A study on Music / Noise Classification

Dong-In Lee

ABSTRACT

In this paper a novel audio classification method based on Entropy Model [1] is presented. By applying this model on Harmonic Product Spectrum [2], a viable classification system that is able to discriminate between music and various ambient noises can be achieved that even performs well in a 'lo-fi' environment with audio signals of poor quality.

1. INTRODUCTION

Gracenote has been providing Music Identification services throughout the world in various forms. One particular service is referred to as Mobile MusicID [3], which enables users to recognize the music around them when they send the ambient audio signal to Gracenote servers through mobile handsets. On the server side, finding possible matches from audio fingerprints that have been either generated on the handset, or from audio submitted by the handset on the server side incurs a significant amount of computational load, which translates into monetary equipment and operational cost. Through monitoring the audio recording submitted to this service from handset users, we have found that some users sent almost silent recordings that have little data to extract a fingerprint from, and the remaining data has a noise-like characteristic (thus simple level thresholding does not function well). Other users frequently submit ambient noise with no discernable music content that could be used by the algorithm for identification. In those cases, it would be desirable to advise the user to record more prominent music audio by, for example, approaching the sound source in order to achieve a positive recognition event. It is further beneficial if we could apprehend query attempts that are bound to fail due to these issues and prevent querying against the recognition service altogether as this would save costs in server operation and also for establishing the network connection between the handset and the service.

With this purpose in mind, a silence detector had been designed in the past, anticipating that this could be solved using a fairly straightforward algorithm. The silence detector simply calculates the energy level for each time frame, which is 3s long. If the average energy level is below a predefined and heuristically determined threshold value the signal is assumed to be silent, and instead of sending the query to the service a proper message is returned to the user. However, in real-world handsets, with the presence of Automatic Gain Controller (AGC) [6], silent sounds are overlaid with system inherent noise from the analog frontend of the handset, causing the overall energy level to become higher than the threshold value. If we raise the threshold value for the silence detection to adapt to this case, however, we will likely also discard music signals which result in energy levels below this adapted threshold value which could potentially still be recognized by the system.

Thus, the essence of the problem rather becomes music versus noise classification challenge in order to combat the influence of the systematic noise in conjunction with the AGC. In addition, since the computing power of mobile handsets is limited, it would not be adequate to apply machine learning techniques with many-dimensional feature vectors. This means that the fewer features we choose, the more feasible the solution will be for real-world implementation. In this paper, some existing features are listed in section 2 which have been investigated for this project. In section 3, a new method for the classification is described.

Subsequently, we present experimental results in section 4. Discussion on the results is presented in section 5, followed by the conclusions in section 6.

2. AUDIO FEATURES FOR CLASSIFICATION

In this section several features which have been used historically for music classification are listed and explained in more detail.

2.1. Zero Crossing Rate

In the context of discrete-time signals, a zero crossing is said to occur if successive samples have different algebraic signs. The rate at which zero crossings occur is a simple measure of the frequency content of a signal. This feature has been used to discern voiced data with unvoiced data. However, it was not feasible to utilize this feature for music versus noise classification since this feature has exhibited little differences between music and noise cases. This might be because music and noise both have unvoiced and voiced parts in them.

2.2. Spectral Roll-Off Point

The Roll-Off is another measure of spectral shape. It is the point in the spectrum where frequency components reside that occurs below some percentage (usually at 85%) of the power spectrum.

2.3. Spectral Centroid

This is the gravity centre of the spectral distribution within a frame. The centroid measures the spectral shape. Higher centroid values indicate higher frequencies.

2.4. Spectral Flatness

Spectral Flatness is a measure of the noisiness or sinusoidal character of a spectrum or a part of it. It is computed by the ratio of the geometric mean to the arithmetic mean of the energy spectrum value. A high spectral flatness indicates that this would sound similar to white noise, and the graph of the spectrum would appear relatively flat and smooth. A low spectral flatness indicates that the spectral power is concentrated in a relatively small number of bands - this would typically sound like a mixture of sine waves, and the spectrum would appear "spiky".

2.5. Auto-Correlation

This feature is a mathematical representation of the degree of similarity between a given time series and a lagged (i.e. time-shifted) version of itself over successive time intervals. It is the same as calculating the correlation between two different time series, except that the same time series is used twice - once in its original form and once lagged one or more time periods. In average, music data exhibits a higher auto-correlation value compared to noise data.

