JTIM : Jurnal Teknologi Informasi dan Multimedia
p-ISSN : 2715-2529
e-ISSN : 2684-9151
https://journal.sekawan-org.id/index.php/jtim/

Features Extraction on Cleft Lip Speech Signal using Discrete Wavelet Transformation

Siti Agrippina Alodia Yusuf [1,*], Muhammad Imam Dinata[2]

[1] Universitas Muhammadiyah Mataram; siti.agrippina@ummat.ac.id
[2] Universitas Muhammadiyah Mataram; imam.dinata@ummat.ac.id
* Correspondence: siti.agrippina@ummat.ac.id

Sitasi: Yusuf, S. A. A.; Dinata, M. I. (**2024**). Features Extraction on Cleft Lip Speech Signal using Discrete Wavelet Transformation. JTIM: Jurnal Teknologi Informasi Dan Multimedia, 6(2), 161-167. https://doi.org/10.35746/jtim.v6i2.545

Received : 19-06-2024
Revised : 14-07-2024
Accepted : 15-07-2024

Abstract: Cleft is one of the most common birth defects worldwide, including in Indonesia. In Indonesia, there are 1,596 cleft patients, with 50.53% having a cleft lip and palate (CL/P), 24.42% having a cleft lip (CL), and 25.05% having a cleft palate (CP). Individuals with clefts encounter difficulties with resonance and articulation during communication due to dysfunctions in the oral and nasal cavities. This study investigates various types of mother wavelets as feature extractors for cleft speech signals. Five different mother wavelets, namely Symlet order 2, Reverse Biorthogonal order 1.1, Discrete Meyer, Coiflet order 1, and Biorthogonal order 1.1 are analyzed. This work aims to find the best type of mother wavelet. The extracted features are statistical features, such as mean, median, standard deviation, kurtosis, and skewness. The dataset used in this study consists of 200 sound signals from 10 individuals with cleft conditions and 10 normal volunteers. To assess the performance of the extractor, classification is performed using K-Nearest Neighbor (KNN) and K-Fold cross-validation. The experimental results indicate that the Reverse Biorthogonal order 1.1 mother wavelet achieves the highest accuracy compared to other types of mother wavelet, where the accuracy is 93%, with sensitivity and specificity of 94% and 92%, respectively.

Keywords: cleft speech signal, wavelet transform, mother wavelet, KNN, K-Fold

1. Introduction

Cleft, which can manifest as cleft lip (CL), cleft palate (CP), or both (CL/P), is among the most frequent birth defects. The global prevalence is estimated at one in every four million newborns annually [1]. In Indonesia, the prevalence of CL and CP remains high. Data indicates that 1,596 individuals are affected by this condition, with 50.53% having CL/P, 24.42% having CL, and 25.05% having CP. Gender-wise, 55.95% are male and 44.05% are female [2]. Individuals with clefts experience difficulty in forming resonance and articulation during communication due to velopharyngeal dysfunction [3]. This dysfunction occurs when the mouth and nasal cavities fail to effectively produce and retain air within the oral cavity during speech. Consequently, people with cleft exhibit different vocal bursts, formant transitions, and spectral characteristics compared to those without the condition. Additionally, every utterance produced is likely to be distorted due to unintended nasalization in vocal production [4].

Speaker recognition is a component of biometric identification, encompassing modalities such as fingerprint, face, and iris recognition. Within the realm of artificial intelligence (AI), speech recognition encompasses several complex tasks, including speech recognition itself, speech segmentation, and speaker recognition [5]. Individuals with cleft palate experience difficulties in communication, particularly with bilabial sounds like /b/, /m/, and /p/, which require both lips to meet [6]. In speech recognition systems, feature extraction is a crucial process. Features are attributes of the speech signal that are extracted and utilized to represent it. The characteristics of the speech signal are influenced by morphological features and the speaker's habits. Morphological features include the size, vocal fold structure, and length of the vocal tract, which contribute to the speaker's unique characteristics. Habitual characteristics are influenced by factors such as education, upbringing, personality, and parental influence [7]. Various methods are commonly employed for feature extraction from speech signals, including Discrete Wavelet Transformation (DWT), Mel-Frequency Cepstral Coefficient (MFCC), Gammatone Frequency Cepstral Coefficient (GFCC), Perceptual Linear Prediction (PLP), and Power Normalized Cepstral Coefficient (PNCC).

