Poetic Meter Classification Using i-vector-MTF Fusion

Rajeev Rajan[1], *Aiswarya Vinod Kumar* [1] † , *Ben P. Babu,* [2]

[1] College of Engineering, Trivandrum, Kerala,India
[2] Rajiv Gandhi Institute of Technology, Kottayam, India

rajeev@cet.ac.in

Abstract

In this paper, a deep neural network (DNN)-based poetic meter classification scheme is proposed using a fusion of musical texture features (MTF) and i-vectors. The experiment is performed in two phases. Initially, the mel-frequency cepstral coefficient (MFCC) features are fused with MTF and classification is done using DNN. MTF include timbral, rhythmic, and melodic features. Later, in the second phase, the MTF is fused with i-vectors and classification is performed. The performance is evaluated using a newly created poetic corpus in Malayalam, one of the prominent languages in India. While the MFCC-MTF/DNN system reports an overall accuracy of 80.83%, the i-vector/MTF fusion reports an overall accuracy of 86.66%. The performance is also compared with a baseline support vector machine (SVM)-based classifier. The results show that the architectural choice of i-vector fusion with MTF on DNN has merit in recognizing meters from recited poems.

Index Terms: meter, poem, fusion, timbral, i-vector

1. Introduction

Poetry has a long history, dating back to the Sumerian Epic of Gilgamesh. Early poems evolved from folk songs such as the Chinese Shijing, or from a need to retell oral epics, as with the Sanskrit Vedas, Zoroastrian Gathas, and the Homeric epics. The musical elements such as rhythm, meter, and sounds are utilized systematically at sonic and typographical levels to write poems [1]. *Each meter(vrtta) defines a sequence of syllable types (stressed or unstressed) for each line of a poem so that the poem follows a rhythm when read out* [2]. The poet has the freedom to choose any of the accepted meter to compose his work. The poetic form had total dominance in the oral tradition due to the ease of committing verses to memory. Strict meter and repetition of sounds in the form of prasa are prevalent in the Indian poetic form to aid oral transmission [2]. In the Western perspective, "poetry" is considered as an art which is spoken using pitch contours which are more-or-less typical of normal, natural speech. However, in many traditions [3], especially in Indian tradition, poetry is actually "sung" not spoken, so that the poem follows a rhythm when reading out [4].

In Indian poetry, each syllable of a word is classified as either a laghu (short syllable, 'U') or a guru (long syllable, '−') [2]. The ordered sequences of syllables of three units in groups of 1, 2 or 4 lines determine the rules for each meter. Eight such sequences can be formed in a tri-syllabic structure. For instance, two lines of a poem written in Kaakali meter is shown in Figure 1. We can observe that 8 tri-syllabic sequences are distributed in two lines. Besides, each tri-syllable consists of two long syllables and one short syllable as seen in Figure. Machine translation of poetry [5], aesthetic and emotional perception study can potentially be benefited from automatic metrical analysis. Most of the previous works rely on orthographic, syntactical and lexical features in poetic classification [6, 7].

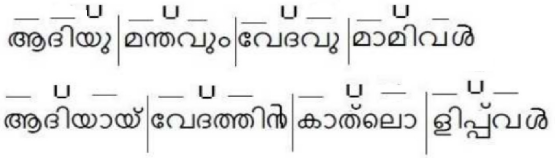

Figure 1: *Two lines of a poem written in Kaakali meter. Sequences of laghu and guru of are also shown.*

Although there has been significant work in rhythm estimation, there has been relatively little work for the estimation of meter from recited poems. Most of the previous approaches rely on syllabic structures rather than the acoustic cues computed from the audio file. The segmented syllables are classified into long/short syllables based on features like zero-crossing rate, PARCOR coefficients, and the temporal duration as part of the automatic estimation of poetic meters in [8]. A set of parameters like acoustic duration, prosodic, rhythmic and stylistic features are also employed for poem comparison in the analysis of English and Italian poetry [9]. Meter classification in traditional Malay poetry is addressed using two sets of experiments, with themes and shape structure in [7]. Statistical methods and word-stress patterns can be effectively used to analyze, generate, and translate rhythmic poetry in [10]. Proposed work focuses on acoustic cues instead of lyrics to identify the meter of the given poem. The experiments reported are the extension the poetic meter estimation approaches discussed in [11, 12]

