Query by Humming Music Information Retrieval using DNN-LSTM based Melody Extraction and Noise Filtration

Andreas Novian Dwi Triastanto
School of Electrical Engineering and Informatics
U-CoE AI-VLB
Institut Teknologi Bandung
Bandung, Indonesia
andreas.novian@students.itb.ac.id

Rila Mandala
School of Electrical Engineering and Informatics
U-CoE AI-VLB
Institut Teknologi Bandung
Bandung, Indonesia
rila@itb.ac.id

2022 5th International Conference on Information and Communications Technology (ICOIACT) | 978-1-6654-5140-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOIACT55506.2022.9972121

***Abstract*— Search engine technology has become a daily necessity. In terms of music search, the most effective and natural way to perform music search is to hum the song or so-called query by humming. In an experiment to improve the performance of a query by humming system, there are two things that can be done, namely using better melody extraction technique and noise filtration. Better melody extraction technique is needed so that the representation of the melody obtained from the query reflects more on the change of pitch while noise filtration is needed so that this system can work well in the real world where the queries tend to have natural noise. This research will try to implement the technique of DNN-LSTM based melody extraction and noise filtration with Fourier series decomposition and spectral subtraction in query by humming system. The system built will be compared with a query by humming system that uses melody extraction with PRAAT and without noise filtration. The results of this study indicate that the performance of the system that uses melody extraction with PRAAT and without noise filtration is still better in terms of the mean reciprocal rank value, top 1/3/5/10 hit ratio, and the required processing time.**

Keywords*— *melody extraction, music information retrieval, noise filtration, query by humming

I. Introduction

Search engine technology has become a daily necessity. Nowadays, information searches are most commonly performed using text or image queries. For most needs, these two types of queries are sufficient for the search engines to provide the required information. However, for music search, the two types of queries are not suitable. Ghias et al. stated that the most effective and natural way to perform music search is to hum the song you want to search or so-called query by humming [1]. Some recent studies on this topic have managed to achieve accuracy values above 80% with search times below 5000ms, as in [2], [3]. Meanwhile, other studies measured the performance of this system with the mean reciprocal rank and got the best score of 0.59, as in [4], [5], [6].

To build a query by humming system that has better accuracy and can be used in the real world, there are two things that can be done, namely using better melody extraction techniques and performing noise reduction in queries. A better melody extraction technique is needed so that the representation of the melody obtained from the query reflects more on the change of pitch so that when the melody is used in the query search process to the database, the results are expected to be better. Noise reduction needs to be performed on queries recorded in real-world conditions where there is often noise originating from the surrounding environment and from problematic audio recording devices. The noise reduction process performed on the queries used in the query by humming system is expected to produce better queries to be used in the query search process in the database.

Research on different filtering techniques and neural networks methods for reducing noise on speech signals, as in [7], [8], conclude that the filtration method is better at reducing noise in speech signals than the neural network method which is more complex and requires a longer execution time. Recent research on noise filtration techniques using Fourier series decomposition and spectral subtraction, as in [9], [10], has resulted in the conclusion that the proposed technique has better results than several other techniques, namely Wiener filter and spectral subtraction, and has been proven to be used to produce a robust system in the speaker identification domain.

Query by humming is part of the research domain of music information retrieval. Several recent studies in this domain have focused on melody extraction techniques to get the best melody representation of a musical recording. This technique is also one of the techniques needed to build a query by humming system. One of the most recent studies on this topic uses a multi-task learning approach based on deep neural network (DNN) and long short term memory (LSTM), as in [11], [12], [13], [14]. This technique takes a learning approach for two tasks at once, namely pitch estimation and voicing detection. This approach is proven to produce higher accuracy and better generalizability.

In this research, a query by humming system will be built using DNN-LSTM based melody extraction and noise filtration with Fourier series decomposition and spectral subtraction. Matching algorithm that will be used is the unified algorithm, as in [4], which is proven to be able to solve two main problems in humming query matching, namely the problem of differences in musical scales and tempo. The unified algorithm is also simpler and faster because it uses precomputed inverted index and edit distance to measure similarity between query and music in database.

