Experiments in Whispered Speech Recognition Based on Wavelet Transformation

1st Branko R. Marković
Department of Computer and Software Engineering
University of Kragujevac, Faculty of Technical Sciences Čačak,
Čačak, Serbia
brankomarko@yahoo.com & https://orcid.org/0000-0003-3924-307X

2nd Đorđe Damnjanović
Department of Computer and Software Engineering
University of Kragujevac, Faculty of Technical Sciences Čačak,
Čačak, Serbia
djordje.damnjanovic@ftn.kg.ac.rs & https://orcid.org/0000-0002-1796-7707

***Abstract*—In this paper the results of normal and whispered speech recognition based on DWT (Discrete Wavelet Transformation) are presented. The feature vectors are obtained based on Daubechies sub-band energy. The experiments are performed using a part of the Whi-Spe database (one female and one male speaker). A back-end of the ASR system is based on DTW (Dynamic Time Warping) algorithm. The following scenarios are analyzed: normal/normal, whisper/whisper, normal/whisper and whisper/normal in the speaker dependent mode. The results confirmed the usefulness of Wavelet transformation in speech recognition.**

***Keywords*—DWT, Daubechies, Whispered speech, Speaker dependent mode, DTW, Speech recognition**

I. Introduction

The recognition of the whispered speech is an actual topic nowadays. In the last two decades, more and more researches are focused on the different speech modes [1]. The whisper is used in different occasions and it has some specific characteristics compared to normal speech [2], [3]. This kind of speech is interesting for researchers because many tools already developed for the recognition of normal speech can be applied to whisper too.

In order to experiment with the whisper, different techniques and algorithms are used. The main goal is to create the best front-end which will convert the inputs (audio files) into sets of digital feature vectors. These vectors are input in a back-end system which is responsible for training and testing, and for a final decision on what is an entry. There are different preprocessing techniques to create the feature vectors. Some of them are based on cepstral coefficients [4] and they are widely used (MFCC, LFCC, PLPCC, GFCC, etc.). They are based on different filters which are distributed properly to capture some important features of the signal (mostly the energy). The Fast Fourier Transformation (FFT) is a standard way of getting the frequency characteristics of the input signal. But, besides the FFT, the Wavelet transformation also can be used. It has some advantages compared to the FFT (covering both frequencies and time scales). This transformation is used by many researches for normal speech [5]. In addition to all mentioned above, DWT is used in this paper as main algorithm for feature extraction. It is of great importance to see and compare results when the features are obtained with some other mention methods and those obtained with wavelets. From the aspects of wavelet families, in this particular case, Daubechies wavelet family is used.

The back-end systems are based on standard methods and often use the following: DTW [6-7], HMM (Hidden Markov Models) [8], NN (Neural Networks) [9], SVM (Support Vector Machine), etc. For this research DTW method is used.

As a dataset, the part of the Whi-Spe database is used [10]. Two speakers from this database: one female (denoted as Speaker1) and one male (denoted as Speaker6) are taken into consideration. Also, two subsets of words are used (colors and numbers) and two modes (normal and whisper). The analysis was performed through four scenarios: Normal/Normal, Whisper/Whisper, Normal/Whisper, and Whisper/Normal.

This paper has the following structure: The second part explains the Discrete Wavelet Transform with the family of Daubechies type. In the third part, the Wavelet extraction and testing are described. The fourth part provides results based on the DTW algorithm. Finally, in the Conclusion, the results are summarized and further research is proposed.

II. Discrete Wavelet Transform and Daubechies Wavelet Family

The Discrete Wavelet Transform (DWT) is a mathematical technique used for analyzing and processing signals, particularly in the field of signal and image processing. It allows for the decomposition of signals into different frequency components, which provides a more detailed and localized representation of the signal compared to traditional Fourier analysis [11]. The DWT is often preferred in many applications due to its efficiency in decomposing signals into approximation and detail coefficients [12]. This discrete method is used to segment signals into various detail levels, with each level containing information about different frequency components of the signal [13]. This is particularly useful for signal compression, detecting changes in the signal, time series analysis, and other similar applications. Additionally, the DWT is more computationally efficient compared to continuous wavelet transform, making it a popular choice in practice [14].

