International Journal of Advanced Culture Technology Vol.12 No.3 206-212 (2024)
DOI https://doi.org/10.17703/IJACT.2024.12.3.206

IJACT 24-9-25

Research on Machine Learning Rules for Extracting Audio Sources in Noise

Kyoung-ah Kwon*

Lecturer, Dept. of Global Media, Soong-sil Univ., Korea
E-mail kyounga.kwon@gmail.com

Abstract

This study presents five selection rules for training algorithms to extract audio sources from noise. The five rules are Dynamics, Roots, Tonal Balance, Tonal-Noisy Balance, and Stereo Width, and the suitability of each rule for sound extraction was determined by spectrogram analysis using various types of sample sources, such as environmental sounds, musical instruments, human voice, as well as white, brown, and pink noise with sine waves. The training area of the algorithm includes both melody and beat, and with these rules, the algorithm is able to analyze which specific audio sources are contained in the given noise and extract them. The results of this study are expected to improve the accuracy of the algorithm in audio source extraction and enable automated sound clip selection, which will provide a new methodology for sound processing and audio source generation using noise.

Keywords: *Audio Source Extraction, Noisy Environment, Training Criteria, Sound Generation, Machine Learning*

1. INTRODUCTION

The research on machine learning criteria for extracting sound from noise is as underdeveloped as the research on noise itself. Most of the research focuses on automating the extraction of melodic loops or drum samples from given audio [1] or developing algorithms that can detect specific music samples in sound databases [2], and even when spectrograms are used, they are mostly concerned with improving the quality and intelligibility of speech signals by reducing noise [3-4] or inferring the source of each sound by measuring a set of characteristics of different types of noise signals in an aggregated set of environmental sounds [5].

The purpose of this research is to propose a selection criterion for algorithms to discover audio sources from sound noise. Spectrogram transforms are at the core of many audio signal processing methods, from sound source separation and modification to noise removal, and are used to reconstruct natural sound signals. In this study, we first identified candidate selection criteria that can be universally used to extract sound sources from noise, and then converted and analyzed various types of noise samples into spectrograms to verify the suitability of each candidate as a sound source extraction criterion. Finally, based on the results derived from this series of experiments, we proposed algorithm application criteria and learning methods for each element.

Manuscript received: May 25, 2024 / revised: June 21, 2024 / accepted: September 1, 2024
Corresponding Author: kyounga.kwon@gmail.com
Tel: ***-****-****
Lecturer, Dept. of Global Media, Soong-sil Univ., Korea

2. SELECTION RULES AND EXPERIMENTAL DATA

2.1 Training Data

For the experimental data, we mainly selected original noise and near-noise sources, and also added musical instruments and human voices to verify the ability to analyze non-noise sounds. In addition, we included 100% noise sources, namely white noise, pink noise, and brown noise, to predict how the non-noise values would change and set a baseline accordingly, and we also added sine waves in the second round to understand how the algorithm basically distinguishes between noise and tonal sounds.

Table 1. List of test data

Categorization	Data 1(1st)	Data 2(2nd)	Sound Types
Noise	Noisey	Noise-white	Pure noise
		Noise-pink	
		Noise-brown	
	Bed		Daily noise
	Crackleice		Natural noise
		Osc-saw	Tool sound
Tonal	Airhorn	Osc-sine	Pure tones
		Osc-triangle	Musical instrument
	Hey		Human voice (short)
	Makesomenoise		Human voice (long)

2.2 Candidates for Analysis

The criteria for selecting rules were factors that allow the algorithm to extract notes that are 'balanced for human hearing' and 'within the range of human recognition'. Although the rule selection process did not consider the difference between 'beat' and 'melody,' we added 'Tonal-noisy balance' as a complementary criterion to distinguish between beat and melody, given that 'Roots' are more often utilized in melodies than in beats. The five rules that were selected to train the algorithm, and the reasons for their selection, are listed below.

