A Study on Speaker Identification Approach by Feature Matching Algorithm using Pitch and Mel Frequency Cepstral Coefficients

Barlian Henryranu Prasetio
Interdisciplinary Graduate School of Agriculture and Engineering, University of Miyazaki, 1-1 Gakuen Kibanadai-nishi, Miyazaki-shi, 889-2192, Japan

Keiko Sakurai
University of Miyazaki, 1-1 Gakuen Kibanadai Miyazaki-shi, 889-2192, Japan

Hiroki Tamura
University of Miyazaki, 1-1 Gakuen Kibanadai Miyazaki-shi, 889-2192, Japan

Koichi Tanno
University of Miyazaki, 1-1 Gakuen Kibanadai Miyazaki-shi, 889-2192, Japan

E-mail: barlian@ub.ac.id, sakurai.keiko.u6@cc.miyazaki-u.ac.jp, htamura@cc.miyazaki-u.ac.jp, tanno@cc.miyazaki-u.ac.jp

Abstract

In this paper, we grouping the words based on the speaker in a sequence of speech in a conversation. There are two speakers in each conversation. The first speech assumed spoken by speaker-1. In recognizing the speakers, we use pitch detection and Mel Frequency Cepstral Coefficients feature extraction with 13 filters. Furthermore, we examine the distance of the second speech vector with the first speech vector using the Feature matching algorithm. Previously, we had experimented on each speaker to find out the mean and variance of the Feature matching. Based on the experimental results, the Standard Deviation of Euclidean, Mahalanobis and Manhattan Distance are 0.0383, 0.0254, and 0.0341. Hence, if the Feature matching value deviates is not more than variance value then the speech is assume spoken by speaker-1. Otherwise, the speech assume spoken by speaker-2.

Keywords: Speaker identification, Pitch, MFCCs, Euclidean, Mahalanobis, Manhattan.

1. Introduction

Stress is a mental disorder that occurs in a person due to pressure.[1] Stress is one of emotion. Emotions divided into two types, conscious and unconscious emotions.[2] Conscious emotions are emotions that we can feel like anger, sadness and happiness. Unconscious emotions are emotions that we cannot feel like stress and depression. It means, we cannot know when we are under stress. It makes relatively difficult to recognize stress.

Speech is one of methods that can be recognizing a stress.[3] It focuses on speech features in the frequency domain. Human voice signals have a very high level of variability.[4] A speech signal issued by different speakers produces different speech patterns. It makes a problem when we recognize a stress in a conversation. As we know, the conversation is a sequence of words spoken by more than one speaker. Therefore, in this paper, we

present a simple method of identifying the speaker in a conversation that indicated to have a stress speech on the speaker.

2. Materials and proposed method

2.1. *Proposed method*

In this paper, we used a sequence of words that has been segmented from two-pilot conversation in an Apache helicopter cockpit.[5] This conversation is recorded and collected by the Linguistic Data Consortium (LDC) in the Speech Under Simulated and Actual Stress (SUSAS) database.[6] This conversation indicates the stress condition of the speaker. The speaker identification consists of two stages, feature extraction and feature matching.[7] The process of our proposed method can be seen in Fig. 1.

The Fig. 1 shows that the speech conversations consisting of sequence words. The first speech extracted the frequency fundamentals pitch and Mel Frequency Cepstral Coefficients (MFCCs). Furthermore, we do the same thing in the second speech. We assume that the first speech spoken by the speaker-1. Furthermore, we examine the distance between the first speech and the second speech vector using the feature-matching algorithm (Euclidean[8], Mahalanobis[9] and Manhattan[10]). In the preliminary experiment, we have analyzed the standard deviation of each speaker. Therefore, if the distance between the two vectors is less or equal to the standard deviation, we decide that the second speech spoken by the speaker-1, otherwise spoken speaker-2.

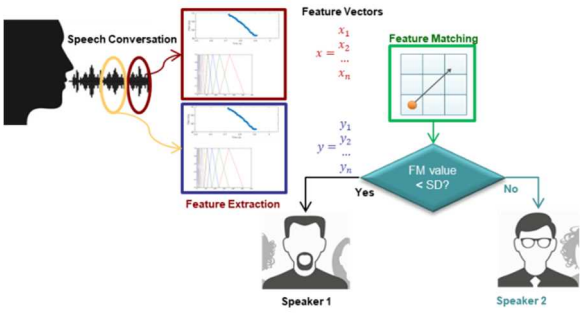

Fig. 1. Proposed method process.

