Wavelet DB44 and MBB Algorithm for Sasak Vowels Recognition

1st Syahroni Hidayat
dept.of computer science
Universitas Bumigora
Mataram, Indonesia
syahroni.hidayat@stmikbumigora.ac.id

2nd Muhammad Tajuddin
dept.of computer science
Universitas Bumigora
Mataram, Indonesia
tajuddin@stmikbumigora.ac.id

3rd Ahmat Adil
dept.of computer science
Universitas Bumigora
Mataram, Indonesia
ahmat.adil@stmikbumigora.ac.id

4th Muhamad Nur
dept.of computer science
Universitas Bumigora
Mataram, Indonesia
insabil@gmail.com

line 1: 5th Andi Sofyan Anas
dept.of computer science
Universitas Bumigora
Mataram, Indonesia
andi.sofyan@stmikbumigora.ac.id

***Abstract*—Ancient manuscripts are original handwriting that is at least 50 years old and has significance for civilization, history, culture, and science. In the preservation of ancient manuscripts, an attempt was made to digitize ancient documents. The hope is that by carrying out this process, the old documents can be used in historical learning and increase the interest of the younger generation in studying history. So far, the attempt to digitize ancient manuscripts is still limited to the storage of manuscripts in the form of digital documents, generally in the way of image files. So, the existing digital manuscripts are still static, while each manuscript has specificity in how to read them. Therefore, this research is a preliminary study for digitizing ancient documents, especially voice-based Sasak language. The vowel sound of Sasaknese speakers used in this research. Wavelet db44 and WPCC used as the feature extractor, and FCM implemented to build the feature reference/model. The accuracy of the recognition is evaluated using the DTW algorithm. From the evaluation obtained that the average recognition accuracy of the system using training dataset is 50% while using testing dataset is 27.14%.**

Keywords—ancient manuscript, Sasaknese, Voice-based digitization, Fuzzy C Means, MBB Algorithm

I. INTRODUCTION

Ancient manuscripts are original handwriting that is at least 50 years old and has significance for civilization, history, culture, and science [1]. Other definitions are written works made directly by stationery and hands, not through mechanical stationery, such as typewriters, printing machines, computers. The writing of the manuscript is done in the past when mechanical writing did not yet exist, and its use was not yet widespread [1].

As a relic of the past, ancient manuscripts can provide information on various aspects of past people's lives such as politics, economics, socio-culture, traditional medicine, earthquake screenings or natural phenomena, human physiology, and so on. Initial information related to this can be found in the contents of the text for everyone to learn. The texts are important, both academically and socio-culturally. The text is a valuable identity, pride, and cultural heritage. In socio-cultural terms, the script contains values that are still relevant to the present life, so that it becomes a responsibility that has been on our shoulders to reveal the 'pearls' contained in it. Ancient manuscripts, besides as cultural documentation, can also be used as teaching objects to take the values and content in them. These values are needed in evaluating the value of goodness that has existed in the past to be applied today.

In the preservation of ancient manuscripts, an attempt was made to digitize ancient manuscripts [2]. The hope is that by carrying out this process, the ancient manuscripts can be used in historical learning and increase the interest of the younger generation in studying history [3]. So far, the attempt to digitize ancient manuscripts is still limited to the storage of manuscripts in the form of digital documents, generally in the way of image files. While preservation of old manuscripts also comes from oral traditions, such as a pepaosan ’in Sasak terms,‘ macapat ’in Javanese terms,‘ mamaca ’in Madura terms and‘ wawacan ’in Sundanese terms [4]. So, the existing digital manuscripts are still static, while each manuscript has specificity in how to read them, one of them is Sasak language.