II. Related Works

The earliest technique used in building query by humming system is done in 1995 using pitch tracking method for feature extraction and string matching to search

query in database. Some research has been done using this matching technique, as in [1], [15], [16]. This matching technique uses three symbols, namely U (up), D (down), and S (same). The U symbol is used when the current note is higher than the previous note, the D symbol is used when the current note is lower than the previous note, and the S symbol is used when the current note is the same as the previous note. This technique is notoriously slow when used with large databases.

There are many researches on query by humming system done using dynamic time warping to measure similarity between query and music in database, as in [2], [17], [18], [19]. This algorithm is proven good to be used to build query by humming systems that can handle large databases but have poor performance when dealing with queries that have different keys. Experiment to combine dynamic time warping and location sensitive hash algorithm have good balance between the retrieval accuracy and retrieval rate while improving overall system performance.

One of the techniques for improving quality of sound signal that has been tried for the speaker identification domain is Fourier series decomposition and spectral subtraction, as in [9]. The way this technique works is that the input voice signal is first segmented into several small frames. Then, each frame is decomposed into N harmonics using the Fourier series. Then, each frame is reconstructed by adding the harmonics to get an estimate of the final frame which is expected to have less noise than the original. This process is repeated until all frames have been processed. After that, spectral subtraction is applied to the reconstructed signal to obtain the final result in the form of an improved sound signal. The reason for framing the sound signal before the Fourier expansion is to obtain more detailed signal results because the smaller the signal scale, the more detailed the results will be obtained from the Fourier expansion.

The DNN-LSTM based melody extraction technique is a composition between two functions, namely pitch estimation and voicing detection [11]. The approach used is a multi-task learning approach that performs these two functions simultaneously. Pitch estimation is responsible for estimating the value of the fundamental frequency of a tone, while voice detection is responsible for identifying the presence or absence of a tone. The output of the pitch estimation function is a fundamental frequency in a certain frequency range, while the output of the voice detection function is binary whether there is a tone or not. The experimental results show that learning for both functions simultaneously can improve accuracy and have better generalizability.

III. System Architecture

A. Noise Filtration

Techniques for reducing noise that have been developed are generally used to reduce noise in signals containing speech, as in [9]. Experiments to apply this technique to a query by humming system have never been done before. In general, the process to perform noise filtration on signals containing speech and signals containing humming queries is the same because both the speech signal and the humming queries signal both originate from human vocals and can distinguish between the main signal and the noise.

Noise filtration with Fourier series decomposition and spectral subtraction has been tested to be able to reduce noise commonly known as white noise. White noise used in other research is Additive White Gaussian Noise (AWGN), which means that the noise is added to the clean sound signal so that the resulting noisy signal tends to have constant and measurable noise [9]. In this study, the humming query used had natural noise generated from various sources, such as background noise, static noise from improper grounding on an audio recording device, and air flow sound from the mouth which was also recorded on the microphone. Therefore, the sound signal generated from the noise filtration process in query humming used in this study may not be better than the results obtained in other study [9].

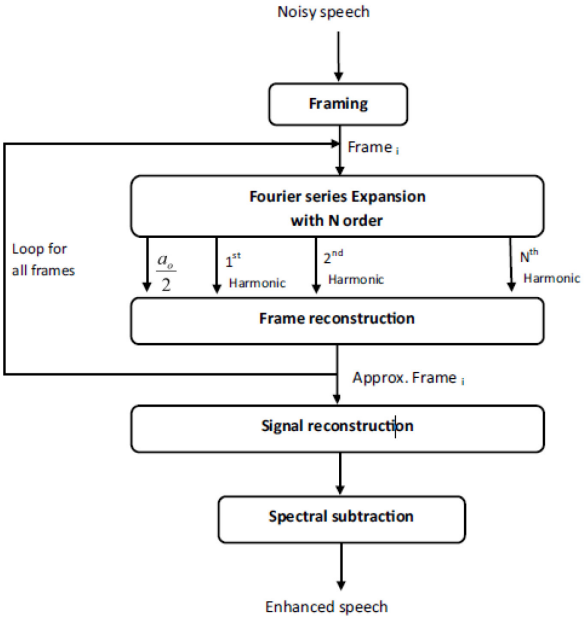

Fig. 1. Noise filtration process with Fourier Series Decomposition and Spectral Subtraction, as in [9]

