Unsupervised Blue Whale Call Detection Using Multiple Time-Frequency Features

Alejandro Cuevas†, Alejandro Veragua†
Dpt. of Electrical Engineering
Universidad de Chile

Sonia Español, Gustavo Chiang
Fundación Meri

Felipe Tobar
Center for Mathematical Modeling
Universidad de Chile

***Abstract*—In the context of bio-acoustic sciences, call detection is a critical task for understanding the behaviour of marine mammals such as the blue whale species (*Balaeonoptera musculus*) considered in this work. In this paper we present an approach to blue whale calls detection from an unsupervised perspective. To achieve this, we use temporal and spectral features of audio acquired with a marine autonomous recording unit. The features considered are 46-dimensional and include the mel frequency ceptrum coefficients, chromagrams, and other scalar quantities; these features were then grouped via two different cluster algorithms. Our findings confirm the suitability of the proposed approach for isolating blue whale calls from other environmental sounds (as validated by a bio-acoustic specialist). This is a clear contribution for the annotation of blue whales calls, where the search for calls can now be performed by analysing the clusters identified instead of the entire recordings, thus saving time and effort for practitioners in bio-acoustics.**

I. Introduction

A. Basics of blue whale's calls

In recent years, passive acoustic monitoring (PAM) has been considered within the study of cetaceans [1], [2] though recording and analysing their acoustic activity; this has been used as a tool for assessing the effect of man-made sounds on such group of mammals [3], [4]. PAM has become a popular resource within bio-acoustics as a detector for cetaceans [5], [6], this is because they are capable of operating over extended periods of time, at day or night, in any weather condition (although their reliability and range decreases when the weather and sea state deteriorate), and in any geographical area. PAM has therefore allowed for extensive monitoring far beyond what has been achieved with visual methods [5], [6].

Within cetaceans, blue whales (*Balaenoptera musculus*) are an endangered species and therefore understanding their behaviour patterns is crucial to design conservations policies; we approach this through analysing their calls. Blue whales calls are mainly characterised for being tonal signals, having a frequency around 10 - 100 Hz, although some vocalisations can reach 400 Hz [7], [8]. Since (i) these frequency ranges are similar to those of the buoy signal and sound of ship's motor, and (ii) the whale call is in the low-frequency part of the spectrum, a common issue within acoustic detectors is that whale calls are overshadowed by external sounds due to their proximity on the spectral domain. This makes the detection challenging specially if only the spectrogram is considered.

†These authors contributed equally

Whales in general produce high-intensity audio signals that can be detected dozens of kilometres away using a single hydrophone [9], however, when the whale is moving farther away from the receptor the signal intensity decreases, thus making the detection challenging. At the same time, if there is a ship close to the measurement point, the sound from the ship will block the low-intensity call [1], [10]. In this sense, the construction of an automatic detector that is able to process a large amount recordings is a direct contribution to whale call detection to depart from manual (human) methods and thus represent an increase processing speed and precision.

Furthermore, it should be noted that the efficiency of acoustic detection schemes varies for different species. This is due to the difference among fundamental frequency, signal intensity, travel direction of the sound and animal behaviour [1].

B. Scope of this study

Recall that the ultimate goal of blue whale call detection is to characterise and understand whales behaviour. We address the call detection problem by analysing submarine audio recordings, and then detecting different types of submarine sounds via clustering; our hypothesis is that the whales calls (having frequency between 10 - 525 Hz) will be isolated into one or more clusters. Our case study considers recordings of the submarine environment obtained with a marine autonomous recording unit (MARU) that was moored to the seafloor (200 metres) near the Guafo island (S43°31.889', W074°26.488') in the south of Chile. Acoustic recordings were acquired between the end of summer and autumn 2012, whereas the data used in this work consisted of 6.5 hours of recordings containing in blue whale calls and environmental sound as ship engines.

Our setting is an audio segmentation one [11]–[13]. In this context, we propose a unsuspervised approach to whale's call detection, where for each section of audio we calculate multiple features of both temporal and spectral nature. Our set of features includes but is not restricted to Mel Frequency Cepstral Coefficients (MFCC) [14] and Chroma features [15]. Then we use various clustering methods to segment the feature space in groups, using Gaussian Mixture Model (GMM) [16] and Density-Based spacial clustering of applications with noise (DBSCAN) [17]. Validation of the algorithm was then made by a bio-acoustic specialist. An example of typical blue

whale's call spectrogram, found in the recordings and validated by a bio-acoustic specialist is shown in Fig.1, where three main parts composing the vocalisation can be identified; notice that for each part the low fundamental frequency is clearly identified and harmonics are also visible.

