changing dynamically; however, since the Mardi Gras data was consistently loud due to the crowd usually being in a moderate to high involvement state (at least for this particular float), and because it was not the same individuals who may tire of cheering continuously, the standard deviations for flux and slope are less extreme. The rather homogeneous nature of the Mardi Gras audio leads to less variation in key low-level features. Just as clustering the two data sets leads to different types of clusters, sub-clustering high crowd involvement leads to different types of subclusters. The values are sufficiently different that simply finding the nearest basketball subcluster is not particularly useful, as very little of the data is partitioned into the focused-involvement basketball subcluster, i.e., subcluster 2 in Figure 5.

15 October 2024 14:10:43

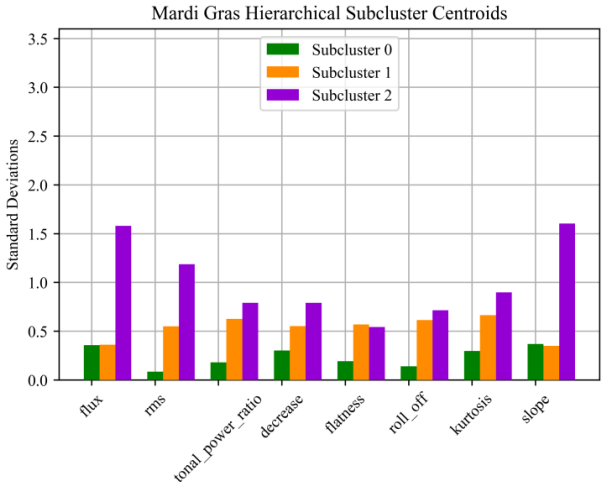

Figure 6. Hierarchical subclusters found using normalized low-level features for the Mardi Gras data.

C. THRESHOLDING WITH FLUX AND SLOPE

While we are not able to use the same approach to the Mardi Gras audio that we used for the basketball data, we are still able to use the results of processing the basketball data to extract some meaningful information from the Mardi Gras data. The most extreme values in the basketball subcluster with focused crowd involvement were the flux and slope. By selecting a threshold value for these low-level features, we identify the most likely instances of focused crowd involvement in the Mardi Gras audio by selecting the half-second intervals where these features are above this threshold. In Figure 7 we have identified in black instances with the highest likelihood of containing focused crowd involvement, based on a threshold for both parameters of 1.8 standard deviations away from a mean feature value of 0.

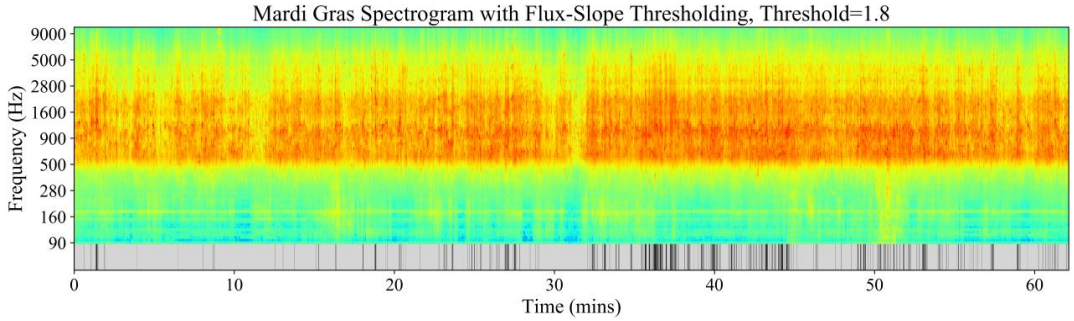

Figure 7. Spectrogram showing the instances with the highest likelihood of containing focused crowd involvement (labeled with black at the bottom).

Different thresholds can be chosen to select a different percentage of the data. A higher threshold value will select less of the data, while lower values include more. Though this approach does not provide us with a discrete partitioning of the data as clustering did with the basketball data, a continuous thresholding is a generalization of the same principle that can overcome the homogeneous nature of the Mardi Gras data. So, while we cannot identify instances of focused crowd involvement in the Mardi Gras data in the same way as in the basketball data, we can still automatically select time intervals where the crowd is particularly engaged.

In Figure 8 we show the energetic average spectrum taken across all instances of focused crowd involvement, compared to the average spectrum of all other instances. As shown, for the instances of focused involvement, there is more energy in the higher frequency bands than at lower frequencies indicating a greater amount of crowd cheering.[16] There is also more total energy present in the focused spectrum, as expected, given

15 October 2024 14:10:43

the higher values of flux. The differences between these spectra support the claim that we have been able to identify instances of focused crowd cheering.

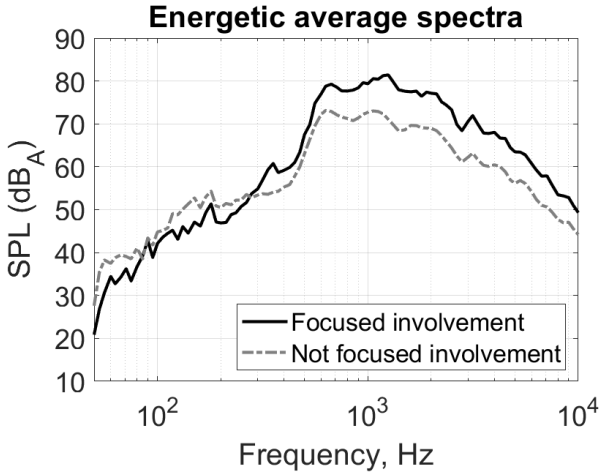

Figure 8. Energetically averaged spectra for focused crowd cheering compared with other instances of cheering.