In the noise filtration process with Fourier series decomposition, the research paper did not stated in their research the best parameters for the number of frames used and the number of harmonics taken for each frame c. That research stated that a clean signal is obtained by taking the first few harmonics which can describe most of the signal with some low frequency noise components, as in Fig. 1. For this study, assuming the sample rate obtained from an average query with a length of 8 seconds is 64000, it then segmented into 640 frames and for each frame the first 10 harmonics will be taken. For the noise filtration process with spectral subtraction, noise samples are needed from each query. In this study, noise samples were taken from the longest unvoiced period for each query. This sample is taken from the part of the query where the fundamental frequency (f0) is not detected using the PRAAT tool.

B. Melody extraction

The output of the DNN-LSTM based melody extraction process is the base frequency for each frame measuring 23 milliseconds. From the observation, it is found that the minimum duration needed for a person to hum a tone is 100

milliseconds or 0.1 second [4], [17]. PRAAT generates the base frequency for each 10 millisecond frame while the DNN-LSTM technique generates the base frequency for 23 milliseconds. From these data, it can be concluded that to hum a tone a minimum of 10 frames from PRAAT and 4-5 frames from DNN-LSTM is required. Only semitones that meet these requirements will be taken for the next process.

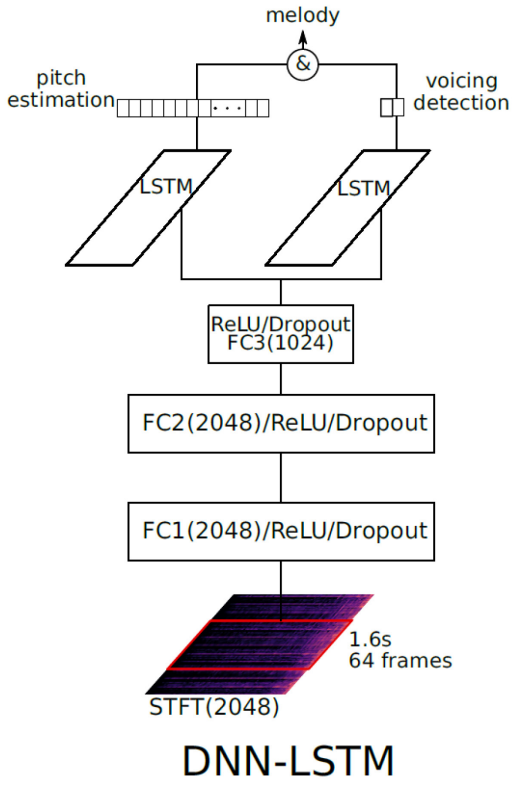

Fig. 2. DNN-LSTM based melody extraction technique, as in [11]

The architecture of this melody extraction system uses a deep neural network (DNN) and a recurrent neural network (RNN) in the form of long short term memory (LSTM). The DNN architecture for both tasks uses three hidden layers with the number of units 2048, 2048, and 1024 with ReLU as a nonlinear function while LSTM is used for each task separately. The model takes several frames from the results of the Short Time Fourier Transform (STFT) on the input audio as an input feature vector and produces an output in the form of two labels for each time lag (timestamp) which is the predicted result of pitch estimation and voicing detection. The two labels are then put together to form a melody prediction in the final stage. An illustration of the architecture can be seen in Fig. 2.

C. *Matching tehnique*

The matching technique that will be used in this research is the unified algorithm [4]. Unified algorithm can solve the problem of differences in pitch that may occur between users, where the same song can be hummed into several different queries because each user uses a different key. This problem can be handled by this technique because it uses the relative distance between pitch and pitch that appear most frequently so that this technique can work well assuming the ups and downs of the pitch used are correct as they should be. In addition, this technique can also solve problems in the form of differences in tempo between users because tempo, which is the time interval between notes, is not considered in the matching process.