The rest of the paper is organized as follows. Section 2 describes the proposed system. Performance evaluation is discussed in Section 3. The analysis of results is given in Section 4. Finally, the paper is concluded in Section 5.

2. Proposed System

The proposed framework is shown in Figure 2. MFCC is fused with MTF and classification is performed using the DNN in first phase. Later, i-vectors are fused with MTF and experiment is repeated. The performances are compared with SVM-based classifier. A detailed description is given in the following sections.

2.1. Feature Extraction

2.1.1. MFCC and I-vectors

MFCCs are widely employed in numerous perceptually motivated audio classification tasks as predictors of perceived similarity of timbre [13]. 20 dim MFCCs are computed with frame-size of 40 ms and frame-shift of 10 ms for the task.

I-vector extraction, the popular feature-modeling technique has been successfully used in many applications like speaker

† Former graduate student, College of Engineering,Trivandrum

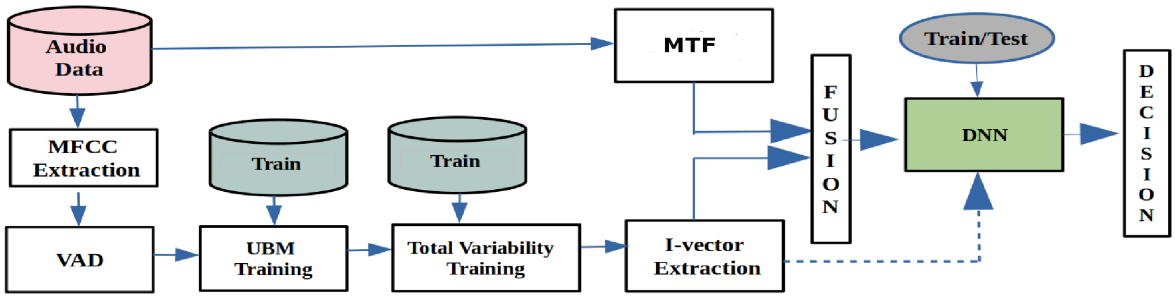

Figure 2: *Block diagram of the proposed system.*

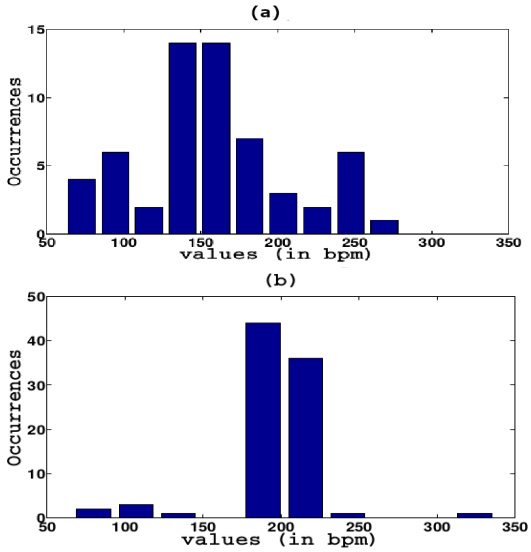

Figure 3: *Beat histogram of poems (a) Druthakaakali (b) Nathonnatha [12].*

and language recognition tasks [14], and audio scene detection [15].

