4th Asia Pacific Conference of the Prognostics and Health Management,
Tokyo, Japan, September 11 – 14, 2023 R02-12

Improving Anomalous Sound Detection by Distance Matrix-Based Visualization of Measurement Flaws

Nobuaki Tanaka, Takeru Shiraga, and Yusuke Itani

Information Technology R&D Center, Mitsubishi Electric Corporation

ABSTRACT

Although recent DNN-based methods have improved the performance of anomalous sound detection systems, it is still difficult to deploy a system in a real environment without performance degradation. This is often due to measurement flaws such as sensor variability, poor setup, or environmental noise. Since such adverse effects are difficult to model by machine learning, a practical approach to this issue is for humans to identify such flaws and correct them. To this end, we propose a method to visualize measurement flaws as a heatmap based on the distance matrix of the samples in the dataset. This method is designed to find unexpected flaws in the measurement process. Using this method, we were able to identify measurement flaws of anomalous sound detection systems in real production lines. The robustness of anomalous sound detection can be improved by correcting the flaws found by our method.

1. INTRODUCTION

Quality inspection is a crucial process in a production line. For mechanical components like motors or actuators, quality inspections often involve human evaluation of their operating sound or vibration. There have been long-standing efforts to automate this human inspection with sensors like microphones or vibration sensors. Approaches for automating inspection tasks with these sensors range from simple methods based on thresholds for sensor signals to more advanced methods employing machine learning [1–4].

In recent years, researchers have introduced deep learning-based methods for these inspection tasks, reporting increased accuracy for various types of machines [5–9]. However, deploying inspection systems in field environments often leads to subpar performance, even with deep learning-based methods. This is mainly due to measurement flaws such as individual sensor differences, improper measurement procedures, inadequate instrument settings, or unforeseen environmental variations.

While recent research has proposed machine learning models that are robust to various data variations [10–13], these measurement flaws cannot be modeled even by such recent methods due to their high uncertainty. In fact, no mention of these measurement flaws is found in the preceding studies. A more practical approach to enhancing the real-world robustness of inspection systems is for humans to identify such measurement flaws and correct them. To achieve this, we believe it is necessary to have an analytical method to identify measurement flaws from the dataset.

To this end, our study aims to explore ways to visualize and quantify the measurement flaws that have an adverse effect, particularly for anomalous sound (vibration) detection tasks. In this paper, we propose a method for visualization and quantification based on distance matrices from sample data. This method is specifically designed to find unexpected flaws in the measurement process by considering additional labels that are expected to be unrelated to the inspection results. The effectiveness of the proposed method on real measurement datasets is reported through several case studies, in which a variety of mechanical components are inspected with anomalous sound (vibration) detection systems. Moreover, we will present an instance where the visualization obtained from the proposed method enabled to refine the inspection process and resulted in an improved accuracy.

2. RERATED WORKS

Machine learning model studies often neglect to mention techniques for identifying adverse effects in the measurement process. We conducted a search for such techniques and discovered numerous relevant studies within the medical and biotech research fields. In these fields, adverse effects during the measurement process are termed “batch effects,” and methods for detecting and correcting these effects are investigated.

To visualize batch effects, researchers in these fields commonly use multivariate analysis such as principal component analysis (PCA) and linear discriminant analysis (LDA) [14–17]. These techniques are indeed effective. They can be partially helpful also in uncovering measurement flaws when inspecting mechanical components. However, several issues arise when applying these methods to such inspection tasks. First, while these techniques excel at revealing the internal structure of a dataset, they do not necessarily expose negative external effects. The internal structure of a dataset encompasses aspects such as the number of clusters and the distances between them. This information is valuable for estimating the difficulty of a task, but often insufficient for pinpointing measurement flaws. Second, it is hard to predict how outliers will appear in these methods. Outliers provide crucial information for identifying measurement flaws, but if the dominant variance in the entire dataset and the position vectors of outliers are orthogonal in the feature space, these methods might not reveal the outliers.

