Emotion Analysis of Textless Audio Features Based on Cat Behaviors

Fangqian Liu

Westa College, Southwest University, Chongqing, 400715, China

2062327167@qq.com

Abstract. The research centered around the emotion analysis of cat's textless audio features. By extracting the audio features of a kitten through dimensions including Spectral Centroid, STFT, and ZCR, etc , the researcher analyzed the commonalities and differences between audios and explored their deeper relationships with emotional mechanisms based on specific behaviors. It provides a reference for the construction of neural network categorization system for textless audio emotion recognition, which has further potential in understanding aphasia patients, babies or other creatures with non-textual communication.

Keywords: textless, audio features, emotion analysis

1 Introduction

Biological populations cannot thrive without communication and interaction between individuals, especially community animals. Like humans, information exchange assists them in their daily behaviors of expressing needs, deterring aggression, seeking help, and conveying friendliness, which helps biological populations to cooperate, compete and develop with each other. Within biological populations, physical communication is mainly based on behavior, facial expressions, and audios. among which audio is the interaction method that is used frequently and has a strong influence.

The speech signal analysis has already gain critical applications nowadays, yet the non-textual sound signals still consist of discerning challenges due to the complexity and groundless ,which requires techniques that can accurately classifying such variability [1],especially for non-textual individuals lacking information convective support, like patients with dysphasia, babies or animals.

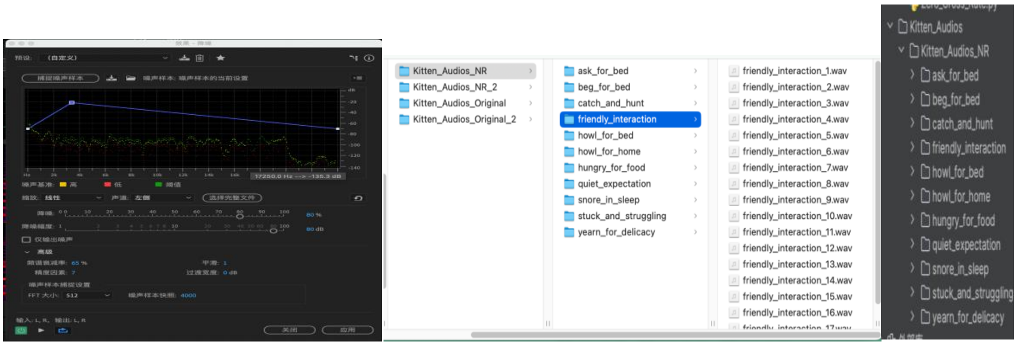
This research is centered around the audios characterization of a 2-month-old kitten, and the signal is disassembled and analyzed from the characteristic dimensions such as the zero-crossing rate, Spectral Centroid, Amplitude Envelope, MFCC coefficients, and STFT (Short-Time Fourier Transform) . From different feature dimensions, researcher explored the audio differentiation and the possibilities of mechanism behind the behaviors regarding the emotions such as joy, anxiety, ferocity, excitement, calmness, etc., so as to provide references and basis for the construction of neural network and deep learning categorization system for text-free audio emotion recognition. The research will be presented through the sections of sample collection, software environment preparation, feature analysis module construction, extraction results and emotion analysis, as well as the conclusion.

CONF-CIAP 2025, January 17-23, Eskisehir, Turkey

DOI 10.4108/eai.17-1-2025.2355247

2 Audio Sample Preprocessing

The raw samples were collected in wav format using a mobile electronic device (with default white noise reduction) , and were then pre-processed by Au software. The researcher imported the recorded audio, listened to and cut down useful kitten speech information, leaving a certain length of speechless white noise background information on both sides, for sample collection and reduction of personalized background noise for each segment.



预设: (自定义)

噪声样本: 噪声样本的当前设置

缩放: 线性

声道: 左侧

降噪: 80 %

降噪幅度: 80 dB

☐ 仅输出噪声

频谱衰减率: 65 %

平滑: 1

精度因素: 7

过渡宽度: 0 dB

FFT 大小: 512

噪声样本快照: 4000

输入: L, R, 输出: L, R

Fig. 1. Audio file processing

Through noise reduction, audio equalization, personalized noise elimination, parameter adjustment and other processing, as it is shown in Fig.1. Under the principle of maintaining original sound, the researchers maximized the elimination of noise interference, so that the audios could reach its original appearance. Based on the kitten's behavior, the collected samples were classified into nine valid audio types, each containing 10-35 speech segments in wav format with a sampling rate of 44,100 Hz.

3 Feature Extraction Module Construction

The researcher configured Anaconda interactive environment and used Python 3.10 on the compilation platform PyCharm 2022 for the construction of the feature extraction algorithms, and introduced third-party libraries including numpy, librosa, matplotlib, etc. for computation, graphing and audio analysis.

The audios in reality are mostly dynamic analog signals changing over time [2]. In order to transform them into suitable data structures for computer transmission, storage, and programming, the researcher specified the number of sampling points to fix the frame length, and divided the signals into short-time unit frames. Additionally, overlaps between frames were set to compensate for partition faults and ensure continuity [2]. This resulted in frame shifts, the difference quantity between the starting points of two neighboring frames.

Zero Pad can overcome the "Fence Effect" and avoids distortions and discontinuities in the spectrum due to signal truncation. The increased data length, which manifests itself as interpolation in the frequency domain, increases the frequency resolution and reduces uncertain peaks and leakage in the spectrum, thus improving the accuracy of spectral analysis. Based on the principle illustrated in Fig.2,the researcher set the frame length to 1024, the frame shift to

512, and the overlap percentage (frame length - frame shift) was set to 50%. If the length of the signal is not divisible by the length of the frame shift, then the signal needs to be zero padded.

