10th International Scientific Conference
Technics, Informatics, and Education – TIE 2024
20-22 September 2024

Session: *Computer Sciences and Information Technology*
Professional paper
DOI: 10.46793/TIE24.074M

The Influence of Daubechies Wavelet Order on Speech Recognition

Branko R. Marković[1]* [0000-0003-3924-307X] and Milan Vesković[1] [0000-0002-7668-4387]
[1] Faculty of Technical Sciences Čačak/Department of Computer and Software Engineering, Čačak, Serbia
* branko.markovic@ftn.kg.ac.rs

Abstract: *This paper will present the results of speech recognition based on different Daubechies wavelet orders. Two speakers (one female and one male) were analyzed in two speech modes: normal and whisper. The patterns are used from the Whi-Spe database. As an input to the recognition system, the Daubechies wavelet feature vectors with different orders were used. As a back-end of the system, the standard Dynamic Time Warping method was considered. The results are given in the form of tables and histograms. They suggest which order of Daubechies is the most convenient for this kind of speech recognition.*

Keywords: *Speech recognition; Discrete Wavelet Transformation (DWT); Daubechies; Whi-Spe database; Dynamic Time Warping (DTW)*

1. INTRODUCTION

Whisper is one of the speech modes that is nowadays very often in use. The usage of mobile telephones is so popular. But, because the speech over the mobile phone can disturb other people around, the speaker usually turns to whisper.

In the last two to three decades this speech mode has in the focus of many researchers. They analyzed different features related to whisper: vocal cord vibration [1], shape of glottis and larynx [2], formant frequency migrations [3, 4] entry levels, signal-to-noise ratio etc.

To make a more realistic mathematical model of whispered speech the researchers used different tools and acoustical features. Hence, the most popular are related to Linear, Mel and Bark scales and their modifications. The whispered speech is transformed into a set of digital vectors and then these vectors are inputs to the system for training and testing.

On the back-end of the Automatic Speech Recognition (ASR) system, there are different tools for recognition. The most popular are DTW (Dynamic Time Warping) [5], HMM (Hidden Markov Models) [6] and ANN/DNN (Artificial Neural Networks/Deep Neural Networks) [7].

In this paper for the acoustical features, the vectors of Discrete Wavelet transformation [8] are used with a specific family: Daubechies wavelet. They are chosen according to the Mel filter bank. For the back-end of this recognition system, DTW is used.

This article is structured in the following way: Section 2 explains Discrete Wavelet Transformation with a focus on the Daubechies wavelet family. Seven different orders of this family were analyzed. Section 3 provides a figure and an explanation of how to extract DWT feature vectors and how to conduct the recognition. Each vector has 12 cepstral coefficients and they are based on 24 frequency subbands. Section 4 gives the results of these experiments. Finally, in Section 5 the conclusion is given with recommendations on what can be done in the future.

2. DAUBECHIES WAVELET FEATURE

The Wavelet Transform (WT) is a mathematical technique used to overcome some shortcomings of the Fourier transform [8]. The Fourier transform gives the signal in the frequency domain but does not provide information about where these frequency components are present in time. With WT the processing signal is cut-off at certain points (in time) and transfered to the frequency domain. Mathematically it can be written as:

$$\gamma(s,\tau)=\int f(t)\psi^{*}_{s,\tau}(t)dt \qquad (1)$$

where *f(t)* is the processing signal and $\psi(t)$ is a ”mother” function of the wavelet. “Mother” wavelet is represented as:

$$\psi_{s,\tau}(t)=\frac{1}{\sqrt{s}}\psi(\frac{t-\tau}{s}) \qquad (2)$$

where s is a scaling factor and τ is a shift parameter.

The Discrete Wavelet Transform (DWT) is a special case of the Wavelet Transform [9]. It is practical for computer applications. It uses a "mother" wavelet in the form:

$$\psi_{j,k}(t)=\frac{1}{\sqrt{s^{j}}}\psi(\frac{t-k*s^{j}}{s^{j}}) \qquad (3)$$

where j and k are integers. In practices, the sampling pattern is dyadic, so the "mother" wavelet is shifted by $k*2^{j}$ and scaled by 2^{j} . Hence, the value of s is 2.

The DWT is efficient in decomposing signals and provides the approximation and detail coefficients. Usage of the DWT is particularly powerful for signal compression, detecting changes in the signal, time series analysis, speech processing etc [8].

The DWT is uses the concept of multi-resolution analysis. An input signal is successively decomposed into many frequency bands or scales. The signal is going through multiple series of high-pass and low-pass filters. As a result, the approximation (low-frequency) and detail (high-frequency) coefficients are obtained and they are used further in signal processing.

