Audio Engineering Society

Convention Paper 10350

Presented at the 148th Convention
2020 June 2-5, Online

Visualization of differences in ear acoustic characteristics using t-SNE

Rei Ominato[1], Shohei Yano[1], Naoki Wakui[1], and Shinnosuke Takamichi[2],

[1] *Yanolab, NIT, Nagaoka College*

[2] *The University of Tokyo*

Correspondence should be addressed to Shohei Yano (syano@nagaoka-ct.ac.jp)

ABSTRACT

Ear acoustic authentication is a biometric authentication technology that recognizes the acoustic characteristics of the ear canal to authenticate users. However, compared to fingerprints, ear acoustic authentication has not been studied sufficiently with regards to the individuality of the acoustic characteristics of the ear canal. Therefore, a study on the visualization of ear canal acoustic characteristic differences using t-distributed stochastic neighbor embedding (t-SNE) which expresses the similarity in high-dimensional space and estimates the similarity in low-dimensional space, was conducted.

1 Introduction

With the progress of digitization of the world, various services, such as shopping on the internet, SNS, net trading, etc., can be used without being subjected to physical restrictions. However, the number of offenses that exploit digital technology, illegal trading by impersonating others, and stealing confidential and personal information is increasing. Personal certification is becoming important to protect personal rights and privacy from such crimes, and realize a safe and secure society. Biometric authentication using different physical features of individuals compared to commonly used passwords, pins, and card keys has the advantage of a low impersonation risk, and no possibility of forgetting or losing passwords. Biometric authentication by fingerprints and face recognition has already been used for immigration review, entrance and exit management of important facilities, etc. Recently, it is being used for logging in terminals, such as in smartphones, and identity verification for online settlements [1]-[3]. Many certifications require authentication operations such as bringing fingers and eyes closer to the sensor while performing authentication. As the frequency of biometric authentication increases, people get annoyed with the authentication operations such as holding the finger over the scanner or gazing at the camera. Biometric authentication is expected to be required without authentication operations. Moreover, in many systems, authentication is performed at the start of the service, making it is difficult to detect "spoofing" in which users are interchanged during the service. Hence, solutions to these problems are required. Therefore, attention is paid to the ear hole (the ear canal) as a new biological information expressing a person's individuality. In the bio metric authentication by the ear hole, nonoperative authentication without requiring a conscious action can be realized by attaching an earphone-type authentication device. In ear acoustic authentication,

biological information can be acquired from both ears. Accuracy improvement is expected by combining two features, but it has not been studied yet.In this research, we confirm that there is a difference in the features of both ears by t-distributed stochastic neighbour embedding (t-SNE). In addition, frequency amplitude characteristics and mel-frequency cepstral coefficients (MFCCs) are calculated for time-series ustic characteristics. In the experiment, we used 50 subjects and the binaural features of 30 measurements.

2 Theory of ear acoustic authentication

2.1 Ear canal acoustic characteristics

The shape of the ear canal, which varies for different people, appears as an acoustic characteristic. The acoustic characteristics of the ear canal can be explained by an air column model. Here, the air column model explains the resonance of sound waves passing between the space filled with a medium (air), and the mechanism behind the fact that a musical instrument, such as a whistle, generates a specific sound. As shown in Fig. 1, an earphone with a built-in microphone is worn on the ear, and the acoustic characteristics of the ear canal are determined. These acoustic characteristics include the characteristics of reflection, diffraction, interference, and resonance depending on the shape of the ear canal and an individual's body. We can apply biometric authentication by treating these characteristics with individual differences as features. It is considered that the characteristic that the acoustic signal passes through the ear canal and is propagated to the eardrum depends on the shape and volume of the ear canal, and the acoustic impedance of the tympanic membrane surface and the likes. In this study, the ear canal transfer characteristic is defined as the characteristic related to the signal transmission between the earphone and the microphone installed at the entrance of the ear canal. The ear canal transfer characteristic becomes an ear canal impulse response (ECIR) in the time domain, and it is represented by an ear canal transfer function (ECTF) in the complex frequency domain. An ECIR can be derived by measuring the acoustic characteristics by an impulse response measurement method using a time stretched pulse signal or maximum length sequence (MLS) signal. The ear canal transmission characteristics include the acoustic characteristics of the ear canal, electroacoustic conversion characteristics of the earphone, acoustoelectric conversion characteristics of the microphone, and characteristics depending on the positional relationship between the microphone and earphone. The ear canal transfer characteristics, together with the head acoustic transfer characteristics, are used for the out-of-head sound image localization technique for creating a sound image at an arbitrary spatial position. In ear acoustic authentication, attention is given to the individuality possessed by the ear canal transfer characteristics, and biometric authentication is performed by applying a feature amount extraction processing to the discriminator.

