You Can Write Numbers Accurately on Your Hand with Smart Acoustic Sensing

Mingshi Chen[1], Panlong Yang[2(✉)], and Ping Li[1]

[1] College of Communications Engineering,
PLA Army Engineering University, Nanjing, China
cms603421@gmail.com, pingli0112@gmail.com
[2] School of Computer Science and Technology,
University of Science and Technology of China, Hefei, China
panlongyang@gmail.com

Abstract. Although smartwatch has drawn many attentions in recent years, small and inconvenient interaction mode limits the prevalence of smartwatches. Writing numbers with hands will naturally extend the input interface for smart watch. In this work, we design a passive acoustic sensing, where smart watches are collecting the ambient sound during writing. First of all, we use the wavelet transformation to mitigate the surrounding noise, and devise the time-frequency figures for AI enabled processing. After that, we apply the CNN(Convolutional Neural Network) model for number recognition, where three layers of convolution and three layers of max pool are incorporated. The number recognition accuracy rate could be above 95% when single person is well trained, and be around 92% when 7 to 9 persons are incorporated.

Keywords: Smartwatch · Wavelet transformation · CNN

1 Introduction

Smart devices have advanced to serve as an inseparable tool and aid for daily life. However, small touchscreen makes the basic selections cumbersome and fallible, and it's inconvenient when taking more complex actions such as typing a long list of phone numbers. For this concern, can we turn the hand-back into a virtual writing plane for interactions with the smart wearable device? Since skin has been applied for a natural extension for interaction [1–4], we can leverage it for operations beyond screen. For instance, we can treat our area of hand-back as a larger interaction surface for writing numbers keys. Such a system can be integrated into the smart wearable devices to enable more convenient operations. Existing work of skin computing and around device interaction either requires dedicated hardware [1–3, 5–8], or instruments the finger with a set of sensors [9–11], limiting their experience of interaction. It is worth nothing that, acoustic sensing is an innovative technology in extending the application scenarios of microphone. For acoustic sensing, it should include the following favorable properties:

L. Wang et al. (Eds.): QShine 2017, LNICST 234, pp. 207–217, 2018.
https://doi.org/10.1007/978-3-319-78078-8_21

- High-accuracy: the traditional input mode is limited by the small screen, especially smart watch. For user-friendly experience and ease of input consideration, people need an input device with high accuracy.
- Adaptability: the input mode should be adaptive to different users and working environments. Especially when considering the personalized users, good performance should be provided in a consistent way.

Unfortunately, there are two intrinsic challenges need to be formally addressed before this inspiring vision could be achieved.

- First, the acoustic signal induced by writing numbers on hand-back is weak. Even worse, the background noise is usually strong, which will possibly lead to errors in number recognition.
- Second, the acoustic features are diverse across persons, even for same person at different time. A stable and sensitive design should be encouraged for ease of imputing when input behaviors are fully respected.

There are two major contributions in our work.

- We present a dual-threshold scheme to deal with the strong background noise for segmentation. The threshold values are carefully selected according to various tests and show satisfiable performance across those scenarios.
- We conduct extensive evaluations to validate our design. Evaluations are made across different volunteers in various scenarios. Specifically, we show effectiveness and accuracy of the number writing behaviors on hand-back. It paves the way for alphabet writing or drawing for future designs.

The rest parts of this paper are organized as follows: First of all, in Sect. 2, we outline the basic design idea and components with working flow illustrations, and then present a comprehensive introduction with technical details for system design and implementation in Sect. 3. Secondly, we demonstrate our design performance with experimental results and analysis in Sect. 4. Finally, we make a conclusion in Sect. 5.

2 System Design

As Fig. 1 shows, our system implementation consists of four primary parts, namely sampling, effective signal segmentation, feature extraction and input recognition.