2.6. Harmonic Product Spectrum (HPS)

If the input signal resembles a musical note, then its spectrum typically consists of a series of peaks, corresponding to the fundamental frequency together with harmonic components at integer multiples of the fundamental. Hence, when we 'compress' the spectrum a number of times (down sampling), and compare it with the original spectrum, we can see that the strongest harmonic peaks line up. The first peak in the original spectrum coincides with the second peak in the spectrum compressed by a factor of two, which coincides with the third peak in the spectrum compressed by a factor of three. Hence, when the various spectrums are multiplied together, the result will form clear peak at the fundamental frequency.

$$\pi_{HPS}(\omega) = \prod_{k=1}^{m} \left| F(k\omega) \right|^2$$

3. DETAILS OF THE PROPOSED METHOD

When a musical instrument is played, the human intervention typically excites specific harmonics. After a while, the acoustic wave eventually disperses and is converted into heat energy. Human perceives the acoustic excitation of the frequency distribution as music, and when the acoustic energy is dispersed to other frequency bands, the sound is no longer noticed as music. HPS is a useful means to represent the human perception of each frequency band. However, the quantification of the inharmonicity is another challenge that needs to be addressed for the distinction of musical signals and non-musical signals.

In this work, HPS is considered as a probability density function of human perception at each auditory scene. Although there is no explicit measure of the inharmonicity, one can quantify the randomness of the HPS distribution by using Shannon entropy [5].

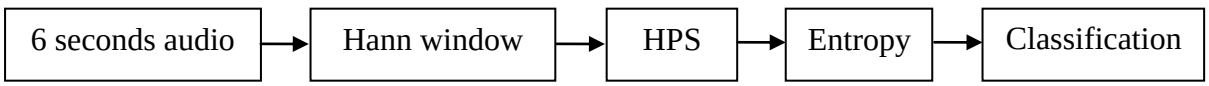
3.1. Entropy

In Shannon's original work, entropy is defined to measure the amount of information that an information channel can transmit, but is equivalently used to measure the uncertainty of probability density functions. We can make use of Shannon entropy to evaluate the randomness of HPS, which enables us to measure the inharmonicity of a given signal.

$$H(X) \equiv -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

where $p(x_i)$ denotes the probability that event x_i happens.

3.2. System Structure



The classification is carried over 6 seconds-length of audio. The audio is first resampled to have 11.025 kHz sampling rate, and then is normalized to have unit energy, and this enables this system to ignore the signal level differences from audio samples. After the normalization, the audio is segmented into overlapping frames. Each frame has a length of 0.185 seconds and is weighted by a hann window with an overlap factor of ¾ (75%). The windowed signal is the input to the HPS. The compression factor R used for HPS is 5. Then, only the values over 200 Hz are considered for further processing.

Since the HPS distribution at each frame is considered as the probability density function for the entropy model, the HPS is always normalized to have their sum as 1. Finally, the entropy of the probability density function at each frame was evaluated to quantify the inharmonicity of the input audio when perceived by the human auditory system.

4. EXPERIMENTS

The noise samples used for the experiments consist of the noise samples from Sony Ericsson and the author's own recording by Nokia N95. All the music samples were recorded through N95 while playing the original music samples from PC speaker. For the recording of test music and noise samples, the same quality setting (8000 Hz, Mono, 16-bit PCM) was used. All the samples used for the explanation is available at [4].

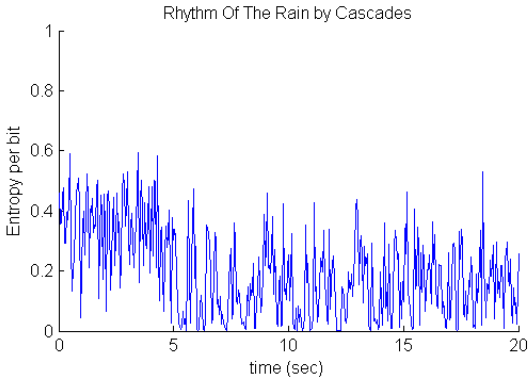


Figure 1. The entropy from *'Rhythm of the Rain'*

Figure 1 shows the entropy of the song 'Rhythm of the Rain'. This song starts with the thunder sound followed by the sound of raindrop. Then, the music band starts playing at around 5 seconds. The raindrop sound lasts until 15 seconds. From this figure, we can see that the entropy values between 0 to 5 seconds period are high since the HPS distribution is close to flat (See Figure 2 left) in that period. When the band starts music, the low entropy values close to 0 are presented since the sounds have harmonicity, and thus the peak value is higher (See Figure 2 right). We can also see that the entropy values between 5 seconds and 15 seconds period which have both raindrop sound and musical sound show similar graph from 15 seconds through 20 seconds where inharmonic raindrop sound has vanished.