The concept of i-vectors which was initially proposed for speaker verification in Dehak *et al.*[16] has become a leading framework in recent days. This approach improves the joint factor analysis by combining the inter and intra domain variability and modeling it in same low dimensional total variability space. I-vector system [16] maps the high dimensional GMM supervector space (generated from concatenating all the mean values of GMM) to low dimensional space called total variability (TV) space. The main idea is to adapt the target utterance GMM from a universal background model (UBM) using the eigenvoice adaption [17]. The target GMM supervector can be viewed as shifted from the UBM. Formally, a target GMM supervector M can be written as:

$$M = m + Tw \tag{1}$$

where m represents the UBM supervector, T is a low dimensional rectangular TV matrix, and w is a standard normal distributed vector. These feature vectors are referred to as identity vectors or i-vectors for short. The feature vector associated with a given recording is the MAP estimate of w. Using training data, the UBM and TV matrix is modeled by expectation maximization (EM). In the E-step, w is considered as a latent variable with normal prior distribution $N(0, I)$. Eventually, the i-vectors will be estimated as the mean of posterior distribution of w, that is [16],

$$w(u) = (I + T^T \Sigma^1 . N(u) . T)^1 T^T \Sigma^1 S(u) \tag{2}$$

where for utterance u, the terms $N(u)$ and $S(u)$ represent zeroth and centralized first order Baum-Welch statistics respectively, and Σ is the covariance matrix of UBM. 10 dim i-vectors(i_{MFCC}) are computed from MFCC in the proposed task.

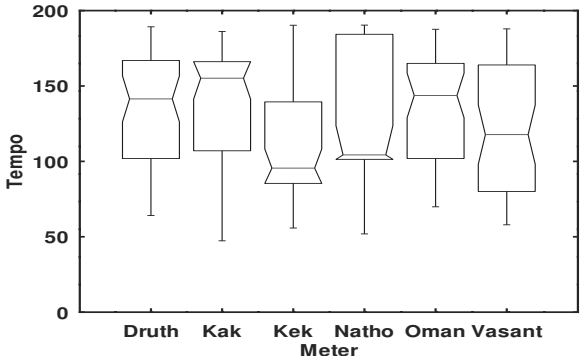

Figure 4: *Tempo distribution in the dataset*

2.1.2. *Musical Texture Features (MTF)*

Musical attributes can be organized into eight different categories, each representing a core concept, namely: dynamics, expressive techniques, harmony, melody, musical form, musical texture, rhythm and tone color (or timbre) [18]. Several audio features have been created and are nowadays implemented in audio frameworks [19]. In our experiment timbral, rhythmic and melodic features are collectively termed as MTF. Five timbral features, namely, spectral centroid, spectral roll-off, spectral flux, zero crossings, and low-energy, are computed [20]. Beat tracking, the extraction of the rhythmic aspects of the musical content has been a topic of active research in recent years[21]. Beat histogram for poems written in DruthaKakali and Nathonnatha is shown in Figure 3. It is worth noting that the dissimilarity in the distribution can potentially be used as an important acoustic cue for the task. Three features, namely, tempo, pulse clarity, event density [22] are computed from the beat histogram. The distribution of tempo, computed from the audio files in the dataset is shown in Figure 4.

High-level melodic features, namely, standard deviation, skewness, and kurtosis, are computed from the melodic pitch.

Table 1: *Meters, considered in the experiment with governing rules*

Sl.No	Meter	Rules or Syntax	Poem/Poet
1	Kakali	8 ganam sequences with three syllables of total 5 mathras.	Ramacharitham/Cheeramon
2	Druthakaakli	It is a variant of Kaakali meter. In Kaakali meter, if we omit first syllable in first and second lines of the poem, it results Druthakakali	Jnapana/Poonthanam
3	Keka	Sequences are in the order 3-2-2-3-2-2. One long syllable should be there in each sequence	Ente Gurunathan/Vallathol
4	Nathonnatha	In first line, eight sequences of two syllables followed by 4 sequences of two syllables in the second line.	Thullal/Kunjan nambiar
5	Omanakuttan	Sequences are in the order 3-2-3-2	Omankuttan-Govindhan/Venmani
6	Vasanthathilakom	If the tri-syllabic sequences are in the order ta-bha-ja-ja-ga, the poem is written in the meter, Vasanthathilakaom	Veena poovu/Kumaranashan

