Classification of Speech Signal based on Feature Fusion in Time and Frequency Domain

1st Domy Kristomo
Magister Teknologi Informasi
Universitas Teknologi Digital Indonesia
Yogyakarta, Indonesia
domy@utdi.ac.id

2nd Fx Henry Nugroho
Rekayasa Perangkat Lunak Aplikasi
Universitas Teknologi Digital Indonesia
Yogyakarta, Indonesia
fx_henry@utdi.ac.id

***Abstract*—The design of a speech recognition system requires a reliable feature extraction process. It has an essential function since a good feature can help to improve the classification rate. Nowadays, the classification of stop consonant is a challenging task, due to the several factors that influence the accuracy of classification. Research that focuses on words formed by stop consonant syllables has not been widely studied by previous local researchers. Feature fusion is one way that can be done in improving the performance of the pattern recognition and classification system. In this paper, we propose three feature sets of the feature fusion by using Discrete Wavelet Transform (DWT) at 7th level decomposition with Daubechies2, Wavelet Packet Transform (WPT) at 4th level decomposition with Daubechies2, Autoregressive Power Spectral Density (AR-PSD), and Statistical method to classify stop consonant word speech signal. According to the experimental results, the classification accuracy for WPT + Statistical, DWT + Statistical, and AR-PSD + Statistical were 94.72%, 92.22%, and 76.38% respectively.**

Keywords—Feature extraction, speech recognition, discrete wavelet transform, wavelet packet transform, auto-regressive power spectral density.

I. INTRODUCTION

Speech recognition technology is currently developing rapidly. The technology known as speech recognition is a technology that converts speech signals into a series of words through an algorithm. In a speech recognition system, the feature generation or feature extraction process plays a very crucial role. Some of the well-known feature extraction methods in speech recognition are Mel Frequency Cepstral (MFCC), Wavelet, Linear Prediction Cepstrum Coefficients (LPCC), and Linear Predictive Coding (LPC).

Stop consonants are an important component of the speech signal. The complex movements in the vocal tract produce the stop sounds. The closing nasal cavity causes a rapid closure or opening, sometimes including the oral cavity. When it is closed, the pressure made is released together with the release of vocal tract closure. In Indonesia, there are six stop consonants (/p/, /t/, /k/, /b/, /d/ and /g/) [1][2]. The study of Indonesian stop consonant features is important to discover their time and frequency domain characteristics. There are some differences in Indonesian and English stop articulation. In Indonesian, /k/, /p/, and /t/ is articulated without aspiration or a release of air. On the contrary, in the "standard" American English those consonants are articulated with aspiration. There is also a place of articulation differences for stops between the two languages. The stops /d/ and /t/ are dental sounds made with a constriction between the tip of the tongue and the upper teeth in the Indonesian language. However, the English pronunciation of /d/ and /t/ is alveolar meaning the blade of the tongue is on the alveolar ridge [1]. Feature extraction of the stops of the syllables has an essential role in the process of classification since a good feature can help to improve the classification rate. The previous research focused on the phonetic characteristics of the Indonesian stop consonant has been conducted by the previous researcher. However, the study that examines the Indonesian stop consonant word features, as well as its classification, has not received much attention from the researchers.