The main focus in DWT is that the "*mother*" wavelet is obtained in discrete steps with the following equation [5], [13]:

$$\psi_{i,j}(t) = \frac{1}{\sqrt{a_0^i}}\psi\left(\frac{t-ja_0^i b_0}{a_0^i}\right) = a_0^{-i/2}\psi\left(a_0^{-i}t - jb_0\right) \tag{1}$$

where a_0 and b_0 are discrete scaling and translation parameters, respectively, $(a_0 > 1)$.

DWT is based on the concept of Multiresolution Analysis, where a signal is successively decomposed into different frequency bands or scales. The signal is passed through a series of high-pass and low-pass filters, resulting in approximation (low-frequency) and detail (high-frequency) coefficients. DWT operates in multiple levels, with each level providing more detailed frequency information. At each level, the signal is subsampled to half its original length, allowing for efficient representation. The original signal can be reconstructed from the approximation and detail coefficients using the inverse DWT, Signal decomposition and reconstruction, using DWT, is presented in Fig. 1 [13-16].

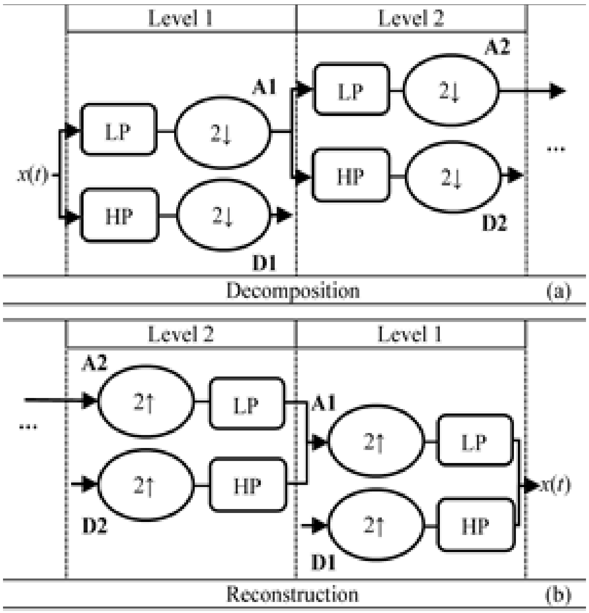

Fig. 1. Block diagram of wavelet transform (a) decomposition and (b) reconstruction.

For the implementation of DWT in applications, different wavelet families can be used. Some of them that are used often are Haar, Daubechies, Symlet, Coiflet, Morlet, biorthogonal, etc [14]. Results presented in this paper are obtained with Daubechies 3 wavelet ('db3' in Matlab software), where number 3 stands for the number of coefficients in the wavelet function.

The Daubechies wavelet family (named after the Belgian mathematician Ingrid Daubechies) is a set of wavelets that are commonly used in the DWT. These wavelets have specific mathematical properties that make them well-suited for signal processing tasks, such as de-noising, compression, feature extraction, and image reconstruction. For example, the Daubechies wavelets have compact support, meaning that they are non-zero over a finite range, making them efficient in terms of computation and storage. Additionally, Daubechies wavelets have good approximation properties, meaning they can accurately represent signals with fewer coefficients than other wavelet families [12]. Fig. 2 presents Daubechies 3 wavelet function.

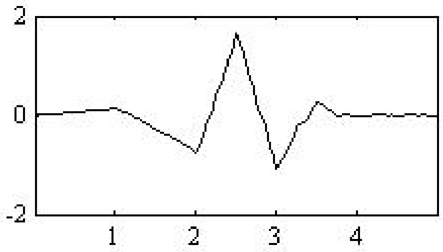

Fig. 2. Daubechies 3 wavelet function.

III. Wavelet feature extraction and testing

The preprocessing algorithm how to obtain wavelet features is given in the block diagram presented in the Fig. 3.