Sasak language is a language spoken by most people of Lombok Island. Sasak language has a variety of dialects, both phonologically, vocabulary, and grammar [5]. Native speaker of Sasak language generally divide into 5 dialects based on words used to refer to "like that" and "like this", such as Kutó-Kuté (North Sasak), Nggetó-Nggeté (Northeast Sasak), Menó-Mené (Central Sasak), Ngenó-Ngené (Middle-Eastern Sasak, Central-West Sasak) and Meriaq-Meriku (South-Central Sasak) [6]. In the phonology of Sasak language, there are eight types of vowels [7]. These eight sounds are represented in the Latin languages as *a, e, i, o* and *u*. Sometimes diacritics are used to distinguish similar sounds Because the amount of these vocal, the researchers focused on discussing vowels in Sasak language. With this research, it can be used as a first step to digitize ancient manuscripts of Sasak language, especially sound-based.

II. METHODOLOGY

The sound of vowels uses in this research is *a, i, u, e, é, o,* and *ó*. The audio recorded from 50 adult speakers, each 25 male, and 25 female speakers. The recording processed at the open area and each speaker uttered the voice sound only once. The International Phonetic Association (IPA) standard of utterance applied in the recording process. The sampling frequency is 16000 Hz, which means the fundamental sound frequency is 8000 Hz. The raw signal saved in *.wav format. There are 350 samples of sound signal dataset uses in this research. Twenty percent of the data set uses as testing, and the others use as the training dataset.

The datasets of sound then denoised to eliminate its noise. The amplitude of each sound signal also made uniform, which is around ± 1 using normalization. The normalized and denoised sound signal then processed to obtain its feature in the feature extraction process.

Wavelet Daubechies db44 implemented in the feature extraction process. The consideration of choosing this wavelet family is based on research on the best wavelets for Indonesian vocal sound signals that have been conducted by Hidayat et al. by applying the cross-correlation method [8].

To obtained the features, the wavelet packet transforms implemented into the raw signal as the decomposition method. The level of decomposition is 6th level. The best level of decomposition for the signal is no more than the 7th level. The higher the level of the decomposition, the characteristics obtained are not representative anymore as the features.

After the decomposition process, its best nodes determined, so the best wavelet packet tree for each vowel signal was formed. Mean Best Basis (MBB) algorithm applied in this process to determine the best wavelet packet tree. This algorithm developed by Galka [9]. MBB algorithm is the improvement of Coifman-Wickerhauser best basis selection algorithm[10]. The research which has been conducted by Hidayat et al. [11] shows that there is a different form of the best wavelet packet tree from the implementation of db44 and db45 in the decomposition process. However, the best wavelet packet tree formed can represent the characteristics of MFCC, refers to the best nodes and the range of frequency of the sound signal. This feature referred to as Wavelet Packet Cepstral Coefficient (WPCC).

The features extracted as the WPCC features is the entropy and the energy of the signal. The features considered based on the research conducted by Diker et al. [12] dan Raj et al. [13]. They have implemented statistic features in their previous study. The entropy and energy included in the statistic features. These features then use in training and testing. Of all the datasets available, a total of 280 datasets used as the training dataset. And the other 70 datasets are used as the testing dataset.

The next step is determining the vector reference or model. The Fuzzy C Means (FCM) algorithm implemented in this step. FCM is generally used to determine several clusters of data sets. However, all data can be members of the other clusters depending on its fuzzy membership function degree. Each cluster has a centroid value. The closest distance between data to the centroid leads the data to be included in a particular cluster member. So, it can be assumed that the centroid is a value which represents all data around it.

In this research, the FCM algorithm applied to the training dataset. Each reference/model will represent the vocal *a, i, u, e, é, o,* and *ó*. So, each 40 dataset training represent as each vowel. The center value gets from FCM algorithm implemented into 40 datasets made as to the reference/model. The FCM only has one cluster as its output. The result of this step saved as the reference/model features. Details of the process shown in Fig. 1.

For the testing process, the steps are similar to the feature extraction steps described earlier. The difference lies in the final part, where the result of feature extraction in the testing process then tested with reference/model features. The details of the testing process shown in Fig. 2.