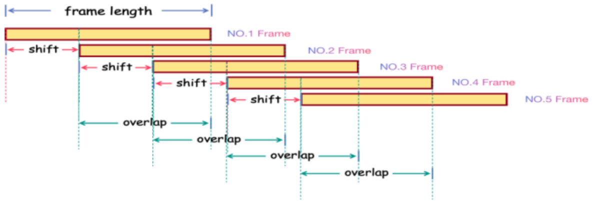


Fig. 2. Schematic diagram of framing principle (original)

Zero-Crossing Rate (ZCR) indicates the occurrences of signal crossings across the axis within a specified time frame [3], and can imply the pitch, periodicity, and energy distribution [3]. It is often used in audio categorization, sound recognition, and music analysis [3]. First, the researcher used a loop as a framework to indirectly determine the starting and ending positions of the current analyzed frame by multiplying frame shift length by the traversed counter plus the frame length. Next, transformations of neighboring sampling points in each frame is summed up, and divided by the length of the frame, according to sign function in ZCR formula. Then ZCR of each frame is appended in a list to get the variation of the whole signal.

$$\mathrm{ZCR} = \frac{1}{2} \sum_{m=n\cdot t}^{n\cdot(t+1)-1} |\operatorname{sgn}(x[n]) - \operatorname{sgn}(x[n+1])|$$

Spectral Centroid is used to indicate the location of “Position Center” in a spectrum [4]. It is often applied in analyzing pitch, timbre and brightness [4], or to distinguish phonemes. A high Spectral Centroid indicates a bright or sharp tone, while a low Spectral Centroid indicates a darker one. To calculate the Spectral Centroid, weighted average frequency of the spectrum is calculated, where the weights are determined by the amplitude of each frequency component. First, a short time Fourier transform (STFT) is computed for each frame to obtain the spectrum. Subsequently, a magnitude weighted average frequency of each frame is calculated according to the formula. x(f) is the magnitude corresponding to different frequencies.

$$\mathrm{SC} = \frac{\sum_{n=1}^{N} f(n) \cdot x(f)}{\sum_{n=1}^{N} x(f)}$$

Amplitude Envelope is a contour of the signal amplitude over time, emphasizing the dynamic characteristics of the signal such as amplitude, strength, or energy changes. Similarly by means of traversed counter and the frame length, the start and ending points of each frame are located sequentially. Max function is used to obtain the maximum value of amplitudes in each frame, which would then be added to the predefined array to get the Amplitude Envelope of the whole signal.

$$\mathrm{AE}_t = \max_{m\cdot t \le n \le m\cdot(t+1)-1} x(n)$$

$$E_n = \sum_{m=0}^{N-1} {x_n}^2 (m)$$

MFCC (Mel Frequency Cepstral Coefficients) is a practical feature index of network training in sound classification tasks [1] by studying the energy distributions corresponding to 13 typical sound frequencies, which provides a representation of the spectral envelope of an audio signal [5]. For transformation from time-domain to a frequency-domain, STFT is applied to all frames [2], where the frequency is converted from Hertz to mel [6]. Then, periodogram method is used to estimate the power spectrum, and the spectrum is filtered with Mel filter banks which simulate human ear's perception proportional to the logarithm of frequencies [2]. The process calculates the energy in each filter, and differentiates the envelope and details, including timbres and pitches, In the Cepstrum analysis of Mel spectrum, Discrete Cosine Transform(DCT) is applied to convert the logarithmic Mel Spectrum in log back into time-domain to obtain the Mel Cepstrum [6]. It is performed on the signal with 26 points to obtain 26 Cepstral Coefficients, and finally the 12 numbers from 2-13 are retained as MFCC features, converting the audio information into multiple sets of feature vectors.

$$\text{FT: } \mathrm{F}(\omega) = \int_{-\infty}^{+\infty} \mathrm{f}(\mathrm{t})\mathrm{e}^{-\mathrm{j}\omega \mathrm{t}}\ \mathrm{dt}$$

$$\text{STFT: } \mathrm{F}(\tau \cdot \omega) = \int_{-\infty}^{+\infty} \mathrm{f}(\mathrm{t})\mathrm{w}(\mathrm{t}-\tau)\mathrm{e}^{-\mathrm{j}\omega \mathrm{t}}\ \mathrm{dt}$$

Short-Time Fourier Transform (STFT) performs the Fourier Transform in short time frames respectively, producing a spectral illustration helpful for audio pattern analysis [7]. It breaks the limitation of conventional FT in non-stationary signal processing, which easily loses dynamic information [8], achieving comprehensive characterization in time-frequency domain [9]. In the research, the total number of sampling points was determined and the signal was first converted into the form of an array. Since the sampling points were taken as real numbers, the total number of them was half of the number of frequency sampling points (containing both imaginary and real numbers). Similarly, the frequency scale on y-axis was plotted, where its range was determined by Nyquist Sampling Theorem that sampling frequency is at least twice the maximum frequency. Finally, the results of FT for each frame were combined to form a two-dimensional time-frequency spectrogram.

4 Results and Discussion

The researcher substituted the audio samples into the computational model. It turns out that different categories of audio samples are irregular in some feature dimensions but follow certain patterns in the others. The following are the features with representative differentiation.

