Advances and Applications in Mathematical Sciences
Volume 21, Issue 9, July 2022, Pages 5517-5530
© 2022 Mili Publications

HOTELLING'S *T*-SQUARE AND BAYESIAN STATISTICS HYBRID AS AN INVESTIGATIVE TOOL

HIMANSHU CHAURASIYA and GIRISH CHANDRA

Department of Electronics and Communication
J. K. Institute of Applied Physics and Technology
University of Allahabad
Prayagraj, U.P. 211002, India
E-mail: hcha.jk@gmail.com
saras.sharda@gmail.com

Abstract

Spectrogram patch segmentation with noise classification is one of the most challenging and complex tasks in acoustic image (feature) analysis environments, especially in deep learning. Although all convolutional neural networks extract speech features (attributes) from the acoustic spectrogram itself, Hotelling's T-square and Bayesian statistics hybrid is used as an exploratory tool for segmentation and classification with better accuracy. To validate this, the paper expands with an experimental analysis. In this analysis, n-gram $(n = 2)$ language model was used for noise modelling and the noise degradation was eliminated with the decision. Frame classification scores of 64.92%, 67.17%, and 69.49% were obtained using this statistical hybrid tool with 10, 20, and 30 frames, respectively. Best performance was achieved with the longest patch.

Introduction

Audio, speech and sound are all branches of acoustics (Talaske, [1]). Acoustic scene analysis (ASA) is in demand nowadays (Imoto, [2]). It focuses on acoustic image analysis. The spectrogram patch is a major/principal component in this analysis. Spectrogram patch segmentation (SPS) with noise classification using this component is one of the most challenging and complex tasks of acoustic feature learning (Latif et al., [3]) environments, especially in deep learning. SPS is shown in Figure 1. While SPS deals with acoustic signal processing, noise classification on the other hand is subject to statistical models such as Gaussian mixing or hidden Markov model (HMM).

2020 Mathematics Subject Classification: 62C10, 62H35, 68T07.
Keywords: Baysian statistics (BS), Convolutional neural network (CNN), Gaussian, Spectrogram patch segmentation (SPS).
Received September 6, 2021; Accepted May 7, 2022

Like a spectrogram, other time-frequency (t-f) representations also exist.

Spectrogram feature learning and its associated analysis play an important role in any representational learning environment such as a convolutional neural network (CNN). SPS with noise classification is a complex and extremely challenging task in CNN (Qian et al., [10]). Although CNN extracts speech features from the acoustic spectrogram itself, Hotelling's T-square and Bayesian statistics hybrid is used as an exploratory tool for segmentation and classification with better accuracy. Prominent statistician Harold Hotelling was the father of Hotelling's T-square statistics (Hotelling, [6]), while Bayesian statistics was indebted to the famous statistician Thomas Bayes (Schoot et al., [4]). Multivariate analysis (or hypothesis testing) cannot be imagined without Hotelling's T-square statistics considerations. It is basically a generalization of Student's t-statistics.

Machine learning subjected dimensionality reduction (Meng et al., [12]) and correlation enrichment are the key features and beauty of this hybrid. As a principal component analysis (PCA) (Jolliffe et al., [9]) tool, it is suitable for large datasets that are becoming increasingly widespread. Canonical correlation analysis (CCA) (Bae et al., [13]) is another most popular multivariate statistical analysis grateful to Hotelling. Unlike Hotelling's T-square statistics, Bayesian statistics provide parameter estimation along with data analysis. Inspired by the increase in data dimensions with sample number requirements and large-scale applications, Bayesian statistics has taken full advantage of the development of new techniques focused on deep learning. Certainly, the future of applied statistics awaits the reception of multidimensional and multivariate statistical data analysis, with Hotelling's T-square and Bayesian statistics-hybrid as an exploratory tool.