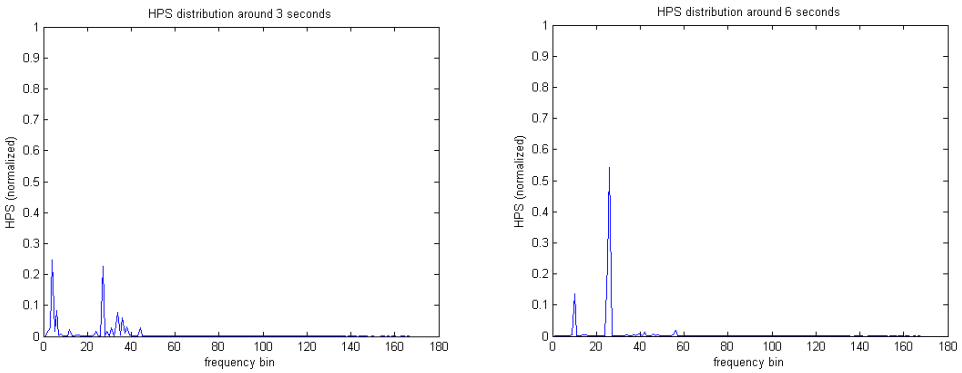


Figure 2. The HPS distribution around 3 seconds (left) and 6 seconds (right)

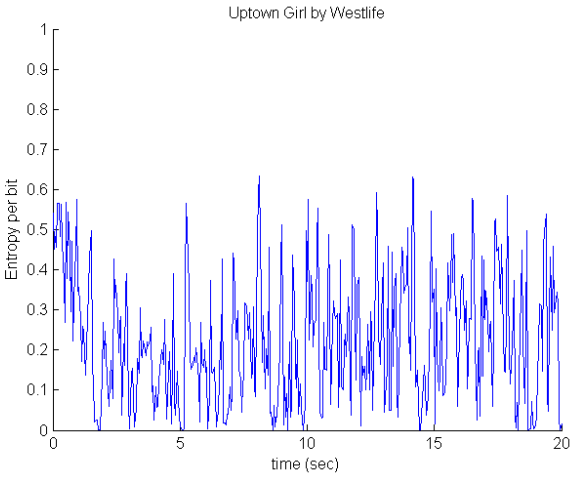


Figure 3. The entropy from *'Uptown Girl'*

Figure 3 is from 'Uptown Girl' which starts with fast drum beating for the first 1 second. Then, singers and band start playing, thus we can see the entropy values close to 0 starting around 2 seconds. The drum beating is loudly repeated during this audio snippet, and we can see that higher entropy values between 0.3 and 0.6 are well synched with drum beats. As to the low entropy values, the main singer starts singing by stressing the words "*up-town-girl*" around 8 seconds, and we can clearly see the three entropy values close to 0 between 8 to 10 seconds in Figure 3.