4.1 Zero Cross Rate Characterization (ZCR) and Emotion Analysis

It can be found in Fig.3 that ZCR of "Snore in sleep" behavior audios are consistently at a very low level (below 0.015), and are always opposite to the amplitude, both in terms of the general trend and the local quantitative magnitude. When ZCR increases, the amplitude decreases. If ZCR goes lower, the amplitude correspondingly goes higher. This may due to the fact that the emotional state of the kitten is under the control of Autonomic Nervous System (ANS) during sleeping state, which operates unconsciously and smoothly, so the dynamic variation and the amplitude present a mutual compensation to achieve the preservation of physiological emotion energy and balanced output.

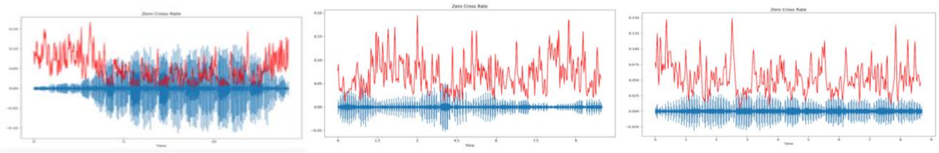


Fig. 3. ZCR of “Snore in sleep” audios

ZCR of "Yearn for delicacy" behavior audios in Fig.4 tend to stay low at 0.025 - 0.05 except for slight local fluctuations. This may relate to the low-frequency "purring" at the kitten’s throat signaling the mood of satisfaction , which serves as a background sound throughout the sample. However, it is occasionally accompanied by higher pitched speech, just like "praise" indicating enjoyment of human beings in an exhilarating mood, which may correspond to localized fluctuations.

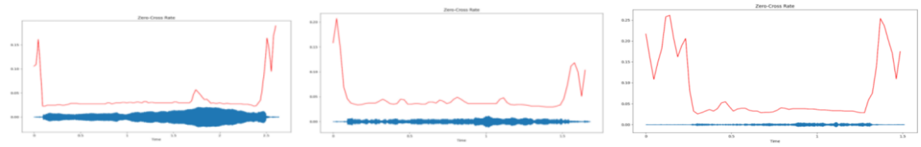


Fig. 4. ZCR of “Yearn for delicacy” audios

ZCR of "Catch and hunt " behavior audios in Fig.5 remain even lower around 0.02-0.025 on an overall basal trend, but the local waveform is extremely unstable and fluctuates at a high frequency, which may symbolize the unstable state of high-energy emotion in deterrence. Furthermore, the amplitude of local fluctuation is incisive, and occasionally accompanied by brief "narrow and prominent" surges and falls, which may be further expressions of aggressive mood with attack.

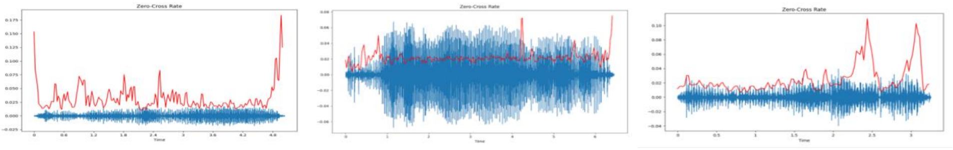


Fig. 5. ZCR of “Catch and hunt” audios

From Fig.6 to Fig.8,　ZCR of those 3 kinds of audios which denote a need are pretty similar. If we enclose the variation with an envelope, a relatively stable but always changing general trend ranging from 0.05 to 0.02 can be found, which may correspond to unsatisfied emotional state of needs. But the changes are not as intense as excitement, or angry ferocious emotions. The frequency and amplitude of localized fluctuations are more moderate in an unbalanced but stable emotional energy.

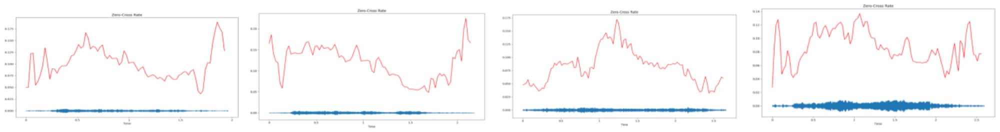


Fig. 6. ZCR of “Howl for bed” audios

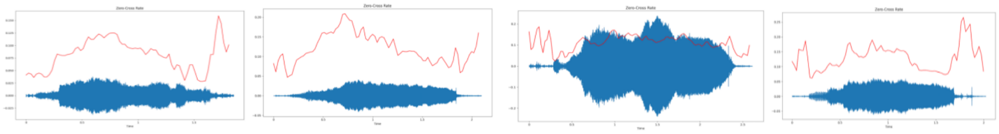


Fig. 7. ZCR of “Howl for home” audios

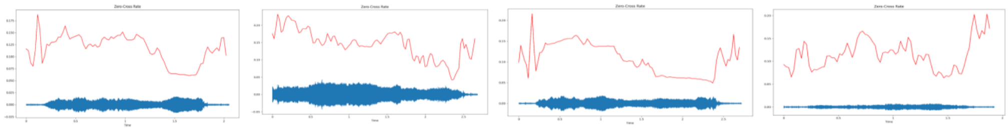


Fig. 8. ZCR of “Hungry for food” audios

Similar to the "requiring behaviors", ZCR of "Quiet expectation" audio in Fig.9 has a general trend fluctuating between 0.05 to 0.175, just like a mountain’s outer contour. The difference is that its duration is generally short-lasting, and the frequency of fluctuations is relatively low even when viewed on the same time scale.