Research on speech recognition in Indonesian or other languages has been carried out by previous researchers using various feature extraction methods, such as Wavelet [3][4][5], MFCC [6][7], Spectrogram [8][9][10][11], and so on. In research [3], the Discrete Wavelet Transform (DWT) method at decomposition levels 2, 4, and 6 with mother Wavelet Daubechies 4, and Symlet 4 is used as a feature extraction method to recognize vowel sounds in Indonesian with Artificial Neural Network (ANN) as a classifier. Based on the experimental results obtained an accuracy of 70.83%. In [4], a feature extraction method based on Wavelet Packet (WP) was used for the recognition of emotional speech in real-time. The performance of the proposed method (WP) is compared with the conventional MFCC method. The experimental results show that the characteristic performance of the proposed method is better than conventional MFCC. In research [5], the improvement of the feature extraction method is proposed by the name of Wavelet Cepstral Coefficients (WCC) where the feature dimensions of WCC are smaller than MFCC. The experimental results show that WCC at DWT decomposition level 5 improves performance by 1.19% for independent speakers and 3.21% for dependent speakers when compared to MFCC. In [6], the MFCC method was combined with the PCA-based feature dimension reduction method for Indonesian speech recognition. The experimental results show that the accuracy without PCA is 86.43%, while the accuracy using PCA increases to 89.29%. In [12], the MFCC method with SVD and PCA characteristic dimensions was used to recognize speech signals. The experimental results showed an increase from 83.57% to 90.71% with the proposed method. In research [7], the MFCC feature generation technique with the K-nearest neighbor (KNN) classification system was used to classify Indonesian syllables. Experimental results with three scenarios show an increase in accuracy of 4% and an improvement in computational time of 0.151 s compared to conventional MFCC. In [8], spectrograms were used to analyze digit speech signals in Indonesian. In this study, it is seen that the spectrogram is similar for certain types of speech digits, as well as the most different form of the

spectrogram (high dissimilarity) to others, namely for the "*empat*" digit. In [9], the Wavelet method was used in the formation of a spectrogram to create a speech recognition system that is robust to noise. In [10], the spectrogram was used to observe the characteristics of three utterances in the Malaysian language, namely: "*satu*", "*dua*" and "*sembilan*". In [11], time-frequency feature extraction and fractal dimensions were used to analyze the intonation of speech signals in Japanese. The characteristics of the signal are observed in the spectrogram. In [13], an analysis of the performance of the Wavelet-based feature extraction method was carried out by comparing several of its mother Wavelets. The experimental results show that Daubechies db44 and db45 are suitable mother wavelets for the introduction of Indonesian vowel speech in which the two mother wavelets are not affected by the gender differences of the speakers. However, there has been no research that combines the characteristics of several methods such as WPT, DWT, AR-PSD, and Time-Frequency domain statistical. for speech signals in Indonesian. The Auto-regressive Power Spectral Density (AR-PSD) technique which is based on Fourier Transform is commonly used for biomedical signals. The reason that underlies the combination of these features is as written in pattern recognition literature that feature extraction is sometimes a problem that is dependent on the object to be studied (dependent task) so that combinations based on the researcher's imagination can allow more discriminant features [14].

In this study, the frequency and time domain feature extraction methods will be used including DWT, WPT, AR-PSD, and time-frequency domain statistical for extraction and classification of speech signals in Indonesian. The type of data used will be more focused on word sound cue data which is formed by stop consonant syllables. This research use frequency domain and time domain extraction method to extract speech signal features. The time domain feature generation technique itself has not been widely used for speech classification, especially in Indonesian. In this study, the performance of the feature extraction method will be compared with the results of its classification accuracy to see which method has the best performance. The combination of characteristics of these methods will also be applied in this study to get even better performance. The specific objective of this research is to obtain accurate features for classifying speech signals in the Indonesian language. Research on speech recognition in Indonesian that uses the characteristics of the frequency and time domain itself has not been widely carried out so that it is hoped that this research can become a new reference in speech recognition research, especially speech recognition in Indonesian. The explanation of the specific specifications related to the research scheme is as follows. This research consists of three main stages, namely pre-processing, feature generation, and classification. At the pre-processing stage, segmentation and normalization of speech signal data were carried out. At the feature generation stage, the speech sound signal feature generation process is carried out to obtain features in the frequency and time domains using the Wavelet, AR-PSD, and Statistical methods. In the final stage, the classification process is carried out to test the performance of the feature extraction method proposed from this study by observing the results of the classification accuracy.

II. METHODOLOGY

A. Speech Data

The word data used in this research were taken and recorded from six male speakers. They were asked to utter an Indonesian stop consonant word. The unvoiced stop consonants include /k, p, t/, the voiced stop consonants include /g, d, b/. The consonants /g, k/ portray velar articulation place, /t, d/ portray the dental articulation place, while /p, b/ portray the labial articulation [15], [1]. Each speaker was asked to utter a word 5 times, so the total data is 6 speakers x 10 utterances x 6 words = 360 utterances as shown in Table I.

