Research on Deep Sound Source Separation

Y. Yang
School of Electronics and Computer Science University of Southampton, UK

H. Li
School of Computer and Information Technology Beijing Jiaotong University, China

ABSTRACT
The cocktail party effect is a fundamental problem in sound source separation, and many researchers have worked to solve this problem. In recent years, the most popular algorithms to solve the problem of sound source separation are Support Vector Machine (SVM), Gaussian Mixture Model (GMM), non-negative matrix factorization (NMF), and Variational Autoencoder (VAE). Especially VAE model showed excellent ability in dealing with the problem of sound separation. In this paper, the β-VAE model, combined with a weakly supervised classification proposed by Karamatli et al., was first reproduced. Since Karamatli’s experiment only completed the connection between sound and words, in order to learn more information about the speaker, this model is used to learn a mapping between sounds and individual speakers and a mapping between sounds and gender. It turns out that the separation results could be obtained by retraining the model after the establishment of the new ‘male’ and ‘female’ labels. His result lays a foundation for the future study of the mapping between individuals and words. When the tag is specific to an individual, more data is needed to support this experiment, and the more data available for training, the better result the model will get.

CCS CONCEPTS
• Computing methodologies → Machine learning; Learning paradigms; Unsupervised learning; Source separation.

KEYWORDS
Sound Separation, VAE, Machine Learning, Neural Network

ACM Reference Format:
Y. Yang and H. Li. 2021. Research on Deep Sound Source Separation. In 2021 13th International Conference on Machine Learning and Computing (ICMLC '21), February 26–March 01, 2021, Shenzhen, China. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3457682.3457741

1 INTRODUCTION
Sound source separation is the process of separating certain audio signals from a combination of signals into a single source. Separating audio signals is a very common problem in audio signal processing and the processing of a mixed sound signal is considerably more complex than a single sound signal, both in terms of effectiveness and accuracy.

A famous example in this field is the cocktail party problem. This scenario states that many people are talking in a determined area, as it is expected in a cocktail party, with a very noisy surrounding environment. A person in such an environment needs to find the voice of a particular person, listen to what they are saying, and subsequently establish communication with them. The human auditory system has the ability to choose one voice or sound over another. This ability helps people to ignore the surrounding noise while focusing their attention on a certain conversation and therefore, get the desired or useful information from someone in a particularly noisy environment. This phenomenon reflects the sophistication of the human auditory system which has been developed through thousands of years, and this complexity highlights why it is a considerably difficult problem to address when this ability is attempted to be reproduced by machines. In order to achieve sound source separation, several scholars have put forward different methods and models, but to a large extent, the problem is still in the stage of development.

Sound source separation for real case scenarios is very common. For example, many applications are currently using the function of acquisition and identification of music to help deaf people find the sound they need to focus on hearing in a noisy environment. Therefore, the objective of this work is to obtain or recover the original source signal from mixed or combined ones through a series of different processes. After getting the separated signal, it is necessary to compare with the source signal to see the quality of separation. In addition, for signals of separation, we want to know more about the source itself, such as the gender of the speaker or, more specifically, who owns the voice. These questions are the research contents of this paper.

1.1 Problem Statement
According to the research of Evangelista et al. (2011), sound source separation refers to the technology of extracting a single sound source signal from a given mixed dataset [1]. The input data contains a combination of mixed signals, which is followed by a procedure that involves the identification of the estimated source signals as this allows for the signals to be superimposed together to achieve redirection. In this case, the number of mixed channels used is usually one, two, or more, however, mixed channels of more than five are rare [2]. There should be no less than two sources of audio for the dataset to be considered as mixed, and the general range goes from two to ten. Even though the number of channels can exceed ten sources, this can be more complicated. The concept of the signal source is somewhat vague. For example, a cello, a viola or a violin can be considered as a separate sound source, but they can also be a part of the sound source of orchestral instruments. In
In this work, two audio sources are randomly combined to produce a mixed channel.

In the development of audio source separation, several models and methods have been proposed which can be generally divided into a supervised and unsupervised learning algorithm. A supervised learning algorithm is a machine learning algorithm that makes predictions using a trained dataset consisting of input values and corresponding output values. The size of the trained dataset determines the predictive ability of the model, that is, the accuracy of sound source separation. Among these methods, the typical ones are the Support Vector Machine (SVM), the Gaussian Mixture Model (GMM), and the Denoising Autoencoder (DAE) [3]. Unsupervised learning algorithm [4] is also a machine learning algorithm that can find some hidden relationships in the unmarked data to get the correct separation of sound signals. Among these methods, the typical ones are Independent Component Analysis (ICA), Sparse Coding, and Non-negative matrix factorization (NMF) [5]. Non-negative matrix factorization and Denoising Autoencoder are popular methods [6, 7] for the separation of sound signals, however, the latter requires a large set of source signal training to ensure its correct performance which, in many cases, is difficult to achieve.

