SERBIAN JOURNAL OF ELECTRICAL ENGINEERING
Vol. 22, No. 3, October 2025, 391-410
UDC: 621.391:517.977.1 **DOI: https://doi.org/10.2298/SJEE250911005M**
Original scientific paper

Exploring Discrete Wavelet Transforms for Bimodal Speech Recognition

Branko Marković[1], Veljko Lončarević[1], Jovan Galić[2]

Abstract: Discrete Wavelet Transforms (DWTs) provide time–frequency representations that are well suited for nonstationary signals such as speech. This study presents a comparison of four wavelet families (Daubechies, Symlets, Coiflets, and Biorthogonal) for bimodal automatic speech recognition across two speech modes (normal and whispered). Experiments use the Whi-Spe database comprising ten speakers (five female and five male). A Dynamic Time Warping (DTW) back-end performs sequence alignment and recognition. Results are reported via summary tables, histograms, and confusion matrices and reveal systematic differences among the wavelet families, identifying the most effective transform for bimodal recognition. These findings provide practical guidance for selecting wavelet-based front ends in whisper-robust automatic speech recognition (ASR) systems.

Keywords: Bimodal speech recognition, Discrete wavelet transformations, Daubechies, Symlets, Coiflets, Biorthogonal, Dynamic time warping.

1 Introduction

Modern automatic speech recognition systems are expected to handle multiple speech modes. In addition to normally phonated speech, these include whispered, soft, loud, and shouted speech. Incorporating such diverse modes into ASR remains a significant challenge. Prior research has proposed classifications of speech modes based on factors such as intensity level, duration and proportion of silence, frame energy distribution, and spectral tilt [1]. Whispered speech, in

[1]Faculty of Technical Sciences Čačak, University of Kragujevac, Čačak, Serbia
branko.markovic@ftn.kg.ac.rs, https://orcid.org/0000-0003-3924-307X
veljko.loncarevic@ftn.edu.rs, https://orcid.org/0009-0007-4296-2709
[2]Faculty of Electrical Engineering, University of Banja Luka, Banja Luka, Bosnia and Herzegovina jovan.galic@etf.unibl.org, https://orcid.org/0000-0002-2487-7136

*An earlier version of this paper was presented at the 12th International Conference on Electrical, Electronic and Computing Engineering (IcETRAN), Čačak, Serbia, June 2025.

Colour versions of the one or more of the figures in this paper are available online at https://sjee.ftn.kg.ac.rs particular, exhibits several distinctive characteristics: reduced energy, altered formant structures for both vowels and consonants [2 – 3], and different articulators' configurations of the vocal folds and glottis [4]. For effective whisper recognition, it is therefore important to identify robust acoustic features that remain reliable during both training and testing. Among the various speech modes, normal and whisper are most frequently studied; their combination and alternation are commonly referred to as "bimodal speech".

The automatic speech recognition system typically consists of two main components: a "front-end", which preprocesses the audio signal, and a "back-end", which performs training and recognition. The front end can employ a variety of algorithms and techniques, with the primary objective of extracting acoustic features that accurately represent the speech signal. Frequency-domain transformations, such as the Fast Fourier Transform (FFT), are commonly applied, and features are often derived using different scales, including linear, Mel, Bark, exponential, or hybrid mappings. Beyond conventional frequency transformations, the Discrete Wavelet Transform provides an alternative approach by capturing both time and frequency-domain characteristics of the signal [5 – 7]. Previous studies have demonstrated that DWT is an effective feature extraction method not only for normally phonated speech [8], but also for whispered speech [9].

At the back end of an ASR system, standard techniques are employed for training and recognition. Commonly used approaches include Dynamic Time Warping [10], Hidden Markov Models (HMM) [11], Support Vector Machines (SVM) [12], Deep Neural Networks (DNN) [13], and Convolutional Neural Networks (CNN) [14]. In the present study, DTW was selected due to its simplicity and proven reliability.