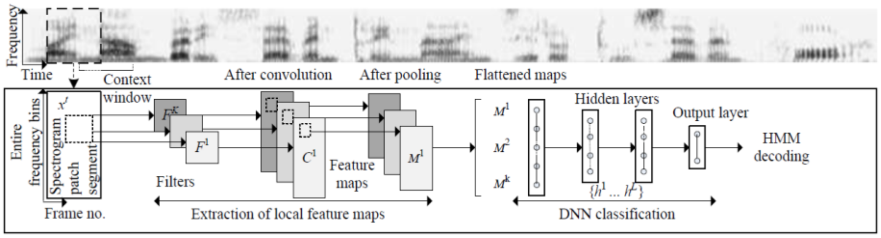

Figure 1. Spectrogram patch segmentation.

Related Work

Long before the world of Bayesian statistics and Harold Hetling's *T*-square, the concept of incidence was introduced as a statistical tool called probability. Bayes's theorem came after the probability tool in this series. Before Hotelling's *T*-square, Bayesian statistics (BS), which was actually necessary and important for any applied statistics, was first revealed by Thomas Bayes in 1763 (Schoot et al., [4]). The oldest (1901) literature relating to the principal component comes from Pearson (Pearson, [5]). In 1933, Hotelling H. was the first to provide the analytical concept of a statistically complex variable related to the principal component (Hotelling, 1933 [6]). 1936 was the year when Hotelling H. proposed the important concept of making a relationship between two sets of observable random variables (Hotelling, 1936 [7]). The degrees of freedom were first recognized by C. F. Gauss in 1821, later developed by W. S. Gossett and R. Fischer as the Student's-t statistical distribution. During 1958, it was Anderson who elaborated on this (Anderson, [8]). The concept of dimensionality reduction was first introduced with Pearson, and further developed with Hotelling H. In addition, in 1992 Rao focused more on modern statistics (as an essential tool) and on Fisher's pioneering work (Rao, [14]). Continuing the reference to dimensionality reduction, the canonical problem relating to high-dimensional data (raw image) was successfully described by J. B. Tenenbaum in 2000 (Tenenbaum et al., [11]). In 2016, Jolliff strengthened this by providing a review on it (Jolliffe et al., [9]). CCA became popular as a form of multivariate analysis when researchers were faced with more complex data, such as spectrograms (or other t-f representations) (Bae E. et al., [13]). This concept continued to evolve with various PCA approaches (Meng et al., [12]).

Segmentation

The pattern of speech segmentation should be optimal, robust and threshold free. For that purpose, the Hotelling's *T*-square (multivariate t-distribution) and Bayesian statistics hybrid is used.

Algorithm

It starts with model selection.

$$BS(A_k) = \log P(I \mid A_k) - \lambda G$$

Where, BS stands for Bayesian statistics,

A_k denotes the model, where $k = 1, 2, \ldots, l$

$\log P(I \mid A_k)$ shows the log-likelihood probability.

$P(I_2, I_2, \ldots, I_t \mid A_k)$ implies maximized likelihood probability.

λ denotes a factor related to penalty.

and G refers to the value of this penalty.

The letter I denotes the image data set,

$$I = I_1, I_2, \ldots, I_N \;\; N \to \infty$$

$$BS(A_k) = \log P(I_1, I_2, \ldots, I_N \mid A_k) - \frac{1}{2} d \log N$$

d denotes the number of independent parameters. These parameters are belonging to model set. N represents cepstral vectors succession.

Considering sequence of cepestral vectors (frame-based),

$$X = x_i \in R^d, i = 1, 2, \ldots, N$$

Suppose audio stream is nested with A_1 and A_2 model.

The first model

$$A_1 : X = \{x_i \mid i = 1, 2, \ldots, N\}$$

or, in the form of independent and identically distributed (i.i.d.) single Gaussian,

$$X = x_1, x_2, \ldots, x_N \sim N(\mu, \Sigma)$$

μ refers to the mean value. Σ denotes the variance.