Dynamic Time Warping (DTW) applied in the testing of the recognition accuracy. The result is a table of recognition result for the training dataset with 280 x 7 matrix size while for the testing dataset is a table with 70 x 7 matrix size. The performance of db44 and WPCC as the feature extractor then evaluated by determining the minimum value of DTW results in the table of recognition.

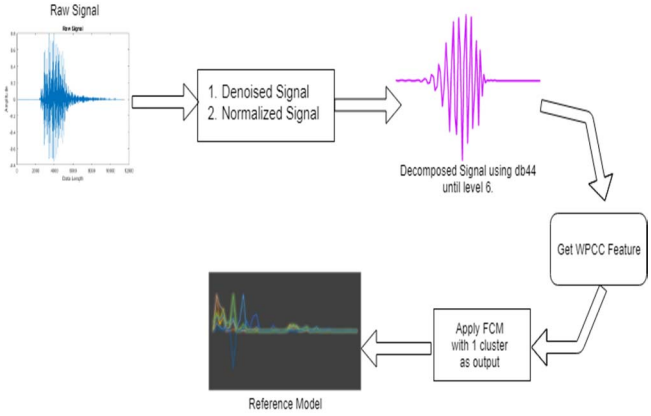

Fig. 1. The training steps of making vector reference/model

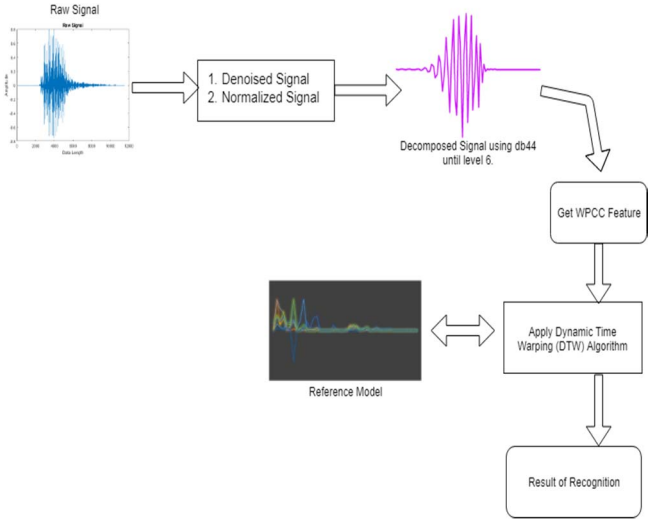

Fig. 2. The testing steps

III. Result and Discussion

From the training of 280 pieces of vowel sound datasets, a vector reference/model formed, as shown in Fig. 3. There are seven vectors characteristic which each represent vowels *a, i, u, e, é, o,* and *ó*. This characteristic vector is a statistical feature formed from the entropy and energy value of vowel signal. This feature is the centroid value obtained from the FCM algorithm with one output cluster.