This study focuses on identifying suitable front-end feature vectors derived from four wavelet families, Daubechies, Symlets, Coiflets, and Biorthogonal, and evaluating their performance in the recognition of both normal and whispered speech. Recognition experiments were conducted under four scenarios: normal/normal, whisper/whisper, normal/whisper, and whisper/normal. For each wavelet family, three different orders were implemented. The selection of wavelets was guided by the Mel scale, using low-pass filters aligned with the corresponding central frequencies.

The experimental data were drawn from the Whi-Spe database [15], which contains recordings from ten speakers (five female and five male). The vocabulary included 14 Serbian numbers, each spoken in two modes–normal and whispered–with ten repetitions per mode. In contrast, a previous study [16] relied on data from only two speakers (one female and one male) and provided a less detailed analysis, particularly regarding phoneme-level confusions and misinterpretations.

The main contributions of this study are as follows:

- An evaluation of the effectiveness of four wavelet families (Daubechies, Symlets, Coiflets, and Biorthogonal), each tested at three orders, for speech recognition.
- An analysis of the most suitable wavelet transformation for whispered speech recognition.
- Recommendations for future directions to extend and improve this line of research.

The paper is structured as follows: Section 2 provides an overview of Discrete Wavelet Transformations, including details on Daubechies, Symlets, Coiflets, and Biorthogonal wavelets. Section 3 describes the feature extraction process and the testing methodology. The results for all scenarios, both match and mismatch, across all transformations and three different orders are presented in Section 4, along with a comparative analysis. Finally, Section 5 offers concluding remarks and suggestions for future research.

2 Discrete Wavelet Transformations

The Discrete Wavelet Transform serves as an effective method for joint time–frequency analysis of signals. In contrast to the Fourier Transform, which reveals only the spectral characteristics, the DWT provides time-localized frequency information by breaking down the signal into a collection of wavelet functions. These wavelets are obtained through scaled and shifted versions of a fundamental function known as the mother wavelet $\psi(t)$, enabling a multi-resolution perspective of the data. Practically, the transform is carried out using two complementary filters: a low-pass filter associated with the scaling function ϕ, and a high-pass filter linked to the wavelet function ψ. This filtering process divides the signal into approximation coefficients (representing low-frequency content) and detail coefficients (capturing high-frequency variations) [17]. The decomposition can be formally described as:

$$W_{\psi}\left[j,k\right]=\frac{1}{\sqrt{2^{j}}}\int_{-\infty}^{\infty}x(t)\psi\left(\frac{t-2^{j}k}{2^{j}}\right)\mathrm{d}t\,. \tag{1}$$

In the context of the Discrete Wavelet Transform, the indices j and k correspond to the dilation (scale) and translation (shift) parameters, respectively. The scaling function ϕ is responsible for producing approximation coefficients that capture the low-frequency structure of the signal, while the wavelet function ψ extracts detail coefficients that highlight high-frequency variations. After each stage of filtering, downsampling is applied, which not only reduces the dimensionality of the data but also improves computational efficiency [18].

This hierarchical or multi-resolution representation makes the DWT particularly effective for analyzing non-stationary signals such as speech. Because speech signals often contain transient features that vary over time, the ability of the transform to provide localized information across different frequency bands is highly valuable. Depending on the application, various wavelet families, such as Daubechies, Symlets, and others, can be selected, each offering distinct advantages in terms of time-frequency localization, symmetry, and the number of vanishing moments [19]. Such adaptability is especially useful in speech recognition scenarios, including the analysis of whispered speech, where high-frequency energy is often diminished.

Among the many wavelet families, the Daubechies wavelets (*dbN*), developed by Ingrid Daubechies, are among the most widely applied. Their popularity stems from three key characteristics: orthogonality, which ensures that scaling and wavelet functions are mutually orthogonal to their shifted versions; compact support, meaning the associated filters are finite in length; and vanishing moments, which allow *dbN* wavelets to exactly represent polynomials up to degree $N-1$. These properties make Daubechies wavelets highly effective in applications such as signal processing, image compression, and numerical computation.