The second model

$$A_2 : X = \{x_i \mid i = 1, 2, \ldots, b\}, b \in (1, N)$$

b shows the time parameter at which changes are occurred for model selection.

$$A_2 : X = x_1, x_2, \ldots, x_b \sim N(\mu_1, \Sigma_1);$$

$$x_{b+1}, x_{b+2}, \ldots, x_N \sim N(\mu_2, \Sigma_2).$$

Where,

$$\{x_i \mid i = 1, 2, \ldots, b\}$$

denotes the number of initial frames.

while,

$$\{x_i \mid i = b+1, b+2, \ldots, N\}$$

denotes the remaining frames. This model does not hold i.i.d. condition

Since, one model uses one Gaussian and the other uses two Gaussians, the difference will be shown as,

$$\Delta BS_b = R(b) - \lambda G$$

R, implies the statics ratio related to maximum-likelihood, and defined as,

$$R(b) = \frac{1}{2}(N \log|\Sigma| - b \log|\Sigma_1| - (N-b)\log|\Sigma_2|)$$

and G is defined as,

$$G = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\log N$$

Therefore,

$$\Delta BS_b =$$

$$\frac{1}{2}(N \log|\Sigma| - b \log|\Sigma_1| - (N-b)\log|\Sigma_2|) - \frac{1}{2}\lambda\left(d + \frac{1}{2}d(d+1)\right)\log N$$

If

$$\Delta BS_b > 0$$

Then segmentation into two parts is possible.

The final decision will be made through,

$$\hat{b} = \arg\max_{1<b<N} \Delta BS_b$$

Where, b denotes maximum-likelihood estimation.

For the two models, Hotelling's T^2 -statistic is reflected as,

$$T^2 = \frac{b(N-b)}{N}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2)$$

Where, T^2 comes from t-distribution. The value of T^2 is distributed according to $N-2$ degrees of freedom (Anderson, [8]).

Symbol Σ represents common covariance matrix.

These statistics are responsible for the likelihood ratio test.

The critical region of this statistic is defined as,

$$T^2 \geq \frac{(N-2)}{N-d-1} F_{d, N-d-1}(\alpha)$$

Where, $d, N-d-1$ refer to degrees of freedom related to point $F_{d,\, N-d-1}(\alpha)$.

Classification

It starts with deep learning, where convolution concepts, pooling and pipeline support CNN (Qian et al., [10]).

Patch input x^t, linear filter F^k (shape $S \times S$), bias b^k and tanh (non-linear function) give convolved feature maps of $i \times j$,

$$C_{ij}^k = \tanh\left(\sum_{m=1}^{S}\sum_{n=1}^{S}(F_{mn}^k x^t_{(i+m-\lfloor \frac{S}{2} \rfloor),\, (j+n-\lfloor \frac{S}{2} \rfloor)}) + b^k\right),$$

where, C_{ij}^k represent the notation of feature maps.

F_{mn}^k, denotes linear filter of order $m \times n$

$x_{\omega,\, \tau}^t$ refers to bin. ω and τ are frequency index and frame index, respectively.

Max pooling gives,

$$M_{ij}^{k} = \max\left(C_{(iP_1:(i+1)P_1),\ (jP_1:(j+1)P_1)}^{k}\right)$$

Where, P_1 and P_2 come from a specific shape, and represent pooling.

Fully connected layers form the rest of the CNN, where sigmoid activation acts with hidden nodes. A typical experimental setup of a CNN is shown in Table 1.

Table 1. A typical experimental setup of a CNN.

A Typical CNN setup	Layer specifications: CNN (1), fully connected (4) and nodes (512)
No. of filters	5, 10, 15
Patch lengths	10, 20, 30
Pooling	1×1
	1×2
	2×1
	2×2
Filter shapes	3×3, 5×5, 7×7

Validation

For verification, SPS with noise classification (SPSNC) framework is used. It is shown in Figure 2. In this frame work any audio acoustic or speech signal is input through microphone, their array or other multi sensor.