Different DTW wavelet families can be used for implementation. The most popular are: Daubechies, Coiflets, Symlets, Biorthogonal, Haar, Morlet, Mexican Hat, etc. used to simulate features such as frequency localization, linear phase characteristic, and orientation of speech signal [10].

The Daubechies wavelet family (introduced by Ingrid Daubechies) is a set of wavelets often used in DWT implementations. It has specific mathematical features that are suitable for signal-processing tasks.

The results presented in this paper are obtained with Daubechies wavelets with 7 different orders used (from 'db2' to 'db8'). In the notation 'db*x*', the number *x* stands for the number of coefficients in the wavelet function.

Figure 1 presents Daubechies wavelet functions with different orders (from 2 to 8) used for this research.

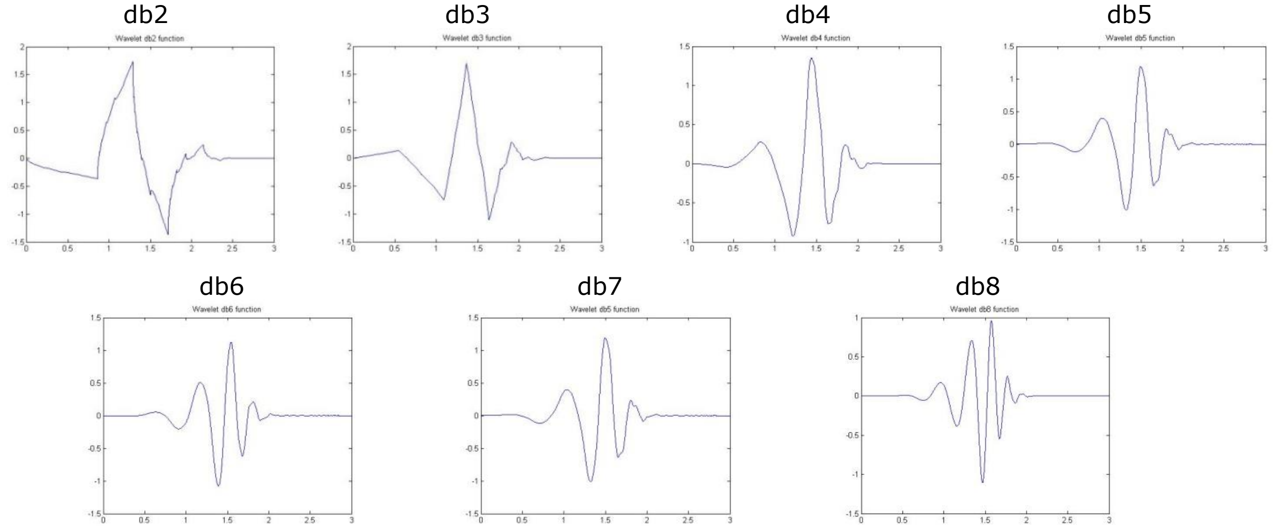

Figure 1. *Daubechies wavelets with different orders ('db2', 'db3', 'db4', 'db5', 'db6', 'db7' and 'db8' respectively)*

3. FEATURE EXTRACTION AND SPEECH RECOGNITION

For these experiments, a part of the Whi-Spe database [11] which contains numbers was used. The database itself contains 10.000 digital speech patterns which represent normal and whisper speech. All patterns are recorded with a frequency 22050 Hz, 16 bits per sample.

Figure 2 depicts a block diagram for feature extraction.

Firstly, the digital speech signal is coming to a block for "Framing/Overlapping". The size of the frame was 192 samples, and the overlap was 50%. The next block is the "Hamming window" and it weights the signal and puts the signal value to zero at the beginning and end of each frame. After that, the "Wavelet transformation" block is applied. For this transformation, the spectrum from 0 Hz to 11025 Hz is divided into 24 subbands following the Mel scale.

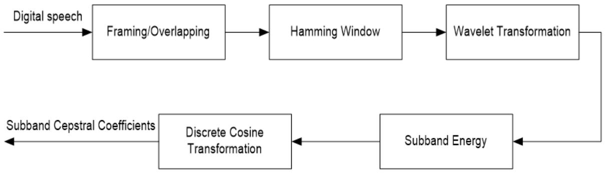

Figure 2. *Block* diagram for feature extraction

Based on a wavelet tree with six decompositions, an energy for each subband is calculated as [12]:

$$S_i = \sum_{mei}[(W_\varphi\, x)(i), m]^2/N_i \qquad (4)$$

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 *ith* subband.

The last step is the application of DCT (Discrete Cosine Transformation). The outputs are Subband Cepstral Coefficients (SBCC). They are obtained by using the following formula:

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

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

Hence, the feature vector which is used for speech recognition contains 12 SBCC elements, and all input signals are transferred into a set of these vectors.