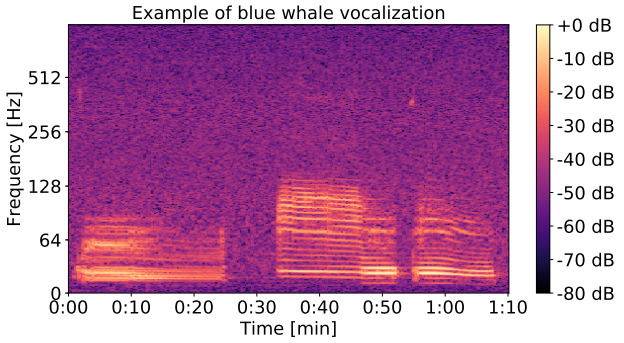

Fig. 1. An example of spectrogram of blue whale's call. Observed the three sections and the low-frequency fundamental component with harmonics.

II. Proposed methodology

Our approach is performed in three stages: First, the available audio recordings are preprocessed and normalised, then the audio signals are divided into small segments of approximately 1 [s] to compute a total of 46 features on those segments. Secondly, clustering is performed on the features to group the audio according their common spectral properties. Thirdly, the result of the clustering is presented using t-SNE, a method for visualising high-dimensional elements. Our aim is that all calls are isolated into one or more clusters, whereas external sounds are represented in the remaining clusters.

A. Time-frequency features considered

For each normalised audio segment, the features extracted and their dimension are shown in Table I, where the top 14 rows are usual scalar-valued features in signal processing analysis both in the time and frequency domains.

TABLE I
Features considered with corresponding dimensions.

Feature	dim	Feature	dim
Zero crossing rate	1	Max frequency	1
Energy entropy	1	Energy	1
Skewness	1	Kurtosis	1
Min	1	Max	1
Range	1	Spectral Centroid	1
Spectral spread	1	Spectral entropy	1
Spectral flux	1	Spectral roll off	1
MFCC	20	Chroma	12

Mel frequency ceptral coefficients (MFCC): Since its introduction in [14], MFCC has been the *de facto* method for obtaining features for speech processing. The ceptrum is defined as the inverse Fourier transform of the logarithm of the power spectral density of a signal, which, due to the Wiener-Khinchin theorem [18], [19], can be interpreted as a log-compressed autocorrelation sequence of the signal:

$$\text{Power ceptrum} = |\mathcal{F}^{-1}\{\log|\hat{X}(\omega)|^2\}|, \quad (1)$$

where $\mathcal{F}^{-1}$ is the inverse discrete Fourier transform and $\hat{X}(\omega)$ is the estimate of the power spectral density. MFCC is an extension of the ceptrum concept that takes into account the energy of Mel-spaced filter banks [20], thus providing information of the signal according to the Mel scale, where different frequency zones have filter banks of different widths to achieve the desired resolution. This is specially useful when describing whale calls, as the calls frequency ranges from 10 to 525 Hz and variable resolution is required.

Projected spectra: Chroma features [15] are an alternative representation of the spectrum of a signal in which the entire spectrum is projected into twelve bins representing the twelve semitones (or chroma) of the musical octave in western music. The chromagram is obtained from the power spectral density and is a reduced-dimensionality representation of the spectrum which will be useful to discriminate among multiple audio sources.

B. Unsupervised learning: Clustering

Segmentation of audio segments (represented by their feature vectors) will be achieved using two clustering methods: DBSCAN [17] and Gaussian mixture model [16]. The rationale behind DBSCAN is that clusters are dense groups of elements, meaning that if a particular element belongs to a cluster, it should be close to a number of other elements in that cluster. The method receives two parameters, $\min_{\text{points}}$ and a radius ϵ, where an element is in a cluster if there is at least $\min_{\text{points}}$ other elements in a radius ϵ around it. It is worth noting that DBSCAN does not need a predefined number of clusters and it is a nonparametric method to group points, where isolated points are labelled togather in cluster 0 (outliers).