1. Dynamics: In music, dynamics refers to the variation in loudness when playing or singing a piece of music, both the difference in volumes and the speed that the difference produces.
2. Roots: The term for the fundamental notes that make up a particular chord, which are the loudest frequencies in a sound with pitch. Since humans recognize notes based on their roots, it is also essential for algorithmic learning.
3. Tonal Balance: Used to distinguish between areas of noise and areas where a particular timbre is emphasized. This involves complex frequency and time analysis, in the same vein as measuring the similarity of tendency to Pink Noise.
4. Tonal-Noisy Balance: This refers to the balance between melodic (tone) and noise in music. This concept plays an important role in contemporary, electronic, and experimental music, where adjusting the balance of tone and noise can be used to adjust the mood of the music and diversify the auditory experience.
5. Stereo Width: This refers to the spatial position (spatial information, that is, spacing) of the sound, and this property is essential because sounds has both temporal and spatial characteristics (unless the

sound source was captured using a mono microphone). In classical and contemporary music, the placement of the instruments also creates spatial differences, so the left and right sides are balanced. This is called Stereo Width Balancing, and if the algorithm can produce music with sound width, it will produce a more natural one. In commercial digital music, the stereo width is determined by how much panning is done, so the general range of availability can be used as a rule for the algorithm's sound selection. (Stereo Width is not tested in this study as it is a post-processing step.)

Table 2. Selected rules and Features of each standard

Features of each standard	Rules
Measurement of sound changes(volume/speed)	Dynamics (w/Transcient)
Tone and noise classification	Tonal Balance
Sound source extraction	Roots, Tonal-Noisy Balance
Spatial characteristics of the sound source	Stereo Width

2.3 Methodology

In this experiment, we utilize the probabilistic YIN (PYIN) algorithm, a modification of the well-known YIN algorithm for fundamental frequency (F0) estimation. The problem with YIN is that it outputs exactly one estimate per frame and cannot rely on different interpretations of the signal in post-processing. PYIN is complemented by outputting multiple pitch candidates with associated probabilities and decoding them to generate an improved pitch track [6]. In particular, when analyzing dynamics, the energy and delta graphs of the RMS were used to analyze three criteria: whether the sound source has a positive, negative, or close to zero change.

3. RESULTS OF RESEARCH

3.1 Dynamics

Dynamics analyzed the RMS Energy and RMS Delta graphs to determine whether the sound source was changing positively, negatively, or close to zero. In <Figure 1>, the blue line corresponds to the RMS (root mean square), or in other words, the actual perceived volume to the human ear. In certain parts of the graph, we can also see that the RMS Energy graph and the Spectrogram show similar trends.

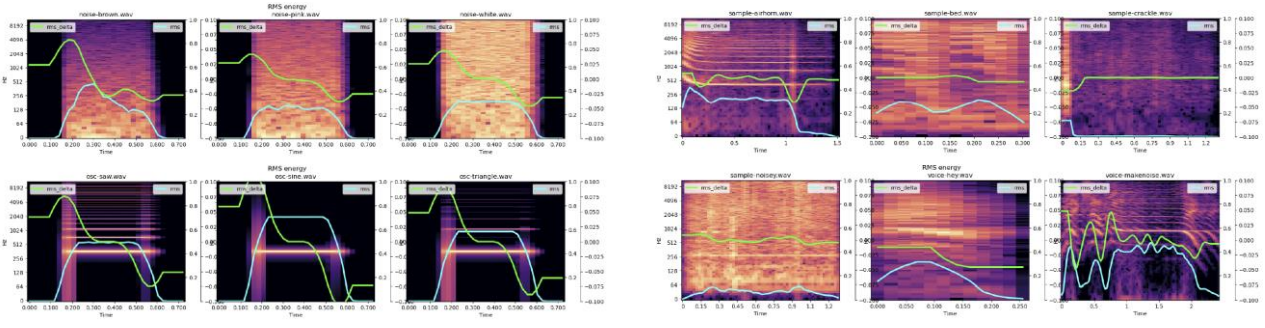

Figure 1. RMS Energy, RMS Delta, and Spectrogram

3.2 Tonal Balance & Roots

In <Figure 2>, the 'root' (light blue line) was generally located between 250 and 800Hz, while the spectral

centroid (light green line), which represents the brightness of the timbre, was mainly located at the top (between 2000 and 4000Hz). Thus, we can see that the root sound is located at the lowest of the different frequencies in the sound source. In the spectrogram where the spectral centroid and the root sound are shown together, the relationship between their positions shows that the spectral centroid is located in the center between the peak of the frequency and the root sound. Of the six files, only the musical instruments and human voice (airhorn, hey, makesomenoise) showed root sounds, while natural, daily, and pure noise (bed, crackleice, noisy) did not show them.