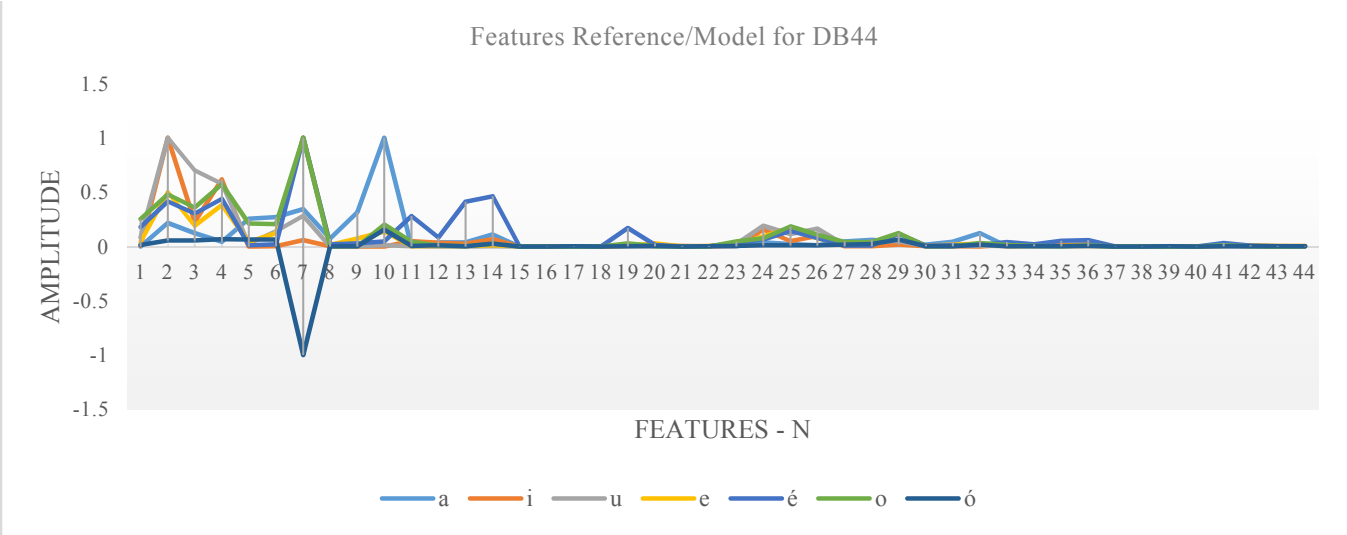

Fig. 3. Vector of feature reference/model from DB44

The characteristic vector obtained then used as a vector reference or model. Furthermore, after the whole vector reference or model purchased, the process is continued by forming the features of the sound testing dataset. In total, there are 70 vocal sound data used as testing datasets. Each voice represented by five male speakers and five female speakers. Each testing data is taken randomly from the dataset. From fig. 3 it can be seen that the variability of the vector reference/model is very high for the first 15 data. Furthermore, the variability begins to decrease in the remaining data. So, visually, this feature vector is quite representative as a vector reference/model.