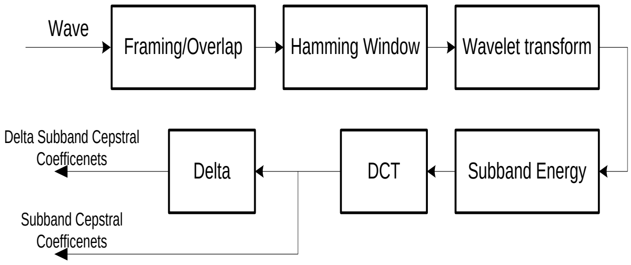

Fig. 3. Block diagram for wavelet feature extraction.

The digital speech signal, which was recorded at sampling rate of 22050 Hz, 16 bits per sample, is coming to a block for framing and overlap. Two sizes of frames are examined: a frame of 256 and a frame of 512 samples. The overlap between frames is 50%.

In the next block, the Hamming window is applied. It puts the signal to zero at the beginning and end of each frame. Then, the Wavelet transformation is performed. The spectrum from 0 Hz to 11025 Hz is divided into 24 subbands. Then, the Wavelet packet transformation is computed for the wavelet tree and subband coefficients are obtained.

The energy for each subband is computed as [5]:

$$S_i = \sum_{mei} [(W_\varphi\, x)(i), m]^2 / N_i \tag{2}$$

where $W_\varphi x$ is the Wavelet packet transformation of x; i – subband frequency index (i= 1,2,3,…L); In this case L=24; N_i - number of coefficients in i^{th} subband.

The next step is to apply DCT (Discrete Cosine Transformation). The outputs are SBC coefficients, and they are obtained by using the following formula:

$$SBC(k) = \sum_{i=1}^{L} \log S_i * \cos(\frac{k(i-0,5)}{L}\pi) \tag{3}$$

where k=1,2,…N (N is the number of SBC coefficients, in this case N=12).

In the end, the Delta coefficients are created. These are the first derivates of SBCs (three neighboring frames are used in these experiments):

$$dSBC_m(i) \approx \mu \sum_{k=-K}^{K} k * SBC_m(i+k) \qquad (4)$$

where μ is the normalization constant, and k is a number of neighboring frames.

Finally, two types of cepstral vectors are obtained:

1) vector which has 12 SBC coefficients, and

2) vector which has 12 SBC + 12 Delta SBC coefficients.

Also, with each type of vector, two types of frames are used: frames with 256 samples (equivalent to 11,6ms) and frames with 512 samples (equivalent to 23,3ms).

In order to make testing, two speakers from the Whi-Spe [10] are chosen: one female speaker denoted as “Speaker1”, and one male speaker denoted as “Speaker6”. Also, two subsets of patterns for these speakers in both modes (normal and whisper) are used: the first subset contains patterns of six colors (in IPA notation: /bela/, /ʒuta/, /tsrna/, /tsrvena/, /plava/, /zelena/) and the second contains patterns of fourteen numbers (/nula/, /jedan/, /dva/, /tri/, /tʃetiri/, /pet/, /ʃest/, /sedam/, /osam/, /devet/, /deset/, /sto/, /hiʎadu/, /million/).

For the back-end of this recognition system the DTW algorithm is used [6]. It’s based on dynamic programming and tries to find an optimal match between two patterns. These patterns are represented with vectors of either SBC or SBC plus Delta coefficients.

The test itself is conducted in the following way: one set of colors for Speaker1 in appropriate mode (normal or whisper) is compared with nine remaining sets of colors (in mode normal or whisper). If the first set is in normal mode and the remaining nine sets are in normal mode – the results are Normal/Normal scenario (further denoted as “Nr/Nr”). In a similar way, other three scenarios are produced: Whisper/Whisper (“Wh/Wh”), Normal/Whisper (“Nr/Wh”) and Whisper/Normal (“Wh/Nr”). It was mandatory to always use the first set of patterns as a reference and to do this for all scenarios (in order to make consistency in comparison). Besides the colors, the same scenarios are produced for numbers. In addition, this is repeated for Speaker6 for different sizes of frames and different types of vectors.

IV. RESULTS

All input audio files are converted into a set of previously explained vectors. For these experiments, 800 audio files are preprocessed. The front-end was developed using the MATLAB toolbox. Then, the DTW back-end system is applied. This back-end is developed using Visual Basic. As a result, an average WRR (Word Recognition Rate) is produced along with a matrix of confusion.