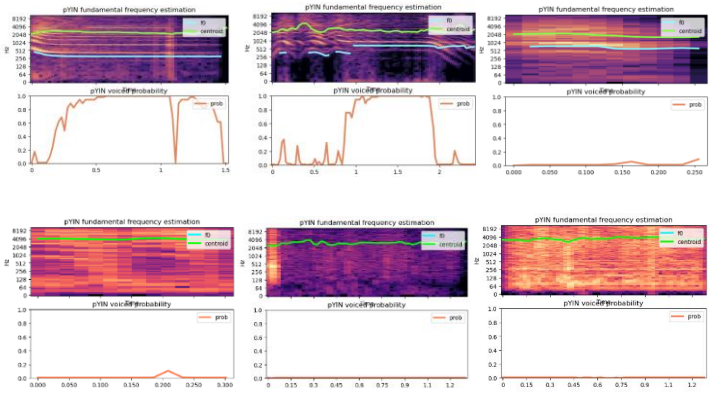

Figure 2. Relationship between spectral centroid and root sound in a tonal balance spectrogram, (top) airhorn, makesomenoise, hey, (bottom) bed, crackleice, noisy

In the second stage, white, pink, and brown noise and sine waves were added to further confirm the difference in uniformity between noise and non-noise, and the presence or absence of roots (<Figure 3>), and Spectro Centroid was measured separately to check the tonal balance.

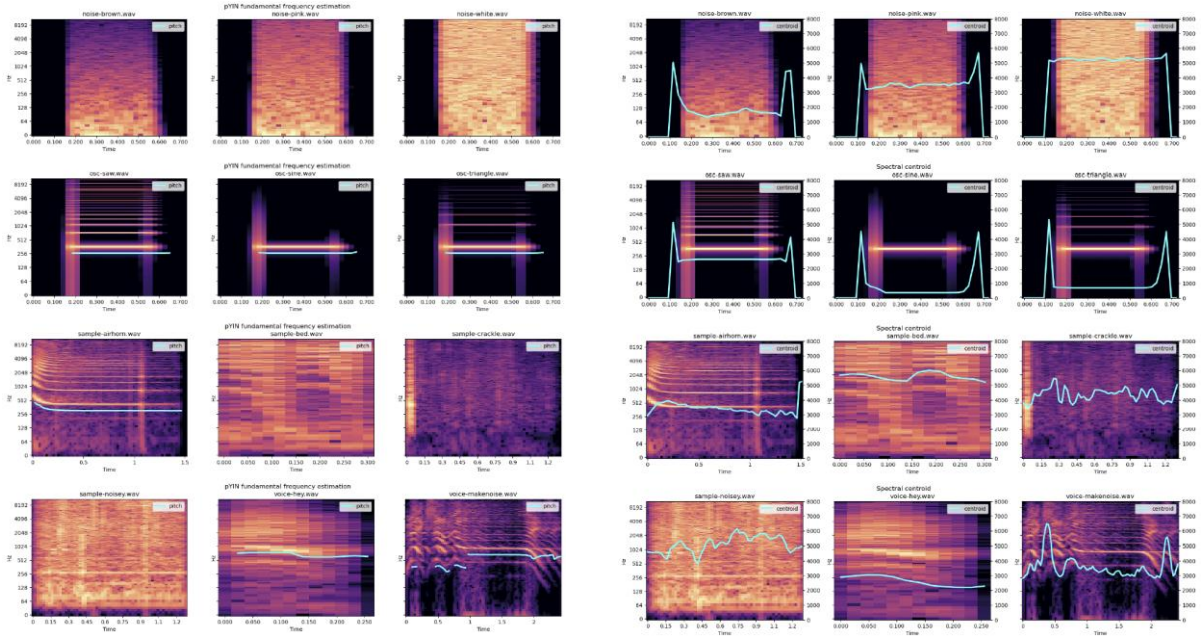

Figure 3. (left) Root sound tracking (light blue line) in noisy and non-noisy sources, (right) Spectral Centroid for Pitch Detection

Voiced probability calculates the probability that a given sound is voiced, with a value of 1 indicating a human voice and a value close to 0 indicating noise [6]. In general, it is known that higher values of Voiced probability are more likely to be tonal. However, in the top two graphs in <Figure 2>, the first spectrogram has a high voiced probability even though it is a musical instrument sound. Also, the clip used in the second spectrogram is composed of human voices from the beginning to the end, but the voiced probability value is close to 1 in the long sections where the voices are sustained, and low in the short sections where the voices are not sustained, so it can be said that the voiced probability value is far from confirming whether it is tonal or not.