TABLE 1. THE RECOGNITION RESULT FOR DB44 USING TESTING DATASET

Vowels	*a*	*i*	*u*	*e*	*é*	*o*	*ó*	Acc. (%)
a	1.62	**1.53**	1.74	1.58	1.65	1.75	1.59	40
	2.09	1.68	1.88	1.78	1.84	2.01	**1.67**	
	2.39	1.89	2.1	2.05	2.07	2.27	**1.85**	
	2.1	1.72	2.01	2.12	2.18	2.28	**1.23**	
	1.98	1.62	1.91	2.18	2.26	2.3	**0.58**	
	1.1	1.75	1.61	1.2	1.49	1.22	2.44	
	1.11	0.8	**0.72**	0.94	1.24	0.9	1.82	
	1.19	1.69	1.49	1.3	1.56	1.2	2.51	
	0.45	1.59	1.61	1.29	1.45	1.32	1.81	
	0.41	1.56	1.61	1.27	1.45	1.29	1.7	
i	1.58	1.83	1.94	1.76	1.91	1.97	**1.49**	10
	1.7	2.14	2.25	1.89	1.96	1.99	**1.49**	
	1.91	2.01	2.33	1.83	**1.58**	1.98	2.17	
	1.7	2.07	2.14	1.79	1.76	1.83	**1.62**	
	1.69	2.08	2.17	1.83	1.9	1.91	**1.5**	
	1.54	0.62	**0.45**	1.21	1.36	1.25	1.61	
	1.58	0.46	**0.32**	1.2	1.38	1.2	1.61	
	1.59	0.65	**0.27**	1.14	1.3	1.13	1.78	
	1.51	**0.46**	0.55	0.87	1.04	0.9	1.84	
	1.75	1.33	1.28	0.96	1.05	**0.85**	2.29	
u	**1.57**	1.72	1.73	1.65	1.83	1.84	1.65	20
	1.7	2.13	2.24	1.89	1.97	1.98	**1.47**	
	1.93	1.79	2.18	2.27	2.39	2.41	**0.73**	
	1.78	1.8	2.21	1.98	2.16	2.17	**1.23**	
	1.67	2.08	2.13	1.77	1.84	1.83	**1.59**	
	1.66	0.86	**0.39**	1.25	1.42	1.22	1.85	
	1.72	0.65	**0.31**	1.28	1.43	1.22	1.8	
	1.67	1.12	0.75	0.74	0.99	**0.68**	2.35	
	1.44	1.12	0.91	0.37	0.77	**0.35**	2.16	
	1.4	1.31	1.23	0.5	0.8	**0.49**	2.1	
e	1.61	1.67	2.05	1.63	1.84	1.84	**1.53**	20
	1.76	1.76	2.19	1.97	2.15	2.15	**1.22**	
	1.94	1.84	2.24	2.29	2.43	2.43	**0.77**	
	1.65	1.72	2.1	1.71	1.92	1.93	**1.46**	
	1.63	1.71	2.09	1.69	1.9	1.9	**1.48**	
	1.24	1.23	0.97	**0.54**	0.86	0.59	2.19	
	1.57	0.9	**0.55**	1.25	1.39	1.24	1.66	
é	1.52	1.52	1.7	1.48	**1.46**	1.62	1.73	30
	1.77	1.77	2.19	1.97	2.14	2.16	**1.23**	
	1.81	1.68	2.02	2.17	2.3	2.29	**0.41**	
	1.65	1.71	2.1	1.71	1.91	1.92	**1.46**	
	1.64	1.7	2.09	1.69	1.89	1.9	**1.48**	
	1.76	1.29	1.03	1	**0.76**	0.95	2.44	
	1.57	0.93	**0.54**	1.18	1.32	1.17	1.79	
	1.37	1.18	1.01	**0.35**	0.68	0.48	2.1	
	1.82	1.62	1.99	2.15	2.26	2.26	**0.49**	
	1.88	1.9	1.81	1.53	**1.34**	1.52	2.43	
o	1.62	1.7	2.08	1.66	1.87	1.87	**1.5**	10
	1.88	1.84	2.27	2.14	2.3	2.31	**1.13**	
	1.9	1.79	2.17	2.26	2.39	2.39	**0.67**	
	1.79	1.77	2.17	1.81	1.99	2.05	**1.54**	
	1.76	1.68	2.11	1.96	2.15	2.13	**1.2**	
	1.37	1.31	1.05	**0.52**	0.87	0.55	2.14	
	1.52	0.98	**0.55**	1.09	1.27	1.07	1.92	
	1.16	1.3	1.06	0.5	0.88	**0.46**	2.14	
	1.55	1.54	1.85	2.09	2.24	2.14	**0.45**	
	0.65	1.46	1.4	1.27	1.49	1.21	1.63	
ó	1.74	1.65	1.91	2.17	2.27	2.28	**0.62**	60
	1.65	1.72	2.11	1.7	1.91	1.92	**1.49**	
	1.84	1.64	1.94	2.17	2.26	2.28	**0.43**	
	1.9	1.69	2.02	2.2	2.31	2.34	**0.55**	
	2.05	2.05	2.45	2.37	2.53	2.56	**1.19**	
	1.44	1.2	0.98	0.44	0.84	**0.42**	2.18	
	1.44	1.22	0.86	0.66	0.97	**0.65**	2.26	
	1.2	1.2	1.08	**0.31**	0.83	0.49	2.09	
	1.65	1.56	2	1.91	2.12	2.01	**1.42**	
	1.49	1.81	1.81	**1.47**	1.58	1.51	2.4	

The process of establishing the feature of dataset testing follows the steps shown in Fig.2. After the characteristic vector obtained, the process continued by classifying the vowel sound with a reference vector. There are two

classification processes applied, first is classification between training datasets and vector reference/models, and second is the classifications between testing datasets and reference vectors. The dynamic time warping (DTW) algorithm applied at this stage. The result of this process is a recognition result, as shown in Table 1. The minimum value in Table 1, marked with red color, used as the closest value between the feature vector and the vector reference/model. It is obtained from the implementation of the DTW algorithm.