Regarding quantification of batch effects, methods like surrogate variable analysis (SVA) [18, 19] and guided PCA (gPCA) [20, 21] have been proposed. However, while SVA is suitable for detecting unknown sources of variation, it is not very useful for improving the measurement process because it cannot quantify the magnitude of the adverse effect of a given known factor. For instance, it is difficult to know whether sensor variability or environmental variability is more dominant. In contrast, gPCA can quantify adverse effects of given factors, but it requires many batches to produce reliable results. For instance, a statistically enough sensors are necessary to quantify the negative effect from the variation of the sensors, but many machine inspection tasks employ only a few sensors.

In this study, we introduce a novel method that addresses the problems above and enables both visualization and quantification of measurement flaws.

3. Proposed Method

3.1. Definition of Terms

In this subsection, definitions for the terms used in the subsequent descriptions are provided.

Main- and Sub-label

A label signifies a specific factor. For instance, the quality of an object being inspected can be considered a label. Its value is typically good or bad, or a numerical value that indicates the degree of abnormality. A main-label represents a factor to be determined during the inspection task. It generally corresponds to the quality (e.g., good or bad) of the inspected object. On the other hand, a sub-label represents a factor that is expected to be unrelated to the main-label. For example, a date and time of the measurement, an ID of the sensor used for the measurement, an ID of the person who took the measurement can be sub-labels. A well measured dataset is one in which the main-label has a clear impact while the sub-labels have a small impact.

Sample

A sample refers to a single segment of sound or vibration waveform data. For instance, a single WAV file containing the sound of a motor running for several seconds corresponds to one sample. Each individual sample is assigned a single main-label and multiple sub-labels.

Table 1. The conditions of the filter bank analysis to extract feature vectors from signal waveforms.

Sample rate	Depends on the dataset
Frame length	1024
Frame shift	512
Window function	Hann
Band-pass filters	Equally spaced overlapping triangular filters
Dimension of vectors	17

Dataset

A dataset comprises numerous samples and their associated main- and sub-labels.

Distance

A distance is a similarity value calculated in a certain manner between two samples. In this paper, to compute the distances, we first extracted a series of filter bank feature vectors from the signal waveform of each sample under the conditions outlined in Table 1. Next, we fitted a Gaussian distribution to the series of the feature vectors for each sample. Consequently, a single sample is represented by a single Gaussian distribution. We quantified the distance between two samples by calculating the Bhattacharyya distance between their distributions. The Bhattacharyya distance is often used to measure the divergence between distributions [22, 23]. This distance measure exhibits symmetry and is zero for identical distributions.

3.2. Visualization

First, suppose the dataset comprises N samples $s_1 \dots s_N$, which will be sorted according to a specific rule. Let p_n represents the n-th sample resulting from this sorting. An $N \times N$ matrix $\mathbf{M}$ is constructed such that the element at the row r and the column c is equal to $d(p_r, p_c)$, where $d(p_r, p_c)$ represents the distance between p_r and p_c. The matrix $\mathbf{M}$ serves as the distance matrix for the entire dataset. If the distances exhibit symmetry, $\mathbf{M}$ becomes a symmetric matrix. Additionally, if the distance between identical samples is zero, the diagonal components of $\mathbf{M}$ will be zero.

Next, the N samples are sorted according to a label corresponding to a factor to be visualized. Suppose, for example, there are three sensors: A, B, and C. If the individual differences between them are to be visualized, the samples are sorted in the order of the corresponding sub-label, the IDs of the individual sensors. As a result, the matrix $\mathbf{M}$ will consist of a 3×3 block matrix separated by the sensor IDs as shown in Fig. 1. The visualization of the factor is done by displaying the resulting matrix $\mathbf{M}$ as a heatmap.

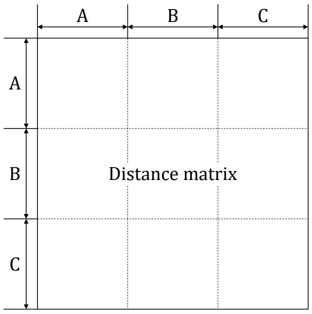
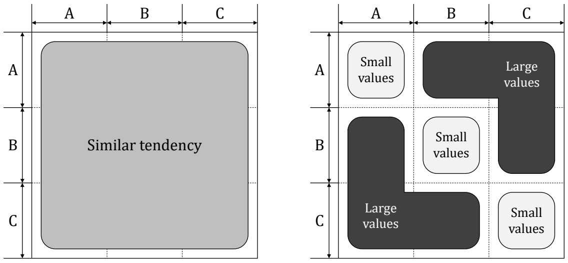


Figure 1. An example distance matrix separated by the sub-label (A, B, and C).