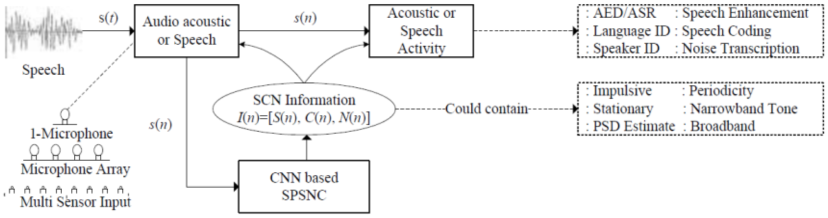
s(t)
Audio acoustic or Speech
s(n)
Acoustic or Speech Activity
: AED/ASR : Speech Enhancement
: Language ID : Speech Coding
: Speaker ID : Noise Transcription
Speech
SCN Information
I(n)=[S(n), C(n), N(n)]
Could contain
: Impulsive : Periodicity
: Stationary : Narrowband Tone
: PSD Estimate : Broadband
1-Microphone
s(n)
Microphone Array
Multi Sensor Input
CNN based SPSNC

Figure 2. SPSNC framework.

In this framework, SCN stands for source (S), channel (C) and noise (N). CNN based SPSNC workflow is represented in Figure 3 (a) as,

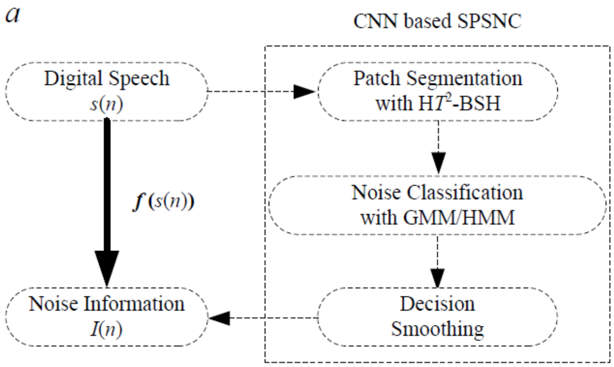

Figure 3 (a)**.** CNN based SPSNC workflow.

The Gaussian noise mixture class set is defined as,

$$S = \{N_1, N_2, \dots, N_\infty\}$$

Where, S denotes an infinite noise space.

or, as a union of noise assignment (A),

$$S = S^{A_1} \cup \dots \cup S^{A_M} \text{ as } M \to \infty$$

Subspaces related to the noise are shown graphically in Figure 3 (b).

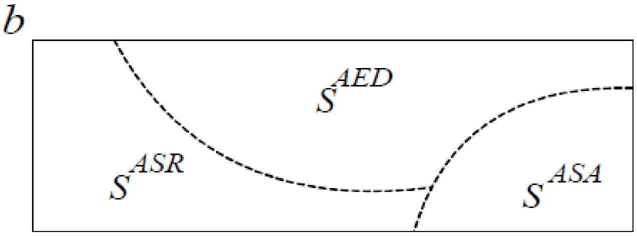

Figure 3 (b)**.** Noise related subspaces.

Where, subspaces include S^{AED} : acoustic event detection, S^{ASR} : automatic speech recognition and S^{ASA} : acoustic scene analysis.

Considering,

$$S^{AED} = \{N_1^{AED}, \ldots, N_{r_{AED}}^{AED}\}.$$

For testing, the conference event was treated as an AED in this framework with three test conditions, where a total of 1.5 h of noise database was collected. MATLAB 2020b was used to calculate 12 dimensional Mel-Fequency Cepstral Coefficient (MFCC), power spectral density (PSD) and others. In addition, same room and same setup was used. The duration of the segment is 1 to 2 seconds. The SPSNC workflow on the spectrogram is shown in Figure 4 (a).

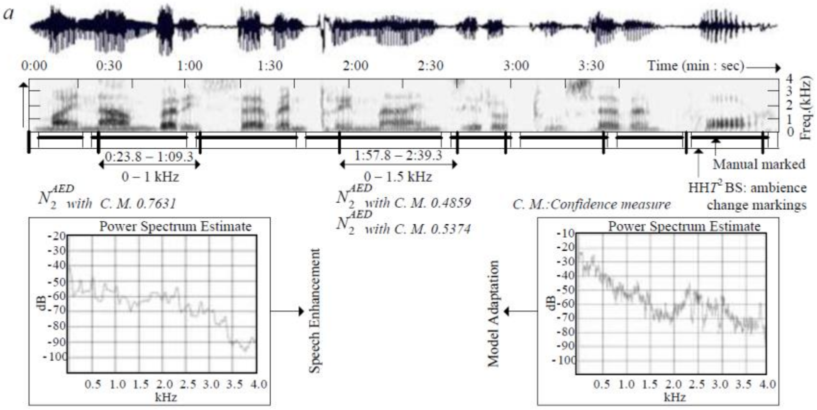

Figure 4 (a)**.** The SPSNC workflow on a spectrogram.