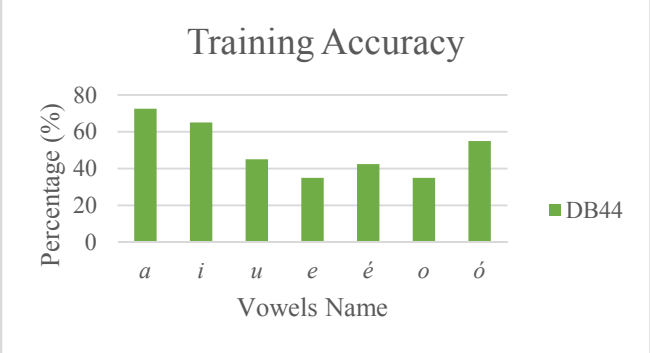

Fig. 4. The recognition accuracy of the training dataset

The accuracy of the system then calculated from the recognition result that has been formed. From the evaluation, the best recognition accuracy using training dataset is 72.5% for vowel *a*, and the lowest recognition accuracy is 35% each for vowel *e* and vowel *o*. While the average value of vocal voice recognition accuracy for training datasets is 50%. This result is clearly shown in Figure 4.

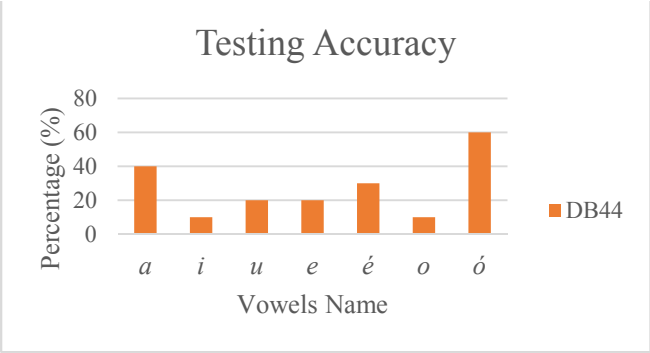

Fig. 5. The recognition accuracy of the testing dataset

The evaluation of the accuracy of the recognition using dataset testing is shown in Figure 5. The figure shows that the best accuracy obtained is 60% for vowel *ó*. And the lowest accuracy value is 10% for each vowel *i* and vowel *o*. So, the average recognition accuracy using a testing dataset is 27.14%. This little recognition accuracy value caused by the Lombard effect on the sound being processed. This Lombard effect is related to the noise contained by sound signals, which causes a decrease in the level of recognition accuracy[14].

Refers to the accuracy of speech recognition, based on gender, it can be concluded that the voice of a female speaker more easily recognized than the voice of a male speaker. It happened because women voice has a higher loudness than men. Besides woman voice frequency is higher than male's voice frequency too. This two-variable are the main component of MFCC since it is built from the energy of these two variables which extracted from the voice signal [15]. Also, the effect of the use of WPCC as the MFCC approach causes the voice quality of female speakers is better than the voice quality of male speakers [16].

IV. Conclusion

From the explanation above, it can be concluded that the application of db44 and WPCC as feature extraction has not provided maximum recognition accuracy. It can be seen from the accuracy of voice recognition using the training dataset that the best average recognition results are 50%. As for using the testing dataset, its best average recognition accuracy is 27.14%. However, the use of db44-WPCC as a feature reference/model still provides high variability. As shown in Fig. 3, the variability of the data in the first 15 characteristics is very high. Also, if it refers to the accuracy of the speech recognition to the gender influence, it can be seen that the voice of a female speaker is easier to recognize than the sound of a male speaker.