For these experiments from the Whi-Spe database [11], two speakers are chosen: one female speaker (denoted as "Speaker1"), and one male speaker (denoted as "Speaker6"). The patterns of these speakers were in both modes (normal and whisper) and contain pronunciation of fourteen numbers (IPA notation): /nula/, /jedan/, /dva/, /tri/, /tʃetiri/, /pet/, /ʃest/, /sedam/, /osam/, /devet/, /deset/, /sto/, /hiʎadu/, /million/. Every word is repeated ten times in both modes (normal and whisper).

At the back-end of the recognition system, the DTW (Dynamic Time Warping) algorithm is applied [5]. It's an old and reliable system and it's based on dynamic programming.

The recognition is conducted in the following way (for example): one set of 14 patterns of "Speaker1" in the appropriate mode (normal or whisper) is compared with nine remaining sets of 14 patterns. If the first set is in the normal mode and the remaining nine sets in the normal mode – the results are called: Normal/Normal scenario ("N/N"). On a similar way, other three scenarios are produced: Whisper/Whisper ("W/W"), Normal/Whisper ("N/W") and Whisper/Normal ("W/N"). To make the experiment consistent, it was mandatory to use the first set of 14 patterns as a reference.

4. RESULTS

The results for these four scenarios are provided in the form of tables and histograms. The results are The Word Recognition Rate (WRR) is presented in percent (%).

Tables 1. to 7. show the results of these experiments.

Table 1. *WRR (%) for Daubechies 'db2' order*

	Speaker1	Speaker6	Average
N/N	83.30	73.81	78.56
W/W	71.43	65.08	68.26
N/W	27.78	32.54	30.16
W/N	24.60	19.05	21.83

Table 2. *WRR (%) for Daubechies 'db3' order*

	Speaker1	Speaker6	Average
N/N	90.48	80.16	85.32
W/W	80.95	76.19	78.57
N/W	30.16	35.71	32.94
W/N	22.22	19.84	21.03

Table 3 *WRR (%) for Daubechies 'db4' order*

	Speaker1	Speaker6	Average
N/N	90.48	82.54	86.51
W/W	79.37	71.43	75.40
N/W	28.57	34.92	31.75
W/N	25.40	20.63	23.02

Table 4. *WRR (%) for Daubechies 'db5' order*

	Speaker1	Speaker6	Average
N/N	90.48	84.92	87.70
W/W	80.95	75.40	78.18
N/W	30.16	31.75	30.96
W/N	25.40	22.22	23.81

Table 5. *WRR (%) for Daubechies 'db6' order*

	Speaker1	Speaker6	Average
N/N	93.65	85.71	89.68
W/W	83.33	76.19	79.76
N/W	31.75	37.30	34.53
W/N	23.02	23.81	23.42

Table 6. *WRR (%) for Daubechies 'db7' order*

	Speaker1	Speaker6	Average
N/N	91.27	90.48	90.88
W/W	84.13	75.40	79.76
N/W	30.16	35.71	32.94
W/N	26.19	26.19	26.19

Table 7. *WRR (%) for Daubechies 'db8' order*

	Speaker1	Speaker6	Average
N/N	92.86	85.71	89.29
W/W	82.54	76.98	79.76
N/W	30.95	35.71	33.33
W/N	25.40	23.81	24.60

From the tables above it is obvious that the female speaker produced better results than the male speaker for the match cases (Normal/Normal and Whisper/Whisper).

Figures 3.-6. describe WRRs in percent for all scenarios and all analyzed Daubechies orders.