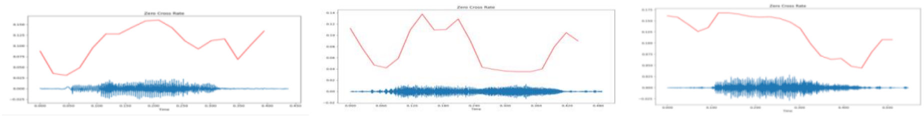


Fig. 9. ZCR of “Quiet expectation” audios

In contrast, the ZCR pattern of "Stuck and struggling" behavior audio in Fig.10 is disorganized, and relevant mode in "Friendly Interaction" in Fig.11 is more varied and irregular.

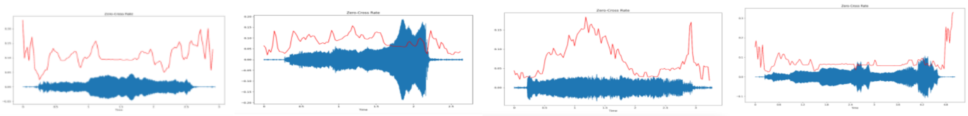


Fig. 10. ZCR of “Stuck and struggling” audios

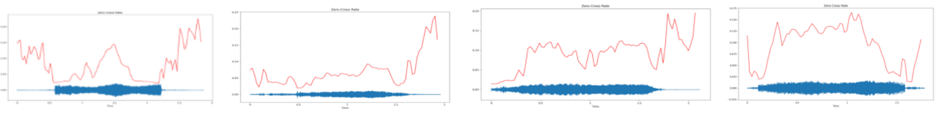


Fig. 11. ZCR of “Friendly Interaction” audios

4.2 Amplitude Envelope Feature and Emotion Analysis

The Amplitude Envelope exhibits fewer distinguishable features, characteristics for each sound type sometimes get confusing. But there still exists some distinguishing and representative features.

Amplitude Envelope of "catch and hunt" behavior audios in Fig.12 shows extremely unstable localized "narrow spike" oscillations, with high frequency as well. The difference is that the overall trend does not show any significant surge, only stays within a high level of 0.01-0.05 amplitude range. According to the signal equation that the energy is the sum of the squares of amplitudes, the amplitude reflects the energy level to some extent. This suggests that the kitten's mood at this time has been maintaining a considerable high energy state and is extremely volatile and fluctuating, possibly corresponding to a high and unstable ferocious emotion.

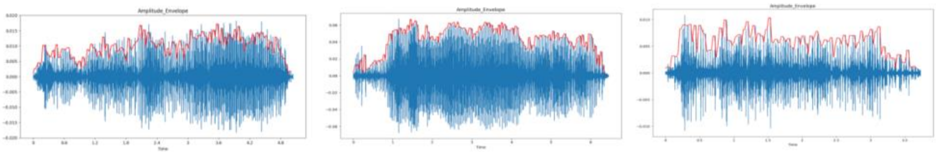


Fig. 12. Amplitude Envelope figure of "Catch and hunt" audios

In the Amplitude Envelopes, audios of demand behaviors from Fig.13 to Fig.15 are still similar in terms of overall trend changes, but there is some differentiation in the local waveforms. Compared to "howl for bed" and "hunger for food" audios, "howl for home" signal has a lower frequency in local oscillations and is more moderate on the same time scale, which may correspond to a calmer and more stable emotional energy. Since "howl for bed" and "hunger for food" are more oriented towards immediate needs for visible goals, while "howl for home" is a kitten's tentative need. Although the kitten tends to back home, staying out and play for another while is acceptable as well. The relatively lower-need emotion accordingly reduces the frequency of the oscillation, so they get moderate and less intense. In terms of amplitude, "howl for home" and "howl for food" are on the order of 0.01, compared to "howl for bed" with 0.001, which corresponds to a more intense-need energy state.