Skewness and kurtosis are computed from the kernel density estimate (KDE) of pitch estimates[23]. Pitch histograms capture the harmonic features of different musical styles. One expects, for instance, that genres with a more complex tonal structure exhibit a higher degree of tonal change and therefore have more pronounced peaks in their histograms than other genres. The similarity in KDE can potentially be used to compute similarity measures of poems, written in same meter.

2.2. Classification Scheme

SVM classifier with a linear kernel is used as a baseline system. Our proposed DNN uses three hidden layers (100 nodes per layer) with Adam optimization algorithm [24]. Relu has been chosen as the activation function for hidden layers and softmax function for the output layer. The network is trained for 500 epochs with a learning rate of 0.002. The optimization is done using Adamax algorithm. The tuning of hyper parameters is performed using 10% of data available.

3. Performance Evaluation

3.1. Dataset

A database is created in a studio environment in Malayalam comprising six meters, namely, Dhruthakakali (Dhruth), Kaakali (Kaak) Keka (Kek), Nathonnatha (Nath), Omanakuttan (Oman) and Vasanthathilakom (Vasant) with 403 audio tracks, covering all the meters. The rules for each meters are explained in Table 1. As an example, for Kakali meter, 5 mathras[1] are required in each sequence and 8 such sequences form two lines of a poem. The poems are sung with both male and female singers with a background drone, tanpura[2]. 60% of the dataset is used for training and 10% is used for validation. A total of 120 files, comprising 30% of data are considered for testing making sure that the same singer does not appear in both training and testing sets.

3.2. Experimental Set-up

Initially, MTFs are fused at feature level (11 dim) with MFCC (20 dim) and classification is done using DNN. Track-level computed MTF are fused with frame-wise computed MFCC averaged across dimensions (20 dim). Later, the experiment is extended with an early fusion of i_{MFCC} (10 dim) and MTF (11 dim). i_{MFCC} and the track-level MTF are computed using Aliźe open source speaker recognition tool kit [25] and MIRToolbox [19] respectively. In the i-vector framework, first, a UBM-GMM model (128 mixture) is built from MFCCs computed from poems other than considered in the experiment. Later, total variability matrix, T is trained using the audio files of the corpus. Baseline-SVM and DNN classifiers are implemented using LibSVM and Keras-TensorFlow, respectively.

[1]Mathra refers to a time-measure with two mathras for long-syllable (guru) and one mathra for short-syllable (laghu)

[2]https://en.wikipedia.org/wiki/Tanpura

Table 2: *Overall accuracy of approaches.*

No	Method	Accr.(%)
1	MFCC + MTF (Fusion) - SVM	81.63
2	MFCC + MTF (Fusion) - DNN	80.83
3	i-vector + MTF(Fusion) - DNN	**86.66**

4. Results and Analysis

The potential of i_{MFCC}-MTF fusion with DNN is analyzed from the results, tabulated in Table 2. From the table, it can be seen that for the baseline SVM, an overall accuracy of 81.66% is reported. The overall accuracy of 80.83% and 86.77% are reported for MFCC-MTF fusion and i_{MFCC}-MTF fusion, respectively. The baseline SVM and DNN on MFCC-MTF fusion give almost similar performance. But, for the i_{MFCC}-MTF fusion, it appears that overall accuracy is improved by 6% and 5% over MFCC-MTF-DNN and SVM respectively. It is already established that the hidden variables (i-vectors) in GMM supervector spaces estimated by factor analysis (FA) provides better discrimination ability and lower dimensionality than GMM supervectors.