Acknowledgment

We want to thank RISTEKDIKTI for funding this research through the Applied Research scheme.

References

[1] U. Saraswati, "Arti dan Fungsi Naskah Kuno Bagi Pengembangan Budaya dan Karakter Bangsa melalui Pengajaran Sejarah," 2017.

[2] M. Tajuddin, Husain, and N. N. Jaya, "Preservasi Naskah Kuno Sasak Lombok Berbasis Digital dan Website," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, pp. 445–454, 2018.

[3] Bermansyah and Y. Antoni, "Digitalisasi Naskah Kuno Dalam Upaya Pelestarian dan Menarik Minat Generasi Muda," *Ganec Swara*, vol. 10, no. 1, pp. 120–127, 2016.

[4] Atisah, "Teks, Konteks, dan Fungsi Pepaosan dalam Tradisi Lisan Lombok," *SAWERIGADING*, vol. 24, no. 1, pp. 73–84, 2018.

[5] P. K. Austin, "Reading the Lontars: Endangered literature practices of Lombok, eastern Indonesia," *Lang. Doc. Descr.*, vol. 8, pp. 27–48, 2010.

[6] P. K. Austin, "Tense, aspect, mood, and evidentiality in Sasak, eastern Indonesia," *Lang. Doc. Descr.*, vol. 11, pp. 231–251, 2012.

[7] F. Seifart, "Orthography development," in *Essentials of Language Documentation*, Berlin: Walter de Gruyter, 2006, pp. 275–300.

[8] S. Hidayat, H. R. P. Negara, and D. T. Kumoro, "Determination of the Optimum Wavelet Basis Function for Indonesian Vowel Voice Recognition," *J. Elektron. dan Telekomun.*, vol. 17, no. 2, pp. 42–47, Dec. 2017.

[9] J. Galka and M. Ziolko, "Mean Best Basis Algorithm for Wavelet Speech Parameterization," in *Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2009, pp. 1110–1113.

[10] R. R. Coifman and M. V. Wickerhauser, "Entropy-Based Algorithms for Best Basis Selection," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 713–718, 1992.

[11] S. Hidayat, Abdurahim, and M. Tajuddin, "Evaluation and design of wavelet packet cepstral coefficient (WPCC) for a noisy Indonesian vowels signal," *J. Phys. Conf. Ser. Pap.*, vol. 1211, no. 012023, 2019.

[12] A. Dİker, Z. Cömert, E. Avci, and S. Velappan, "Intelligent System based on Genetic Algorithm and Support Vector Machine for Detection of Myocardial Infarction from ECG signals," in *26th Signal Processing and Communications Applications Conference (SIU)*, 2018, pp. 0–3.

[13] A. S. R. A, N. Dheetsith, S. S. Nair, and D. Ghosh, "Auto Analysis of ECG Signals Using Artificial Neural Network," in *2014 International Conference on Science Engineering and Management Research (ICSEMR)*, 2014, pp. 1–4.

[14] D. Vlaj and Z. Kacic, "The Influence of Lombard Effect on Speech Recognition," in *Speech Technologies*, no. 977126, 2014, pp. 1998–2001.

[15] M. Sawalha and M. A. M. Abushariah, "The Effects of Speakers ' Gender, Age, and Region on Overall Performance of Arabic Automatic Speech Recognition Systems Using the Phonetically Rich and Balanced Modern Standard Arabic Speech Corpus," in *Proceedings of the 2nd Workshop of Arabic Corpus Linguistics WACL-2*, 2013.

[16] J. S. Mason and J. Thompson, "GENDER EFFECTS IN SPEAKER RECOGNITION," in *Proc. ICSP-93*, 1993, pp. 733–736.