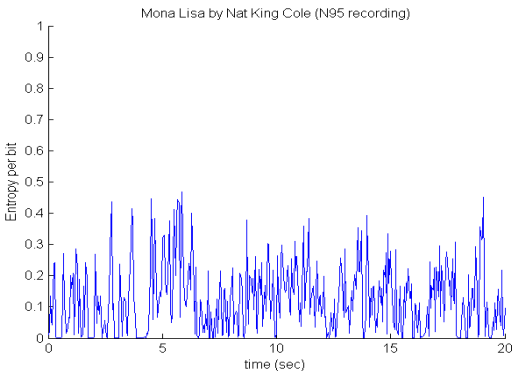


Figure 4 (a) N95 recording

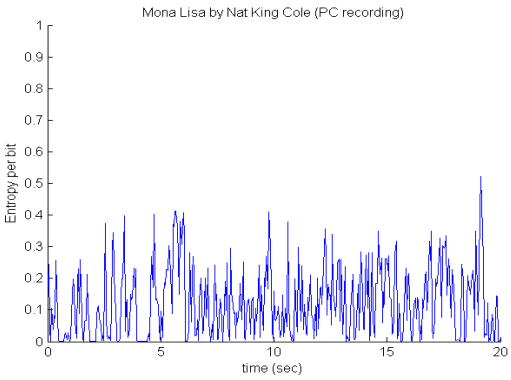


Figure 4 (b) PC recording

Figure 4 (a) is from the N95 recordings of "Mona Lisa" by Nat King Cole, which starts with a guitar plucking repeated 12 times. We can see the entropy values are close to 0 from the start to 5 seconds. After the 12th plucking at around 5 seconds, the dispersed music sound becomes noisy sound until new notes are presented at around 6.5 seconds, which we can verify by seeing the entropy value going down around that time.

Figure 4 (b) is from PC recordings of the same song, and this comparison shows how the AGC in handsets can make a difference when it comes to recording. If we listen to the recording from N95, we can see that the guitar plucking sound level at around 5 seconds abruptly decreases shortly after plucking. This is why the entropy value around 5 seconds in Figure 4 (a) is a little off from 0 whereas this plucking is clearly presented in Figure 4 (b) since there is no AGC in PC. Similar AGC effects are presented at around 19 seconds, which we can see the differences by comparing Figure 4 (a) and (b) at around 19 seconds.

The experiments were done with 6 seconds music and noise recordings from N95. The entropy value graph is sorted in increasing order. Figure 5 shows the unsorted time order entropy on the left and sorted entropy graph on the right from a noise sample recorded in a silent room. Figure 6 shows them for the music 'Green Hornet' by Al Hirt which has a very fast trumpet sounds in it.