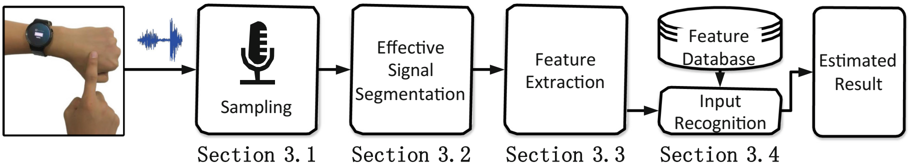


Fig. 1. Working flow

When user is writing on the hand-back, the acoustic signal, generated by the friction between the finger and the surface of hand-back, is captured by the built-in microphone sensor of smartwatches (Sect. 3.1). When it comes to input recognition through acoustic features, there are two points of feasibility:

First: hand-back is almost the closest input position to the smartwatch, except for the screen itself. Therefore, it is possible to capture the acoustic signal of input writing, in terms of the distance, which reduces the interference of ambient noise.

Second: as in the article [14], the acoustic characteristics generated by the desktop writing can identify user input. Similarly, the acoustic signal generated by users writing on the hand-back, even though weakness, it contains sufficient features to input recognition.

Then we segment the collected signal, and the effective writing signal fragments are extracted by analyzing the characteristics, such as short time energy and zero-crossing rate. In order to eliminate the impact of sudden noise, we take full advantage of the build-in gyro sensor in smart device to determine whether the user is in a writing state (Sect. 3.2). In the following, the effective acoustic fragments are subjected to spectral analysis and characteristics extraction. Since We use the convolution neural network (CNN) for classification training and recognition, we convert the features into picture for preservation (Sect. 3.3). Eventually, the characteristics and labels are imported into CNN, train in advance. In the actual writing recognition process, the user directly gets the results of CNN classification, which is done by off-line training and on-line identification (Sect. 3.4).

In our system implementation, the accuracy rate of 0–9 numbers recognition reaches more than 90%.

3 Implementation

3.1 Sampling Process

We use the built-in microphone on smartwatch, the position of which is closest from the hand-back of writing, to do a favourable collection of acoustic signals. In order to facilitate the subsequent segmentation and judgment of effective signal, we also collect gyro sensor data simultaneously. Through the vibration caused by finger sliding, we determine whether the user is in a state of writing. In the implementation, we used the Android smartwatch of HUAWEI WATCH 1, call the AudioTract API to collect the audio, AudioRecord API to record the audio and SensorManager API to collect the gyroscope data.

3.2 Effective Signal Segmentation

The main basis of signal segmentation is that, after filtering and denoising, the short-time energy of effective signal is higher than the ambient signal, so the position of the effective acoustic signal can be found by peak detection. Coupled with the gyro sensor peak detection as auxiliary confirmation, we can accurately split out the writing signal. This module is divided into three steps, respectively, preprocessing, peak detection, segmentation.

Preprocessing: built-in microphone sensor of smartwatch, whose default sampling rate is 44100, can fully collect the surrounding acoustic signal. First of all, we make a wavelet time-frequency analysis to acoustic signal, in favor of voice segmentation and feature extraction. As shown in the Fig. 2, the acoustic signal, in the intermediate frequency of which has a significant effective signal area, and the intensity is remarkable, there has almost no other interference signal in time domain of the entire frequency band. Since the acoustic signal is only used for location in this section, we do a 10 times down-sampling in the process of searching for the position, not only accurately locates the position of effective signal writing fragment, but also reduces the amount of data processing calculations.

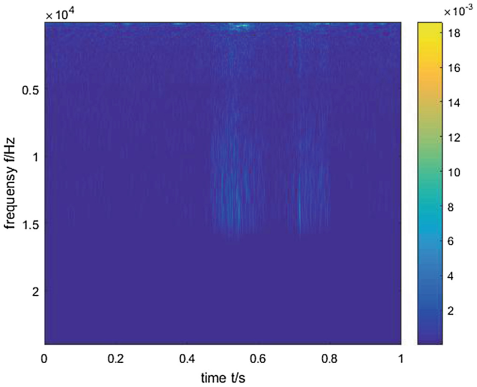


Fig. 2. Wavelet time-frequency figure

Peak detection: we denote amplitude data of the original signal, which has been down-sampled, as x(n). As for the acoustic signal, in the 10–30 ms short time, can be regarded as a quasi-steady state, we split it through short-term energy and zero-crossing rate (the formula is as following). These two features are often used for a voice signal detection, and segmenting effective voice. We set the two thresholds of short-term average energy and zero-crossing as 0.2 and 0.3, based on experiment and experience.