Confusion matrices for MFCC+MTF fusion and i_{MFCC} + MTF fusion with DNN are shown in Tables 4 and 5, respectively. The entries in the table are the number of files. DNN ensures 80% class-wise accuracy for all the classes except for Omanakuttan. It can be seen from Table 5, that the overall results improved. Two meters Keka and Nathonnatha reported classification accuracy of 100%, but for Omanakuttan , class-wise accuracy is slightly decreased. A possible cause for the misclassification errors in Omankuttan is the piece-wise similarity poems, with other meters while rendering. The precision (P), recall (R) and F1-measure (F1) of the all phases can be seen in Table 3. Average precision, recall and F1 measure of 0.84, 0.82 and 0.83 are reported for SVM framework. The same metrics obtained for i_{MFCC}-MTF fusion on DNN are 0.87, 0.87

Table 3: *Precision (P), recall (R), and F1 measure for the three phases*

SL.No	Meter	MFCC+MTF-SVM			MFCC+MTF-DNN			i_{MFCC}+MTF-DNN		
		P	R	F1	P	R	F1	P	R	F1
1	Druthakaakli	0.84	0.55	0.67	0.85	0.85	0.85	0.86	0.90	**0.88**
2	Kakali	0.62	0.80	0.70	0.73	0.80	0.76	0.73	0.80	**0.76**
3	Keka	0.95	0.90	0.92	0.74	0.85	0.79	0.87	1.00	**0.93**
4	Nathonnatha	0.83	1.00	0.91	0.80	0.80	0.80	1.00	1.00	**1.00**
5	Omanakuttan	0.77	0.85	0.81	0.88	0.75	0.81	0.78	0.70	**0.74**
6	Vasanthathilakom	1.00	0.80	0.89	0.89	0.80	0.84	1.00	0.80	**0.89**

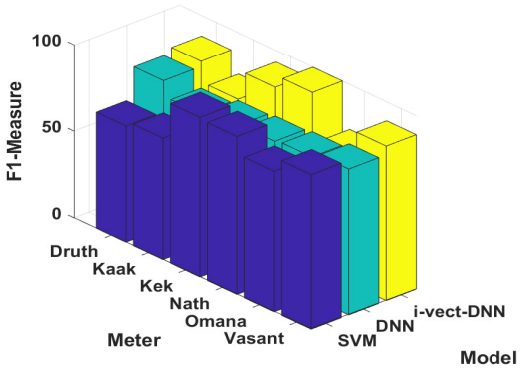

Figure 5: *F1-measure (scaled) for three schemes*

and 0.87 respectively. There results show the potential of DNN on i_{MFCC} over traditional machine learning approaches. The F1 measure of various schemes are shown in Figure 5. The efficacy of i_{MFCC}-MTF fusion can be seen from by comparing methods in Figure 5.

Table 4: *Confusion matrix of MFCC+MTF-DNN framework*

Class	Dhruth	Kaak	Kek	Nath	Omana	Vasant
Dhruth	**17**	2	0	1	0	0
Kaak	0	**16**	1	2	1	0
Kek	0	2	**17**	1	0	0
Nath	0	1	3	**16**	0	0
Oman	1	1	1	0	**15**	2
Vasant	2	0	1	0	1	**16**

The approach in [8] reports an accuracy of 69% for 12 classes, but leads to more syllabification and classification errors. It is observed that a recall of 92% is reported for poem classification (four classes) using stylometric features such as lexical and syntactic features [6]. The performance of the system can potentially be improved by incorporating mesoscale qualities such as phrase and sub-phrase quantifiers, temporal features, bottleneck features from long short-term memory (LSTM) and n-gram identification of rhythmic and melodic motives. Long-range temporal dependencies can be captured by Long short-term memory recurrent neural networks (LSTM-RNNs). The promise of data augmentation methods using deep convolutional generative adversarial networks (DCGAN) can be employed to overcome the data-scarcity during training. MTF used in the experiment is a subset of various timbral, rhythmic and melodic features. Features such as jitter, dynamic features, the salience of the most substantial peak in the beat-histogram and syllable feature can also be used as acoustic cues in the front-end.