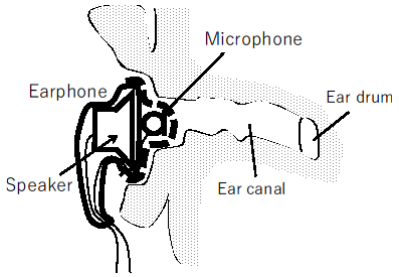

Fig. 1: Theory of ear acoustic authentication

2.2 ECIR measurement method

ECIR is measured using the measurement system of Fig. 2. The canal-type earphone (BOSE Sound True Ultra) is used. As shown in Fig. 3, by installing the

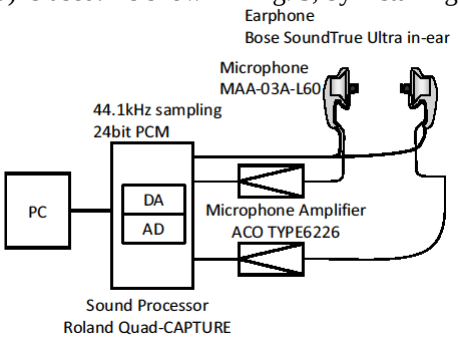

Fig. 2: The measurement system

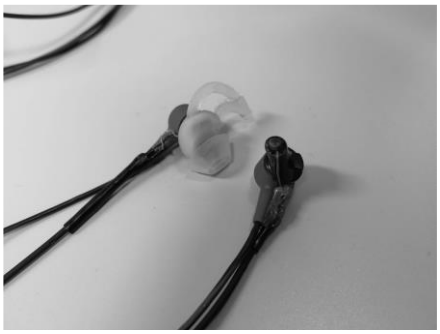

Fig. 3: The special earphone

small microphone on the earphone and forming it as an integral type, the sound collecting portion of the small microphone enters the ear canal when the earphone is attached, and it is possible to collect the sound signal in the ear canal. Using the MLS signal (signal length 2^{14}-1), the measurement signal is reproduced at the ear noise level of about 70 dB (A) and the number of synchronous addition is 5 times. For ECIR data, a minimum phase transformation process using Hilbert transformation is applied. After bandpass filtering (100Hz-22kHz) is performed, normalization processing is performed to set the power of the whole signal to 1.

3 Features

3.1 Frequency amplitude characteristic

The ECIR is f(t). Frequency amplitude characteristic F(ω) is one of the ear canal acoustic characteristics obtained by the FFT of time series data f(t).

$$F(\omega) = 20log \int f(t)e^{-j\omega t}dt \qquad (1)$$

3.2 Mel-frequency cepstral coefficients

The MFCCs features F(m) has M number of mel-filter banks H(ω, m) for the amplitude spectrum X(ω). It is one of the ear canal acoustic characteristics, which is obtained by applying a discrete cosine transformation.

$$S(m) = log\left(\sum_{m=0}^{T-1}|X(\omega)|\,H(\omega,m)\right) \qquad (2)$$

$$F(m) = \sum_{m=1}^{M} logS(m)\,cos\left[\frac{\pi n}{M}\left(m-\frac{1}{2}\right)\right] \qquad (3)$$

3.3 Signal length reduction

f(t) has a signal length of 16,384 for about 0.37ms on the time axis. As the signal is very long, a cutting process is performed from the beginning, and the signal length becomes 4096. After that, bandpass filtering (100Hz ~22kHz) is performed.