Short time energy formula:

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

Zero-crossing rate formula:

$$Z_n = \sum_{m=0}^{N-1} |sgn[x_n(m)] - sgn[x_n(m-1)]| \tag{2}$$

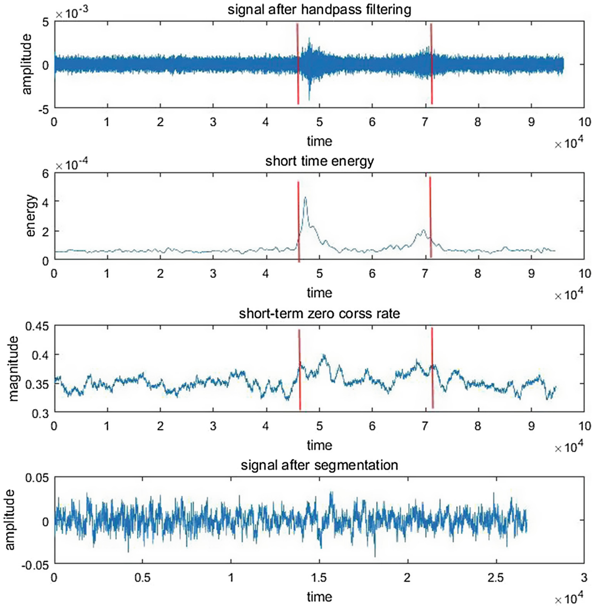


Fig. 3. Segmentation through short time energy and short-term zero cross rate

Where sgn[.] is a sign function, N is the size of window and n is the sequence number.

$$\operatorname{sgn}[x] = \begin{cases} 1, & (x \geq 0) \\ -1, & (x < 0) \end{cases} \tag{3}$$

As we can see from Fig. 3, these two features ensure that we can split out effective signals, but there may be a sudden outbreak of environmental noise is partitioned into an effective fragment, which increases the false error rate. Thus, we introduce the gyro sensor data, and the gyroscope data x, y and z of the three directions are summed to obtain g. Calculating the short term energy of g, and then, we perform peak detection to find the peak position. If the peak position is within one second interval on the middle of the effective signal previously split, it determined the current fragment is an effective writing signal.

Segmentation: with the position of effective signal, we segment effective fragment on original signal, for the following operation.

3.3 Feature Extraction

As we know from the previous, the acoustic signal generated by writing on hand-back, whose frequency distribution is ranged from 5 k to 15 k, and has a time characteristic. After bandpass filtering, the main part is the writing signal, with less IF(intermediate frequency) noise. Since its own acoustic signal of hang-back writing is very weak, we need to find a efficient and complete descriptor, which can characterize it. After experimental comparison, as shown in the Fig. 4, we find that the sound spectrum has a optimum performance of the complete time-frequency characteristic. The other characteristic such as the MFCC(Mel-Frequency Cepstral Coefficients) [15], commonly used in the voice recognition, is not suitable for hand-written IF signal feature extraction. Because it mimics the human ear structure for feature extraction, is more sensitive to the low-frequency signal, in which the medium-high frequency feature information is damaged.

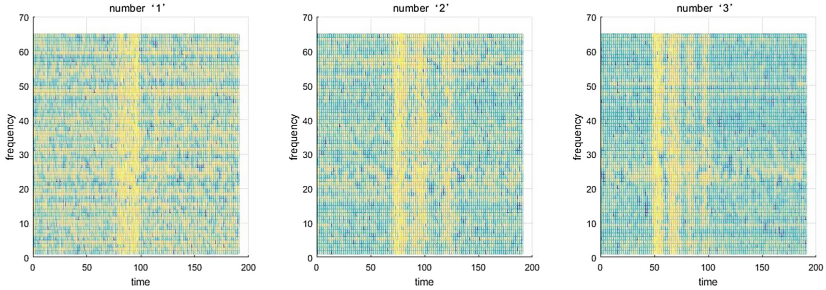


Fig. 4. Spectrograms of different numbers

Thus, we perform a sound spectrum analysis to the effective signal and save it as a spectrogram. We intercept the middle band 64 frames long, the eigenvalue of the whole time, converted to grayscale image for saving.