Extensive research has been conducted on speech signals, yet investigations specifically targeting cleft speech signals remain scarce, particularly for Indonesian speech. The study by [8] categorizes speech signals into cleft and normal classes, comparing various feature extraction methods such as Mel-Frequency Cepstral Coefficients (MFCC), Jitter, Shimmer, the combination of MFCC and Bionic Wave Transformation (BWT), and the energy derived from BWT. The dataset comprised speech samples from 15 individuals with cleft conditions and 15 without, each articulating predefined words commonly used in speech therapy to evaluate speech quality. The system achieved 85% accuracy, 82% sensitivity, and 85% specificity when using the combined MFCC and BWT methods. Another study compared the performance of MFCC and the Pitch Adaptive MFCC (PAMFCC) between cleft and normal speech on the vowels /a/ in the word /papa/, /i/ in the word /pipi/, and /u/ in the word /pupu/ [9]. The findings indicated that PAMFCC was more efficient in feature extraction, yielding accuracies of 83.45% for /a/, 88% for /i/, and 85% for /u/. Further analysis on the same dataset was conducted by [10], focusing on sinusoidal features of words with the consonant-vowel-consonant-vowel (CVCV) structure, specifically "papa," "pipi," and "pupu," extracted from 30 normal and 30 cleft speech samples. Sinusoidal features included normalized harmonic amplitude (NA), harmonic amplitude ratio (HAR), and prominent harmonic frequency (PHF). The accuracy differences between the studies [9] and [10] were minimal.

Another study utilized MFCC to extract features from cleft speech signals, focusing on the recognition of the phoneme /p/ in Indonesian [11]. The research involved three words: "paku," "kapak," and "atap." The dataset comprised samples from 10 individuals with normal speech and 10 individuals with cleft conditions. The study reported accuracies of 74% for "paku," 76% for "kapak," and 75% for "atap". The classification of pathological voices in [12] utilized Wavelet transform with wavelet energy as features, achieving a 93% accuracy rate, showing that wavelet energy is effective in recognizing pathological voices. In [13], energy and statistical features were used for automatic voice disorder classification, employing Haar's four-level decomposition of Stationary Wavelet Transform (SWT) to extract features, achieving 99% accuracy. For detecting hypernasality in speech signals, [14] used statistical features. The study combined various feature extraction methods and compared them. The combination of statistical and energy features resulted in 93% accuracy for utterance recognition and 94% accuracy for subject classification. Most prior research has concentrated on vowels or specific words; thus, there is a need for studies focusing on bilabial sounds. Moreover, studies on cleft signals in the Indonesian language are still limited. Consequently, this research will analyze words containing bilabial sounds using DWT, the aim of this work is to find the most suitable mother wavelet to work on cleft speech signals.

2. Materials and Methods

The study comprises four stages as illustrated in Figure 1. The first stage involves the collection of speech signal data, followed by preprocessing of the signals to enhance their quality. Subsequently, feature extraction is performed on the signals, and the final stage is classification.