4 t-distributed stochastic neighbor embedding

t-SNE is a kind of dimension reduction and is known to be effective for nonlinear data. Dimension reduction generally means that while maintaining the relationship between data consisting of N high-dimensional vectors X = (x_1, x_2, ..., x_N), it is possible to grasp the relationship between data points by drawing a scatter diagram of data consisting of low-dimensional vectors Y = (y_1, y_2, ..., y_N). Among these dimension reduction methods, among these dimension reduction methods, t-SNE specializes in visualization because it is an image in which high-dimensional data is arranged on a two-dimensional plane while maintaining its relationship. The feature of t-SNE is that the proximity between two points is represented by a probability distribution. In t-SNE, we consider a normal distribution centered on the reference point x_i. First, we define posterior probabilities $p_{j|i}$ representing the proximityof point x_j to point x_i.

$$p_{j|i} = \frac{exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k=1}^{N} exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right) - 1} \qquad (4)$$

σ_i indicates the standard deviation corresponding to the i-th data. Based on this, σ_i defines the target probability $p_{i|j}$.

$$p_{ij} = \frac{p_{j|i} - p_{i|j}}{2} \quad (5)$$

Therefore, p_{ij} is larger at points closer to x_i, and smaller at points farther from x_i. Next, consider the probability q_{ij} representing the closeness of the points y_i and y_j after the dimension reduction. This corresponds to x_i and x_j before dimension reduction. Proximity after dimension reduction is also expressed by probability distribution, but not by normal distribution, but byt-distribution with one degree of freedom.

$$q_{ij} = \frac{\frac{1}{1 + \|y_i - y_j\|^2}}{\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\frac{1}{\|y_k - y_l\|^2}\right) - n} \quad (6)$$

The position of the point y_i after the dimension reduction is obtained by calculating the Kullback-Leibler divergence of the probability distribution p_{ij} before the dimension reduction and the probability distribution q_{ij} after the dimension reduction and minimizing this. Let this Kullback-Leibler divergence amount be a loss function C.

$$C = \sum_{N}^{i=1}\sum_{N}^{j=1} p_{ij} log \frac{p_{ij}}{q_{ij}} \quad (7)$$

By using the t-distribution, which has a heavier tail than the normal distribution, the point y_i and the pointy_j after the dimension reduction can be located farther away if the data of the points x_i and x_j before the dimension reduction are some distance apart. In other words, by using the t-distribution, the positional relationship between data points before and after dimensionality reduction is such that close data is close and far data is farther. The minimization of the loss function C is performed by the gradient descent method using the gradient expressed by the following equation,which is obtained by partially differentiating the loss function C with y_i.

$$\frac{\partial C}{\partial y_i} = 4\sum_{N}^{j=1} \frac{(p_{ij} - q_{ij})(y_i - y_j)}{1 + \|y_i - y_j\|^2} \quad (8)$$

We used scikit-learn, a python machine learning library, to implement t-SNE.

5 Experimental Methods

We used the ECIR (30 times/person) of both ears for 50 subjects (male and female in their 10s to 40s), visualizing the data for every 5 subjects. We visualize the difference between the features of both ears.

5.1 Visualization of time series data

The signal length is cut from the beginning to 128 in order to get a place with a lot of change. This time series data is visualized using t-SNE.

5.2 Visualization of Frequency amplitude characteristic

The frequency bands were visualized in three categories: (a) 250Hz-1.5kHz, (b) 1.5kHz-16 kHz, and (c) 16kHz-22kHz. In sound source direction perception, frequency band (b) contains information important for hearing. (a) and (c) have low response levels and are difficult to perceive perceptually, but are buried in noise and are difficult to observe. However, differences between users that help with ear acoustics authentication may also be included in frequency bands (a) and (c).

5.3 Visualization of MFCCs

20, 30, …,120-dimensional MFCC is visualized by t-SNE.