Between the two previous classifications of learning algorithms, there is another type of method called weak supervised learning algorithm. This algorithm does not require a highly trained set of labels and is more specific than abstract unsupervised learning. Therefore, it can also be used in sound source separation [8].

Autoencoder (AE) is an artificial neural network which can learn the efficient representation of input dataset through unsupervised learning. The task of the autoencoder is to input and output the training data after simple learning. This may sound trivial, but when constraints are added to the neural network in different ways, this task can be extremely difficult. This process can be applied to the pre-training of deep neural network.

As the encoding mode of AE is fixed, it is difficult to tackle any challenges, so a new model called Variational Autoencoder (VAE) [9] was created. This model adds “Gaussian noise” to the encoder results based on the autoencoder, so that the decoded results can be robust to noise. The additional loss generated in the process is a regular term for the encoder, assuming that every encoder comes out with a zero mean. Besides the VAE model, there are many changes and optimizations, with one example being β-VAE. In particular, Karamatlı et al. (2019) proposed a weakly supervised learning approach based on a β-VAE that performs as well as those VAE modeled on each class.

To learn more about the speaker, an idea of establishing a voice to speech mapping. This can be said by anyone. As the sound signal is very abstract, some processing is needed before the system outputs just a "yes", which can be said by anyone.

All the articles quoted [10] in Google academic. In [11] this model is used to conduct clustering research. In [12–14] Weak Label was further studied through this model. In [15] optimizing the author’s experimental process.

In the article by Karamatlı, his team conducted two experiments, one comparing the performance of their proposed model with other models, and the other comparing the performance of the model with a different number of classes. The results show that the proposed model performs well in separation, but only so. To address this limitation, a new attempt is made in this paper based on the experiments of Karamatlı et al. In this study we first reproduced the model proposed by Karamatlı et al. and ran the code to determine if the performance was good as described in their study. Since this model only learns a mapping between sounds and words but does not learn a mapping between sounds and individual speakers. To learn more about the speaker, an idea of establishing a voice to gender mapping came up. Gender is input into the model as a new label for training. The success of this step lays the foundation for the establishment of a speaker and voice mapping.

2 METHODOLOGY AND DATA

2.1 Proposed Methodology

Figure 1 shows the VAE working principle for this work. This principle can be divided into two types: the ones that have the source signals for supervision (Figure 1A) and the ones that do not have access to the source signals for supervision but only for their classes (Figure 1B). The parameters are explained in Table 1.

In the first case, the function $G_{i|x}$ represents the encoding process, while the function $F_{i|y}$ represents the decoding process. The source signal is encoded and decoded to generate a new corresponding signal, and then the difference between the source signal and the newly generated signal is compared.

In the second case, because we can only know the class of Source signal, a parameter $h$ is added. When $h=0$, this means that the sound signal of this class is not included in the mixed dataset. On the other hand, when $h=1$, this means that the sound signal of this class is included. However, the second case requires the mixed data to hold different classes. Then, based on the second case and assuming that $Z$ obeys some distribution, since VAE is the generation model, it is necessary to ensure that the new $X_i$ is generated from $Z_i$, so one VAE is assigned to each class.

Because we use additive components to decompose the mixed audio, this model is very similar to the NMF algorithm. The difference between the two algorithms is the following: in NMF, components are represented using a linear template, while in this model, components are represented using a non-linear and more expressive neural network model. Furthermore, in this model, each encoder only focuses on the influence factors of the source signal related to itself and does not consider the influence factors related to other source signals in the mixed audio. The decoder also learns to reconstruct the source signal related to it.