Formally, the construction of Daubechies wavelets relies on two central functions. The scaling function $\phi(t)$, also referred to as the "father wavelet," generates the low-frequency approximation of the signal, while the wavelet function $\psi(t)$, known as the "mother wavelet," captures the detail components. The mathematical form of the scaling function is given by:

$$\phi(t)=\sum_{k} h_{k}\phi\left(2t-k\right), \tag{2}$$

where h_k represents the coefficients of the low-pass filter, obtained from the refinement relation of the scaling function. The corresponding wavelet function $\psi(t)$ is subsequently defined as:

$$\psi(t)=\sum_{k} g_{k}\phi\left(2t-k\right), \tag{3}$$

where g_k denotes the high-pass filter coefficients, derived from the low-pass filter through the alternating sign relationship:

$$g_k=\left(-1\right)^k h_{N-k-1}\,. \tag{4}$$

The label *dbN* refers to Daubechies wavelets of order *N*, where the order specifies the number of vanishing moments. This property determines how effectively the wavelets can represent polynomials, with higher orders enabling more accurate approximation of smooth functions.

A related family, known as Symlets (*symN*), was introduced as a refinement of Daubechies wavelets. The primary motivation behind Symlets is to enhance the symmetry of the wavelet functions while preserving their orthogonality [20]. Similar to Daubechies wavelets, the parameter N designates the order of the function. Symlets maintain compact support and the same number of vanishing moments as their Daubechies counterparts, but their improved symmetry reduces phase distortions, making them advantageous for various signal processing tasks. This balance is achieved by carefully modifying the filter coefficients h_k so that asymmetry is minimized, though the functions are not perfectly symmetric.

Another widely used family is the Coiflets (*coifN*), which extend the concept of vanishing moments further. Unlike *dbN* and *symN*, Coiflets are designed so that both the scaling function and the wavelet function possess vanishing moments [21]. In terms of symmetry, they lie between Daubechies and Symlets, more symmetric than *dbN* but less than *symN*. A distinguishing characteristic of Coiflets is their longer filter length: while *dbN* and *symN* use filters of length $2N$, Coiflets require $6N-1$ coefficients. This increased length provides smoother reconstructions and makes them particularly well-suited for numerical analysis and approximation problems.

$$\int_{-\infty}^{\infty} x^m \phi(x)\mathrm{d}x = 0, \quad \text{for } m \text{ in } 0,1,\ldots,N-1\,. \tag{5}$$

Biorthogonal wavelets extend the concept of orthogonal wavelets such as Daubechies, Symlets, and Coiflets by relaxing the strict requirement of orthogonality in order to achieve perfect symmetry and linear phase. These properties are particularly important in applications like image compression [22], including standards such as JPEG 2000, as well as in general signal processing.

The key idea of biorthogonal wavelets is the use of two distinct sets of filter banks, one for decomposition and another for reconstruction. During the analysis stage, the signal is convolved with a low-pass filter and a high-pass filter, followed by downsampling, to generate approximation and detail coefficients. In contrast, the synthesis stage applies a different pair of low-pass and high-pass filters, combined with upsampling and convolution, to accurately reconstruct the original signal. The principle of biorthogonality ensures that the analysis and synthesis functions form mutually inverse bases, which guarantees perfect reconstruction even when the filters are asymmetric or differ in length. This flexibility allows for independent control of vanishing moments in the decomposition and reconstruction processes, while the use of symmetric, linear-phase filters reduces reconstruction artifacts.

Fig. 1 illustrates several wavelet families of order 4 that were applied in this study: Daubechies, Symlets, Coiflets, and Biorthogonal. The figure highlights that Daubechies wavelets are inherently asymmetric, Symlets increase symmetry without sacrificing compact support, Coiflets achieve greater symmetry and provide vanishing moments for both scaling and wavelet functions, while biorthogonal wavelets obtain the highest degree of symmetry by employing separate functions for decomposition and reconstruction.