2.2. *Feature extraction*

2.2.1. Pitch

The Pitch is the fundamental frequency of the vocal cord vibration (called F_0) followed by formants bandwidth at higher frequencies.[11] Typically, the male voice pitch is around 85-155 Hz and the female is about 165-255 Hz.[12]

$$F_0 = \frac{1}{2L}\sqrt{\frac{\sigma}{\rho}} \tag{1}$$

Where: L is length of vocal folds
σ is longitudinal stress
ρ is Tissue density

2.2.2. Mel Frequency Cepstral Coefficients (MFCCs)

The MFCC can use as a characteristic vector to represent human sound and musical signals. The sound analysis on Mel-Frequency based on the perception of human hearing,[13] because the human ear has observed to function as a filter at a certain frequency. The filters have a frequency response forming a triangle, and the space between their bandwidth determined by a constant mel-frequency interval.[14]

The triangular filters applied to computing filter banks.[15] We used 13 filters to extract frequency bands on a Mel-scale to the power spectrum. The model of filter bank on a Mel-Scale can be express as:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \le k \le f(m) \\ \frac{2(f(m+1)-k)}{f(m-1)-f(m)}, & f(m) < k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{2}$$

Where: m is the number of filters we want, $f()$ is the list of $m+2$ Mel-spaced frequencies.

The frequency band of the filter[16] shows in Table 1.

Table 1. The frequency band of the filter.

Filters	Passband Edges (Hz)
Filter1	[133 267]
Filter2	[200 333]
Filter3	[267 400]
…	…
Filter12	[867 999]
Filter13	[933 1071]

2.3. *Feature Matching Algorithm*

2.3.1. Euclidean Distance

The Euclidean distance between points x and y is the length of the line segment connecting ($\overline{xy}$). In Cartesian coordinates, if $x= (x_1, x_2,..., x_n)$ and $y = (y_1, y_2,..., y_n)$ are two points in Euclidean *n-space*, then the distance (d_{ED}) from x to y, is given by the Pythagorean formula:[17]

$$d_{ED}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (4)$$

2.3.2. Mahalanobis Distance

The Mahalanobis distance is a measure between two samples point. The Mahalanobis distance between a vector x and y with covariance σ is[18]

$$d_{MD}(x,y) = \sqrt{\sum_{i=1}^{n}\left(\frac{x_i - y_i}{\sigma_i}\right)^2} \quad (3)$$

2.3.3. Manhattan Distance

The distance between two points measured along axes at right angles. In a plane with x at (x_1, x_2) and y at (y_1, y_2), it is:[19]

$$d_{MH}(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad (5)$$

3. Results and Discussions

In this work, we used three conversations between two speakers. Therefore, we have from six speakers feature vector data. Furthermore, we calculate the standard deviation for each speaker in each distance algorithm, as follows:

- Euclidean Distance = 0.0383,
- Mahalanobis Distance = 0.0254,
- Manhattan Distance = 0.0341.

This standard deviation value is the threshold for determining speaker identification.

In the features extraction, each speech extracted from its pitch and MFCC features. The feature vector sample from feature extraction can be seen in Fig. 2.

Then on, we calculated the vector distance both of speech vector. Feature matching of each speech in the first conversation can be seen in Fig. 3.

Finally, we compare the number of speeches that have grouped with the actual number of speeches for each distance algorithm to determine the accuracy of our proposed method. The accuracy based on its features extraction. The system accuracy can be seen in Fig. 4.

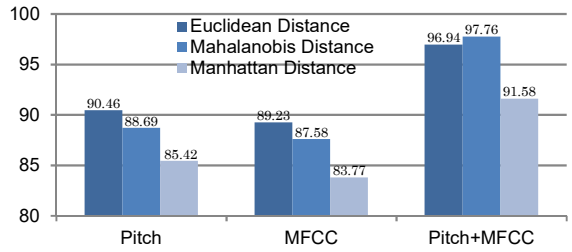

Fig. 4. The system accuracy.

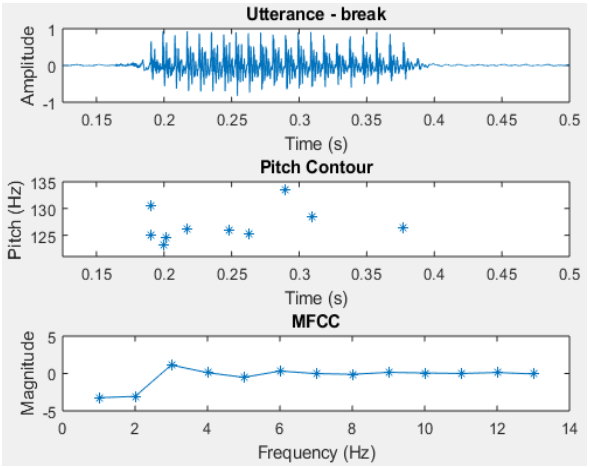

Fig. 2. The feature vector sample. The utterance is "*break*". Pitch features is 10 feature vectors and 13 feature vectors for MFCC

In Fig. 4 can be seen that in general the accuracy reaches above 80%. Accuracy for multi-feature extraction is better than single features, above 90%.