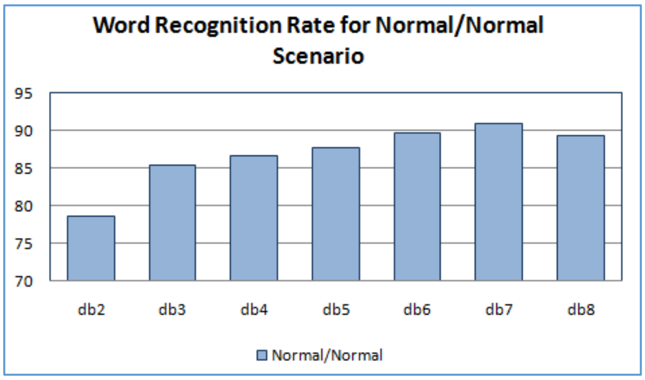

Figure 3. *Results for Normal/Normal scenario*

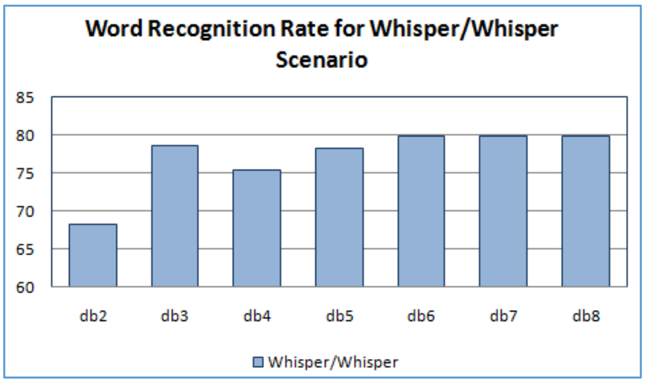

Figure 4. *Results for Whisper/Whisper scenario*

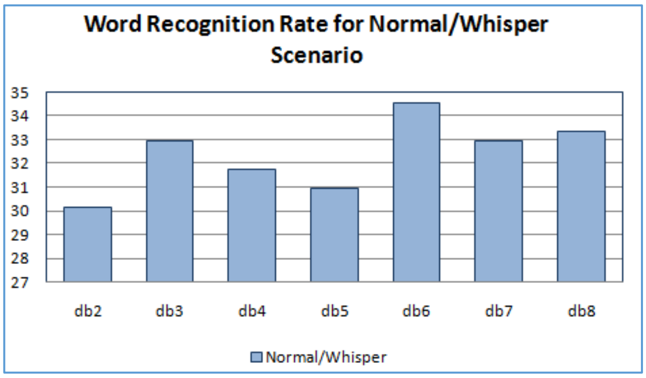

Figure 5. *Results for Normal/Whisper scenario*

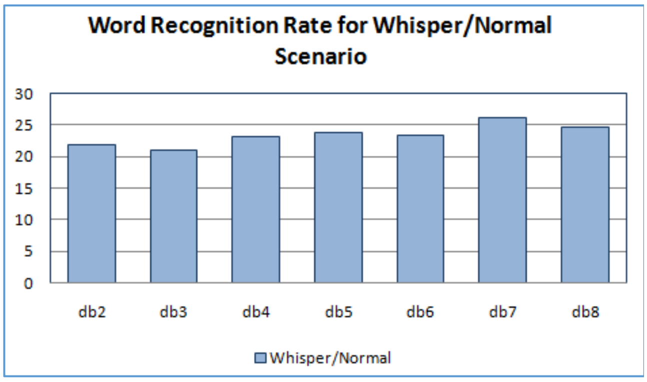

Figure 6. *Results for Whisper/Normal scenario*

Based on the histograms above it's obvious that 'db7' gives the best results for N/N (90.88%), W/W (79.76%) and W/N (26.19%) scenarios. After that 'db6' is a candidate for the second place (gives the best result for N/W (34.53%) scenario), but also 'db8' is very close.

5. CONCLUSION

The results of this research can be summarized in a few sentences:

- Usage of different Daubechies orders can impact the recognition of speech.
- The higher order consumes more computer's time.
- The experiments with orders from 2 to 8 suggest the best is to use 'db7'.

Further research may be focused on including all speakers and all patterns for the Whi-Spe database and make this result more reliable. Also, including vectors with delta coefficients, delta-delta coefficients and applying the cepstral normalization techniques (i.e. CMS) [13] the results for mismatch scenarios (N/W and W/N) should be improved. The research can incorporate speaker-independent scenarios and different methods on the back-end of ASR (i.e. HMM, DNN etc.) and also may take in consideration noisy environments.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Catford, J.C. (1977). *Fundamental problems in phonetics,* Edinburgh: Edinbourgh University Press.

[2] Ito, J.T., Takeda, K., Itakura, F. (2005). *Analysis and Recognition of Whispered speech,* Speech Communication, pp. 129-152.

[3] Jovičić, S.T. (1988). *Formant feature differences between whispered and voiced sustained vowels,* ACUSTICA - Acta Acustica, 84(4), pp. 739-743.

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

[5] Marković, B., Galić, J., Grozdić, Đ. . Jovičić, S. T. (2013). *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

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

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

[8] Mallat, S. (2008). *A Wavelet Tour of Signal Processing*, Third Edition. Academic Press.

[9] Rioul, O., Vetterli, M. (1991). *Wavelets and signal processing*, IEEE Signal Processing Magazine, vol. 8, no. 4, pp. 14–38, Oct. 1991.

[10] Van Berkel, M. (2010). *Wavelets for Feature Detection; Theoretical background,* Eindhoven University of Technology, Department of Mechanical Engineering, Eindhoven, Literature study, Mar. 2010.

[11] Marković, B., Jovičić, S.T., Galić, J., Grozdić, Đ. (2013). *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.

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

[13] Grozdić, Đ., Jovičić, S., Šumarac-Pavlović, D., Galić J., Marković, B. (2017). *Comparison of Cepstral Normalization Techniques in Whispered Speech Recognition,* Advances in Electrical and Computer Engineering, Vol. 17. Number 1, 2017, pp 21-26.