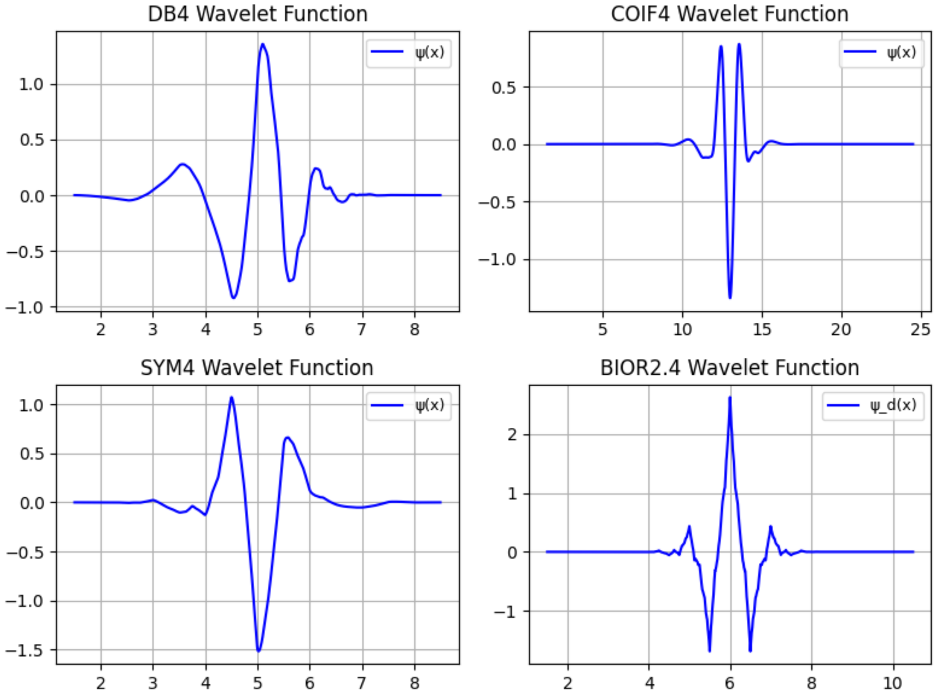


Fig. 1 – *Different wavelet families with order 4 (DB4, COIF4, SYM4 and BIOR2.4).*

3 Feature Extraction and Testing

Fig. 2 presents a block diagram illustrating the fundamental steps involved in generating acoustic features. The input to this system is the Whi-Spe database [15], which comprises 10,000 audio recordings captured in two modes: normally phonated speech and whisper. For the purposes of these experiments, a subset containing Serbian numerals spoken by all ten speakers was selected. All recordings were sampled at 22,050 Hz with a resolution of 16 bits per sample.

The processing begins with the “Framing/Overlap” stage, where the audio signal is segmented into frames of 192 samples with a 50% overlap. Each frame is then passed through a Hamming window to reduce spectral leakage. In the subsequent stage, various wavelet transformations are applied, including Daubechies, Symlets, Coiflets, and Biorthogonal wavelets, each evaluated with an appropriate order. The transformations are implemented using MATLAB’s wavedec() function with the arguments ‘db4’, ‘sym4’, ‘coif4’, and ‘bior2.6’, respectively. During this transformation, the frequency spectrum ranging from 0 Hz to 11,025 Hz is partitioned into 24 subbands based on the Mel scale (**Table 1**).

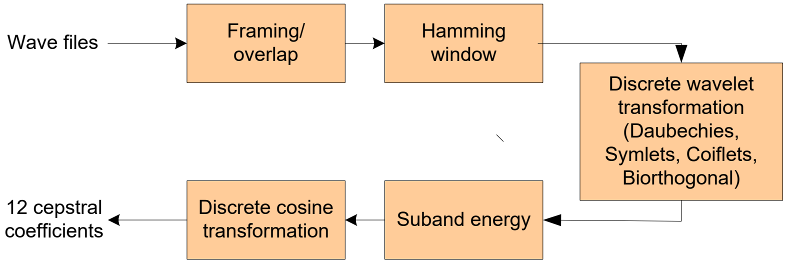

Wave files
Framing/ overlap
Hamming window
Discrete wavelet transformation (Daubechies, Symlets, Coiflets, Biorthogonal)
Suband energy
Discrete cosine transformation
12 cepstral coefficients

Fig. 2 – *Block diagram for feature extraction.*

Table 1

The frequency ranges of 24 filters [*in* Hz].