As the sound signal is very abstract, some processing is needed to present it visually. Short-Time Fourier transformation (STFT) [16] is often used in the spectral analysis of sound signals, which
Table 1: Parameter interpretation in Figure 1

| Parameter | Description |
|-----------|-------------|
| X         | Magnitude spectrogram of mixture audio. |
| ϕ         | Neural network parameter. |
| K         | Number of class. |
| g_{ϕ_k}(X) | The neural network encoder pair for the k^{th} source class. |
| Z_K       | Potential variable of k^{th}. |
| f_{ϕ_k}(Z_K) | The neural network decoder pair for the k^{th} source class. |
| S_K       | Magnitude spectrogram of separated audio. |
| S_1       | Magnitude spectrogram of original audio. |
| h_K       | Whether class K exists. If h=0, does not exist; if h=1, exists. |
| X̂         | Magnitude spectrogram of reconstructed mixture audio. |

presents the sound information in the form of pictures to facilitate the observation of the frequency spectrum signals changing along the time. Poisson’s distribution [17] is very common in the analysis of magnitude spectrograms. Using the β-VAE model assume a Poisson distribution:

\[
p_\theta(X^{(i)}|z) = PO(X^{(i)}|\lambda = \hat{X}^{(i)})
\]  

(1)

To train this model, the loss function is:

\[
L(\cdot) = -\beta \frac{1}{2} \sum_{k,j}(1 + \log(\sigma_{k,j}^{(i)})^2 - \mu_{k,j}^{(i)})^2 - \sigma_{k,j}^{(i)})^2
\]

\[+ D_{GKL}(\hat{X}^{(i)} \parallel \hat{X}^{(i)})
\]  

(2)

Where K stands for category and J stands for a potential unit. Because \(p_\theta(z)\) and \(q_\phi(z|x)\) are based on the gaussian hypothesis, the first item KL is the divergence. This is a pre-constraint on the potential units because it encourages independence not only between the potential cells but also between the source signals. This constraint provides an effective regularisation effect.

Finally, a product neural network (CNN) [18] is employed, which can be used for a spectrum diagram with a size of \(T \times F\), where \(T\) represents time and \(F\) represents frequency. The encoder structure is shown in Table 2, where Convolutional represents the convolutional layer, Fully-Connected represents the full connection layer, and Gauss represents the Gauss potential output layer. The whole encoder-decoder network is therefore symmetric.

2.2 Data

The data used in this paper is the Speech Commands Dataset (SCD) [19, 20], which contains 35 labels and a total of 105,829 audio data. The audio comes from different speakers, and each piece of audio content is a commonly used word that is the label to which the audio belongs. The specific label contents of data and the number of pieces of data contained in each label are in Table 3. This dataset is published under the Creative Commons BY 4.0 license [21].

For privacy reasons, all contributors need to sign an agreement before recording that allows others to use the audio data. When collecting contributors’ audio data, the collector asks them to be in a room alone during the recording, and not to have any personal information in the background. In addition, contributors are not required to enter any personally identifiable information, such as name or gender, during the recording process. Because contributors are required to record in a separate room, which is usually quiet, the clarity of the audio data will be high. Such data are too ideal to exist in real life. More natural audio should have some noise in the background. So the collector created some artificial noise and mixed it with audio data to get closer to real life data.
Table 2: Encoder Network Architecture

|                  | Convolutional 1 | Convolutional 2 | Convolutional 3 | Fully-Connected | Gauss |
|------------------|-----------------|-----------------|-----------------|-----------------|-------|
| Filters/units    | 128             | 128             | 256             | 512             | 128   |
| Filter size      | 1×F             | 4×1             | 4×1             | -               | -     |
| Stride           | 1×1             | 2×1             | 2×1             | -               | -     |

Table 3: Labels of Speech Commands Dataset

| Word       | Number | Word   | Number | Word   | Number |
|------------|--------|--------|--------|--------|--------|
| Backward   | 1,664  | Happy  | 2,054  | Sheila | 2,022  |
| Bed        | 2,014  | House  | 2,113  | Six    | 3,860  |
| Bird       | 2,064  | Learn  | 1,575  | Stop   | 3,872  |
| Cat        | 2,031  | Left   | 3,801  | Three  | 3,727  |
| Dog        | 2,128  | Marvin | 2,100  | Tree   | 1,759  |
| Down       | 3,917  | Nine   | 3,934  | Two    | 3,880  |
| Eight      | 3,787  | No     | 3,941  | Up     | 3,723  |
| Five       | 4,052  | Off    | 3,745  | Visual | 1,592  |
| Follow     | 1,579  | On     | 3,845  | Wow    | 2,123  |
| Forward    | 1,557  | One    | 3,890  | Yes    | 4,044  |
| Four       | 3,728  | Right  | 3,778  | Zero   | 4,052  |
| Go         | 3,880  | Seven  | 3,998  |        |        |