TABLE I. THE DATA USED IN THIS RESEARCH

Data		
Indonesian words	***Translation***	***Utterances***
Gigit	Bite	60
Duduk	Sit	60
Papan	Board	60
Bibit	Seed	60
Tutup	Closed	60
Kakak	Older sibling/cousin	60
Total		360

Fig. 1 shows the plot of the speech sample of the word '*kakak'* in the time domain.

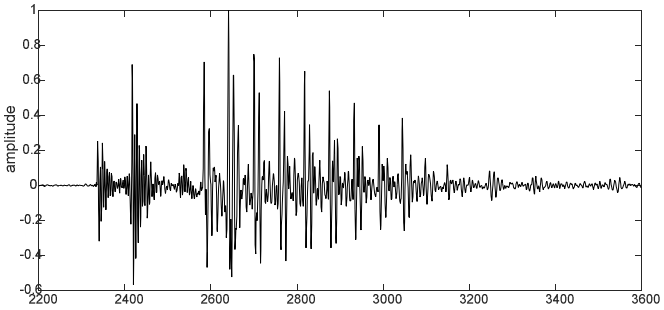

Fig. 1. The speech sound signal of the word '*kakak*'.

B. Time Domain Features

In this study we used five times domain features are as follows:

1) Mean

This is a measure of average intensity and a very simple feature, defined as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \quad (1)$$

2) Standard Deviation

This is a measure of variation or dispersion of a set of data values, defined as

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (2)$$

3) Skewness

This is a measure of the skewness of a signal, defined as

$$Sk = \frac{E(x-\mu)^3}{\sigma^3} \quad (3)$$

4) Kurtosis

This is a measure of the peak of signal intensity distribution, defined as

$$\alpha 4 = \frac{m_4}{m_2^2} \quad (4)$$

5) Entropy

This is a measure of randomness, defined as

$$H = -\sum_{i=0}^{N} Pi.\log(pi) \quad (5)$$

C. Frequency Domain Features

1) Autoregressive Power Spectral Density (AR-PSD)

Spectral features are quite common in speech recognition and audio recognition/classification. In this study, we used AR-PSD based on the Yule-Walker AR algorithm. From the spectral envelope curve of AR-PSD, we derived features of Δf (interval frequency of f1 and f2 with threshold 0,1), f0 (central frequency of AR-PSD), magnitude, skewness, mean, kurtosis, std. deviation, and entropy. The AR model in P order and the estimation of AR-PSD can be written in Eq. 6 and Eq. 7 below:

$$x_{pp}(t) = -\sum_{k=1}^{p} a_k x_{pp}(t-k) + e(t) \quad (6)$$

$$P_{AR}(f) = \frac{T\sigma_w^2}{|1+\sum_k^p a_k e^{-2\pi f k T}|^2} = T\sum_{m=1}^{C-1} r_{xx} e^{-2\pi f m k T} \quad (7)$$

D. Wavelet Domain Features

1) Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT) is a technique that can be utilized in analyzing the temporal and spectral properties of non-stationary signals like audio, based on the time-frequency multi-resolution property of wavelet transform. The Daubechies wavelets are a set of orthogonal wavelets that define a DWT and are distinguished by a maximum number of vanishing moments for a given support. The Daubechies wavelet of class D-2N can be formulated as:

$$\psi(x) := \sqrt{2}\sum_{k=0}^{2N-1} (-1)^k h_{2N-1-k}\varphi(2x-k) \quad (8)$$

A signal in the frequency domain will be obtained after the transformation process by using DWT. The moving average feature was calculated from each of the twenty samples after squaring the DWT transformed signal until the maximum sample of the signal magnitude.

Fig. 2 shows the tree structure of DWT used in this research. The number of sub-bands in the DWT can be calculated with the formula n+1, with n is the level decomposition. This study used DWT with 7th level decomposition, so the number of sub-bands is 7+1 = 8 sub-bands.