The unified algorithm first performs a search using an inverted index structure to match the n-gram pattern, namely relative pitch 4-grams (RP4G), 3-grams (RP3G), and 2-grams (RP2G). When the desired result has not been found, the mode normalized frequency (MNF) algorithm is applied to measure the similarity of the pattern using the edit distance technique. This technique is a pattern matching method that uses Euclidian distance measurements where distance is the minimum number needed to add, delete, or replace a string so that it match with another string. The smaller the distance value obtained, the more similar the query to music in the database is. The unified algorithm eliminates the need for query transpose which in other studies is needed to overcome the problem of differences in the basic tone between queries and databases. This can be achieved because in the n-gram pattern matching algorithm what is seen is the relative distance between tones while in the MNF algorithm this difference is handled by changing the tone representation into letters that are relative to the tone that occurs most often (mode).

IV. Experiments and Analysis

A. *Experimental scenario*

The experimental scenario to be carried out is as follows:

1. The base-line query by humming system uses PRAAT to extract melody, without using noise filtration, and uses a unified algorithm to perform the matching.

2. The query by humming system uses DNN-LSTM based melody extraction, without using noise filtration, and uses a unified algorithm to perform the matching.

3. The query by humming system uses PRAAT to extract melodies, uses Fourier series decomposition and spectral subtraction to perform noise filtration, and uses a unified algorithm to perform the matching.

4. The query by humming system uses DNN-LSTM based melody extraction, uses Fourier series decomposition and spectral subtraction to perform noise filtration, and uses a unified algorithm to perform the matching.

B. *Dataset*

Dataset used in this study is taken from Music Information Retrieval Evaluation eXchange (MIREX) forum. There are two datasets provided in this forum, the first dataset is Roger Jang's MIR-QBSH corpus which consists of 4,431 queries with 48 music in the database while the second dataset is the IOACAS corpus which consists of 759 queries with 298 music in the database. In order to compare the experimental results of the two datasets, a third dataset was created which is a smaller version of the IOACAS corpus with 340 queries and 48 music in the database.

C. *Experiments results*

To measure the quality of the query by humming system, two kinds of measurement methods will be used, namely top 1/3/5/10 hit ratio and mean reciprocal rank (MRR). The experiments results for each dataset can be seen in Fig. 3, Fig. 4, and Fig. 5.