5.4 Count features mixed with other clusters

As shown in Fig. 4, one graph of the visualization results shows a scatter plot of the characteristics of the right and left ears of five subjects. User: 1R and User: 1L denote User 1's right ear and User 1's left ear, respectively. The number of users who have both ears in a cluster are counted by looking at the graph of features which indicates that visualization is as useful as the biometric data. In Fig. 4, one of the features of

User: 3L is mixed with User: 3R cluster. Therefore, you can see that there is one user with similar characteristics of both ears. The percentage of mixed clusters was calculated.

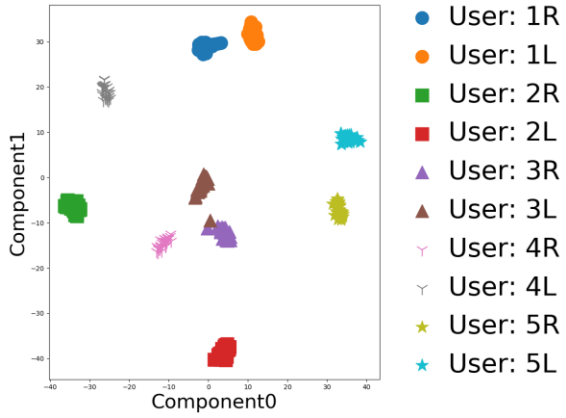

Fig. 4: Example of visualization result

6 Result

6.1 Visualization result of time series data

Fig. 5 shows a part of the visualization results of time-series data.

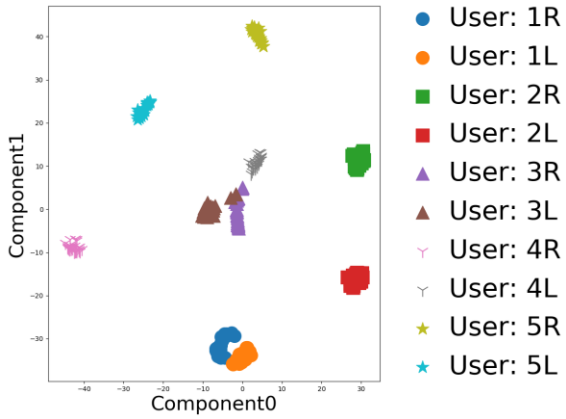

Fig. 5: Visualization result of time series data

The visualization showed that the time-series data was valid as biometric data because clusters were formed in each ear.

6.2 Visualization result of Frequency amplitude characteristic

Figures 6, 7, and 8 show the visualization results of (a) 250Hz-1.5kHz, (b) 1.5kHz-16kHz, and (c) 16kHz-22kHz.

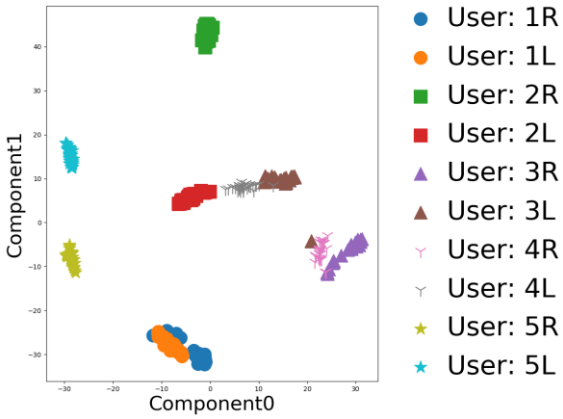

Fig. 6: Visualization result of (a)250Hz-1.5kHz

The visualization showed that the (a) was valid as biometric data because clusters were formed in each ear.