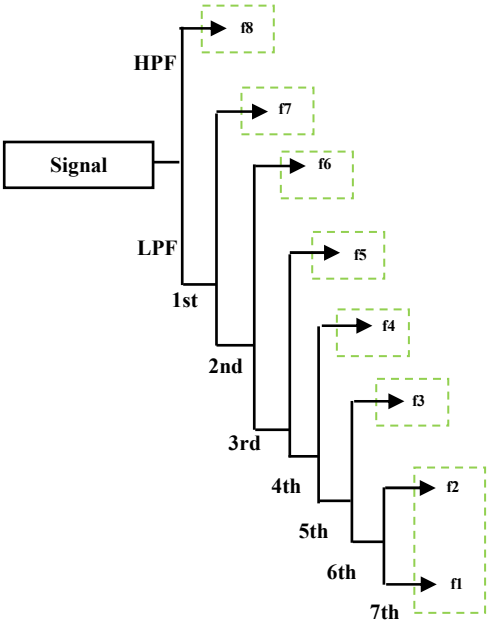

Fig. 2. The tree structure of DWT with 7th level decomposition.

2) Wavelet Packet Transform

The wavelet transform (WPT) is a wavelet transform where the discrete-time (sampled) signal is passed through a low-pass filter and high-pass filter. WPT provides more and way better frequency resolution features of voice signals by breaking down lower (approximate) and higher (detailed) frequency bands resulting in an adjusted double tree structure.

Fig. 3 shows the tree structure of WPT used in this research. The number of sub-bands in the WPT can be calculated with the formula 2^n, with n is the level decomposition. This study used WPT with 4th level decomposition, so the number of sub-bands is $2^4 = 16$ sub-bands.

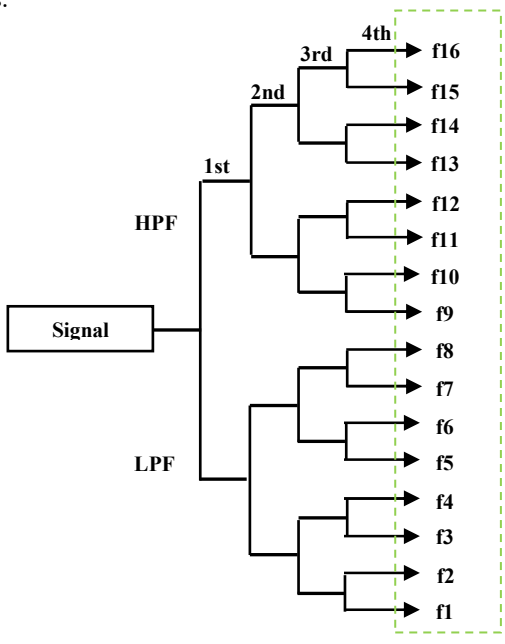

Fig. 3. The tree structure of WPT with 4th level decomposition.

E. Multi-layer Perceptron (MLP)

The last step of this research is the classification process. We used Multi-layer Perceptron (MLP) or Artificial Neural Network (ANN) in the classification process. MLP or ANN is an information processing system that has certain performance characteristics such as biological neural networks. MLP can be trained to perform pattern classification. MLP possesses several more strengths than HMM and GMM. It is commonly recognized that MLP is more effective to be applied in displaying the non-linear mapping [16]. In this research, 10-fold cross-validation is used to validate the classifier result.

III. Result And Discussion

In this section, the performance of the proposed feature set of feature fusion was evaluated. According to the research method used in this research, the purpose of this research is to find the best feature set of feature fusion that can be used to classify the Indonesian stop consonant words.

A. Feature Extraction

The first step which is the feature extraction or feature generation step has a purpose to obtain the signal characteristic in the form of the differences between the one word with another word. This research uses a feature generation method based on WPT, AR-PSD, and time-frequency domain statistical. This study proposes three feature sets, namely WPT + Time Domain Statistics (Feature Set 1) for a total of 30 features, DWT + Time Domain Statistics (Feature Set 1) for a total of 29 features, and AR-PSD + Time-Frequency Domain Statistics (Feature Set 3) for a total of 13 features. There are differences with previous studies where statistical calculations were carried out on the frequency domain signal after the signal was transformed by DWT [17]. But in this study (Feature Set 1 and Feature Set 2) statistical methods were applied to the signal in the time domain.