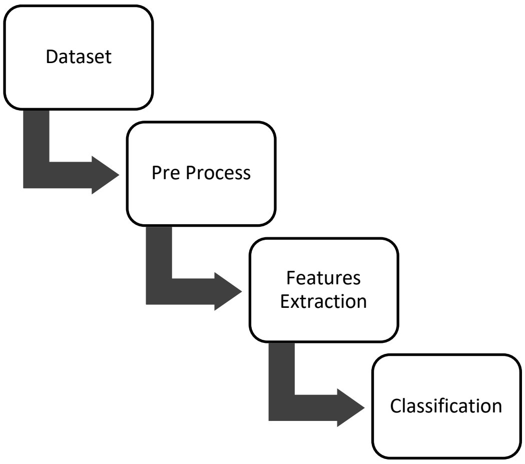

Figure 1. Proposed method

2.1. *Dataset*

The data collection was conducted by recording the voices of 20 volunteers, comprising 10 volunteers with cleft conditions (including CL, CP, and CL/P) and 10 normal volunteers. Each volunteer was asked to pronounce the word "Lampu," repeated 10 times. The selection of the word was based on the presence of the bilabial phonemes /p/ and /m/. A total of 200 voice signals were obtained and stored in *.wav format. The speech signal data were divided into two groups: 100 labeled as normal and 100 labeled as cleft.

2.2. *Pre-process*

At this stage, the recorded speech signals will undergo quality enhancement. The quality improvement is achieved by reducing noise present in the signals. The method employed is pre-emphasis with a coefficient of $\alpha = 0.97$. This technique enhances the quality of the signal at high frequencies. The mathematical formulation for this method is expressed in Equation (1):

$$y[n] = x[n] - \alpha * x[n-1] \quad (1)$$

Where y[n] represents the output signal, *x[n]* represents the input signal, α denotes the pre-emphasis coefficient, and *x[n-1]* represents the previous signal. After enhancing the signal quality, the next step is to normalize the signal to align the magnitudes, ensuring that the highest magnitude in the signal is 1. The implementation of signal normalization is represented by Equation (2), where *Snorm* is the output signal, *s[n]* is the original signal and *max(|x[n]|)* represents the absolute value of the signal:

$$Snorm[n] = \frac{s[n]}{\max(|s[n]|)} \quad (2)$$

2.3. *Features Extraction*

In this stage, the characteristics or features of the sound signal are extracted. One popular feature extraction method is Discrete Wavelet Transform (DWT). This method divides the signal into two channels: approximation channel (A) and detail channel (D). Channel A is obtained from a low-pass filter (LPF), while channel D is obtained from a high-pass filter (HPF). In the subsequent levels of decomposition, signal A will be further divided into A2 and D2, and so forth up to the desired level. The application of 2-level DWT decomposition is illustrated in the Figure 2.

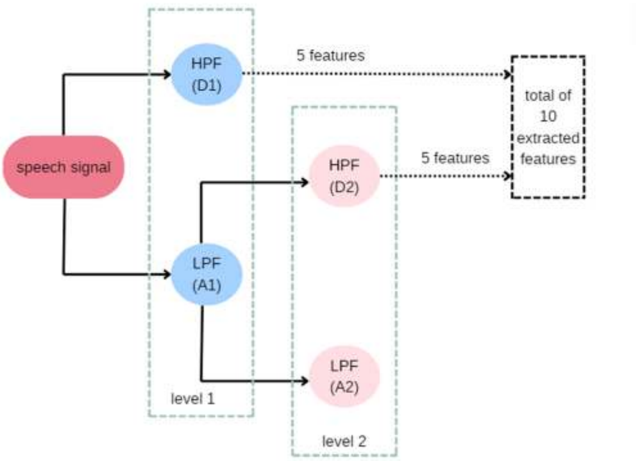

Figure 2. DWT 2-level decomposition

Equations 3 and 4 represent the mathematical forms that are used to compute the LPF and HPF.

$$A[n] = \sum_{k}^{\infty} = -\infty\ s[k]g[2n-k] \tag{3}$$

$$D[n] = \sum_{k}^{\infty} = -\infty\ s[k]h[2n-k] \tag{4}$$

In this study, the sound signal will be decomposed into 2 levels using several types of mother wavelets, such as Coiflet of order 1, Reverse Biorthogonal 1.1, Discrete Meyer, Biorthogonal, and Symlet of order 2. The selection of these mother wavelet types is based on previous research where these types have achieved high accuracy in speaker recognition [7]. After decomposing the signal using each type of mother wavelet, features will be extracted from channels D1 and D2. The extracted features include statistical characteristics such as mean, median, standard deviation, kurtosis, and skewness. A total of 10 features will be obtained for each signal. The selection of these features is based on research [14], [15] showing their effectiveness in achieving high accuracy in classification processes.