Table 5: *Confusion matrix of i_{MFCC}+MTF-DNN framework.*

Class	Dhruth	Kaak	Kek	Nath	Oman	Vasant
Dhruth	**18**	0	0	0	2	0
Kaak	1	**16**	2	0	1	0
Kek	0	0	**20**	0	0	0
Nath	0	0	0	**20**	0	0
Oman	1	4	1	0	**14**	0
Vasant	1	2	0	0	1	**16**

Using a measure of melodic structure in music, Menninghaus et al. show that individual poems feature distinct and text-driven pitch and duration contours, just like songs and other pieces of music [26]. Poems impose higher prosodic regularity on language by virtue of implementing special metrical patterns. The study reveals many musical properties of poems and proves the close link between poetry and music. As different from Western tradition, many traditions follow singing style in poetry, such as Al-Taghrooda [27], Hausa [28] and Palestinian poetry [3]. Al-Taghrooda poems are composed and recited by men travelling on camelback through desert areas of the United Arab Emirates and the Sultanate of Oman. Short poems of seven lines or less are improvised and repeated between two groups of riders often as antiphonal singing. Generally, the lead singer chants the first verse, and the second group responds. It is worth mentioning that no language-specific features are used in our experiment. As part of the extended work, annotations can also be made available in the dataset for further study on poetry. The proposed work is relevant to music information retrieval community since the theory and the methods can potentially be extended beyond Indian poetry.

5. Conclusion

Poetic meter classification using the fusion of i-vectors and MTF using DNN is addressed. The efficacy of i-vectors on meter estimation is manly investigated in the work. The systematic evaluation is done using six meters from poetic corpus in Malayalam. While the MFCC-MTF fusion experiment using DNN resulted in an overall accuracy of 80.83%, the i_{MFCC}-MTF resulted in an overall accuracy of 86.66%. It shows an improvement of 5% over baseline SVM. The results show the potential of the i-vector/MTF frmework in poetic meter classification task.

6. References

[1] H. R. Tizhoosh, F. Sahba, and R. Dara, "Poetic features for poem recognition: A comparative study," *J. Pattern Reco. Research*, vol. 3, no. 1, pp. 24–39, 2008.

[2] A. Namboodiri, P. Narayanan, and C. Jawahar, "On using classical poetry structure for Indian language post-processing," *in Proc.*

of Int. Conf. on Document Analysis and Reco., vol. 2, pp. 1238–1242, 2007.

[3] D. Sbait, "Debate in the improvised-sung poetry of the palestinians," *Asian Folklore Studies*, vol. 52, no. 1, 1993.

[4] L. Morgan, R. K. Sharma, and A. Biduck, "Croaking frogs: A guide to Sanskrit metrics and figures of speech," *Createspace Independent Publishing Platform*, 2011.

[5] D. Genzel, J. Uszkoreit, and F. Och, "Poetic statistical machine translation: Rhyme and meter," *in Proc. of the Conf. on Empirical Methods in Natural Lang. Proces.*, pp. 158–166, 2010.

[6] G. Rakshit, A. Ghosh, P. Bhattacharyya, and G. Haffari, "Automated analysis of bangla poetry for classification and poet identification," *in Proc. of 12th Int. Conf. on Natural Language Processing*, pp. 247–253, 2015.

[7] N. Jamal, M. Mohd, and S. A. Noah, "Poetry classification using support vector machines," *J. of Computer Science*, vol. 8, no. 9, pp. 1441–1446, 2012.

[8] S. Hamidi, F. Razzazi, and M. P. Ghaemmaghami, "Automatic meter classification in Persian poetries using support vector machines," *in Proc. of IEEE Int. Conf. on Signal Processing and Information Technology*, pp. 563–567, 2009.