A Gaussian mixture model (GMM) is a latent variable model that assumes all data points come from a mixture of finite number of multivariate Gaussians, each one with its own mean vector and covariance matrix. The latent variable is the probability of a point to have been generated by a given Gaussian component. GMM can be understood as an generalisation of K-Means [21], since K-Means considers isotropic Gaussians only whereas and GMM incorporates the covariance structure of the data to the clustering.

C. Visualisation

Given that the feature space is 46-dimensional, once the clustering stage is performed a dimensionality reduction method will be used to visualise the data and provide intuition into the found clusters. In this work, we consider the t-distributed stochastic neighbour embedding method (t-SNE) [22], a non-linear dimensionality-reduction algorithm that constructs a probability distribution over pairs of high-dimensional objects and then projects them onto a lower-dimensional space, where similar points will have high probability of being near. The low-dimensional space will be 2- or 3-dimensional and can therefore be plotted.

III. Case Study: *Balaeonoptera Musculus*

In this study, data were obtained from hydrophones placed in nautical buoys. Having a hydrophone in a fixed place, instead of having it fixed to the whale such as a D-Tag [23], has a key advantage: The passive movement of the water does not disturb as much as having the instrument attached to a whale, where sudden movements and water splashes saturates the hydrophone.

A. Preprocessing

For this study, we used 3 files of marine audio recordings: One containing blue whale's calls, one containing background sounds, and containing one far ship engine sounds. Each recording was 900-second long (45 minutes in total), resampled at 2 kHz and converted to mono (by averaging both channels). Then, the recordings were standardised individually, then, the entire set was normalised again to obtain unit variance. Finally, the available dataset was divided into shorter, overlapped frames of approximately 1 [s] (2048 data points), with 30% overlapping, to calculate the 46 features mentioned in Section II-A. We emphasise that although we know where the whale calls are in the data, our training approach is fully unsupervised, and the labels are only used to validate the segmentation obtained.

B. Training

Clustering was applied in the feature space, where for DBSCAN the heuristics chosen for hyperparameters is that $\min_{\text{points}}$ was set equal dimension of feature space minus two (i.e., 44), and the radius (ϵ) equal to the mean distance to the $\min_{\text{points}}$ neighbour. For GMM, an unconstranied covariance matrix was used to produce a general model, and the number of components was set by inspection from 3 to 8 components, where the methods performed consistently.

After clustering, the segments of audio corresponding to points in the same cluster were grouped in the same audio file maintaining their order in time, thus facilitating the validation performed by the bio-acoustic specialist. Recall that t-SNE was used to show the prototypes (centres)found via clustering.

C. Experimental results

DBSCAN, using the aforementioned heuristic with $\min_{\text{points}} = 44$ and $\epsilon = 2.716$, yielded three clusters shown in Fig. 2. The outliers found, marked in green, are spread across the low dimensionality projection, it is precisely in this cluster where all the calls were grouped together with some ship engines passing close to the hydrophone. The reason the calls were considered as outliers is that the distance between elements that were not whale calls is, in average, smaller than the distances between whale's calls, as calls are formed by different sound structures—in simpler terms, calls are too distant to one another to form clusters under DBSCAN. The segmentation in time made by DBSCAN, shown in Fig. 3, reveals that calls were grouped with non-call audio, since we know that there are no calls after the first third of the data. The use of DBSCAN found some of the calls but not because of their harmonic structure but for being an outlier in the frequency domain, however, DBSCAN gives an intuition to set the number of clusters for GMM: at least three, since the outliers can still have undiscovered structure. An example of audio labelled in the same cluster as the calls is shown in Fig.4 where it can be seen other elements besides the whale calls.

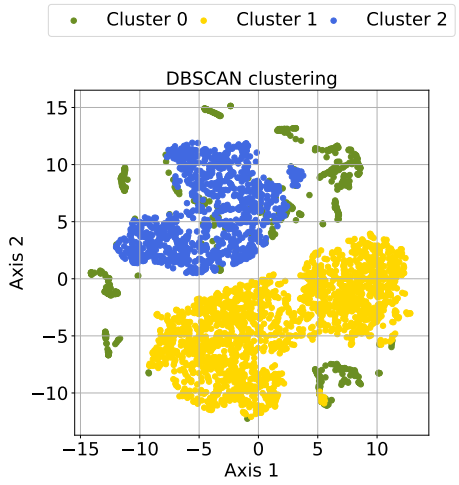

Fig. 2. t-SNE projection of feature samples clustered by DBSCAN (3 clusters). The cluster 0 (outliers) contains both the whale calls and a some ship engine sounds, cluster 1 contains submarine background noise and cluster 2 the rest of the ship sounds.