2.4. Classification

The classification stage is the step to categorize each signal based on the extracted features. In this stage, the classification method used is K-Nearest Neighbors (KNN). KNN is a method that is based on comparative learning [16], which operates in two steps: training and testing. In this study, the distance measure used to assess proximity is Euclidean distance. The mathematical form of Euclidean distance is shown in Formula 5.

$$d_{st} = \sqrt{\sum_{j-1}^{n} (x_{sj} - y_{tj})^2} \tag{5}$$

Where x and y represent the data whose distance needs to be measured, and n is the number of dimensions of each data point. To ensure accuracy variance, K-Fold Cross Validation is also implemented; this method divides the data into K datasets, where the first dataset is used as the training set and the subsequent datasets are used as the testing sets. Then, the second dataset is used as the training set, and the remaining datasets are used as the testing sets, this process continues until K datasets are processed. The number of folds used in this study is 3 and 7.

3. Result and Discussion

In this study, the proposed method is applied to a dataset of speech signals obtained from 20 volunteers, comprising 10 volunteers with cleft lip and/or palate conditions (CL, CP, and CL/P), and 10 volunteers with normal conditions. Each volunteer uttered the word 'Lampu' (lamp) 10 times, resulting in a total of 200 speech signal samples. The acquired speech signals are first enhanced in quality using pre-emphasis with a coefficient $\alpha = 0.97$. This process aims to improve signal quality at higher frequencies. Subsequently, each signal undergoes decomposition up to level 2 using DWT. During this process, statistical features of the signals are extracted. Five types of statistical features are extracted: mean, median, skewness, standard deviation, and kurtosis. These features are obtained from channels D1 and D2 of the signal, representing 10 features for each signal. Additionally, several types of mother wavelets such as Coiflet of order 1, Reverse Biorthogonal 1.1, Discrete Meyer, Biorthogonal, and Symlet of order 2 are used in the decomposition process. After feature extraction is completed, the features from each signal are fed into the K-Nearest Neighbors (KNN) classifier for classification. The classification results for each type of mother wavelet are shown in the following tables.

Table 1. Accuracy result on K-Fold 3.

Wavelet Family	Accuracy	Sensitivity	Specificity
Coif1	0,77	0,76	0,78
Rbiol1.1	0,91	0,93	0,89
Dmey	0,80	0,79	0,82
Bior	0,90	0,93	0,88
Sym2	0,75	0,75	0,77

Table 2. Accuracy result on K-Fold 7.

Wavelet Family	Accuracy	Sensitivity	Specificity
Coif1	0,77	0,77	0,76
Rbiol1.1	0,93	0,94	0,92
Dmey	0,81	0,81	0,81
Bior	0,90	0,91	0,89
Sym2	0,76	0,78	0,75

The performance accuracy of various mother wavelets on K-Fold 3 is illustrated in Table 1. The wavelet Reverse Biorthogonal 1.1 achieves the highest accuracy at 91%, followed closely by Biorthogonal at 90%. Additionally, experiments conducted with K-Fold set to 7 show an improvement in recognition accuracy, with Reverse Biorthogonal 1.1 achieving the highest at 93%, followed by Biorthogonal as shown in Table 2. The excellent performance of Reverse Biorthogonal 1.1 is due to its capability to manage both symmetric and asymmetric signal properties, which is advantageous for addressing irregularities in cleft speech signals. Furthermore, Biorthogonal and Reverse Biorthogonal are proven to be advantageous in extracting features on certain phonemes [17]. The application of the DWT method facilitates multiresolution analysis of speech signals, effectively capturing low-frequency components associated with vocal fold vibrations and high-frequency components related to formant transitions and consonant articulation. This multiresolution analysis is essential for cleft speech signals, which show significant deviations in spectral and temporal characteristics due to velopharyngeal dysfunction. Additionally, the selection of the fold number in the model significantly impacts classification results. Wavelets like Reverse Biorthogonal 1.1 exhibit notable performance enhancement with an increase in fold number, suggesting that evaluation with more folds provides a more precise representation of model performance. In contrast, wavelets such as Coiflet 1 and

Biorthogonal show consistent performance regardless of the number of folds, indicating strong stability in their performance.