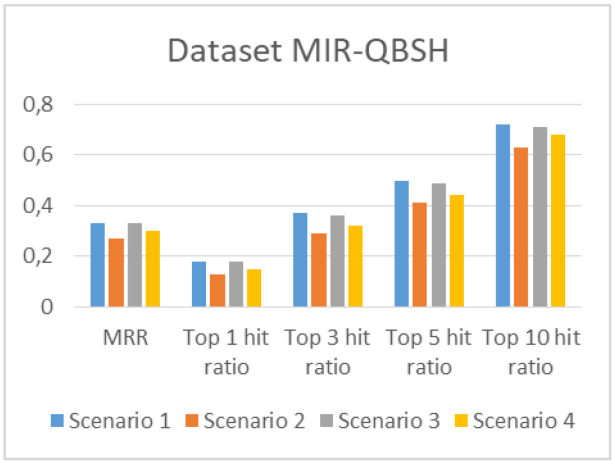

Fig. 3. Experiments results for MIR-QBSH dataset in each scenario

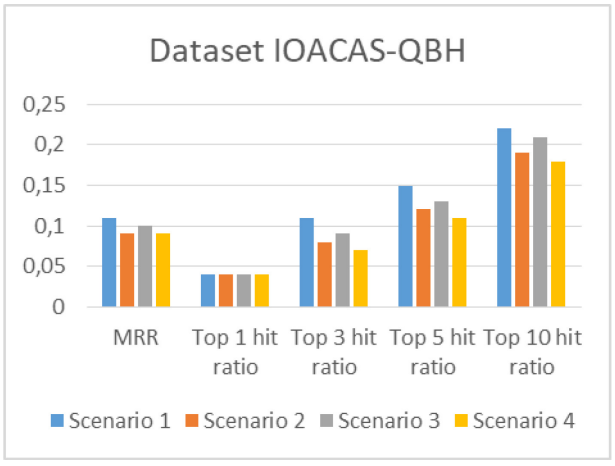

Fig. 4. Experiments results for IOACAS-QBH dataset in each scenario

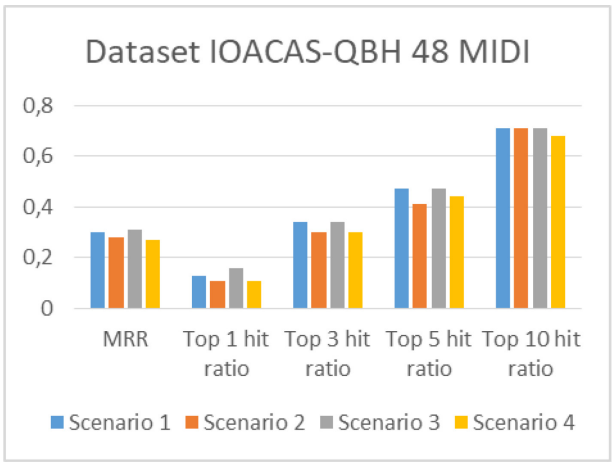

Fig. 5. Experiments results for IOACAS-QBH 48 MIDI dataset in each scenario

Based on the results of experiments that have been carried out, in general there is no experiment that gives better results than the base-line system. Better results are only shown by scenario 3 on the IOACAS-QBH 48 MIDI dataset which can be seen in Fig. 5. In the graph, it can be seen that the MRR value increased by 0.01 points and the top 1 hit ratio value increased by 0.03 points compared to the base-line system. The comparison of the time of each process in each experimental scenario for each dataset can be seen in Fig. 6, Fig. 7, and Fig. 8. In terms of the execution time of each process, there is a decrease in the matching time for the IOACAS-QBH and IOACAS-QBH 48 MIDI datasets which can be seen in Fig. 7 and Fig. 8. But on the other hand, a very significant increase in time can be seen in the melody extraction process in scenarios 2 and 4 in each dataset.