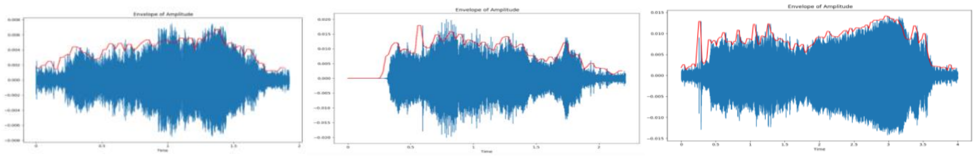


Fig. 13. Amplitude Envelope figure of "Hungry for food" audios

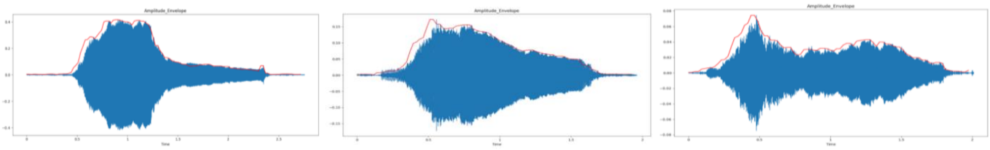


Fig. 14. Amplitude Envelope figure of "Howl for home" audios

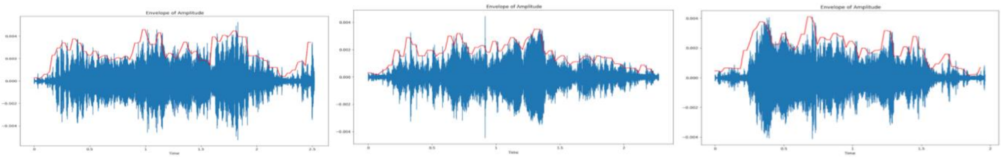


Fig. 15. Amplitude Envelope figure of "Howl for bed" audios

4.3 Spectral Centroid Feature and Emotion Analysis

Spectral Centroid variation of the 3 demanding audios from Fig.16 to Fig.18 are similarly diverse in their overall trends without being regular. However, a closer look reveals that their distribution ranges are clearly differentiated. Spectral Centroid of "howl for bed" fluctuates nicely around 1000-3000 Hz, and rarely crosses the boundary; The lowest Spectral Centroid of "howl for home" is still around 1000Hz, but the highest point mostly peaks at 4000-5000Hz, or even reach a level of 6000-7000Hz; Spectral Centroid of "hunger for food" only focuses on the fluctuation between1500-3250Hz, and there is generally no transgression as well.

On the whole, Spectral Centroid variation of "howl for home" is generally greater than "hunger for food", and "hunger for food" is greater than "howl for bed". In reality, "howl for home" signals the speech audience at a greater distance, which may be the reason why kittens raise the overall spectrum. Relatively higher level in "hungry for food" illustrates excitement and impatience while waiting for food as well. Transmission distance, excitement and impatience are both possible reasons for higher Spectral Centroid. They may be important considerations and trade-offs in designing classifiers.

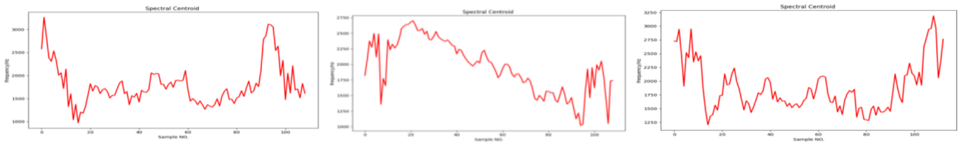


Fig. 16. Spectral Centroid variation of “Howl for bed” audios

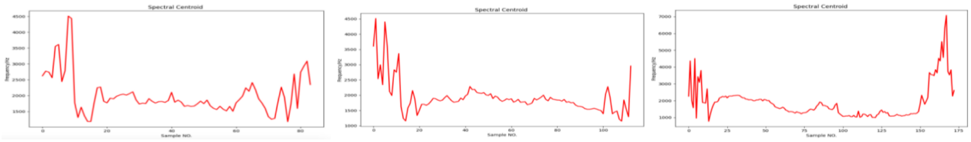


Fig. 17. Spectral Centroid variation of “Howl for home” audios

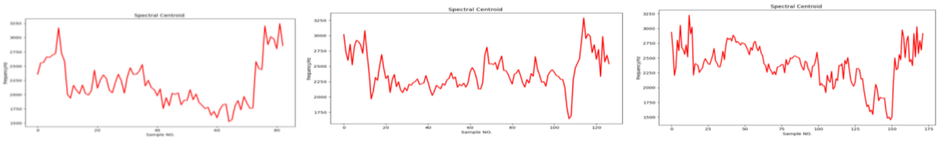


Fig. 18. Spectral Centroid variation of “Hungry for food” audios

Spectral Centroid of "catch and hunt" behavior audios in Fig.19 seems always at a horizontal line with a extremely low frequency of 500 Hz, which is more likely to show a sense of authority and hefty power, and correlates with the ferociousness and readiness in hunting. The sudden step to the mid-frequency range still with high-frequency fluctuations may stand for the real-life emotional expression of further deterrence. The localized waveform in the whole figure, which are extremely unstable fluctuations of considerably high frequency and amplitude in the shape of "narrow, thin, and sharp", may correspond to the precarious, high-energy emotional state of aggression that the cats maintain when they see bird-like objects, such as feathery or tissue balls. Analogously, this aggressive mood peaks in the surge of the "narrow, thin and pointed" shape in the general trend, which may correspond to the actual impetuous aggressive behavior.

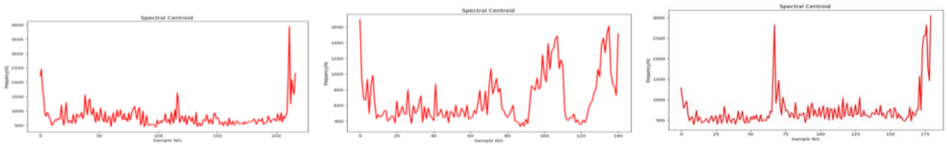


Fig. 19. Spectral Centroid variation of “Catch and hunt” audios

The most prominent Spectral Centroid variation of "Stuck and struggling" behaviour audios in Fig.20 is the "chunky" localized mild fluctuations at the range of 1000-2000 Hz, which may associate with the unrelieved anxiety of a kitten that is trapped by an object. In addition, the Spectral Centroid shows two kinds of rises at irregular intervals, one is a "gentle climbing" rise, which corresponds to the kitten's “call for help” behavior concerning anxiety and pleading emotions. One is "sharp surge" rises. This happens when the kitten struggles further, indicating a release of irritable stress and anger.