3.4 Feature Matching

In this paper, the structure of CNN is shown in Table 1. In addition to the input and output layers, the middle layer consists of three layers of convolution and three layers of the pool, the core sizes are 11 * 11, 5 * 5 and 3 * 3, respectively. Our framework was inspired by AlexNet [13], published in 2012, which obtained Imagenet best results in current year. AlexNet [13] is improvement of LeNet [12], which is the first neural network method of handwriting numeral recognition, emphasises more on the role of the whole connection layer. It adds the dropout layer, to prevent over-fitting, and reduce the number of weights. In the course of the experiment, we randomly divide the data into 8:2, the former is the training set, the latter is the test set, and calculate the final accuracy. Our recognition accuracy rate of single hand-back numbers writing reach 96% or more, even adopting multi-person data, the accuracy rate is more than 92%.

4 Performance Evaluation

In order to fully verify the performance of our proposed algorithm, we conduct a comprehensive experiment in real environment.

Table 1. Structure of CNN

Layer	Name	Configuration
1	Image input	64 × 64 × 3 images with ‘zerocenter’ normalization
2	Convolution	64 11 × 11 convolutions with stride [1 1] and padding [2 2]
3	ReLU	ReLU
4	Normalization	Cross channel normalization with 5 channels per element
5	Max pooling	3 × 3 max pooling with stride [2 2] and padding [0 0]
6	Convolution	128 5 × 5 convolutions with stride [1 1] and padding [2 2]
7	ReLU	ReLU
8	Normalization	Cross channel normalization with 5 channels per element
9	Max pooling	3 × 3 max pooling with stride [2 2] and padding [0 0]
10	Convolution	256 3 × 3 convolutions with stride [1 1] and padding [2 2]
11	ReLU	ReLU
12	Max pooling	3 × 3 max pooling with stride [2 2] and padding [0 0]
13	Dropout	50% dropout
14	Fully connected	256 fully connected layer
15	ReLU	ReLU
16	Fully connected	10 fully connected layer
17	Softmax	Softmax
18	Classification output	Cross-entropy

4.1 Experiment Setup

We set up the experiments in the lab, dormitory and canteen, where people often appear, and the noise level gradually increased. We implement our algorithm in the HUAWEI WATCH I with android 4.3 OS, by which we collect the user’s acoustic signals of hand-back writing. The smartwatch collects the writing signal and transmits the effective signal to the server, while, the server sends the result back to the watch terminal after processing. We achieve off-line processing, on-line identification.

We invited 10 volunteers (7 males, 3 females, evenly distributed at different ages), each of whom writes 50 times of each number, with total 5000 acoustic samples.

4.2 Various Experiments

Average accuracy of each number recognition: We first evaluate the average recognition accuracy of different numbers. We let volunteers wear smart-watch in their comfortable environment for hand-back writing. For purpose of marking labels, we require the volunteers write each number repeatedly at least 50 times. Then, we put collected signals into the algorithm, and statistics recognition accuracy of each number, the results shown in Fig. 5. As we can see that the overall accuracy rate is 90%, and the accuracy rates of some numbers, such as 4, reach 98%. It is 6 and 9, of which the lowest accuracy rates are only 88%. For a more deeply analysis of the differences in accuracy among numbers, we calculate the confusion matrices about recognition accuracy. As shown in the Fig. 6, the number 6 is easily mistaken for 1.

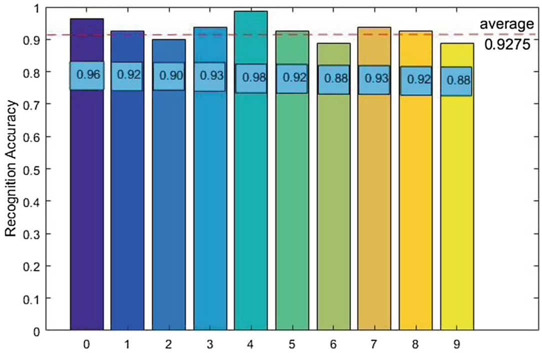
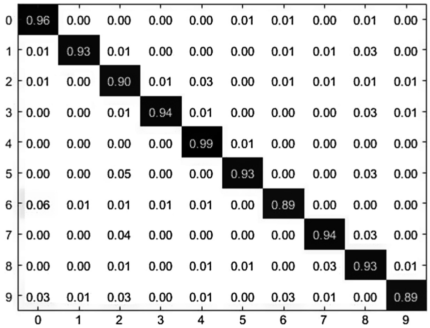