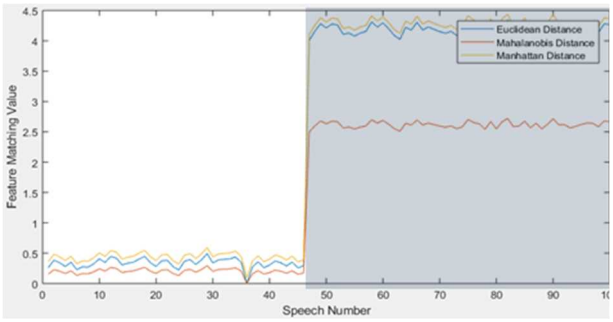

Fig. 3. The distance value on feature matching process of each speech in the first conversation.

Euclidean distance is better for single feature extraction. Mahalanonbis distance is better on multi-features.

4. Conclusion

The experiment involves three conversations between two speakers. The system accuracy rate calculated from the number of words clustered in both speakers by comparing pitch feature, MFCC feature and combination of pitch and MFCC with self-calculation. The experimental result shows that Euclidean Distance is better for single feature extraction and Mahalanobis Distance is better for multi-features extraction.

Acknowledgements

Thank you for Tamura Laboratory that has supported us in this study. Thank you to LDC that has given us the SUSAS Data.

References

1. N. Schneiderman, G. Ironson and S. D. Siegel, Stress And Health: Psychological, Behavioral, and Biological Determinants, *Annu Rev Clin Psychol,* 2005(1) (2005) 607-628.
2. A. K. Tiwari, Non-Verbal Communication-an Essence of Interpersonal Relationship at Workplace, *SMS Varanasi Management Insight,* 11(2) (2015) 109-114.
3. N. R. Prakash and J. Kaur, A Study on Physiological Parameters Used To Monitor Stress in Experimentally Induced Stimuli, *International Journal of Computer Science and Information Technologies,* 6(6) (2015) 5244-5246.
4. M. Latinus and P. Belin, Human voice perception, *Current Biology,* 21(4) (2011) 143-145.
5. J. H. L. Hansen, Composer, *SUSAS Transcripts LDC99T33.* [Sound Recording] (Philadelphia: Linguistic Data Consortium. 1999).
6. J. H. L. Hansen, Composer, *SUSAS LDC99S78. Web Download.* [Sound Recording] (Philadelphia: Linguistic Data Consortium. 1999).
7. S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal and R. Wang, Speaker identification features extraction methods: A systematic Review, *Expert Systems With Application* 2017(90) (2017) 250-271.
8. A. . S. Thakur and N. Sahayam, Speech Recognition Using Euclidean Distance, *Proc. International Journal of Emerging Technology and Advanced Engineering,* 3(3) (2013) 587-590.
9. G. Aradilla, J. Vepa and H. Bourlard, Using Posterior-Based Features in Template Matching for Speech Recognition, in *Proc. Interspeech ICSLP* (Pittsburgh, Pennsylvania, 2006).
10. M. D. Malkauthekar, Analysis of euclidean distance and Manhattan Distance measure in face recognition, in *Proc. Third International Conference on Computational Intelligence and Information Technology* (Mumbai, India, 2013).
11. N. Peleg, Homepage of Audio-Video Compression Course, 2 2010. [Online]. Available: http://cs.haifa.ac.il/~nimrod/Compression/Speech/S1Basics2010.pdf. [Accessed 23 3 2018].
12. N. Zivic, *Modern Communications Technology, Oldenbourg* (De Gruyter Studium, 2016).
13. L. Muda, . M. Begam and I. Elamvazuthi, Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques, *Journal of Computing* 2(3) (2010) 138-143.
14. A. V. Bhalla and S. Khaparkar, Performance Improvement of Speaker Recognition System, *International Journal of Advanced Research in Computer Science and Software Engineering* 2(3) (2012) 82-87.
15. H. Fayek, Haytham Fayek, 21 4 2016. [Online]. Available: http://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html. [Accessed 21 3 2018].
16. L. R. Rabiner and S. Ronald W, *Theory and Applications of Digital Speech Processing* (Upper Saddle River: Pearson, 2010).
17. H. Anton, in *Elementary Linear Algebra (7th ed.)* (John Wiley & Sons, 1994), pp. 170-171.
18. R. Wicklin, SAS Blog, The DO Loop , 15 2 2012. [Online]. Available: https://blogs.sas.com/content/iml/2012/02/15/what-is-mahalanobis-distance.html. [Accessed 24 10 2018].
19. P. E. Black, xlinux, NIST, 31 5 2006. [Online]. Available: https://xlinux.nist.gov/dads/HTML/manhattanDistance.html. [Accessed 21 10 2018].