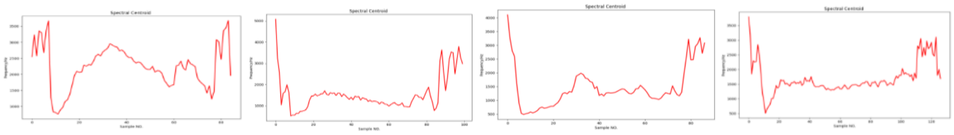


Fig. 20. Spectral Centroid variation of “Stuck and struggling” audios

Spectral Centroid of "snore in sleep" audios in Fig.21 is mainly ranging from 1000 to 3500 Hz, and the amplitude and frequency are notably high, with extremely dense fluctuations, which is distinct.

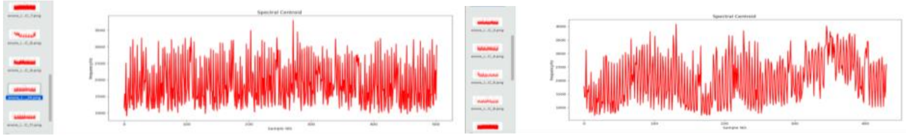


Fig. 21. Spectral Centroid variation of “Snore in sleep” audios

Spectral Centroid of "Yearn for delicacy" behavior audios in Fig.22 is generally based on persistent low-frequency segments. However, it is occasionally punctuated by moderate surges. When cats are comfortable and contented, they tend to make a low "purring" vibration with their throat vocal cords, and a persistent low-frequency background should be a quantitative expression of "purring". The medium level surges may relate to the cat's excitement and joy in "enjoying food". Since most of the time, the cat distracts part of its attention and energy to feed, so the surge is mostly at medium level. Occasionally, high frequency surges occurred between feedings, which may be a further expression of the excitement and joyful emotion under "delicacy".

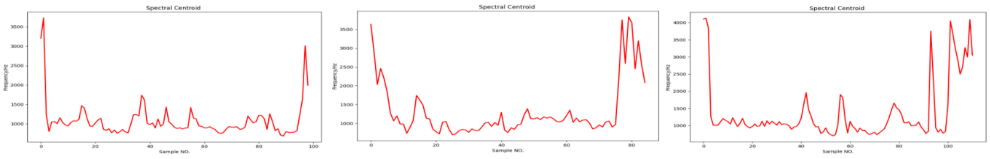


Fig. 22. Spectral Centroid variation of "Yearn for delicacy" audios

4.4 MFCC Matrix Results

MFCC results from Fig.23 to Fig.31 are presented as a graphical form of Meier frequency cepstrum coefficient matrix, with 13 rows on the vertical axis, representing 13 special frequency bands that are closely related to the pitch and timbre of sound. The horizontal axis represents the sequent frames, the vertical columns divided by the frames stand for eigenvectors which is corresponding to different frequency bands of audio signal, and the colors reflect the distribution of coefficient energies.

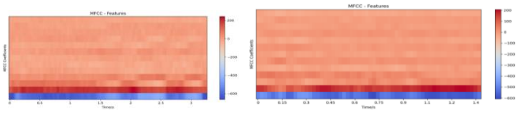


Fig. 23. MFCC of "Catch and hunt" audios

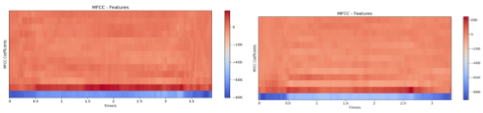


Fig. 24. MFCC of "Friendly Interaction" audios

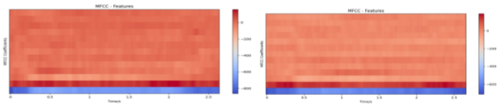


Fig. 25. MFCC of "Howl for bed" audios

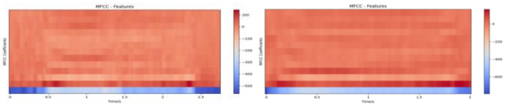


Fig. 26. feature of "Howl for home" audios

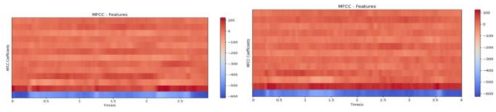


Fig. 27. MFCC of "Hungry for food" audios

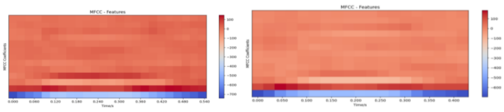


Fig. 28. MFCC of "Quiet expectation" audios

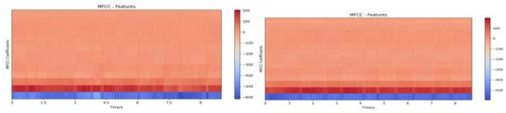


Fig. 29. MFCC of "Snore in sleep" audios

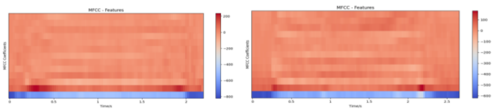


Fig. 30. MFCC of "Stuck and struggling" audios

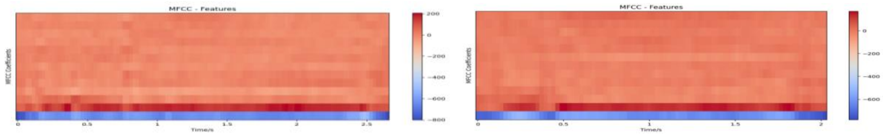


Fig. 31. MFCC of "Yearn for delicacy" audios

From the results, it can be seen that different types of audios in MFCC matrix, are somewhat differentiated. The energy distribution coefficients calculated in the matrix are more used as feature vectors or matrices that are substituted into the deep learning model for constructing and training. In neural network classification models, MFCC is usually used in combination with other features such as frame-level energy, time and frequency cepstrums to improve recognition accuracy.