3 EXPERIMENTAL EVALUATION AND RESULTS

3.1 Reproduce Karamatlı’s Experiment

It was found that the weakly supervised learning method based on β-VAE and used by Karamatlı et al. could be classified only with the class tags of the audio source while maintaining the same separation effect as the supervised learning method. Thus, to verify this method, it is important to reproduce the experiment showcased in Karamatlı’s paper. The author states that the audio files underwent preprocessing at a sampling rate of 16kHz and this was then changed to 8kHz. Since audio files are recordings from different people, it is inevitable that there will be large and small sounds, hence, the root mean square (RMS) requires normalisation. The RMS of each audio file is normalized to the RMS of the training set, where the RMS of the training set is set as 0.067.

The original file contains 35 labels, hence, to reduce the workload, Karamathony selects 10 labels from number 0 to number 9 for training. Since the data in the dataset are all independent and relatively clear audio sources, the first step to separation is to randomly generate mixed audio sources from these signal sources. Hence, in Karamatlı’s experiment, the mixed audio requires random generation from at least three categories and the number of categories should not exceed 10. The study generated 15,000 training data, 1875 test data, and 1875 validation data. At the end of the training, 20 examples were randomly generated, and each example contained 12 files. Among them, there are two pictures representing the audio spectrum. These two pictures are irrelevant to the results; thus, they will not be discussed in further depth in this research. The remaining ten files are mixed audio signal, separated audio signal 1, separated audio signal 2, source audio signal 1 before mixing, source audio signal 2 before mixing, as well as the magnitude spectrogram corresponding to these five audio signals.

The more similar the two frequencies are, the more similar their distributions will be, and the more closely related their spectrum diagrams will be. Therefore, before playing the separated audio, we can preliminarily judge the separation result by the spectrum diagram of the audio. If the images are very similar, the separation effect should be relatively ideal; in contrast, if the images are very different, the separated audio has a big difference from the audio before mixing, and the separation effect is not good. Since the audio file cannot be shown in this paper, the frequency spectrum of the audio is shown in Figure 2. Figure 2A highlights the good separation results and Figure 2B shows the less ideal separation results.

After three attempts, it was clear that the separation results were all sound. Hence, this indicated that the Karamatlı’s model is able to complete the separation task efficiently. Following on from this, Karamathically completed two experiments, including comparing the efficiency with AE and generating mixed audio with a different number of classes. The results were similar to the ones obtained by the author. Therefore, code reproduction was complete.

One of the problems of this experiment is that the experimental results only include the audio separation results and the spectrum diagram, and the only way to judge whether the separation results are good is by listening to the audio and comparing the spectrum diagram. There is no specific algorithm for calculating model accuracy. To test the accuracy of this model, we generated 1,000 random results and listened to each one. In 46 of these 1,000 cases, the separation is a little fuzzy. In this case, the training accuracy of this model under this set of training data is 95.4%, but this accuracy is not representative.
Figure 2: Compare of magnitude spectrogram of separated signal and source signal

Figure 3: The magnitude spectrogram of 0, 1 and 5. The signals in the figure are from the same speaker, 0 and 1 are quite similar, and 5 is quite different from these two.

In this model, Karamatlı clearly indicated that at least three labels should be selected when generating mixed audio, however, actually, two different pieces of audio can mix a piece of data. Thus, the number of labels set to 2, and then run Karamatlı’s code with randomly selecting three groups of labels, (0,1), (1,5) and (3,5) training data. The results showed that the two labels could also complete the experiment, however, the separation effect could not be guaranteed. Among the three groups of data, the separation results of (1,5) and (3,5) are relatively good, but the separation effect of (0,1) is very poor. Moreover, the separated audio is essentially the same as the mixed audio. Trying to modify some of the parameters did not help the separation of 0 and 1 to a great extent. Specifically, under the label of 0 and 1, the speaker is the same person, which highlights that each one participated in the 0 to record have completed 1. Although the generation of the mixed audio is random, the same people stated that the two words outweigh the risk of other tags, since the sound source separation relies on a difference in the voice frequency distribution. Therefore, in this case, the same speaker voice signals are too similar and the result in the case of limited data is inferior to the larger other mixed sound source gap data. This is also the reason why the two labels (0,1) are not ideal. Figure 3 shows the difference among magnitude spectrogram of 0, 1 and 5.