The Reverse Biorthogonal 1.1 wavelet offers substantial advantages in speech signal processing by effectively eliminating noise without removing speech information, unlike conventional thresholding methods. Reverse Biorthogonal 1.1 employs multistage convolution with wavelet filters in both high and low pass frequency bands of the speech signal, ensuring noise removal in each band individually while maintaining speech integrity. This method yields superior results compared to traditional thresholding techniques like Donoho and Johnstone thresholding and the Birge-Massart thresholding strategy[18]. Nevertheless, these results underscore the effectiveness of Reverse Biorthogonal 1.1 in analyzing cleft speech signals and highlight the importance of fold selection in thoroughly evaluating classification models.

4. Conclusion

In this research, cleft speech signals were examined using the Discrete Wavelet Transform (DWT) with various mother wavelets to determine the most suitable one for these types of signals. The experiments revealed that the highest accuracy was obtained with the Reverse Biorthogonal wavelet of order 1.1, followed by the Biorthogonal wavelet. The accuracies recorded were 93% and 90%, with sensitivities and specificities of 94% and 92% for Reverse Biorthogonal 1.1, and 90% and 89% for Biorthogonal, respectively. Due to data constraints, this study does not encompass all types of cleft conditions and is limited to a few statistical features. Moreover, the study did not compare the performance of DWT with other feature extraction methods like MFCC, PLP, or PNCC. Future studies should increase the dataset size, include more types of cleft signals, explore additional features, and compare the effectiveness of each feature extraction method. This will provide a more comprehensive understanding of DWT's strengths and capabilities in processing cleft speech signals.

Acknowledgment: The gratitude is extended to Universitas Muhammadiyah Mataram for their support in this research project through internal funding grants.

References

[1] C. I. Alois and R. A. Ruotolo, "An overview of cleft lip and palate," *JAAPA Off. J. Am. Acad. Physician Assist.*, vol. 33, no. 12, pp. 17–20, 2020, doi: 10.1097/01.JAA.0000721644.06681.06.

[2] U. Elfiah, K. -, and S. Wahyudi, "Analisis Kejadian Sumbing Bibir dan Langit: Studi Deskriptif Berdasarkan Tinjauan Geografis," *J. Rekonstr. Dan Estet.*, vol. 6, no. 1, p. 34, 2021, doi: 10.20473/jre.v6i1.28230.

[3] F. R. Larangeira *et al.*, "Speech nasality and nasometry in cleft lip and palate," *Braz. J. Otorhinolaryngol.*, vol. 82, no. 3, pp. 326–333, 2016, doi: 10.1016/j.bjorl.2015.05.017.

[4] P. N. Sudro, R. K. Das, R. Sinha, and S. R. Mahadeva Prasanna, "Significance of Data Augmentation for Improving Cleft Lip and Palate Speech Recognition," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2021 - Proceedings*, 2021, pp. 484–490.

[5] M. F. Mridha, A. Q. Ohi, M. M. Monowar, Md. A. Hamid, Md. R. Islam, and Y. Watanobe, "U-Vectors: Generating Clusterable Speaker Embedding from Unlabeled Data," *Appl. Sci.*, vol. 11, no. 21, p. 10079, Oct. 2021, doi: 10.3390/app112110079.

[6] E. Trianingsih, U. Hasanah, S. Lestariana, A. Setyaningrum, and N. D. Adzkia, "Gangguan Berbahasa pada Remaja Usia Delapan Belas Tahun Akibat Bibir Sumbing : Perspektif Fonologi," vol. 3, no. 1, 2023, doi: https://doi.org/10.20884/1.iswara.2023.3.1.7206.