In the Gaussian mixture model (GMM) classification, random noises are represented by r.

Here, $s(i, j)$ denotes class score $s = [i^{th}$ frame block, noise model $\lambda_j]$. The class of noise in GMM is shown in Figure 4 (b).

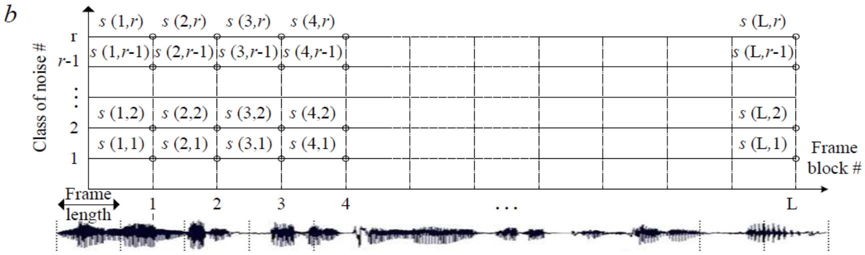

Figure 4 (b)**.** Class of noise in GMM.

Noise assignments related to long-term include A1: idle fan or air conditioning noise (door closed and no movement), A2: Spectators walk without smart phone babbles and lock the door, A3: Spectators walk with smart phone babbles and lock the door, A4: Conference inauguration and door closed, A5: Conference inauguration and the door open, A6: Conference inauguration, door open (half-way), A7: spectators move door open (half-way), and AX: other.

While, short-terms include SR: smart phone-ring, SS: spectator sneeze, PL: personal laugh, PT: person-talk, FR: fart-relax, and IM: impulse. These short-term noises can persist into long-term noises. In this noise evaluation, a 15-ms shift with a 15-ms frame is considered for bigram ($n = 2$) noise modeling.

Results

The average PSD of A is obtained with a scatter plot as shown by the star in Figure 5 (a).

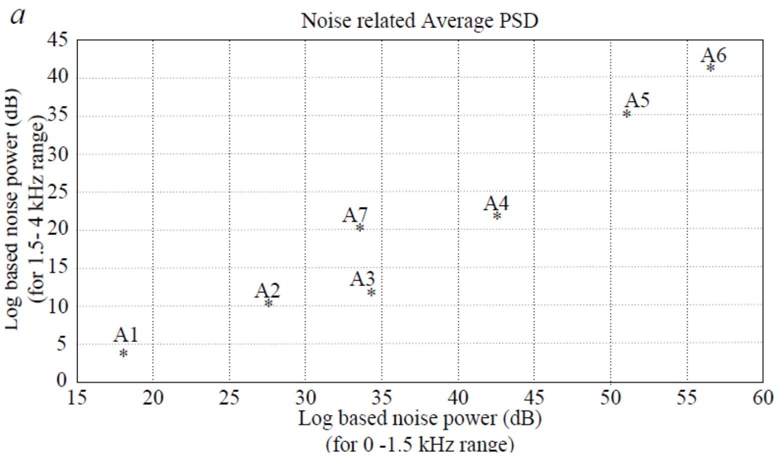

Figure 5 (a)**.** A scatter plot of A.

A1 to A7 transition probabilities are effectively shown in Figure 5 (b). These are mainly achieved with 1.5 hours of complete training.