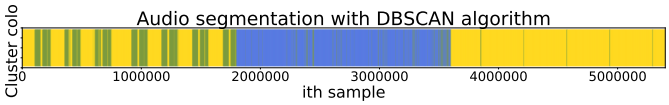

Fig. 3. Time segmentation using labels obtained from DBSCAN. Colour code follows Fig. 2

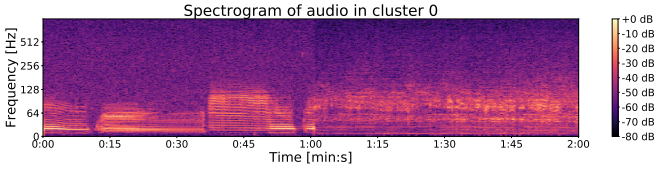

Fig. 4. 120 [s] spectrogram for cluster 0 assigned by DBSCAN. We can see calls up to 1:00 and non-calls from 1:00.

GMM was first trained using three components (GMM-3) as there are (theoretically) three sound sources: background noise, ship's motor and blue whale's calls—recall that this was confirmed by DBSCAN. The hard assignment of GMM is shown in Fig. 5, as in similar way with DBSCAN, the two large groups seen in the t-SNE projection are grouped together, the calls are identified in the blue cluster together with ship motor sounds. Notice that as GMM forces the size of the clusters so as to fit all the data into the given number

of clusters, thus grouping points that may not be sufficiently similar due to a poor choice of the number of clusters. The time segmentation for GMM is shown in Fig. 6 where most of the whale's call is the same cluster as the large section marked in blue. Both for DBSCAN and GMM with three components, Table II shows the number of elements in each cluster, the total of data points and duration per cluster. Note that, as opposed to DBSCAN, GMM groups most of the observations in two clusters. An example of audio assigned to cluster 2 is shown in Fig. 7, where both calls and non-calls can be identified.

TABLE II
AUDIO SEGMENTATION FOUND BY THE CLUSTERING ALGORITHMS WITH 3 CLUSTERS.

Cluster id		0	1	2
DBSCAN	samples	675	1 934	1 158
	audio samples	1 036 155	2 793 562	1 706 769
	duration [s]	518.08	1 396.78	853.38
GMM	samples	1.774	1.927	66
	audio samples	2 564 897	2 778 611	100 728
	duration [s]	1 282.45	1 389.30	50.36

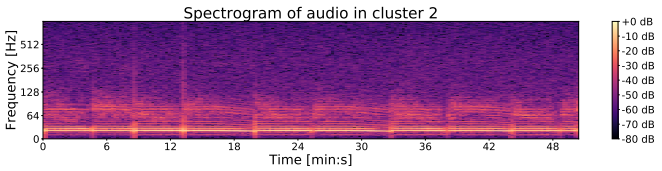

Fig. 7. 120 [s] spectrogram for cluster 2 assigned by GMM with three components.

Finally, GMM was implemented with eight components (GMM-8) and the result is shown in Fig. 8, where the outliers and the largest cluster of GMM-3 were split in different clusters. The time segmentation shown in Fig. 9 reveals that most of the blue whale's calls are in individual clusters. The number of elements in each cluster in the GMM-8, as well as the total of the data points and duration is shown in Table III, where the blue whale calls are in clusters 2, 3, 4, 6 and 7, each cluster with a different call structure. Where the tree main structures of a call are in clusters 3, 4 and 7.

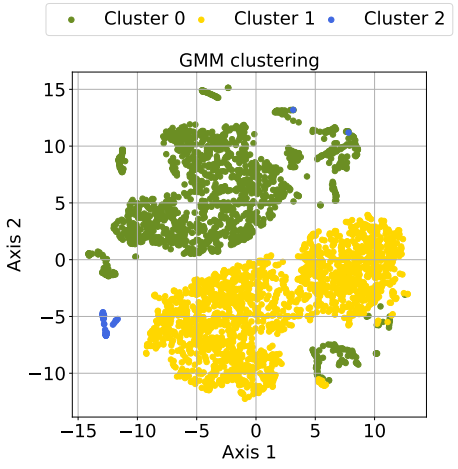

Fig. 5. t-SNE projection of feature space of partition found by GMM with 3 components. The cluster 0 is submarine background, cluster 1 contains calls and ship sounds, and cluster 2 contains a last part of the calls.