Filter 1	Filter 2	Filter 3	Filter 4	Filter 5	Filter 6
0-86.13	86.13-172.27	172.27-344.53	344.53-459.37	459.37-574.21	574.21-689.06
Filter 7	**Filter 8**	**Filter 9**	**Filter 10**	**Filter 11**	**Filter 12**
689.06-918.74	918.74 - 1148.42	1148.42-1378.13	1378.13-1550.39	1550.39-1722.66	1722.66 - 2067.20
Filter 13	**Filter 14**	**Filter 15**	**Filter 16**	**Filter 17**	**Filter 18**
2067.20-2411.72	2411.72-2765.25	2765.25-3445.31	3445.31-4134.38	4134.38-4478.91	4478.91-4823.44
Filter 19	**Filter 20**	**Filter 21**	**Filter 22**	**Filter 23**	**Filter 24**
4823.44-5512.50	5512.50-6201.56	6201.56-6890.63	6890.63-8268.25	8268.25-9646.88	9646.88-11025

In the next stage, the energy of each of the 24 subbands is computed [8], where the wavelet tree structure includes seven levels of decomposition.

$$S_i = \sum_{mei} [(W_\varphi x)(i), m]^2 / N_i \,, \tag{6}$$

$W_\varphi x$ represents the wavelet packet transformation of x,

i – frequency index for each subband *(*i = 1, 2, 3, …, 24*)* and

N_i – is a number of coefficients in the i^{th} subband.

Finally, a Discrete Cosine Transform (DCT) is applied, producing subband cepstral coefficients (SBCC). These coefficients are derived using (7):

$$\mathrm{SBCC}(k)=\sum_{i=1}^{L}\log S_i \cos\left(\frac{k\left(i-1/2\right)}{L}\pi\right), \tag{7}$$

k=1, 2,…, N (where N is a number of subband cepstral coefficients; N = 12).

As shown in the feature extraction block diagram, the outputs are vectors containing 12 subband cepstral coefficients. Each audio file is thus represented as a sequence of vectors, with each vector comprising 12 elements. These feature vectors served as input to the back-end ASR system for speech recognition.

The experimental vocabulary was derived from the Whi-Spe database and consisted of 14 Serbian numerals, transcribed using the International Phonetic Alphabet (IPA) as follows: /nula/, /jedan/, /dva/, /tri/, /tʃetiri/, /pet/, /ʃest/, /sedam/, /osam/, /devet/, /deset/, /sto/, /hiʎadu/, and /million/. For each numeral, there are ten pronunciations in both speaking modes (normal and whisper). In this study, speech samples from all ten available speakers were used: five female speakers (labeled “Speaker1” to “Speaker5”) and five male speakers (labeled “Speaker6” to “Speaker10”). Hence, for these experiments, a total of 10 (speakers) × 10 (pronunciations) × 2 (modes) × 14 (numerals) = 2,800 speech samples were used.

For the speech recognition task, the back-end of the ASR system employed the Dynamic Time Warping method [10]. DTW, based on dynamic programming, determines the optimal alignment between two patterns by identifying the path that yields the best match (Fig. 3).

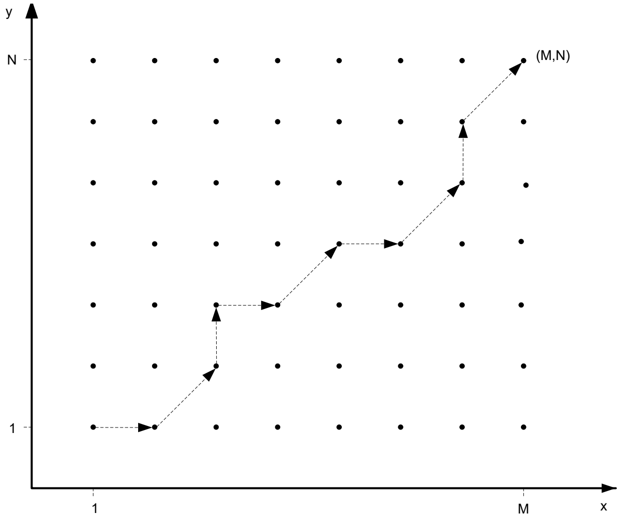


Fig. 3 – *Finding an optimal path with DTW algorithm.*

The recognition procedure was carried out as follows: one set of 14 patterns (corresponding to the vocabulary of numerals) from a given speaker was designated as the reference set. This reference was then compared against the remaining nine sets of 14 patterns. When a reference set of normal speech was compared with the other nine normal speech sets, the scenario was denoted as N/N (Normal/Normal). Analogously, three additional scenarios were defined: W/W (Whisper/Whisper), N/W (Normal/Whisper), and W/N (Whisper/Normal). For consistency across experiments, the first set of 14 patterns was always selected as the reference.