4. CONCLUSION

Basketball crowd noise is fundamentally different from Mardi Gras parade crowd noise, in part because of the type of event to which the crowd is responding. For example, while basketball audio includes booing and distraction noise, the Mardi Gras audio does not. In addition, the Mardi Gras data is rather homogeneous whereas the basketball data has noticeably discrete events. The difference in crowd types leads to different feature values, which affect clustering techniques. The same approach to both crowd types will not necessarily give the same results, as differences in the distributions of both spectral levels and low level features affect clustering results. However, there appear to be underlying principles that do generalize from one type of crowd to the other.

By adjusting the approach to account for the differences in the distribution of the various features, we can ultimately identify periods where the crowd is likely to show focused involvement. Validation of results for this particular dataset is a challenging task, since the Mardi Gras crowd is almost constantly cheering and screaming (as confirmed by auralization of the recordings). While a supervised learning approach could have provided an easier framework for validation, the unsupervised approach undertaken is useful as we can circumvent the need for time-consuming manual labeling. Indeed, the unsupervised approach described here could lead to a real-time classification of crowd noise, which is necessary to be able to reward crowds for their cheering efforts.

In this case, using unsupervised machine learning allowed us to experiment with different sets of features, and different methods of partitioning the data, which aided in discovering one way to detect high crowd engagement. From our findings, a simple thresholding approach using just the spectral slope and flux can identify time periods with a high likelihood of containing focused crowd involvement. While not infallible, the results demonstrate that this approach can identify focused cheering, rather than just particularly loud crowds. This process could be expanded further to analyze audio recordings taken with a stationary microphone along the parade route, as well as at other sporting and recreational events.

ACKNOWLEDGMENTS

The authors gratefully acknowledge mentored research funding from the College of Physical and Mathematical Sciences at Brigham Young University.

15 October 2024 14:10:43

REFERENCES

[1] Basili, Roberto, Alfredo Serafini, and Armando Stellato. “Classification of musical genre: a machine learning approach.” In Proceedings of the International Conference on Music Information Retrieval (ISMIR) (October 2004).

[2] Radhakrishan, R., Ziyou Xiong, A. Divakaxan, and Y. Ishikawa. "Generation of Sports Highlights Using a Combination of Supervised & Unsupervised Learning in Audio Domain." Fourth International Conference on Information, Communications and Signal Processing. 2003. doi:10.1109/icics.2003.1292595.

[3] Jurafsky, Dan, and James H. Martin. Speech and Language Processing. Upper Saddle River, NJ: Prentice Hall, Pearson Education International, 2014.

[4] S. Davies and D. Bland, "Interestingness Detection in Sports Audio Broadcasts," 2010 Ninth International Conference on Machine Learning and Applications, Washington, DC, 2010, pp. 643-648, doi: 10.1109/ICMLA.2010.99.

[5] Baillie, Mark, and Joemon M. Jose. "An audio-based sports video segmentation and event detection algorithm." In 2004 Conference on Computer Vision and Pattern Recognition Workshop, pp. 110-110. IEEE, 2004.

[6] Salamon, Justin, and Juan Pablo Bello. "Unsupervised Feature Learning for Urban Sound Classification." 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015. doi:10.1109/icassp.2015.7177954.

[7] M. Dundar, Q. Kou, B. Zhang, Y. He and B. Rajwa, "Simplicity of Kmeans Versus Deepness of Deep Learning: A Case of Unsupervised Feature Learning with Limited Data," 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, 2015, pp. 883-888, doi: 10.1109/ICMLA.2015.78.

[8] T. Heittola, A. Mesaros, T. Virtanen and M. Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 8677-8681, doi: 10.1109/ICASSP.2013.6639360.

[9] Brooks A. Butler, Katrina Pedersen, Mylan R. Cook, Spencer G. Wadsworth, Eric Todd, Dallen Stark, Kent L. Gee, Mark K. Transtrum and Sean Warnick. Classifying crowd behavior at collegiate basketball games using acoustic data, Proceedings of Meetings on Acoustics 35, 055006 (2018); https://doi.org/10.1121/2.0001061.

[10] Lerch, Alexander. An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics. Hoboken, NJ: IEEE Press, 2012.

[11] Logan, Beth. "Mel Frequency Cepstral Coefficients for Music Modeling." In Proceedings of the International Symposium on Music Information Retrieval (ISMIR), Plymouth, Oct. 2000.

[12] McKinney, Martin F., and Jeroen Breebart, "Features for Audio and Music Classification." In Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR), Baltimore, Oct. 2003.

[13] Tzanetakis, George, and Perry Cook, "Musical Genre Classification of Audio Signals." Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293-302, 2002. DOI: 10.1109/TSA.2002.800560.

[14] Roux, Maurice. "A comparative study of divisive hierarchical clustering algorithms." arXiv preprint arXiv:1506.08977 (2015).

[15] De Silva, Pavani Y., Chiran N. Fernando, Damith D. Wijethunge, and Subha D. Fernando. "Recursive Hierarchical Clustering Algorithm." International Journal of Machine Learning and Computing 8, no. 1 (2018).

[16] Navvab, Mojtaba, Gunnar Heilmann, and W. SULISZ Dennis. "Crowd noise measurements and simulation in large stadium using beamforming." In Eleventh International IBPSA Conference. 2009.

15 October 2024 14:10:43