When the first type of vectors is used (vector which has 12 SBC coefficients) the results for all scenarios (Nr/Nr, Wh/Wh, Nr/Wh and Wh/Nr) and two sizes of frames (256 or 512 samples) are presented in Tables I-IV.

TABLE I. WORD RECOGNITION RATE FOR SPEAKER1 - COLORS

Frame/Scenario	256 [%]	512 [%]
Nr/Nr	94,44	88,89
Wh/Wh	59,26	55,56
Nr/Wh	53,70	50,00
Wh/Nr	46,30	48,15

TABLE II. WORD RECOGNITION RATE FOR SPEAKER6 - COLORS

Frame/Scenario	256 [%]	512 [%]
Nr/Nr	75,93	75,93
Wh/Wh	75,93	72,22
Nr/Wh	37,04	40,74
Wh/Nr	33,33	31,48

TABLE III. WORD RECOGNITION RATE FOR SPEAKER1 - NUMBERS

Frame/Scenario	256 [%]	512 [%]
Nr/Nr	85,71	84,92
Wh/Wh	80,16	63,49
Nr/Wh	32,54	21,43
Wh/Nr	28,57	34,92

TABLE IV. WORD RECOGNITION RATE FOR SPEAKER6 – NUMBERS

Frame/Scenario	256 [%]	512 [%]
Nr/Nr	77,78	60,32
Wh/Wh	73,81	58,73
Nr/Wh	30,16	29,37
Wh/Nr	25,40	22,22

As can be seen from the tables, for Speaker1 results are better than for Speaker6 for almost all scenarios including both colors and numbers. Also, the WRR for colors is better than for numbers which is expectable because there were six colors vs. fourteen numbers. The match scenarios (Nr/Nr and Wh/Wh) have given better results than mismatch scenarios (Nr/Wh and Wh/Nr).

From the aspects of the frame’s size, the results are in most cases better when the frame is shorter (256 samples).

In Figures 4 and 5 the average Word Recognition Rates for both speakers, all scenarios, and for a frame size of 256 samples are presented.

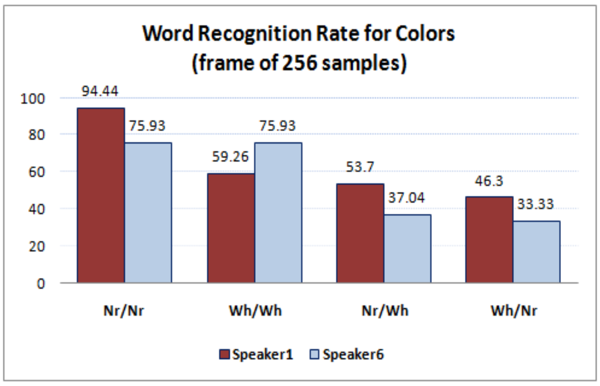

Fig. 4. Word Recognition Rate for Colors (vector of 12 SBC coefficients)

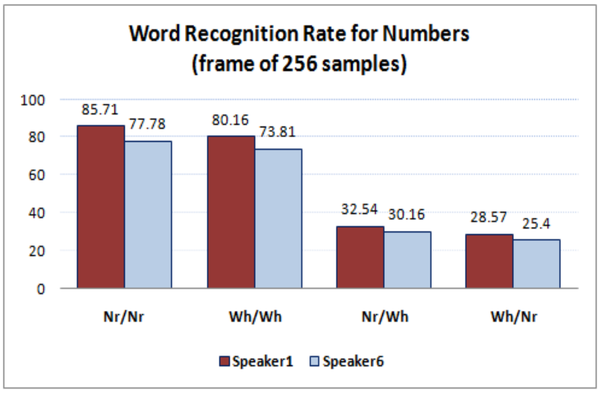

Fig. 5. Word Recognition Rate for Numbers (vector of 12 SBC coefficients)

Similarly, when the second type of vectors is used (vector which has 12 SBC plus 12 Delta SBC coefficients) the results for all scenarios (Nr/Nr, Wh/Wh, Nr/Wh and Wh/Nr) and two sizes of frames (256 samples and 512 samples) are presented in Tables V-VIII.