4.5 Representative features of STFT (Short-time Fourier Transform)

STFT time-frequency atlas is mainly used as neural networks feature vector training set. The results are selected to display representative features, where the horizontal axis displays the sequent frames, the vertical axis shows the frequency bands, and the color represents the intensity or energy.

The distinction of the 3 demand audios, as seen in the STFT time-frequency mapping from Fig.32 to Fig.34, is clearly demonstrated. "Howl for home" has a wider range of high-frequency bands, while "Hungry for food" has more intensity and energy in the higher frequency bands. The change to higher frequencies is characterized by a gradual break in "howl for bed" and a stepped disappearance of the break in "Hungry for food".

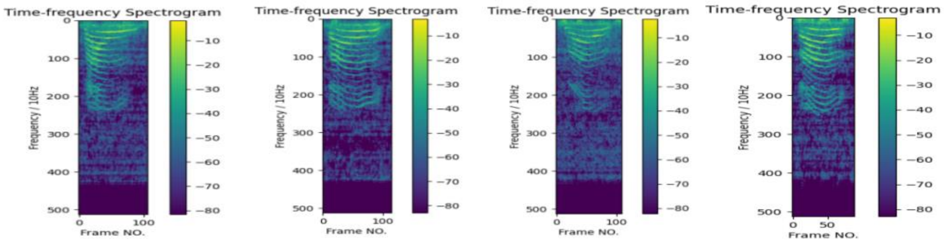


Fig. 32. STFT figures of “Howl for bed” audios

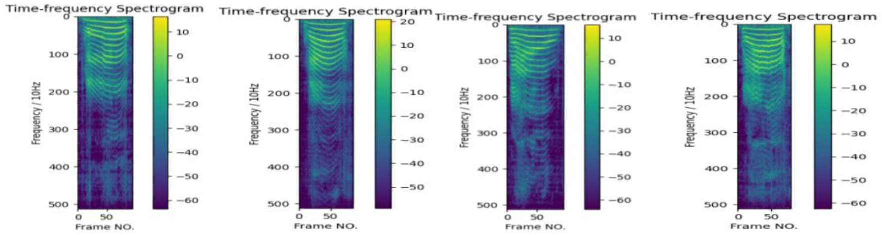


Fig. 33. STFT figures of “Howl for home” audios

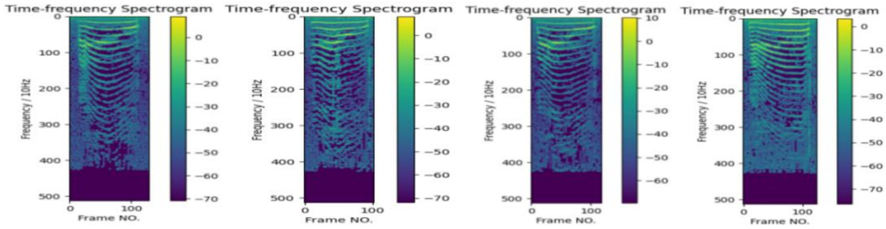


Fig. 34. STFT figures of “Hungry for food” audios

“Yearn for delicacy” behavior audios in Fig.35 and Fig.36 have a wide, relatively even distribution in the high frequency level, consistent with a sustained emotional energy state of high pitch and joy.

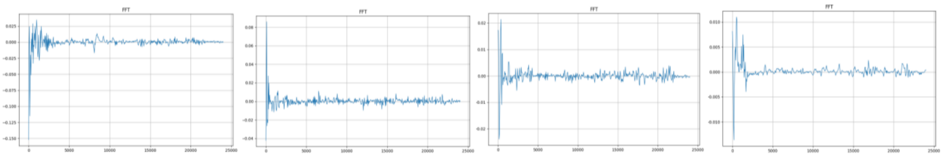


Fig. 35. FFT figures of “Yearn for delicacy” audios

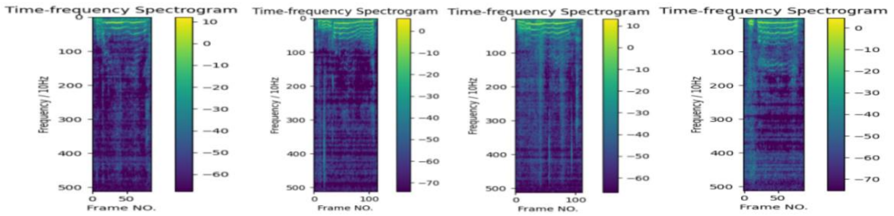


Fig. 36. STFT figures of “Yearn for delicacy” audios

"Snore in sleep" shown in Fig.37 is well aligned, with almost all the effective energy distribution concentrated in low frequency instead of the high ones, which corresponds to the complementary nature of the color intensities in the first and second dimensions of MFCC matrix, where the energy is mainly determined by low frequencies, corresponding to a conservative and stable emotional state.