(a) (b)

Figure 2. Tendencies of distance values in the distance matrix.

Fig. 2 (a) shows an example of a case where there is little variation among the three sensors. In this case, each component of the 3×3 block matrix separated by the sensor IDs has a similar range of values. Thus, no special pattern appears in the heatmap. Fig. 2 (b), on the other hand, shows an example of a case where there is clear variation among the three sensors. In this case, the diagonal components of the 3×3 block matrix have small average values because they consist of distances obtained from the same sensors. The off-diagonal components have large values because they are composed of distances obtained from different sensors. As a result, a checkerboard-like pattern with the boundaries that match the 3×3 block matrix appears in the heatmap. Thus, the presence of a checkerboard pattern indicates that the factor has an impact on the dataset. In the absence of such a pattern, the factor has no or small impact.

If the main-label has an impact on the data set, this is a good sign because the measurement process is capturing the difference between good and bad samples. If any of the sub-labels have an impact, there may be unexpected flaws in the measurement process.

The above is the visualization procedure for a single label. If there are more than two labels to be considered, there are several ways to arrange them. One way is to set a priority for each label. In this way, the samples are first sorted by the label with the highest propriety. Samples that are in the same order are sorted by the label with the next highest propriety. We used this method for the visualization examples shown in the following section. Another way is to sort the samples by the single label to be visualized, while samples that are in the same order are sorted randomly.

Table 2. The details of the datasets.

#	Target	Number of samples	Sample rate	Duration of a sample
1	Actuator for door mirror	669 in total 654 good samples, 15 bad samples	40000 Hz	2.3 sec
2	Fishing reel	280 in total 150 good samples, 130 bad samples	40000 Hz	5.5 sec
3	Actuator for air conditioner	1088 in total 1021 good samples, 67 bad samples	25000 Hz	4.4 sec

3.3. Quantification

The visualization method above can be interpreted as follows: if a dataset can be classified by a certain label, then that label has an impact on the dataset. With this idea, the visualization method above can be used to quantify the magnitude of the adverse effects of certain factors.

There are several ways for quantification. In this paper, we use the difference between the average distance of the entire dataset and the average distance within certain clusters, with reference to LDA. When a distance matrix is generated for a label L as described above, the average distance of the diagonal components of the block matrix can be interpreted as the within-class variance $V_{intra}(L)$. Similarly, the average distance of the off-diagonal components can be interpreted as the between-class variance $V_{inter}(L)$. The difference of these values can be expected to represent the possibility of classification like the score function of LDA. We therefore quantify the impact of the factor corresponding to the label L with the following formula:

$$E(L) = V_{inter}(L) - V_{intra}(L) \tag{1}$$

The higher this value, the greater the impact of the corresponding factor on the data set. This value allows the magnitude of the adverse effect of each factor to be compared.

If impact of a given sub-label is greater than that of the main-label, the inspection task will not go well. In this case, the measurement procedure must be corrected to reduce the effect of such a problematic factor.

4. CASE STUDIES

In this section, we report some of the results of our proposed method as applied to several datasets from real production lines. Details of the datasets presented below are shown in Table 2.

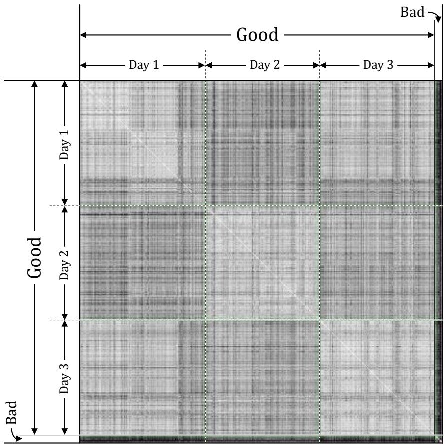


Figure 3. The heatmap obtained from the dataset #1.