B. Classification

Table II shows the stop consonant words classification result by using time domain and frequency domain features. The result shows that the classification accuracy for the feature set based on WPT, DWT, and, AR-PSD were 94.72%, 92.22%, and 76.38%, respectively. The classification accuracy indicates that the fusion features of DWT 4th level decomposition with Daubechies 2 (25 features) and Time Domain Statistical (5 features) show better performance than the feature set based on DWT and AR-PSD. This is because the WPT structure decomposes the voice signal in a balanced way in both parts of the frequency band, namely the low frequency (approximate) and high frequency (detail) which allows better resolution of the feature frequency, while the DWT structure decomposes only the low frequency. The AR-PSD method has lower accuracy than DWT and WPT because the number of features is less, namely 13 features, besides that AR-PSD uses a Fourier transform-based method that transforms sound signals from the time domain to the frequency domain, the Fourier transform is not suitable if applied to the analysis of non-stationary cues such as speech sound cues, and is more suitable for analysis of stationary cues such as medical cues. Our previous research also shows that the accuracy of word classification is better than syllable classification, this is because the word segmentation signal is longer so that the resulting signal is more unique (having discriminant properties) when compared to short segmented syllables.

TABLE II. The Result of Stop Consonant Words Classification by Using DWT, WPT, and AR-PSD Features

Feature Extraction Method	*Classification Accuracy (%)*
Feature Set 1: WPT 4th level decomposition with Daubechies 2 (25 features) + Time Domain Statistical (5 features) = 30 features	94.72
Feature Set 2: DWT 7th level decomposition with Daubechies 2 (24 features) + Time Domain Statistical (5 features) = 29 features	92.22
Feature Set 3: AR-PSD (3 features) + Frequency Domain Statistical of AR-PSD (5 features) + Time Domain Statistical of AR-PSD (5 features) = 13 features	76.38

Table III shows the classification result of stop consonant words of our previous research by using feature set of WPT, WPT + Singular Value Decomposition (SVD), and WPT + Karhunen-Loeve (KL) [18]. The classification accuracy results of our current research are not as high as previous research, even with a greater number of features than previous research. The other factor is the difference in the amount of data, our current research uses 360 data while our previous research uses 300 data. Also, there are several factors such as shift variance in the segmentation process due to the manual segmentation process of the data, the possibility of different lengths of the data, and the noise factor of the data. So further research is still needed.

TABLE III. The Result of Stop Consonant Words Classification by Using WPT, WPT+SVD, and WPT+KL [18]

Feature Extraction Method	*Classification Accuracy (%)*
WPT (25 features)	95
WPT + SVD2	95.67
WPT + KL (18 features)	89

IV. Conclusion

This paper presents the implementation of feature extraction and classification of the Indonesian stop consonants word using feature fusion of DWT, WPT, AR-PSD, and Time-Frequency domain statistical method. The results find that this approach can be used for the classification of the speech sound signal. The experimental results show that the classification accuracy for Feature Set 1 (WPT + Time Statistical), Feature Set 2 (DWT + Time Statistical), and Feature Set 3 (AR-PSD + Time-Frequency Statistical) were 94.72%, 92.22%, and 76.38% respectively. The classification accuracy results of our current research are lower than the previous research, even with a greater number of features and data so that further research is needed to improve its performance.

Acknowledgment

The authors would like to thank DRPM Kemenristekdikti/BRIN and LLDIKTI Region V Yogyakarta for the support in this research with the research scheme of "*Penelitian Dosen Pemula*" under the contract number 3278.18/LL5/PG/2021.

REFERENCES

[1] F. L. Hardjono and R. A. Fox, "Stop Consonant Characteristics: VOT and Voicing in American-Born-Indonesian Children's Stop Consonants," The Ohio State University, 2011.

[2] H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A. M. Moeliono, "Tata Bahasa Baku Bahasa Indonesia (Indonesian Grammar)," *Balai Pustaka, Jakarta, Indones.*, 2003.