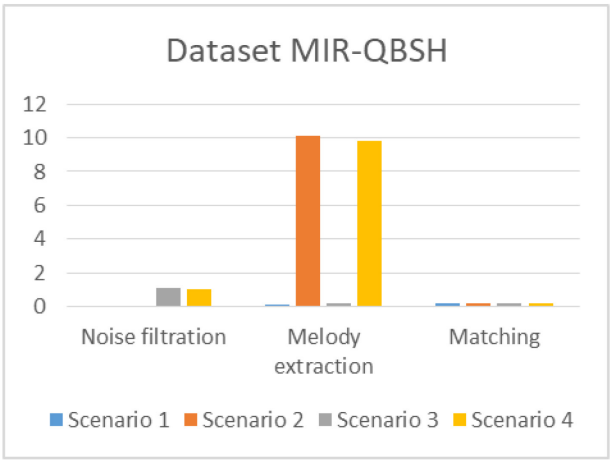

Fig. 6. Comparison of process time for MIR-QBSH dataset in each scenario

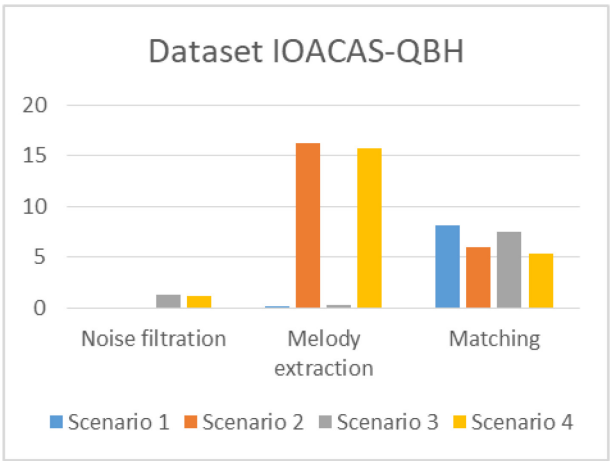

Fig. 7. Comparison of process time for IOACAS-QBH dataset in each scenario

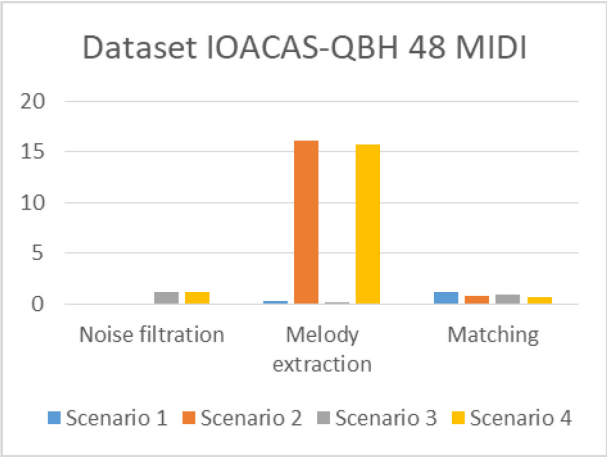

Fig. 8. Comparison of process time for IOACAS-QBH 48 MIDI dataset in each scenario

From the results of experiments that have been carried out, it can be seen that the query by humming system that uses melody extraction with PRAAT and without noise filtration produces the best performance when compared to system that use DNN-LSTM melody extraction, noise filtration with Fourier series decomposition and spectral subtraction (FSD-SS), and the combination of DNN-LSTM melody extraction and noise filtration with FSD-SS. From these results, it can be concluded two things: melody extraction with PRAAT alone is quite effective to be used in a query by humming system without the need to use noise filtration, especially FSD-SS, and DNN-LSTM melody extraction is not as good as melody extraction with PRAAT to be used in a query by humming system. Some of the reasons why these two things can happen are as follows:

Melody extraction with PRAAT use an algorithm called autocorrelation function [20]. In this research, the algorithm used is proven to be able to extract the fundamental frequency of noisy signals quite well. Meanwhile, in the paper that explains DNN-LSTM melody extraction technique, there is no prove of the ability of the algorithm to perform fundamental frequency extraction on noisy signals [11].

Noise filtration, especially spectral subtraction, has the possibility to remove valuable information from a signal. In this study, it was shown that most of the signal could be lost due to errors in noise sampling, so it is possible that in the process there are parts of the signal that should be able to help in the melody extraction process but are lost along with the noise.

Melody extraction with PRAAT uses the autocorrelation technique, which means that the prediction of the fundamental frequency in a frame is carried out based on the calculation of the fundamental frequency value on the frame with a delayed copy of itself. In other words, this technique does not depend on any training data to do its job. Meanwhile, melody extraction with DNN-LSTM uses training data to conduct training on an artificial neural network model so that it is possible to overfitting the model in the training data.

V. Conclusion and Future Works

From the experiments that have been carried out, it can be concluded that melody extraction with DNN-LSTM and noise filtration with FSD-SS did not make better query by humming system compared to system that use melody extraction with PRAAT and without noise filtration. However, noise filtration can improve the performance of system that use melody extraction with DNN-LSTM. There is a very significant increase in time when using melody extraction with DNN-LSTM compared to PRAAT. For further experiment, a system that can detect the quality of a query before the next process can be built to find out whether the noise reduction process needs to be done for the query or not.

Acknowledgment

The authors would like to express our gratitude to School of Electrical Engineering and Informatics, Institut Teknologi Bandung and University Center of Excellence on Artificial Intelligence for Vision, Natural Language Processing & Big Data Analytics (U-CoE AI-VLB) for the support in facility and funding.

References

[1] A. Ghias, J. Logan, D. Chamberlin, dan B. C. Smith, "Query by humming," 1995, hal. 231–236, doi: 10.1145/217279.215273.

[2] S. Zhou, Z. Zhao, P. Shi, dan M. Han, "Research on matching method in humming retrieval," in *Proceedings of 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference, ITOEC 2017*, 2017, vol. 2017-Janua, hal. 516–520, doi: 10.1109/ITOEC.2017.8122349.

[3] W. Zhao, "A high efficient music retrieval algorithm based on content," in *Proceedings - 2016 8th International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2016*, 2016, hal. 12–15, doi: 10.1109/ICMTMA.2016.12.