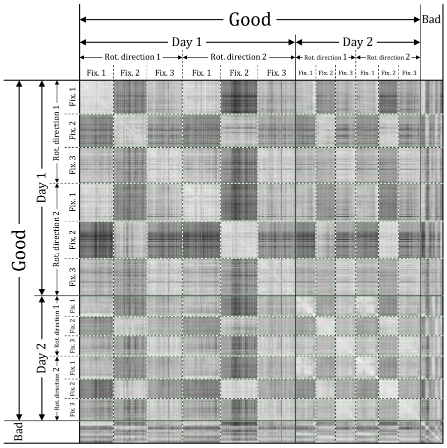


Figure 5. The heatmap obtained from the dataset #3.

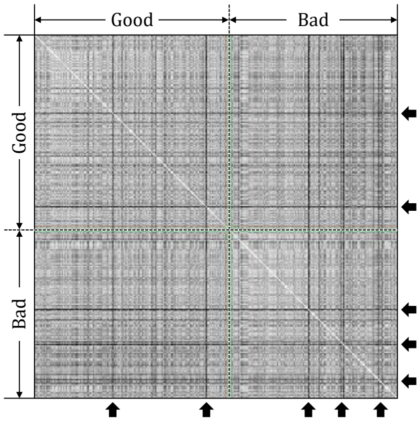


Figure 4. The heatmap obtained from the dataset #2. The allows on the right and bottom sides of the heatmap indicate outliers.

4.1. Actuator for door mirror (dataset #1)

First, a heatmap obtained from a relatively well measured dataset #1 is shown in Fig. 3. This dataset was obtained by measuring the vibration of actuators with a vibration sensor. In this heatmap, the samples are ordered first by the main label (good or bad) and then by the sub-label corresponding to the date of recording. Regarding the colors in the heatmap, white represents zero. The distance increases as the color approaches black.

The region on the heatmap separated by the main-label is colored close to white for the region of “good” and close to black for the region of “bad” (note that the narrow areas on the right and bottom are the “bad” areas). This is a good sign, since it implies that the main-label has a large impact on the dataset. On the other hand, each region separated by the sub-label (the date recorded) has a diluted checkerboard pattern, which might be a bad sign. Overall, however, the main-label has a stronger impact. Thus, the inspection system for these actuators showed a reasonably good performance.

4.2. Fishing reel (dataset #2)

Next, an example of a heatmap obtained from a relatively poorly measured dataset #2 is shown in Fig. 4. This dataset was obtained by measuring the vibration of fishing reels with a vibration sensor. In this heatmap, the samples are ordered by the main-label (good or bad).

In this case, we can see that the inspection task will be difficult because the boundaries of the main-label do not form a clear checkerboard pattern. To achieve high inspection accuracy for this dataset, a model capable of complex modeling, such as a Gaussian mixture model or a neural network, was needed. Furthermore, even with a well-tuned model that could handle the complexity, it was difficult to achieve practical accuracy.

It is also noteworthy that the outliers in the dataset are represented as straight lines in the heatmap. Methods such as PCA do not always visualize outliers, but our method is better in terms of outlier detection because it can always visualize outliers.

From this, we learned that the measurement setup for this task needed to be reviewed, such as where to mount the sensor on the object or how to secure the object.

4.3. Actuator for air conditioner (dataset #3)

Fig. 5 shows a heatmap for a dataset where the measurement failed completely. This dataset was obtained by measuring the vibration of fishing reels with a vibration sensor. The samples are ordered first by the main-label and then by the sub-labels: the date recorded, the rotation direction of the actuator, and the individual fixture used to secure the object.

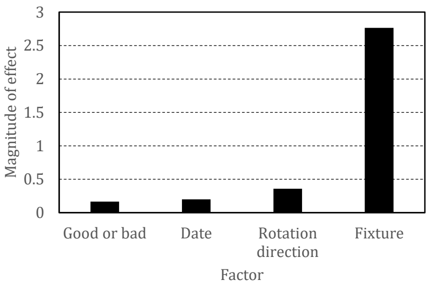


Figure 6. The quantification results from the dataset #3.