Fig. 5. Average accuracy of each number recognition



	0	1	2	3	4	5	6	7	8	9
0	0.96	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.00
1	0.01	0.93	0.01	0.00	0.00	0.00	0.01	0.01	0.03	0.00
2	0.01	0.00	0.90	0.01	0.03	0.00	0.01	0.01	0.01	0.01
3	0.00	0.00	0.01	0.94	0.01	0.00	0.00	0.00	0.03	0.01
4	0.00	0.00	0.00	0.00	0.99	0.01	0.00	0.00	0.00	0.00
5	0.00	0.00	0.05	0.00	0.00	0.93	0.00	0.00	0.03	0.00
6	0.06	0.01	0.01	0.01	0.01	0.00	0.89	0.00	0.00	0.00
7	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.94	0.03	0.00
8	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.03	0.93	0.01
9	0.03	0.01	0.03	0.00	0.01	0.00	0.03	0.01	0.00	0.89

Fig. 6. Confusion matrix among different digits

Average accuracy in different scenarios: To prove the strong environmental adaptability of our algorithm, we perform it in lab, dormitory and canteen, with the increasing noise level of environment, which were 45, 60 and 80 respectively. Similarly, each scenario, where we repeat 0–9 each for 100 times. From the Fig. 7, even in

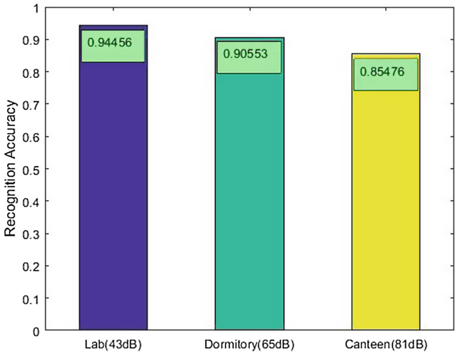


Fig. 7. Average accuracy in different scenarios

the most noisy place, canteen, our recognition accuracy rate still reaches 85%, that shows the excellent performance of our algorithm.

Average accuracy with different users: To evaluate the robustness of our algorithm to different users with write differences. We invite 10 volunteers, in the laboratory, repeatedly writing 50 times of each number. After that, the average writing accuracy of each person is calculated, of which the results are shown in Fig. 8. It can be seen that our algorithm performs well among different users although the accuracy of different users varies in the average accuracy rate of 95% fluctuation.

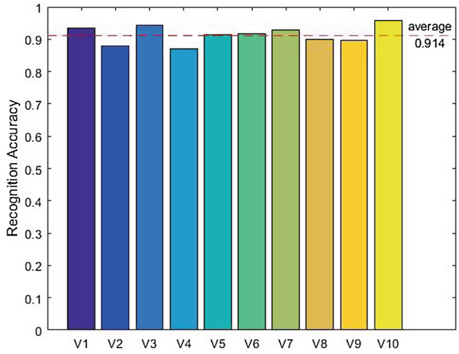


Fig. 8. Average accuracy with different users

The impact of training instances on recognition accuracy: The accuracy of CNN recognition depends on the effect of training, which definitively lies on the instances of training. So we put the acoustic samples into CNN, record statistics accuracy of recognition at different training times. As shown in the Fig. 9, as the increasing of training times, the recognition accuracy grows positively. But after 16 with reaching the peak of 94.5%, there has been a slight decrease in the rate of accuracy. Therefore, we set the number of training as 16 in the subsequent experiment process.