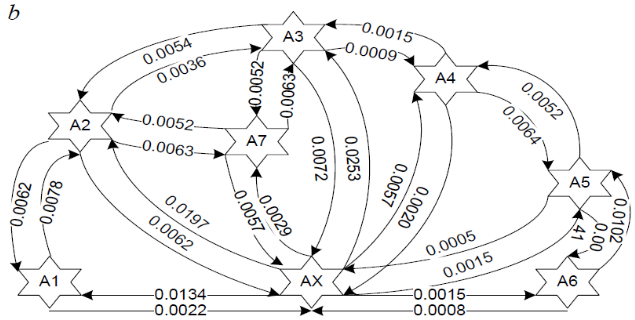

Figure 5 (b)**.** A1 to A7 transition probabilities.

CNN verifies frequency and time pooling with speech spectrogram patches.

Frame classification scores of 64.92%, 67.17%, and 69.49% are obtained using this statistical hybrid tool with 10, 20, and 30 frames, respectively. This is shown in Figure 5 (c).

Best performance is achieved with the longest patch.

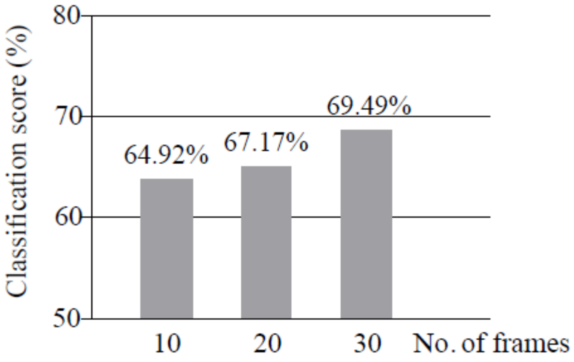

Figure 5 (c)**.** Frame classification scores.

Table 2. A typical comparison with other recent results.

Input	Methods	ACC(%)
MFCC (Geiger et al., [15])	NN	39.0
LLD (Geiger et al., [15])	NN	50.0

MFCC (Geiger et al., [15])	SVM	64.0
LLD (Geiger et al., [15])	SVM	64.0
EBP (Lee et al., [16])	GMM	60.0
CF (Lee et al., [16])	GMM	68.0
MFCC (Stowell et al., [17])	HMM	61.52
MFCC (Bae S. H. et al., [18])	CNN	64.12
MFCC (Mesaros et al., [19])	CNN	64.90
Mel Spec. (Schindler et al., [20])	CNN	64.14
MFCC (Ding [21])	CNN	59.80
MFCC (Wang et al., [22])	CNN	63.60
Mel Spec. (Morató et al., [23])	CNN*	66.20
Mel Spec. (Lee et al., [24])	CNN	67.12
SPSNC	CNN	69.49

***Separable**

CF: Combined features, EBP: Event-based pooling, LLD: Low level descriptor, Spec.: Spectrogram, NN: Neural network, SVM: Support vector machine.

A typical comparison with other recent results is shown in Table 2. This high performance of accuracy is indicating a widespread use of this technique in the ASA of speech quality, as well as in speech intelligibility. Intelligibility is nothing but a dimension of speech quality.

Conclusion

An obvious conclusion is that Hotelling's T-square and Bayesian statistics-hybrid, as an exploratory tool, analyzes deep features present in any complex image such as a spectrogram. In addition, it effectively helps in generating segmentation with noise classification using dimensionality reduction. Thus, as a tool, it will be helpful for advanced multivariate analysis in the near future.

References

[1] R. Talaske, Applying Lindsay's acoustical wheel to architectural acoustics, J. Acoust. Soc. Am. 126(4) (2009), 2268-2268.

[2] K. Imoto, Introduction to acoustic event and scene analysis, Acoust. Sci. Tech. 39(3) (2018), 182-188.

[3] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir and Bjorn W. Schuller, Deep representation learning in speech processing: challenges, recent advances, and future trends, (2020) 01-25. arXiv preprint arXiv:2001.00378.

[4] R. V. D. Schoot, S. Depaoli, R. King et al., Bayesian statistics and modelling, Nat Rev Methods Primers 1(1) (2021), 01-26. doi:10.1038/s43586-020-00001-2.