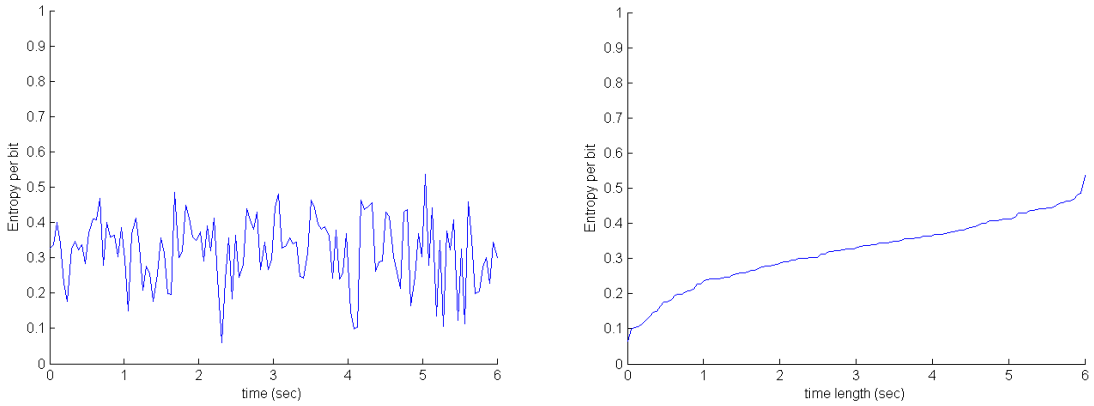


Figure 5. The entropy from a noise sample

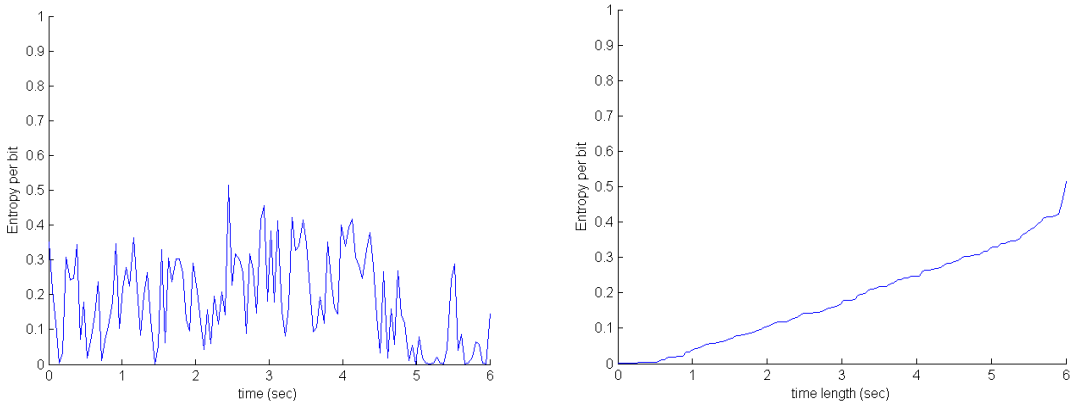


Figure 6. The entropy from a music sample

Different tendency is shown in the sorted entropy graph between Figure 5 and 6. First of all, y-intercept is higher in Figure 5, suggesting that the noise sample has little low valued entropy, which is quite reasonable given that this noise signal does not have harmonic components. Also, the entropy value of Figure 6 increases slowly compared to that of Figure 5, which means that the music signal has an abundance of harmonic sounds in it. Thus, we can generally say that the area between the sorted graph and x-axis from origin to 1 second will have smaller value in Figure 6. Based on this tendency, a classifier function was defined as follows:

$$c(x) = \begin{cases} music & if \quad \sum_{i=1}^{i=\frac{N}{6}} Entropy_{increasingly\ sorted}(i) < threshold \quad where\ N\ is\ the\ number\ of\ samples \\ noise & otherwise \end{cases}$$

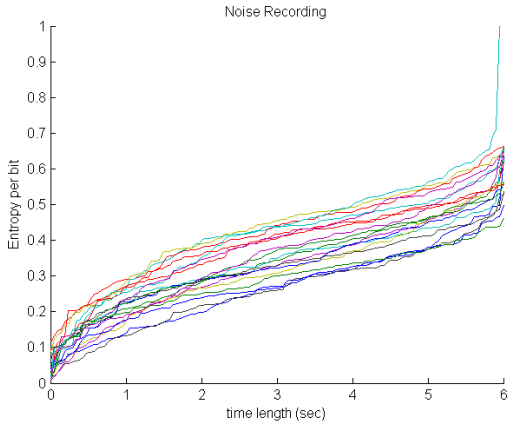


Figure 7 (a) Entropy from Noise samples

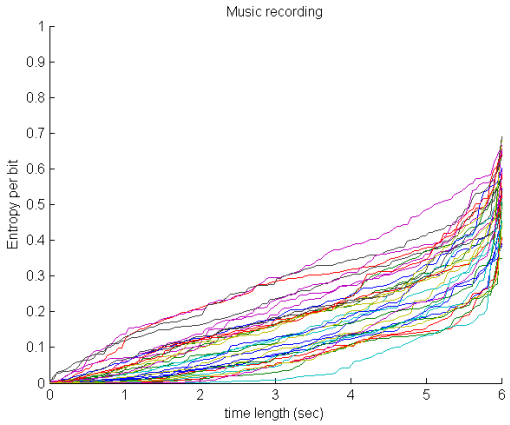


Figure 7 (b) Entropy from Music samples

5. DISCUSSION

Figure 7 shows experimental results with music and noise. Noise sources include silent room, wind, rain, stepping on the street, wave on the shore, and moving cars in distance.

In Figure 7 (a), it is noticeable that the entropy values are increasing rapidly from the origin to 1 second. The sky blue color graph is from one of the noise samples from Sony Ericsson, and it has reached to entropy 1, the maximum entropy. It turned out that this sample has some amount of total silence (amplitude 0) in it, and it explains the presence of maximum entropy.

In Figure 7 (b), we can see that most music samples show exponential-shaped graph, which grows very slowly at the origin as we expected. We can also see that some music samples show entropy tendency close to that from noise. This can be explained by some examples.