4 Results

Fig. 4 summarizes accuracy across four mother wavelets—Daubechies (Db4), Coiflets (Coif4), Symlets (Sym4), and Biorthogonal (Bior26)—under four phonation mode pairings: N/N, W/W, N/W, and W/N. Two consistent trends emerge: (i) within-mode evaluation (N/N, W/W) substantially outperforms cross-mode evaluation (N/W, W/N), and (ii) Coif4 is the most accurate wavelet family in every condition.

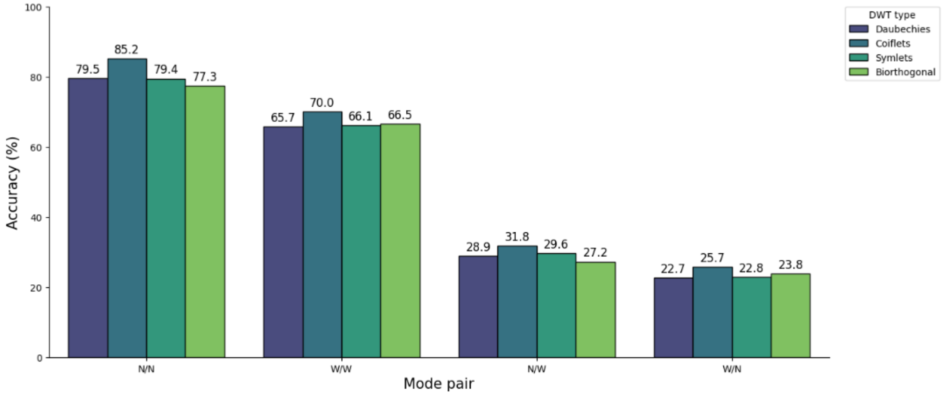


Fig. 4 – *Accuracy by mode pair and DWT type.*

Averaging over wavelets, N/N achieves 80.4% whereas W/W achieves 67.1%, i.e., whispering yields a ~13.3-point absolute reduction relative to normal speech (range 10.8–15.2 points across wavelets). This is consistent with the spectral/temporal shifts introduced by whisper phonation (reduced voicing, altered formant structure), which degrade template similarity even when modes are matched.

When the phonation mode differs between the two sides of the comparison, accuracies drop sharply. Relative to their respective within-mode baselines, cross-mode accuracies retain only ~36.5% of N/N performance and ~35.4% of

W/W performance on average. In absolute terms, the N/N→N/W drop averages ~51.0 points, and W/W→W/N averages ~43.3 points. These performance drops highlight a significant domain mismatch between normal and whisper speech in the current feature space. Cross-mode accuracy is consistently higher for N→W (mean 29.4%) than for W→N (mean 23.8%), by ~5.6 points on average (per-wavelet gaps 3.4–6.8). This asymmetry suggests that the transformation from normal to whisper alters cues more severely (or less reversibly) than the converse for the classifier's decision surface, or that the whisper representations are less compatible with normal exemplars. Across all mode pairings, coif4 is uniformly best, exceeding the next-best wavelet by ~2–8 points depending on condition (e.g., +5.7 vs db4 for N/N, +4.3 for W/W, +2.9 for N/W, +3.0 for W/N). Averaged across the four mode pairs, overall accuracies are: coif4 53.2%, sym4 49.5%, db4 49.2%, bior26 48.7%. Coiflets' superior performance is plausible given their near-symmetry and higher number of vanishing moments for both scaling and wavelet functions, which can enhance time–frequency localization and reduce phase distortion—properties that appear advantageous for these speech tokens.