[5] K. Pearson, On lines and planes of closest fit to systems of points in space, Phil. Mag. 2 (1901), 559-572. doi:10.1080/14786440109462720.

[6] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Edu. Psychol. 24(6) (1933), 417-441. doi:10.1037/h0071325.

[7] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936), 321-377. doi:10.1093/Biomet/28.3-4.321.

[8] T. W. Anderson, An introduction to multivariate statistical analysis, Wiley New York, (1958).

[9] I. T. Jolliffe and J. Cadima, Principal component analysis: a review and recent developments, Philosophical transactions, Series A, Mathematical, Physical, and Engineering Sciences 374(2065) (2016), 20150202. doi:10.1098/rsta.2015.0202.

[10] Y. Qian, M. Bi, T. Tan and K. Yu, Very deep convolutional neural networks for noise robust speech recognition, IEEE/ACM Trans, on Audio, Speech and Lang. Proc. 24(12) (2016), 2263-2276. doi:10.1109/taslp.2016.2602884.

[11] J. B. Tenenbaum, V. D. Silva and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290(5500) (2000), 2319-2323.

[12] C. Meng, O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami and A. C. Culhane, Dimension reduction techniques for the integrative analysis of multi-omics data, Briefings in Bioinformatics 17(4) (2016), 628-641. doi:10.1093/bib/bbv108.

[13] E. Bae, J. W. Hur, J. Kim, J. S. Kwon, J. Lee, S. H. Lee and C. Y. Lim, Multi-group analysis using generalized additive kernel canonical correlation analysis, Scientific Reports 10(1) (2020), 12624. doi:10.1038/s41598-020-69575-x.

[14] C. R. Rao and R. A. Fisher, The founder of modern statistics, Statist. Sci. 7(1) (1992), 34-48. doi:10.1214/ss/1177011442.

[15] J. T. Geiger, B. Schuller and G. Rigoll, Large-scale audio feature extraction and SVM for acoustic scene classification, 2013 IEEE workshop on applications of signal processing to audio and acoustics (2013), 01-04. doi:10.1109/waspaa.2013.6701857.

[16] K. Lee, Z. Hyung and J. Nam, Acoustic scene classification using sparse feature learning and event-based pooling, 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (2013), 01-04. doi:10.1109/waspaa.2013.6701893.

[17] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange and M. D. Plumbley, Detection and classification of acoustic scenes and events, IEEE Trans. on Multimedia 17(10) (2015), 1733-1746. doi:10.1109/tmm.2015.2428998.

[18] S. H. Bae, I. Choi and N. S. Kim, Acoustic scene classification using parallel combination of LSTM and CNN, Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 Hungary, Budapest, September 3 (2016), 01-05.

[19] A. Mesaros, T. Heittola and T. Virtanen, Assessment of human and machine performance in acoustic scene classification: DCASE 2016 case study, 2017 IEEE workshop on applications of signal processing to audio and acoustics (2017), 319-323. doi:10.1109/waspaa.2017.8170047.

[20] A. Schindler, T. Lidy and A. Rauber, Multi-temporal resolution convolutional neural networks for acoustic scene classification, DCASE 2017 Munich, Germany, November, 16 (2018), 01-05. arXiv preprint arXiv:1811.04419v1.

[21] B. Ding, Low-complexity acoustic scene classification using simple CNN, Tech. Rep., DCASE 2019 Challenge, (2019) 01-05.

[22] H. Wang, D. Chong and Y. Zou, Acoustic scene classification with multiple decision schemes, Tech. Rep., DCASE 2020 Challenge (2020), 01-04.

[23] I. M. Morató, T. Heittola, A. Mesaros and T. Virtanen, Low-complexity acoustic scene classification for multi-device audio: analysis of DCASE 2021 Challenge systems, arXiv (2021), 01-05. doi:10.48550/arxiv.2105.13734.

[24] S. Lee, M. Kim, S. Shin, S. Baek, S. Park and Y. Jeong, Ensemble-guided model for performance enhancement in model-complexity-limited acoustic scene classification, Appl. Sci. 12(44) (2022), 01-15. doi:10.3390/app12010044.