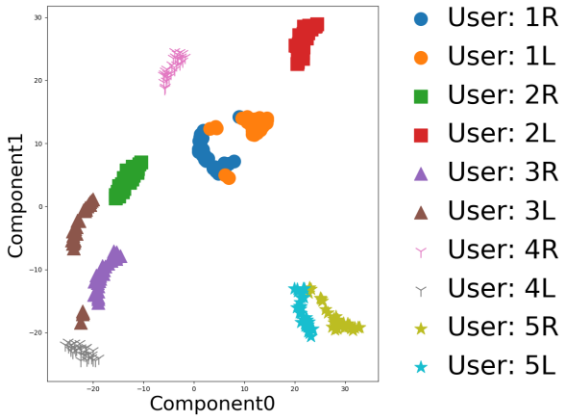

Fig. 7: Visualization result of (b)1.5kHz-16kHz

The visualization showed that the (b) was valid as biometric data because clusters were formed in each ear.

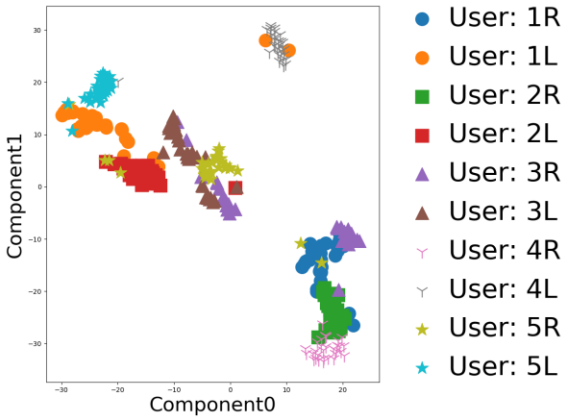

Fig. 8: Visualization result of (c)16kHz-22kHz

Visualization did not show that (c) was valid biometric data because no clusters formed in any ear. These results show that the low frequency band is valid as biometric data, but the high frequency band is not.

6.3 Visualization result of MFCCs

Fig. 9 shows a part of the visualization results of MFCCs.

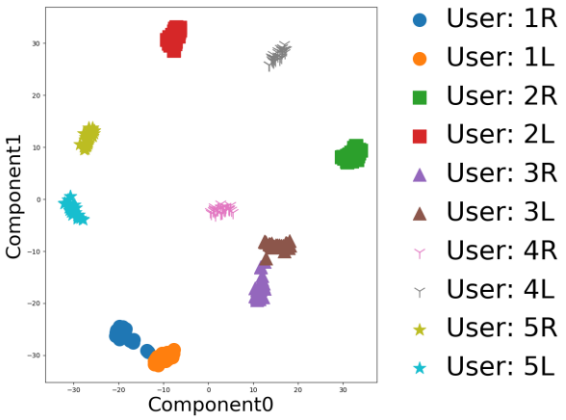

Fig. 9: Visualization result of MFCCs

In all dimensions, MFCCs were shown to be useful as biometric data.

6.4 Percentage of clusters mixed.

The MFCCs showed similar results in all dimensions, so we summarized them.

Table. 1: Percentage of clusters mixed

	Percentage [%]
time series data	16
(a)250Hz-1.5kHz	46
(b)1.5kHz-16kHz	34
MFCCs	20

The results in Table. 1 show that the time series data and MFCCs have a low rate of mixing of clusters.

7 Summary

In this research, we used t-SNE to visualize the differences in the characteristics of both ears. The percentage of mixed clusters was calculated. Visualization revealed that the high frequency side of the frequency amplitude characteristics was not effective as biometric data. From the calculation result of the ratio, it was found that the time series data and the MFCC did not have similar features of both ears compared to the frequency amplitude characteristics.

References

[1] M. Mizoguchi and M. Hara, "Fingerprint/ palmprint matching identification technology," NEC Technical Journal, vol. 5, pp.18--22, 2010.

[2] H. Imaoka, "Face recognition research: Beyond the limit of accuracy," The IAPR 2014 Biometrics Lecture, The 2014 International Joint Conference on Biometrics (IJCB 2014), 2014.

[3] T. Koshinaka, O. Hoshuyama, Y. Onishi, R. Isotani, and M. Tani, "Speech/acoustic analysis technology-its application in support of public solutions," NEC Technical Journal, vol. 9, no. 1, pp. 82–85, 2015.