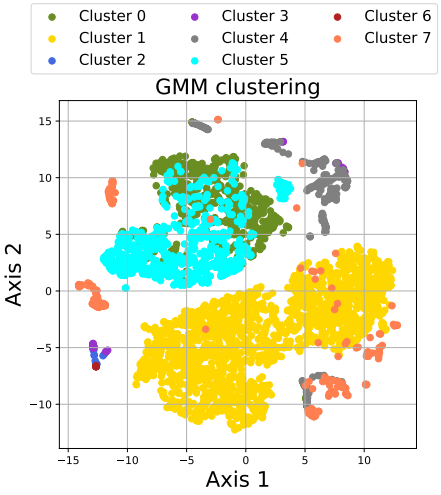

Fig. 8. t-SNE projection of feature space by partition found by GMM with 8 components.

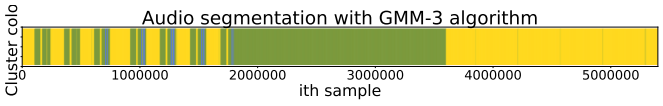

Fig. 6. Time segmentation using labels obtained with GMM with 3 components.

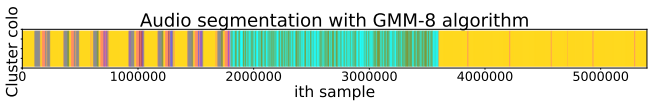

Fig. 9. Time segmentation using labels obtained with GMM with 8 components.

TABLE III
AUDIO SEGMENTATION FOUND BY THE GMM WITH 8 COMPONENTS.

cluster id	samples	audio samples	duration
0	626	1 028 053	514.02
1	1 880	2 710 030	1355.01
2	18	31 329	15.66
3	47	77 806	38.90
4	264	396 147	198.07
5	670	1 087 415	543.70
6	14	21 907	10.95
7	248	386 134	193.06

It is also worth noting that the main components of blue whale calls were associated to individual clusters consisting only in that part of the call, whereas background noise, ship's motors and other unknown sources were isolated in the remaining clusters. Advised by the bio-acoustic specialist, we identified the clusters that contained "parts of calls". Then, we post processed the outcome of GMM-8 by combining all the clusters with parts of calls in *meta-cluster* 1 and the rest of the samples in *meta-cluster* 0. An example of two minutes of meta-cluster 1 (containing only calls) is shown in Fig. 10, where multiples calls are stacked, but without the silence between the first and second part shown in Fig.1. Fig.11 shows *meta-cluster* 0, i.e., the combination of clusters that are not calls, where it can be seen the transition from background noise to ships engine sound at second 45.

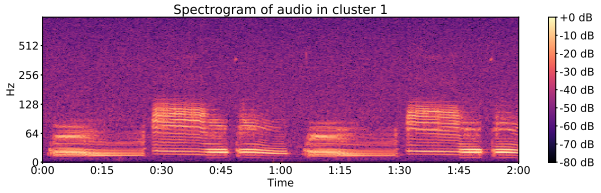

Fig. 10. 120 [s] spectrogram for meta-cluster 1 assigned by GMM with eight components: this meta-cluster contains all the original clusters containing calls.

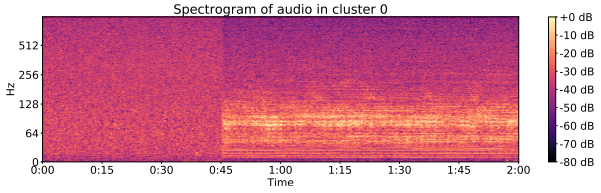

Fig. 11. 120 [s] spectrogram for meta-cluster 0 assigned: this meta-cluster contains all the original clusters which are not calls.

IV. DISCUSSION AND FURTHER STEPS

Our experimental results have shown that submarine audio recordings can be separated to find blue whale calls using (i) multiple time-frequency features and (ii) clustering in an unsupervised manner, where posterior analysis showed that MFCC and Chroma were the most influential of the features. Out of the clustering methods used, GMM with 8 components yielded the best results, being able to separate the audio in the true original sources and find one cluster for each main component of a blue whale call. The proposed framework represents a practical contribution for bio-acoustics, where annotation of whale calls is simplified due to the clustering performed: The bio-acoustic specialist can now focus on each cluster and does not need to analyse the entire recording.