TABLE V. WORD RECOGNITION RATE FOR SPEAKER1 - COLORS

Frame/Scenario	256 [%]	512 [%]
Nr/Nr	94,44	96,30
Wh/Wh	70,37	68,52
Nr/Wh	55,56	62,96
Wh/Nr	57,41	50,00

TABLE VI. WORD RECOGNITION RATE FOR SPEAKER6 - COLORS

Frame/Scenario	256 [%]	512 [%]
Nr/Nr	77,78	83,33
Wh/Wh	77,78	77,78
Nr/Wh	38,89	40,74
Wh/Nr	40,74	37,04

TABLE VII. WORD RECOGNITION RATE FOR SPEAKER1 - NUMBERS

Frame/Scenario	256 [%]	512 [%]
Nr/Nr	92,06	85,71
Wh/Wh	89,68	73,02
Nr/Wh	34,92	26,19
Wh/Nr	32,54	37,30

TABLE VIII. WORD RECOGNITION RATE FOR SPEAKER6 - NUMBERS

Frame/Scenario	256 [%]	512 [%]
Nr/Nr	85,71	72,22
Wh/Wh	80,95	66,67
Nr/Wh	36,51	38,10
Wh/Nr	26,19	23,02

The results are much better now than with the first type of vectors (12 SBC coefficients). In most cases, the results are better for the vector’s frame, which has 256 samples. This is especially obvious for numbers.

Figures 6 and 7 depict the WRRs for both speakers, all scenarios, and the size of 256 samples per frame.

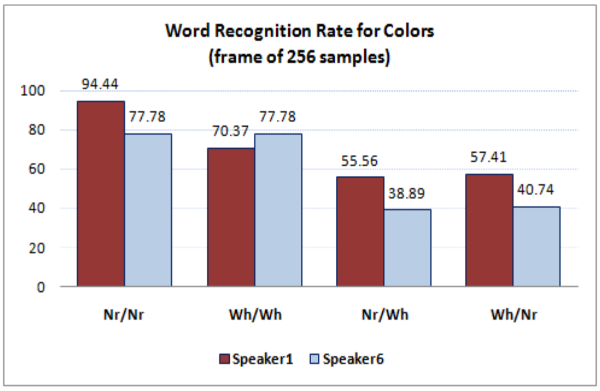

Fig. 6. Word Recognition Rate for Colors (vector of SBC + Delta coefficients)

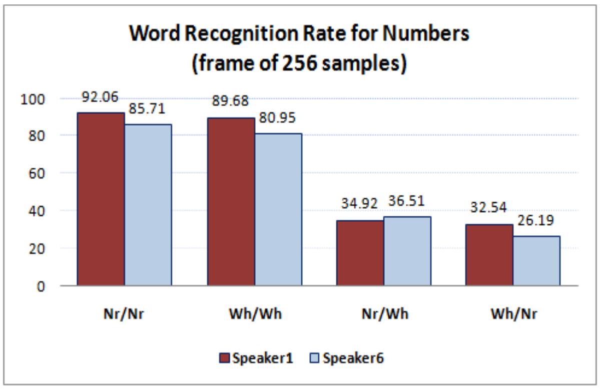

Fig. 7. Word Recognition Rate for Numbers (vector of SBC + Delta coefficients)

V. CONCLUSION

As a general conclusion, the following can be emphasized:

- The best results are obtained with match scenarios: Normal/Normal and Whisper/Whisper. For colors, they were 96,30% and 77,78% respectively and for numbers, they were 92,06% and 89,68% respectively.
- For mismatch scenarios the results are modest and the best was 57,41% (Whisper/Normal) for colors and 38,10% (Normal/Whisper) for numbers.
- For frames of 256 samples, results are better.
- Adding Delta coefficients, the recognition is improved in all cases and it was from a few percent to 13% (for example Wh/Wh scenario for colors).

Further analysis may be focused on improving mismatch scenarios by adding some normalization techniques and enhancing the robustness of vectors by adding Delta-Delta coefficients. Also, different wavelet families could provide rather different results.