The comparison results between the traditional Autoencoder model and the $\beta$-VAE model presented are shown in the Figure 4A. Ten other labels were selected to test changes in Source to Distortion Ratio (SDR), Source to Interference Ratio (SIR) and Source
3.2 New experimental attempts

The label used by the Karamati in the experiment is the label given in the data, which are essentially certain common words. This separation result can satisfy the part of finding the desired word within the noisy environment in terms of the cocktail party effect. What if we want to learn more about the speaker? Gender is a good choice.

The data is divided into two labels, male and female. The first step is to deal with the file names in terms of training as in the previous experiment. In this first step, 5000 pieces of mixed audio data were generated for training, and the results from this process indicated that the voices of the females were clearer than the voices of the males. Upon adding the data in the label for the second time, 10000 pieces of mixed audio training data received, and the separation results were very clear for both males and females. Although only two labels are used here to generate mixed audio, due to the large gap between male and female voices in terms of timbre, frequency, and other aspects, there is no unsatisfactory situation like mixing 0 and 1 as what was observed in Experiment 1. Due to the fact that the voice range of women is generally wider than that of men, as highlighted in Figure 5, the distribution of women’s voices is broader than that of men. Therefore, when the data was trained less, men’s voices were easily mixed with women’s forehead voices, thereby, resulting in the mixing of women’s voices when separating men’s voices. This result proves that the Karamati’s code, with some modifications and by searching for certain data, can help to solve the cocktail party issue.

Relabel by gender helps narrow the gap, however, it still cannot be proven specifically by whom. For example, one need to hear Jack say yes to him, and the ‘yes’ that separated may be said by Jack or Amy. The idea was to train the model with a name tag to see if it could still do the same job. In the data set, the number of the same speaker remains identical, and the audio files with the same number are searched for and then placed in a folder. This file is then provided with a new label name. Since contributors are not forced to count when recording, this results in speakers with more audio, speakers with less audio, and even a minimum number of...
one audio data. Therefore, four speakers with more audio as the label were chosen and named as ‘Amy’, ‘Jack’, ‘Mary’ and ‘Tom’. Since the audio was contributed by the same speaker, their audio files have the same name and need to be renamed when they are organised into a folder.

The results show that this model can be used to separate the voice of the designated person, but the effect is not ideal. A large reason for the unsatisfactory separation results is related to the lack of data. Even if more audio from all four speakers is selected, the average amount of audio under each tag becomes only 100. This is far too small in comparison with the 3000-4000 pieces of data under each tag in the original data.

4 CONCLUSION

In this paper, a series of sound separation experiments were conducted under different labels. It was proved that the $\beta$-VAE model and the weakly supervised learning model performed very well in separating common words. Meanwhile, the model could also be used to separate individual speakers’ voices and male and female voices, although the separation effect still needs to be improved. The experimental results show a particular risk when only two labels are selected to generate mixed audio. If the pronouncement of the two labels or the speaker’s voice is very close, the separation result will be affected. The more labels use, the more mixed data can generate to avoid this problem slightly. When the two audio frequencies are very similar, the model’s performance is not as expected, and you can continue to adjust the parameter optimization problem.

When individual speakers were used as the label, the generated results were not generic due to the lack of data. On the one hand, follow-up research could be continued on the one hand, data could be added, and on the other hand, parameters could be modified. When gender is used as a label, the effect is better than before because more data were used. However, to increase the experimental data, we copied and pasted many copies of an audio file, which may lead to an overfitting during training. In general, this experiment can help the model solve the cocktail party problem better.

In the future, more data can be found to solve the problem of insufficient data. When enough data is available, more rigorous and general conclusions can be drawn after training this model. Secondly, as for the similar audio, you can consult more information, adjust the model parameters, and continuously optimize it. Besides, both the Karamatlı’s experiment and my experiment used audio data that contained only one word and was one second in length. In my opinion, since the model can complete the mapping between sound and individual speakers, more kinds of data can be tried, such as phrases or sentences, because real-life conversations are very long. Finally, some research can be done on letting the model calculate the accuracy automatically. All these attempts may help improve this model’s accuracy and effect, which can be tried later.

ACKNOWLEDGMENTS

I want to express my sincere gratitude to Dr. Christine Evers — my supervisor — for her great ideas, constructive advices and useful critics. I also want to thank her for her open and friendly nature that made me feel welcome for any problem at any time. Then I would like to thank my second examiner Dr. Yonne Howard for listening to me patiently about my project and giving me some advice.