[7] S. Hidayat, M. Tajuddin, S. A. A. Yusuf, J. Qudsi, and N. N. Jaya, "Wavelet Detail Coefficient As a Novel Wavelet-Mfcc Features in Text-Dependent Speaker Recognition System," *IIUM Eng. J.*, vol. 23, no. 1, pp. 68–81, 2022, doi: 10.31436/IIUMEJ.V23I1.1760.

[8] M. Golabbakhsh *et al.*, "Automatic identification of hypernasality in normal and cleft lip and palate patients with acoustic analysis of speech," *J. Acoust. Soc. Am.*, vol. 141, no. 2, pp. 929–935, Feb. 2017, doi: 10.1121/1.4976056.

[9] A. K. Dubey, S. R. M. Prasanna, and S. Dandapat, "Pitch-Adaptive Front-end Feature for Hypernasality Detection," in *Interspeech 2018*, ISCA, Sep. 2018, pp. 372–376. doi: 10.21437/Interspeech.2018-1251.

[10] A. K. Dubey, S. R. M. Prasanna, and S. Dandapat, "Sinusoidal model-based hypernasality detection in cleft palate speech using CVCV sequence," *Speech Commun.*, vol. 124, pp. 1–12, Nov. 2020, doi: 10.1016/j.specom.2020.08.001.

[11] A. Anggoro, S. Herdjunanto, and R. Hidayat, “MFCC dan KNN untuk Pengenalan Suara Artikulasi P,” *Avitec*, vol. 2, no. 1, pp. 13–19, 2020, doi: 10.28989/avitec.v2i1.605.

[12] S. E. Shia and T. Jayasree, “Detection of pathological voices using discrete wavelet transform and artificial neural networks,” in *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, Srivilliputhur: IEEE, Mar. 2017, pp. 1–6. doi: 10.1109/ITCOSP.2017.8303086.

[13] A. Shrivas *et al.*, “Employing Energy and Statistical Features for Automatic Diagnosis of Voice Disorders,” *Diagnostics*, vol. 12, no. 11, p. 2758, Nov. 2022, doi: 10.3390/diagnostics12112758.

[14] A. Mirzaei and M. Vali, “Detection of hypernasality from speech signal using group delay and wavelet transform,” in *2016 6th International Conference on Computer and Knowledge Engineering (ICCKE)*, Mashhad, Iran: IEEE, Oct. 2016, pp. 189–193. doi: 10.1109/ICCKE.2016.7802138.

[15] S. A. A. Yusuf and N. Sulistianingsih, “Ekstraksi Fitur Sinyal EKG Myocardial Infarction menggunakan Discrete Wavelet Transformation,” *TEKNIMEDIA*, vol. 4, no. 1, pp. 38–44, Jun. 2023, doi: https://doi.org/10.46764/teknimedia.v4i1.96.

[16] M. J. Al Dujaili, A. Ebrahimi-Moghadam, and A. Fatlawi, “Speech emotion recognition based on SVM and KNN classifications fusion,” *Int. J. Electr. Comput. Eng. IJECE*, vol. 11, no. 2, p. 1259, Apr. 2021, doi: 10.11591/ijece.v11i2.pp1259-1264.

[17] E. Z. Engin and Ö. Arslan, “Selection of Optimum Mother Wavelet Function for Turkish Phonemes,” *Int. J. Appl. Math. Electron. Comput.*, vol. 7, no. 3, pp. 56–64, Sep. 2019, doi: 10.18100/ijamec.556850.

[18] K. Daqrouq, I. N. Abu-Isbeih, O. Daoud, and E. Khalaf, “An investigation of speech enhancement using wavelet filtering method,” *Int. J. Speech Technol.*, vol. 13, no. 2, pp. 101–115, Jun. 2010, doi: 10.1007/s10772-010-9073-1.