3.3 Tonal-Noisy Balance

To measure the tonal-noisy balance, we used two metrics: spectral flatness, which measures the similarity to white noise, and spectral contrast [7], which measures the intelligibility of the sound source. Since spectral flatness measures how close the sound is to the noise and spectral contrast measures how far away it is from the noise, using these two metrics together allows us to experiment with contrast. In <Figure 4>, the spectral flatness values of the six samples were around 0.3 for sources close to noise, and between 0.0 and 0.2 for tonal sources such as musical instruments and human voices.

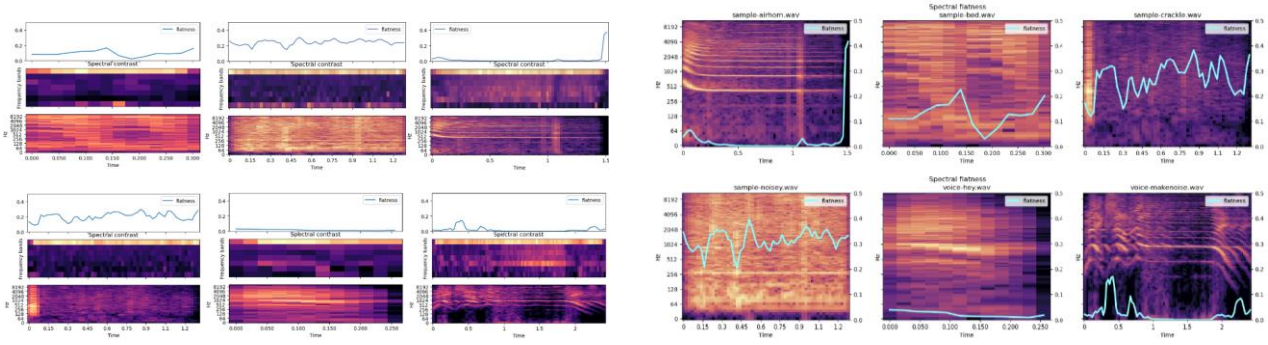

Figure 4. (left) Spectral Flatness and Spectral Contrast spectrograms – 1st step, (right) Spectral Flatness – 2nd step

Because Contrast is a metric that contrasts values, it requires different analysis than other metrics. Since Spectral Contrast estimates "by comparing the mean energy in the top quantile (peak energy) to that of the bottom quantile (valley energy)" [7], what we can get from a Spectral Contrast comparison (which is also explained in the documentation) is that a contrast value indicates a narrow band signal. In the text where Spectral Contrast was first proposed, it was used to create an algorithm to automatically distinguish between classical and pop, and it was claimed that the higher the contrast, the closer to pop. Empirically, classical music is more likely to have lower contrast because it emphasizes harmony and involves more instruments, while pop is more likely to have higher contrast because it involves fewer instruments. As a result, it seems that Spectral Contrast alone can be used to measure Tonal-noisy Balance.

3.4 Application and Algorithm Learning Method

Based on the above experiments, we derived the algorithm application criteria and learning method for each rule as follows.

Table 3. Criteria and learning methods for applying algorithm

Rules	Application	Learning Method

Dynamics	(1) prioritize sound sources with fast attack and decay, (2) slow attack and decay, and (3) constant sustain. * Especially, when the clip is about one beat long, a decrease in the average volume of the near-frequency or overall frequency is appropriate.	(1) Since it is necessary to check the velocity difference of the attack and that of the decay at the same time to distinguish the change of sound, it is better to have the selection focus on the beginning of the sound rather than the end of. - Decay tendency (loud and then quieter than quiet and then louder) should be evenly distributed. - Prioritize those with short attacks and decays - Check the largest volume and then smaller volume (2) ADSR: A model that quantifies Envelope (Envelope: describes how sounds and music change over time) [8]. - 4 parameters of Envelope: Attack - Decay - Sustain - Release
Roots	(1) Separate into low/mid/high frequencies (2) Filtering out 'root' when it is located in the ultra-low or ultra-high range	(1) Find a root note in the main melody and select it if it is between 300Hz and 2000Hz. (2) Find the strongest frequency (e.g., in piano, "Middle C" (440Hz) is a root note)
Tonal Balance	(1) Discard clips where 'root' is not measured (2) Pink Noise tendency measurement criteria	(1) For Pink Noise tendency, discard if it deviates too much from the norm (2) For Tonal Balance, human follow-up is needed
Tonal-Noisy Balance	(1) Tonal weight up to 0.2	(1) Measure the Noisy Balance ratio (2) Filter out clips with random, holistic frequencies, represented by White Noise (3) Learn not to consider 100% noise (white noise) and 100% tonal (sine wave)
Stereo Width	(1) Sound sources recorded with a mono microphone do not have stereo information, so no criteria need to be applied. (2) No need to consider if aiming for mono output. (being able to be covered by post-processing)	Measure the amount of dispersion in a phase oscilloscope trajectory