[3] N. Amalia, A. E. Fahrudi, and A. V. Nasrulloh, "Indonesian Vowel Recognition Using Artificial Neural Network Based On the Wavelet Features," *Int. J. Electr. Comput. Eng.*, vol. 3, no. 2, pp. 260–269, Apr. 2013, DOI: 10.11591/ijece.v3i2.2325.

[4] Y. Huang, A. Wu, G. Zhang, and Y. Li, "Extraction of adaptive wavelet packet filterbank-based acoustic feature for speech emotion recognition," *IET Signal Process.*, vol. 9, no. 4, pp. 341–348, 2015, DOI: 10.1049/iet-spr.2013.0446.

[5] T. B. Adam, M. S. Salam, and T. S. Gunawan, "Wavelet Cesptral Coefficients for Isolated Speech Recognition," *TELKOMNIKA Indones. J. Electr. Eng.*, vol. 11, no. 5, pp. 2731–2738, 2013, doi: 10.11591/telkomnika.v11i5.2510.

[6] A. Winursito, R. Hidayat, and A. Bejo, "Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition," *2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018*, vol. 2018-Janua, pp. 379–383, 2018, doi: 10.1109/ICOIACT.2018.8350748.

[7] R. Hidayat and A. Winursito, "Improving accuracy of isolated word recognition system by using syllable number characteristics," *Int. J. Technol.*, vol. 11, no. 2, pp. 411–421, 2020, doi: 10.14716/ijtech.v11i2.3678.

[8] M. Fachrie and M. Fachrie, "Robust Indonesian Digit Speech Recognition using Elman Recurrent Neural Network," no. January 2015, 2018.

[9] S. Badiezadegan and R. C. Rose, "A wavelet-based data imputation approach to spectrogram reconstruction for robust speech recognition," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, no. May 2014, pp. 4780–4783, doi: 10.1109/ICASSP.2011.5947424.

[10] M. A. A. Zulkifly and N. Yahya, "Relative spectral-perceptual linear prediction (RASTA-PLP) speech signals analysis using singular value decomposition (SVD)," *2017 IEEE 3rd Int. Symp. Robot. Manuf. Autom. ROMA 2017*, vol. 2017-Decem, pp. 1–5, 2017, doi: 10.1109/ROMA.2017.8231833.

[11] M. Phothisonothai, Y. Arita, and K. Watanabe, "Extraction of Expression from Japanese Speech based on Time-Frequency and Fractal Features," 2013, pp. 1–5.

[12] A. Winursito, R. Hidayat, A. Bejo, and M. N. Y. Utomo, "Feature Data Reduction of MFCC Using PCA and SVD in Speech Recognition System," *2018 Int. Conf. Smart Comput. Electron. Enterp. ICSCEE 2018*, 2018, doi: 10.1109/ICSCEE.2018.8538414.

[13] S. Hidayat *et al.*, "Implementation of Cross-correlation in Selecting the Best Wavelet Basis Function for Indonesian Vowel Voice Recognition System," no. April, p. 2018, 2018.

[14] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. 2008.

[15] R. P. Sharma, O. Farooq, and I. Khan, "Wavelet based sub-band parameters for classification of unaspirated Hindi stop consonants in initial position of CV syllables," *Int. J. Speech Technol.*, vol. 16, no. 3, pp. 323–332, Sep. 2013, doi: 10.1007/s10772-012-9185-x.

[16] P. Kral, "Discrete Wavelet Transform for automatic speaker recognition," in *2010 3rd International Congress on Image and Signal Processing*, Oct. 2010, pp. 3514–3518, doi: 10.1109/CISP.2010.5646691.

[17] D. Kristomo, R. Hidayat, and I. Soesanti, "Feature Extraction and Classification of the Indonesian Syllables Using Discrete Wavelet Transform and Statistical Features," 2016.

[18] D. Kristomo and Y. Kusnanto, "Dimensionality Reduction of Speech Signals using Singular Value Decomposition and Karhunen-Loeve," no. Conrist 2019, pp. 78–84, 2020, doi: 10.5220/0009432200780084.