The heatmap in Fig. 5 shows a clear checkerboard pattern for the sub-labels, indicating a significant negative impact on the measurement. In addition, there is no clear difference between the regions for good and bad samples. In fact, no machine learning models worked at all on this dataset.

Fig. 6 shows quantification results of the factors corresponding to the sub-labels available in the dataset. These values were obtained with the proposed method described in the section 3. They indicate how much the factors affect the dataset. This figure shows that all the factors other than the main-label (good or bad) have more impact than the main-label. Thus, the inspection task will not go well. The individual differences in the fixtures used to secure the inspection target have a particularly large impact. Therefore, to improve this situation, it is necessary to first address the individual differences in the fixtures.

4.4. Creation of a user interface

From the case studies above, we concluded that the proposed method is effective in finding measurement flaws.

To make the method usable by non-experts, we created a user interface that automatically performs these visualizations and quantifications. This interface allows users to easily generate heatmaps and bar-charts of the quantification results from the dataset. Fig 7 shows the schematic diagram of the interface. Using this interface, the user can specify how the samples are sorted by the priorities of the labels. In this section, we report the usefulness of the interface to improve the accuracy of inspection tasks.

A non-expert user obtained a heatmap shown in Fig. 7 using the interface. The user's dataset consists of vibration signal samples of another type of actuators for air conditioner. The dataset contains 30 samples in total, 20 good samples, and 10 bad samples. These samples were measured with two different vibration sensors.

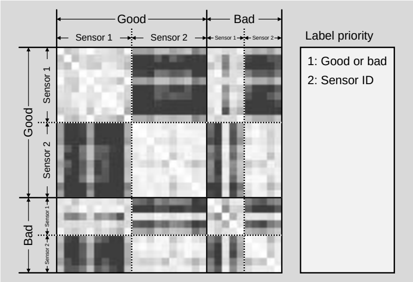


Figure 7. The schematic diagram of the interface. The heatmap in the figure is from the vibration dataset of actuators for air conditioner.

At first, the inspection system for the actuators with a machine learning model did not work well. The user thereby checked the heatmap and found that the heatmap shows a large variability between the two sensors. To address this variability, the user trained two distinct models for each sensor. With these models, the accuracy of the task improved enough to replace the human evaluation.

5. Discussion

In the preceding sections, we have been able to demonstrate the effectiveness of our proposed method. In all cases, our method provided useful information for improving the measurement process on the datasets where machine learning models could not achieve satisfactory performance.

Normally, this type of information cannot be obtained from 2D scatter plots visualized with conventional methods such as PCA or LDA. This is mainly because the data variance caused by measurement flaws is often buried within the overall data variance, making it difficult to detect visually.

Moreover, unlike methods such as SVA, our proposed approach can quantify the extent to which specific factors affect the measurement process. This makes it more useful for pinpointing the causes of performance degradation. Furthermore, unlike gPCA, which requires a significant number of batches because it performs PCA based on the average vector of each batch, our proposed method provides reliable results even with a limited number of batches (e.g., only two batches: good and bad). This is because our method calculates statistical measures from the many samples within each batch. This avoids the need for a large number of batches.

However, in the dataset #2, although several outliers were visualized as measurement flaws, the reasons for these outliers remain unclear. This is due to the lack of assigned sub-labels in this case, leaving only main labels to use. To prepare for such a situation, we believe that all applicable sub-labels should be assigned as much as possible during data collection.

6. Conclusion

We proposed a method to visualize and quantify the measurement flaws in anomalous sound detection tasks. Through several case studies, we confirmed the effectiveness of our proposed approach. Furthermore, we presented a case where the accuracy of anomalous detection system was improved with the proposed method.

References

[1] S. Nandi, H. A. Toliyat, and X. Li, “Condition monitoring and fault diagnosis of electrical motors—A review,” *IEEE Trans. Energy Convers.*, vol. 20, no. 4, pp. 719–729, 2005.