4. CONCLUSION

When we first designed this experiment, we aimed to derive a specific sound source from the noise, but as we separated the reference music into units for training the algorithm, we realized that we needed to use

not only the refined source, but also the noise--that is, the noise could be used as a beat or high-pitch sound. Since the root sound is mainly needed for melodic work, and the beat (rhythm) is composed independently of the root sound, the possibility of utilizing noise sources especially for beat work increases. In this context, among the selection criteria proposed in this study, "root" can be used not only as a criterion for extracting the sound source from the noise for the algorithm to use, but also as a criterion to distinguish whether the source should be used for melody or beat.

When experimenting with Tonal-Noisy Balance, we encountered cases where the spectral flatness of the sound source was not significantly different due to fragmentation, especially in footage shot with a normal camera. Since pitch requires low spectral flatness, that is, low similarity to white noise, to exist, the algorithm tends to select clips that do not differ significantly in spectral flatness. As a result, Tonal-Noisy Balance, which was initially adopted as a criterion to distinguish between beats and melodies, became meaningless for distinguishing beats from melodies as it resulted in selecting only clips with low spectral flatness when selecting melody sources. The algorithm may also use only clips with low spectral flatness to reconstruct the sound source, making it meaningless to compare the differences between each clip. In other words, we had to reconsider whether or not to use the criterion that was initially intended to filter out white noise by comparing its similarity to white noise, as we realized during the course of our experiments that this criterion could not be used to encompass other types of noise, such as brown noise or pink noise. Furthermore, when the algorithm reconstructs the sound source, matching clips based on similarity increases the likelihood of poor matching, which raises the need to exclude Tonal-Noisy Balance from the rule or apply a sample slicing method other than Pitch Detection. This may be an issue with footage captured with a standard camera, so it may be worth retesting with footage captured and ingested with professional equipment.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (RS-2023-00240479)

REFERENCES

[1] D. Pituk, *Automatic audio sample finder for music creation: Melodic audio segmentation using DSP and machine learning*, Master Thesis. KTH Royal Institute of Technology, Stockholm, Sweden, 2019.

[2] Van Balen, Jan., *Automatic recognition of samples in musical audio*, Master Thesis. *Barcelona: Universitat Pompeu Fabra*, 2011.

[3] D. Zhiyao, G. J. Mysore, and P. Smaragdis, “Speech enhancement by online non-negative spectrogram decomposition in nonstationary noise environments,” *Thirteenth annual conference of the international speech communication association*, 2012.

[4] D. Jonathan, H. D. Tran, and H. Li, “Spectrogram image feature for sound event classification i n mismatched conditions,” *IEEE signal processing letters* , Vol. 18, No. 2, pp. 130-133, 2010.

[5] K. Peerapol, C. Lursinsap, and T. Raicharoen, “Very short time environmental sound classificati on based on spectrogram pattern matching,” *Information Sciences*, Vol. 243, No. -, pp. 57-74, 2013.

[6] M. Mauch and S. Dixon, “PYIN: A fundamental frequency estimator using probabilistic threshol d distributions,” *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, pp. 659-663, 2014, doi: 10.1109/ICASSP.2014.6853678.

[7] J. Dan-Ning, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai, “Music type classification by spectra l contrast feature,” in *Multimedia and Expo, 2002, ICME’02, Proceedings, 2002 IEEE Internatio nal Conference on*, vol. 1, pp. 113-116, IEEE, 2002.

[8] V. Mark, *The Synthesizer: A Comprehensive Guide to Understanding, Programming, Playing, an d Recording the Ultimate Electronic Music Instrument*, Oxford University Press, Incorporated, 20 14, p. 152.