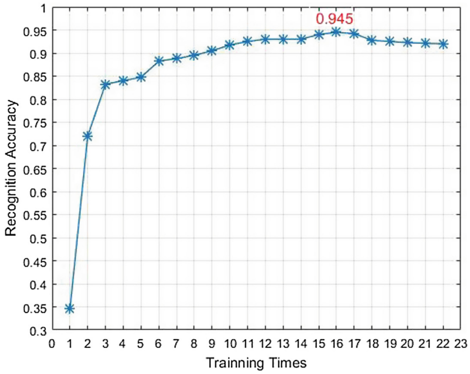


Fig. 9. Recognition accuracy with different training instances

5 Conclusion and Future Work

In this paper, we innovate the way we used for gestures recognition, which is directly based on the one-dimensional acoustic feature. Acoustic characteristics of two-dimensional information, time domain and frequency domain, is completely extracted and converted into images, combined with CNN, which has excellent performance in image classification, achieving fantastic results. In the identification of 0–9 numbers, we achieved an average accuracy of 92%, in a quiet environment, it even reached a 96% accuracy rate. Thus, not only the ability to identify numbers is demonstrated, but the possibility of discerning alphabets is also verified.

In the future work, we will add the experiment and verification of 26 alphabets. Let hand-back written changes people's input experience, and achieves innovation of wearable equipment.

Acknowledgement. This research is partially supported by 2017YFB0801702, National key research and development plan, NSFC with No. 61772546, 61632010, 61232018, 61371118, 61402009, 61672038, 61520106007, China National Funds for Distinguished Young Scientists with No.61625205, Key Research Program of Frontier Sciences, CAS, No. QYZDY-SSW-JSC002, and NSF OF Jiangsu For Distinguished Young Scientist: BK20150030.

References

1. Harrison, C., Tan, D., Dan, M.: Skinput: appropriating the body as an input surface. In: Sigchi Conference on Human Factors in Computing Systems, pp. 453–462 (2010)
2. Weigel, M., Lu, T., Bailly, G., Oulasvirta, A., Majidi, C.: iSkin: flexible, stretchable and visually customizable on-body touch sensors for mobile computing. In: ACM Conference on Human Factors in Computing Systems, pp. 2991–3000 (2015)
3. Huang, D.Y., Chan, L., Yang, S., Wang, F., Liang, R.H., Yang, D.N., Hung, Y.P., Chen, B. Y.: Digitspace: designing thumb-to-fingers touch interfaces for one-handed and eyes-free interactions. In: CHI Conference on Human Factors in Computing Systems, pp. 1526–1537 (2016)
4. Zhang, Y., Zhou, J., Laput, G., Harrison, C.: Skintrack: using the body as an electrical waveguide for continuous finger tracking on the skin. In: CHI Conference on Human Factors in Computing Systems, pp. 1491–1503 (2016)
5. Kratz, S., Rohs, M.: Hoverflow: exploring around-device interaction with ir distance sensors. In: International Conference on Human-Computer Interaction with Mobile Devices and Services, p. 42 (2009)
6. Hansen, J.P., Biermann, F., Jonassen, M., Lund, H., Agustin, J.S., Sztuk, S.: A gaze interactive textual smartwatch interface. In: ACM International Joint Conference, pp. 839–847 (2015)
7. Xiao, R., Laput, G., Harrison, C.: Expanding the input expressivity of smart-watches with mechanical pan, twist, tilt and click, pp. 193–196 (2014)
8. Perrault, S.T., Lecolinet, E., Eagan, J., Guiard, Y.: Watchit: simple gestures and eyes-free interaction for wristwatches and bracelets. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 1451–1460 (2013)
9. Chen, K.Y., Lyons, K., White, S., Patel, S.: uTrack: 3D input using two magnetic sensors. Springer (2015)

10. Chan, L., Liang, R.H., Tsai, M.C., Cheng, K.Y., Su, C.H., Chen, M.Y., Cheng, W.H., Chen, B.Y.: FingerPad: private and subtle interaction using fingertips. In: ACM User Interface Software and Technology Symposium, pp. 255–260 (2013)
11. Chen, K.Y., Patel, S., Keller, S.: Finexus: tracking precise motions of multiple fingertips using magnetic sensing. In: CHI Conference on Human Factors in Computing Systems, pp. 1504–1514 (2016)
12. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems, pp. 1097–1105 (2012)
14. Zhang, M., Yang, P., Tian, C., Shi, L., Tang, S., Xiao, F.: Soundwrite. In: The International Workshop, pp. 13–17 (2015)
15. https://en.wikipedia.org/wiki/Mel-frequency_cepstrum