Future work includes filterbanks specially designed for the range of frequency of interest, as MFCC uses Mel scaled filterbanks and Chroma is based on the western musical scale. Within the choice of features, the following question also arises: Is it possible to avoid the design of time-frequency features and rely on fully-automatic feature discovery? The answer to this might be the use of autoencoder neural networks [24], where a compressed representation of the signal (or its spectrum) can be learnt and then used to perform the clustering stage. Finally, using other methods of clustering, such as the Bayesian Gaussian mixture or the Dirichlet processes [16], may allow us to infer the posterior distribution over the clustering parameters, including the number of clusters.

ACKNOWLEDGEMENTS

This work was partially supported by Conicyt projects PAI-82140061 and Basal-CMM.

REFERENCES

[1] D. K. Mellinger, K. M. Stafford, S. Moore, R. P. Dziak, and H. Matsumoto, "Fixed passive acoustic observation methods for cetaceans," *Oceanography*, vol. 20, no. 4, p. 36, 2007.

[2] A. K. Stimpert, W. W. Au, S. E. Parks, T. Hurst, and D. N. Wiley, "Common humpback whale (megaptera novaeangliae) sound types for passive acoustic monitoring," *The Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 476–482, 2011.

[3] L. J. May-Collado and D. Wartzok, "A characterization of guyana dolphin (sotalia guianensis) whistles from costa rica: The importance of broadband recording systems," *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 1202–1213, 2009.

[4] S. Kimura, T. Akamatsu, S. Li, L. Dong, K. Wang, D. Wang, and N. Arai, "Seasonal changes in the local distribution of yangtze finless porpoises related to fish presence," *Marine Mammal Science*, vol. 28, no. 2, pp. 308–324, 2012.

[5] W. Richardson, C. Greene, C. Malme, and D. Thomson, "Marine mammals and noise academic press," *San Diego, CA*, 1995.

[6] W. Au and M. Hastings, "Principles of marine bioacoustics. series: Modern acoustics and signal processing," 2008.

[7] W. C. Cummings and P. O. Thompson, "Underwater sounds from the blue whale, balaenoptera musculus," *The journal of the Acoustical Society of America*, vol. 50, no. 4B, pp. 1193–1198, 1971.

[8] S. J. Buchan, R. Hucke-Gaete, L. Rendell, and K. M. Stafford, "A new song recorded from blue whales in the corcovado gulf, southern chile, and an acoustic link to the eastern tropical pacific," *Endangered Species Research*, vol. 23, no. 3, pp. 241–252, 2014.

[9] J. Barlow and B. L. Taylor, "Estimates of sperm whale abundance in the northeastern temperate pacific from a combined acoustic and visual survey," *Marine Mammal Science*, vol. 21, no. 3, pp. 429–445, 2005.

[10] C. W. Clark, W. T. Ellison, B. L. Southall, L. Hatch, S. M. Van Parijs, A. Frankel, and D. Ponirakis, "Acoustic masking in marine ecosystems: intuitions, analysis, and implication," *Marine Ecology Progress Series*, vol. 395, pp. 201–222, 2009.

[11] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in neural information processing systems*, 2009, pp. 1096–1104.

[12] T. Theodorou, I. Mporas, and N. Fakotakis, "An overview of automatic audio segmentation," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 6, no. 11, p. 1, 2014.

[13] J. X. Zhang, J. Whalley, and S. Brooks, "A two phase method for general audio segmentation," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 2009, pp. 626–629.

[14] 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, 1980.

[15] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer International Publishing, 2015.

[16] K. P. Murphy, *Machine learning : a probabilistic perspective*, ser. Adaptive computation and machine learning. MIT Press, 2012.

[17] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[18] N. Wiener, "Generalized harmonic analysis," *Acta mathematica*, vol. 55, no. 1, pp. 117–258, 1930.

[19] A. Khintchine, "Korrelationstheorie der stationären stochastischen prozesse," *Mathematische Annalen*, vol. 109, no. 1, pp. 604–615, 1934.

[20] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[21] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA., 1967, pp. 281–297.

[22] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[23] M. P. Johnson and P. L. Tyack, "A digital acoustic recording tag for measuring the response of wild marine mammals to sound," *IEEE journal of oceanic engineering*, vol. 28, no. 1, pp. 3–12, 2003.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.