[2] W. Zhou, T. G. Habetler, and R. G. Harley, “Bearing condition monitoring methods for electric machines: A general review, in *Proc. IEEE Int. Symp. Diagnostics Electr. Mach., Power Electron. Drives*, Sep. 2007, pp. 3–6.

[3] Z. Feng, M. Liang, and F. Chu, “Recent advances in time-frequency analysis methods for machinery fault diagnosis: A review with application examples,” *Mech. Syst. Signal Process.*, vol. 38, no. 1, pp. 165–205, 2013.

[4] S. Riaz, H. Elahi, K. Javaid, and T. Shahzad, “Vibration feature extraction and analysis for fault diagnosis of rotating machinery—A literature survey,” *Asia Pacific J. Multidiscip. Res.*, vol. 5, no. 1, pp. 103–110, 2017.

[5] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, “Optimizing acoustic feature extractor for anomalous sound detection based on neyman-pearson lemma,” in *Proc. EUSIPCO*, 2017, pp. 698–702.

[6] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, “Unsupervised Detection of Anomalous Sound based on Deep Learning and the Neyman-Pearson Lemma,” *IEEE/ACM Trans. on Audio Speech and Language Processing*, pp.212–224, 2019.

[7] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, “ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection,” in *Proc. WASPAA*, Oct. 2019, pp. 313–317.

[8] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, “MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection,” in *Proc. DCASE*, Oct. 2019, pp. 209–213.

[9] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, Y. Kawaguchi, “Anomalous Sound Detection Based on Interpolation Deep Neural Network,” in *Proc. ICASSP*, May. 2020, pp. 271–275.

[10] G. Wichern, A. Chakrabarty, Z.-Q. Wang, and J. Le Roux, “Anomalous sound detection using attentive neural processes,” in *Proc. WASPAA*, 2021, pp. 186–190.

[11] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, “ToyADMOS2: another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions,” *arXiv preprint arXiv:2106.02369,* 2021.

[12] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, “Mimii dg: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task,” *arXiv preprint arXiv:2205.13879*, 2022.

[13] S. Venkatesh, G. Wichern, A. Subramanian, and J. Le Roux, “Disentangled surrogate task learning for improved domain generalization in unsupervised anomalous sound detection,” DCASE2022 Challenge, Tech. Rep., 2022.

[14] J. Luo *et al.*, “A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data,” *Pharmacogenomics J.* (2010) 10, pp. 278–291.

[15] C. Lazar *et al.*, “Batch efect removal methods for microarray gene expression data integration: a survey,” *Brief. Bioinform.* vol. 14, no. 4, pp. 469–490, 2012.

[16] D. J. McCarthy, K. R. Campbell, A. T. K. Lun, and Q. F. Wills, “Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R,” *Bioinformatics*, 33(8), 2017, pp. 1179–1186.

[17] L. H. Nguyen and S. Holmes, “Ten quick tips for effective dimensionality reduction,” *PLoS Comput. Biol.* 15(6): e1006907, 2019.

[18] J. T. Leek and J. D. Storey, “Capturing heterogeneity in gene expression studies by surrogate variable analysis,” *PLoS Genet.* 3(9): e161, 2007.

[19] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, “The sva package for removing batch effects and other unwanted variation in high-throughput experiments,” *Bioinformatics* vol. 28, no. 6, 2012, pp. 882–883.

[20] S. E. Reese *et al.*, “A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis,” *Bioinformatics* vol. 29, no. 22, 2013, pp. 2877–2883.

[21] J. M. Franks, G. Cai, and M. L. Whitfield, “Feature specific quantile normalization enables cross-platform classification of molecular subtypes using gene expression data,” *Bioinformatics*, vol. 34, no. 11, 2018, pp. 1868–1874.

[22] C. H. You, K. A. Lee, and H. Li, “An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition,” *IEEE Signal Process. Lett.* vol. 16, no. 1, pp. 49–52, 2009.

[23] N. T. Vu *et al.*, “A first speech recognition system for Mandarin-English code-switch conversational speech,” in *Proc. ICASSP*, Mar. 2012, pp. 4889–4892.