[9] R. Delmonte, "A computational approach to poetic structure, rhythm and rhyme," *in Proc. of the First Italian Conf. on Computational Linguistics*, pp. 144–150, 2014.

[10] E. Greene, T. Bodrumlu, and K. Knight, "Automatic analysis of rhythmic poetry with applications to generation and translation," *in Proc. of the Conf. on Empirical Methods in Natural Lang. Proces.*, vol. 4, no. 10, pp. 524–533, 2010.

[11] R. Rajan and A. A. Raju., "Deep neural network based poetic meter classification using musical texture feature fusion," *in Proc. of 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain,*, vol. 52, no. 1, 2019.

[12] R. Rajan and A. A. Raju, "Poetic meter classification using acoustic cues," *in Proc. of Int. Conf. on Signal Processing and Communications (SPCOM), Bangalore, India, 2018"*, pp. 31–35, 2018.

[13] G. Richard, S. Sundaram, and S. Narayanan, "An overview on perceptually motivated audio indexing and classification," *in Proc. of the IEEE*, vol. 101, no. 9, pp. 1939–1954, 2013.

[14] J. Zhong, W. Hu, F. Soong, and H. Meng, "DNN i-vector speaker verification with short, text-constrained test utterances," *in Proc. of the Ann. Conf. of the Int. Speech Communication Association, Interspeech*, pp. 1507–1511, 08 2017.

[15] B. Elizalde, H. Lei, and G. Friedland., "An i-vector representation of acoustic environments for audio-based video event detection on user generated content," *in Proc. of IEEE Int. Symposium on Multimedia*, pp. 114–117, 2013.

[16] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing,*, vol. 19, pp. 788–798, 2011.

[17] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. on Speech and Audio Processing*, vol. 13, pp. 345–354.

[18] R. Panda, R. Malheiro, and R. P. Paiva, "Musical texture and expressivity features for music emotion recognition," *in Proc. of ISMIR*, pp. 383–391, 09 2018.

[19] O. Lartillot, P. Toiviainen, and T. Eerola, "A matlab toolbox for music information retrieval," *in Preisach C., Burkhardt H., Schmidt-Thieme L., Decker R. (eds) Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg*, pp. 261–268, 2008.

[20] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," *in Proc.of the 26th Annual Int. ACM Conf. on Research and development in information retrieval*, pp. 282–289, 2003.

[21] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Tran. on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[22] O. Lartillot, T. Eerola, P. Toiviaine, and J. Fornari, "Multi-feature modeling of pulse clarity:design, validation, and optimization," *in Proc. of the 9th Int. Conf. on Music Information Retrieval.*, pp. 521–526, 2008.

[23] Y.-C. Chen, "A tutorial on kernel density estimation and recent advances," *Biostatistics and Epidemiology*, vol. 1, no. 1, pp. 161–187, 2017.

[24] G. Dahl, "Deep learning approaches to problems in speech recognition, computational chemistry, and natural language text processing," PhD Dissertation, University of Toronto, 2015.

[25] J.-F. Bonastre, F. Wils, and S. Meignier, "Aliźe, a free toolkit for speaker recognition," *in Proc. of the Ann. Conf. of the Int. Speech Communication Association,Interspeech*, vol. 1, pp. 737–740, 01 2005.

[26] W. Menninghaus, V. Wagner, C. Knoop, and M. Scharinger, "Poetic speech melody: A crucial link between music and language," *PloS one*, vol. 13, no. 11, p. 1:5, 2018.

[27] "Al-taghrooda, traditional bedouin chanted poetry," *https://en.unesco.org/silkroad/silk-road-themes/intangible-cultural-heritage/al-taghrooda-traditional-bedouin-chanted-poetry,Accessed on 06 May 2020*.

[28] "Hausa poetry and songs," *http://aflang.humanities.ucla.edu/language-materials—/hausa-poetry-song/". Accessed on 06 May 2020*.