REFERENCES

[1] G. Evangelista, S. Marchand, M. Plumbley and E. Vincent, “Sound Source Separation”, DAFX. Digital Audio Effects, pp. 551-588, 2011. Available: 10.1109/9781199912988.ch14
[2] A. Kumar, “Audio Source Separation with Deep Learning”, Medium, 2020. [Online]. Available: https://towardsdatascience.com/audio-source-separation-with-deep-learning-e725089267.
[3] M. Pimpale, S. Therese, and V. Shinde. A survey on: Sound source separation methods. International Journal of Computer Engineering In Research Trends (IJCERT), 3, 2016.
[4] T. Virtanen, "Unsupervised learning methods for source separation in monaural music signals" in Signal Processing Methods for Music Transcription, New York:Springer, pp. 267-296, 2006.
[5] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization", Proc. 13th Int. Conf. Neural Inf. Process. Syst., pp. 555-561, 2001.
[6] F. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription”, Proc. IEEE Workshop Appl. Signal Process. Audio, Acoust., pp. 177-180, 2003.
[7] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders", Proc. IEEE Global Conf. Signal Inf. Process., pp. 1265-1269, Nov. 2017.
[8] N Zhang, J Yan and Y Zhou, "Weakly supervised audio source separation via spectrum energy preserved Wasserstein learning", Proc. 27th Int. Joint Conf. Artif. Intell., pp. 4574-4580, 2018, [Online] Available: https://doi.org/10.24963/ijcai.2018.636.
[9] D. Kingma and M. Welling, ‘Auto-encoding variational bayes” arXiv preprint arXiv:1312.6114, 2013.
[10] F. Karamatlı, A. T. Cemgül and S. Kirbız, ‘Audio Source Separation Using Variational Autoencoders and Weak Class Supervision’, in IEEE Signal Processing Letters, vol. 26, no. 9, pp. 1349-1353, Sept. 2019. doi: 10.1109/LSP.2019.2929440.
[11] K. Lim, X. Jiang and C. Yi, ‘Deep Clustering With Variational Autoencoder’, in IEEE Signal Processing Letters, vol. 27, pp. 231-235, 2020, doi: 10.1109/LSP.2020.2963328.
[12] Q Kong, Y. Wang, X. Song, Y. Cao, W. Wang and M. D. Plumbley, ‘Source Separation with Weakly Labeled Data: an Approach to Computational Auditory Scene Analysis’, ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 101-105, doi: 10.1109/ICASSP40776.2020.9053396.
[13] F. Pishdadian, G. Wichern and J. L. Roux, ‘Learning to Separate Sounds from Weakly Labeled Scenes’, ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 91-95, doi: 10.1109/ICASSP40776.2020.9053055.
[14] a F. Pishdadian, G. Wichern and J. Le Roux, ‘Finding Strength in Weakness: Learning to Separate Sounds With Weak Supervision’, in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2386-2399, 2020, doi: 10.1109/TASLP.2020.3031405.
[15] H. D. Do, S. T. Tran and D. T. Chau, ‘Speech Source Separation Using Variational Autoencoder and Bandpass Filters’ in IEEE Access, vol. 8, pp. 156219-156231, 2020, doi: 10.1109/ACCESS.2020.3019495.
[16] a A. Elbir, H. O. İlhan, G. Serbes and N. Aydın, ‘Short Time Fourier Transform based music genre classification’, 2018 Electric Electronics, Computer Science, Biomedical Engineering’s Meeting (EBBT), Istanbul, Turkey, 2018, pp. 1-4, doi: 10.1109/EBBT.2018.8391437.
[17] H. L. Zhang, J. Liu, T. Li, Y. Xue, S. Xu and J. Chen, ‘Extracting sample data based on poison distribution’, 2017 International Conference on Machine Learning and Cybernetics (ICMLC), Ningbo, China, 2017, pp. 374-378, doi: 10.1109/ICMLC.2017.8108950.
[18] W. Hsu, Y. Zhang and J. Glass, ‘Support vector machines for gender classification’, Proc. Interspeech, pp. 1273-1277, 2017.
[19] P. Warden, ‘Speech commands: A dataset for limited- vocabulary speech recognition’, 2018.
[20] Speech commands dataset version 2. [Online]. Available: http://download.tensorflow.org/data/speech_commands_v0.02.tar.gz
[21] a Creative commons international attribution international 4.0 license, 2018. [Online]. Available: https://creativecommons.org/licenses/by/4.0/