[4] V. Makarand dan K. Parag, "Unified algorithm for melodic music similarity and retrieval in query by humming," in *Advances in Intelligent Systems and Computing*, 2018, vol. 673, hal. 373–381, doi: 10.1007/978-981-10-7245-1_37.

[5] C. C. Wang dan J. S. R. Jang, "Improving query-by-singing/humming by combining melody and lyric information," *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2015, doi: 10.1109/TASLP.2015.2409735.

[6] S. Gulati, J. Serra, dan X. Serra, "An evaluation of methodologies for melodic similarity in audio recordings of Indian art music," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2015-Augus, hal. 678–682, 2015, doi: 10.1109/ICASSP.2015.7178055.

[7] G. K. Rajini, V. Harikrishnan, M. Jasmin Pemeena Priyadarisini, dan S. Balaji, "A research on different filtering techniques and neural networks methods for denoising speech signals," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 9 Special issue 2, hal. 503–511, 2019, doi: 10.35940/ijitee.I1107.0789S219.

[8] M. Erseven dan B. Bolat, "Regression-based speech enhancement by convolutional neural network," in *26th IEEE Signal Processing and Communications Applications Conference, SIU 2018*, 2018, doi: 10.1109/SIU.2018.8404639.

[9] A. I. Siam, H. A. El-khobby, M. M. A. Elnaby, H. S. Abdelkader, dan F. E. A. El-samie, "A Novel Speech Enhancement Method Using Fourier Series Decomposition and Spectral Subtraction for Robust Speaker Identification," *Wirel. Pers. Commun.*, no. 0123456789, 2019, doi: 10.1007/s11277-019-06453-4.

[10] S. Mavaddaty, S. M. Ahadi, dan S. Seyedin, "A novel speech enhancement method by learnable sparse and low-rank decomposition and domain adaptation," *Speech Commun.*, 2016, doi: 10.1016/j.specom.2015.11.003.

[11] Z. Cao, X. Feng, dan W. Li, "A multi-task learning approach for melody extraction," *Lect. Notes Electr. Eng.*, vol. 635, hal. 53–65, 2020, doi: 10.1007/978-981-15-2756-2_5.

[12] S. Kum, C. Oh, dan J. Nam, "Melody extraction on vocal segments using multi-column deep neural networks," in *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016*, 2016.

[13] R. M. Bittner, J. Salamon, S. Essid, dan J. P. Bello, "Melody extraction by contour classification," in *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015*, 2015.
[14] H. Park dan C. D. Yoo, "Melody extraction and detection through LSTM-RNN with harmonic sum loss," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017, doi: 10.1109/ICASSP.2017.7952660.
[15] R. J. McNab, L. A. Smith, I. H. Witten, C. L. Henderson, dan S. J. Cunningham, "Towards the digital music library: Tune retrieval from acoustic input," in *Proceedings of the ACM International Conference on Digital Libraries*, 1996.
[16] L. Tao, H. Xianglin, Y. Lifang, dan Z. Pengju, "Query by humming: Comparing voices to voices," in *Proceedings - International Conference on Management and Service Science, MASS 2009*, 2009, hal. 5–8, doi: 10.1109/ICMSS.2009.5305356.
[17] R. A. Putri dan D. P. Lestari, "Music information retrieval using Query-by-humming based on the dynamic time warping," *Proc. - 5th Int. Conf. Electr. Eng. Informatics Bridg. Knowl. between Acad. Ind. Community, ICEEI 2015*, hal. 65–70, 2015, doi: 10.1109/ICEEI.2015.7352471.
[18] P. Koirala, M. Chapagain, N. Pantha, dan N. B. Adhikar, "Effects of Auto Tuning and Pitch Normalization on Query by Humming," vol. 01, no. 02, hal. 1–6, 2018.
[19] J. Li, J. Han, Z. Shi, dan J. Li, "An efficient approach to humming transcription for query-by-humming system," in *Proceedings - 2010 3rd International Congress on Image and Signal Processing, CISP 2010*, 2010, doi: 10.1109/CISP.2010.5646801.
[20] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, 1993.