ACKNOWLEDGMENT

This study was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, and these results are parts of the Grant No. 451-03-66/2024-03/200132 with University of Kragujevac – Faculty of Technical Sciences Čačak.

References

[1] C. Zhang, J.H.L. Hansen, “Analysis and classification of Speech Mode: Whisper through Shouted,” Interspeech 2007, 2007, pp. 2289-2292

[2] S.T. Jovičić, “Formant feature differences between whispered and voiced sustained vowels,” ACUSTICA - Acta Acustica, 84(4), 1998, pp. 739-743.

[3] S.T. Jovičić, Z.M. Šarić, „Acoustic analysis of consonants in whispered speech,” Journal of Voice, 22(3), 2008, pp. 263-274.

[4] J T. Ito, K. Takeda, F. Itakura, “Analysis and Recognition of Whispered speech,” Speech Communication, 2005, pp. 129-152.

[5] R. Sarikaya, B. L. Pellom, and J. H. L. Hansen, “Wavelet packet transform features with application to speaker identification.,” in IEEE Nordic signal processing symposium, Denmark, 1998, pp. 81–84.

[6] B. Marković, J. Galić, Đ. Grozdić, S. T. Jovičić, “Application of DTW method for whispered speech recognition”, Speech and Language 2013, 4th International Conference on Fundamental and Applied Aspects of Speech and Language, Belgrade, October 25-26, 2013.

[7] Branko R. Marković and Jovan Galić, „Whispered Speech Recognition Based on DTW algorithm and μFCC feature”, IcETRAN 2021, Proceedings of 8th International Conference of Electrical, Electronic and Computer Engineering, AKI 1.1, September 8-10, 2021, Ethno Village Stanišići, Republic of Srpska, pp. 37-40, ISBN 978-86-7466-894-8.

[8] J. Galić, S.T. Jovičić, Đ. Grozdić and B. Marković, HTK-Based Recognition of Whispered Speech, A. Ronzhin et al. (Eds.): *SPECOM 2014*, LNAI 8773, Springer International Publishing Switzerland 2014, 251 (2014).

[9] Đ.T. Grozdić, B. Marković, J. Galić, S.T. Jovičić (2013). „Application of Neural Networks in Whispered Speech Recognition“, TELFOR Journal, Vol. 5, No. 2, 2013, pp. 103-106.

[10] B. Marković, S.T. Jovičić, J. Galić, Đ. Grozdić,: Whispered Speech Database: Design, Processing and Application, 16th International Conference, TSD 2013, I. Habernal and V. Matousek (Eds.): TSD 2013, LNAI 8082, Springer-Verlag Berlin Heidelberg, pp. 591-598

[11] A.N. Akansu, W.A. Serdijn and I.W. Selesnick, “Emerging applications of wavelets: A review”, *Physical communication*, vol. 2, no. 1, pp.1-18, 2010.

[12] S. Mallat, “A Wavelet Tour of Signal Processing”, Third Edition. Academic Press, 2008.

[13] Đ. Damnjanović, D. Ćirić, Z. Perić, “Wavelet-Based Audio Features of DC Motor Sound”, FACTA UNIVERSITATIS, Series: Electronics and Energetics, 34(1), 2021, pp. 71–88, doi: doi.org/10.2298/FUEE2101071D.

[14] M. van Berkel, “Wavelets for Feature Detection; Theoretical background”, Eindhoven University of Technology, Department of Mechanical Engineering, Eindhoven, Literature study, Mar. 2010.

[15] Đ. Damnjanović, D. Ćirić, “Usage of Wavelet De-noising for Estimation of Room Impulse Response Truncation Time”, in Proceedings of the 5th International Conference 158 on Electrical, Electronic and Computing Engineering “IcETRAN 2018”, Palić, Serbia, Jun. 2018, pp. 565–570

[16] H. Zaynidinov, U. Juraev, S. Tishlikov, J. Modullayev, “Application of Daubechies Wavelets in Digital Processing of Biomedical Signals and Images”, In: Intelligent Human Computer Interaction, Lecture Notes in Computer Science, vol 14531. Springer, Cham., 2023, https://doi.org/10.1007/978-3-031-53827-8_19