Music such as “Don’t let me be misunderstood” by Animals or “Mickey” by Toni Basil have only percussion instruments for the first 6 seconds, and the algorithm presented in this work relies on harmonicity/inharmonicity information through entropy model. Thus, this algorithm classifies the music as noise since the HPS distribution of the music would be flat, close to uniform distribution, and the entropy of this distribution will be higher than that of normal harmonic music. Actually, the percussion sound itself is noisy; it is our perception that recognizes them as music. Nonetheless, this algorithm can still work for percussion sounds if there is some amount of harmonic sound. Figure 8 shows this situation. The first 8 seconds of “Don’t let me be misunderstood” only consist of percussion instruments. This percussive sound lasts throughout the music. When the electric guitar sound emerges at around 8 seconds, we can see low entropy values originated from the guitar sound are also being presented. At around 23 seconds, the guitar starts playing the main melody. The guitar sound hits its maximum loudness around 28 seconds, and this is shown as a low entropy values at the graph. Shortly after 40 seconds, other non-percussive instruments start playing.

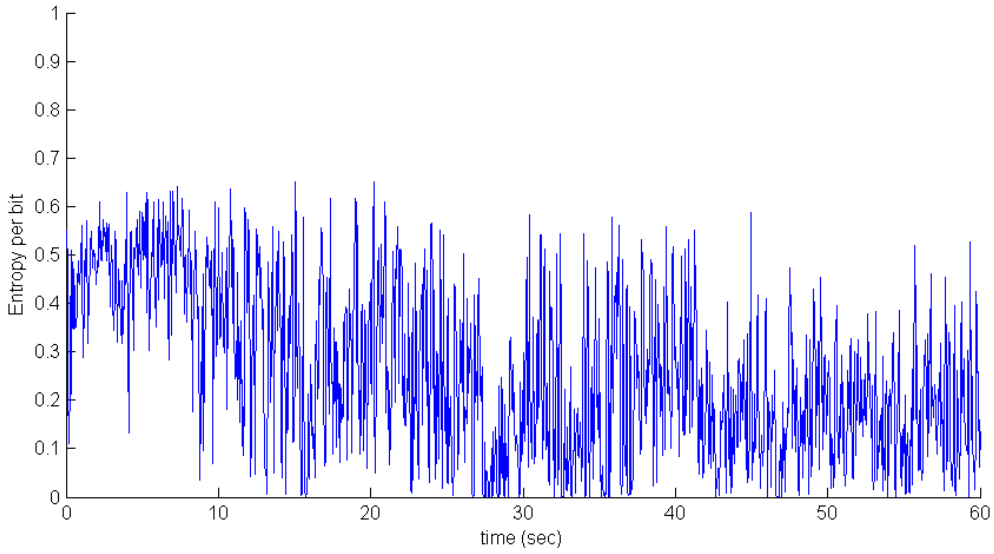


Figure 8. The entropy from Don't let me be misunderstood

Harmonic Product Spectrum itself is prone to fail at determining the fundamental frequency of the signal when the inharmonicity of the signal increases. This is because one high peak in HPS distribution with harmonic signal will be spreading into several low peaks when the inharmonicity increases. The details of this effect for piano sound can be found at [7].

Although the HPS has the limitation when it comes to picking up the fundamental frequency of the signal, if only the harmonic characteristics are presented in the signal, this will result in some peaks in the HPS distribution. Thus, the HPS combined with entropy works well even with the mixture of harmonic and inharmonic signals.

6. CONCLUSION

In this paper, we have presented a novel music/noise classification method based on entropy model with Harmonic Product Spectrum. Experimental evaluation shows that this method can work with reasonable classification power based on the amount of harmonicity. More elaborated music/noise classification would require incorporating onset detection for percussive sounds.

7. REFERENCES

[1] T. M. Cover and J. A. Thomas, Elements of Information Theory. New York: Wiley, 1991.
[2] "Pitch Determination of Human Speech by the Harmonic Product Spectrum, the Harmonic Sum Spectrum and a Maximum Likelihood Estimate," Proceedings of the Symposium on Computer Processing in Communications, Vol. XIX, Polytechnic Press: Brooklyn, New York, (1970), pp. 779-797.
[3] http://www.gracenote.com/business_solutions/mobileMusic/
[4] See project at https://ccrma.stanford.edu/~joshua79/application/

[5] SHANNON, C.E., & WEAVER,W. (1949). The mathematical theory of communication. Urbana, IL: University of Illinois Press.
[6] http://www.nap.edu/openbook.php?record_id=10094&page=281 Memorial Tributes: National Academy of Engineering, Volume 9 (2001) page 281
[7] http://www.speech.kth.se/prod/publications/files/qpsr/1994/1994_35_1_135-144.pdf