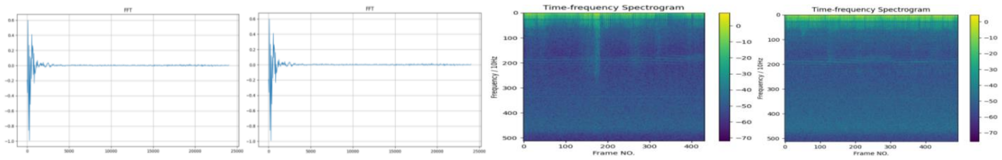
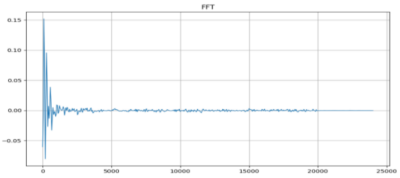


Fig. 37. FFT and STFT figures of “Snore in sleep” audios

"Catch and hunt" and "Stuck and struggling" are similar on FFT transform according to Fig.38 and Fig.39. In the high-frequency range, except for minor fluctuations, they are nearly straight. Obvious differences are shown in STFT time-frequency spectrograms in Fig.40 and Fig.41. They both decrease abruptly from low to high frequencies, but "catch and hunt" tends to decrease mildly like a "flat and straight brush", however, "Stuck and struggling" is a "rippling" decrease with a slow swoosh.



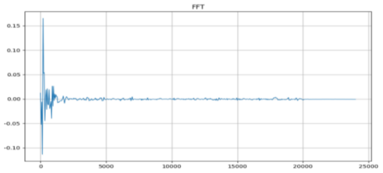


Fig. 38. FFT of Catch and hunt” audios **Fig. 39.** FFT of “Stuck and struggling” audios

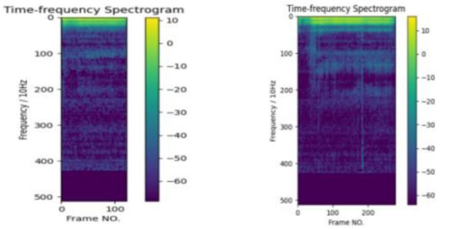


Fig. 40. STFT of Catch and hunt" audios

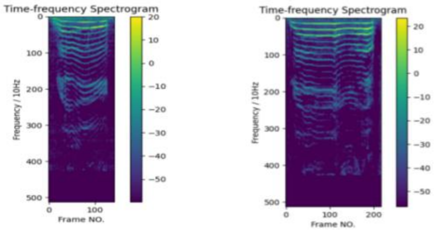


Fig. 41. STFT of "Stuck and struggling" audios

5 Conclusion

The research focuses on textless audio emotional analysis based on the behaviors of kitten. Researcher extracted effective audio features of different dimensions including Spectral Centroid, STFT, amplitude, and ZCR, and explored their potential relations with emotions concerning joy, anger, ferocity, excitement, balanced contentment, anxiety, pleading, stress, irritation, and anticipation, which provides references for weighting parameter setting, combination of features for emotion recognition in neural networks and deep learning of classification structures.

However, due to limited experimental conditions, there still exists space for improvement, such as expanding the sample size, for better recognition tolerance to more complex signals; or exploring more detailed and differentiated features, to distinguish samples with high approximation; or becoming more personalized for universality, to suit the application for different individuals and oral habits. It is hoped that in the future, the technology can be developed to be more refined and mature to assist information interactions without text, behavior or other expression support.

References

[1] R. Ahuja, V. Solanki, V. Khullar and L. Kumar, "Classification of Non-Speech Sound Signals: An Approach of Machine Learning with MFCC Feature Extraction," 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT), Greater Noida, India, 2024, pp. 1-5, doi: 10.1109/ICEECT61758.2024.10738971.

[2] Santoso, T. A. Sardjono and D. Purwanto, "Optimizing Mel-Frequency Cepstral Coefficients for Improved Robot Speech Command Recognition Accuracy," 2024 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, 2024, pp. 284-289, doi: 10.1109/iSemantic63362.2024.10762627.

[3] S. Singhal et al., "Audio Based Machine Fault Diagnosis using Hybrid Feature Extraction and Ensemble Learning," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10724147.

[4] G. Fu, Y. Jiang, H. Li, L. Ling and W. Wei, "Research on Loose Wedge Detection Method of Generator Slot Based on Acoustic Feature," 2024 IEEE 14th International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), Copenhagen, Denmark, 2024, pp. 517-522, doi: 10.1109/CYBER63482.2024.10749636.

[5] D. Prabakaran and S. Sriuppili, "Speech processing: MFCC based feature extraction techniques - An investigation", Journal of Physics: Conference Series, 2021.

[6] F. T. Al-Dhief, N. M. Abdul Latiff, N. N. N. A. Malik, M. M. Baki, N. A. Muhammad and M. A. Abbood Albadr, "Investigating Fast Learning Network for Voice Pathology Detection," 2024 IEEE 7th International Symposium on Telecommunication Technologies (ISTT), Langkawi Island, Malaysia, 2024, pp. 108-113, doi: 10.1109/ISTT63363.2024.10750772.

[7] N. Steinmetz and N. Balal, "Remote Speech Decryption Using Millimeter-Wave Micro-Doppler Radar," 2024 IEEE International Conference on Microwaves, Communications, Antennas, Biomedical Engineering and Electronic Systems (COMCAS), Tel Aviv, Israel, 2024, pp. 1-5, doi: 10.1109/COMCAS58210.2024.10741985.

[8] Y. Wang and D. Lai, "A small sample conventional circuit breaker fault diagnosis method based on SWT-STFT and double flow CNN-SVM," 2024 IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific), Xi'an, China, 2024, pp. 126-131, doi: 10.1109/ITECAsia-Pacific63159.2024.10738700.

[9] Sun Xinwei, Ji Aimin, Du Zhantao et al., "Diagnosis method for variable speed fault of rolling bearings in high-speed train gearbox [J]", Journal of Harbin Institute of Technology, vol. 55, no. 01, pp. 106-115, 2023.