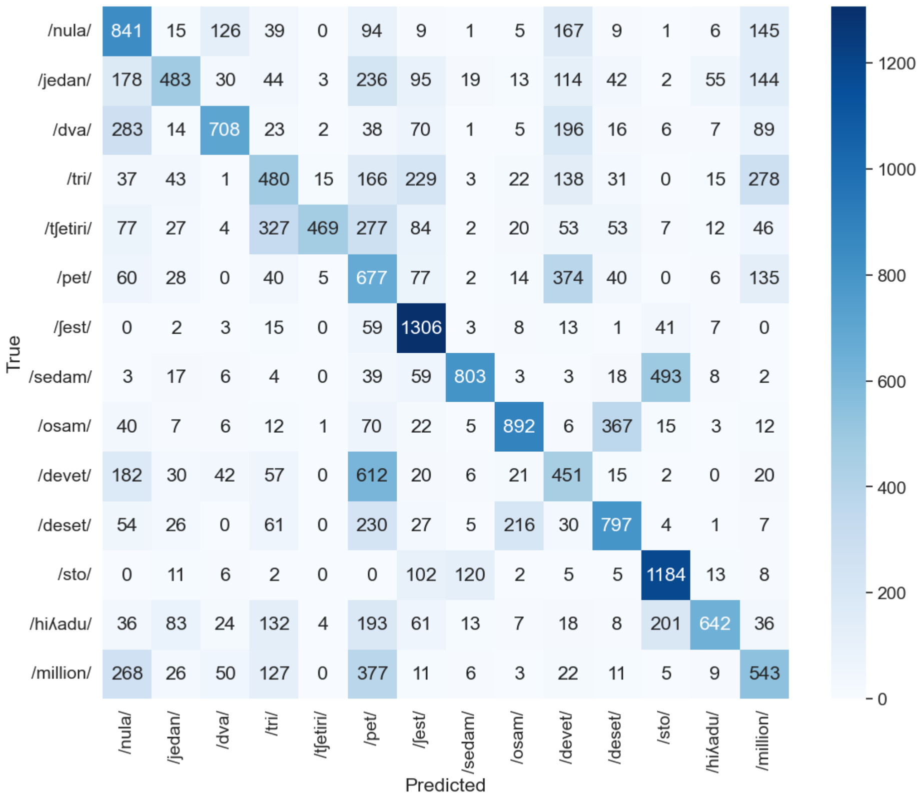
As shown in **Table 2**, Coif4 wavelet achieved the highest performance in the N/N mode, with precision at 86.8%, F1 score at 85.5% and accuracy at 85.2%, indicating stable recognition when training and testing on normal speech.

Table 2
Coif4 Classification Report.

	N/N	W/W	N/W	W/N
Accuracy	85.2%	70.0%	31.8%	25.7%
Precision	86.8%	73.0%	37.4%	34.5%
F1 Score	85.5%	70.3%	29.9%	20.2%

Performance dropped moderately in W/W (accuracy = 70.0%), though precision (73.0%) remained slightly higher, suggesting minor over-prediction of certain classes. Cross-mode conditions showed substantial degradation: N/W accuracy fell to 31.8% with a relatively higher precision of 37.4%, while W/N accuracy reached only 25.7%, despite precision being more than a third higher (34.5% vs. 25.7%). This disparity indicates that, although some predicted matches are correct, the model misses a large proportion of actual targets in cross-mode scenarios.

Fig. 5 summarizes 20,412 classification trials over 14 labels and exposes a strongly non-uniform error pattern with clear, concentrated failure modes. Overall accuracy is 10,276 / 20,412 = 50.34%, so about half of all items are correctly labeled. Importantly, the dataset is balanced on the true side: every true label appears 1,458 times (each row sums to 1,458), which means the ground-truth sampling is uniform and the observed performance differences are not explained by unequal true-class frequency. By contrast, the distribution of predicted labels is highly uneven: predicted totals range from 499 (the fewest predictions of /tʃetiri/) to 3,068 (the most predictions of pet). This asymmetry has two consequences: (1) classes that the model over-predicts (e.g. /pet/) accumulate many false positives and therefore low precision, and (2) classes that the model rarely predicts (e.g. /tʃetiri/) can have high precision but low recall.



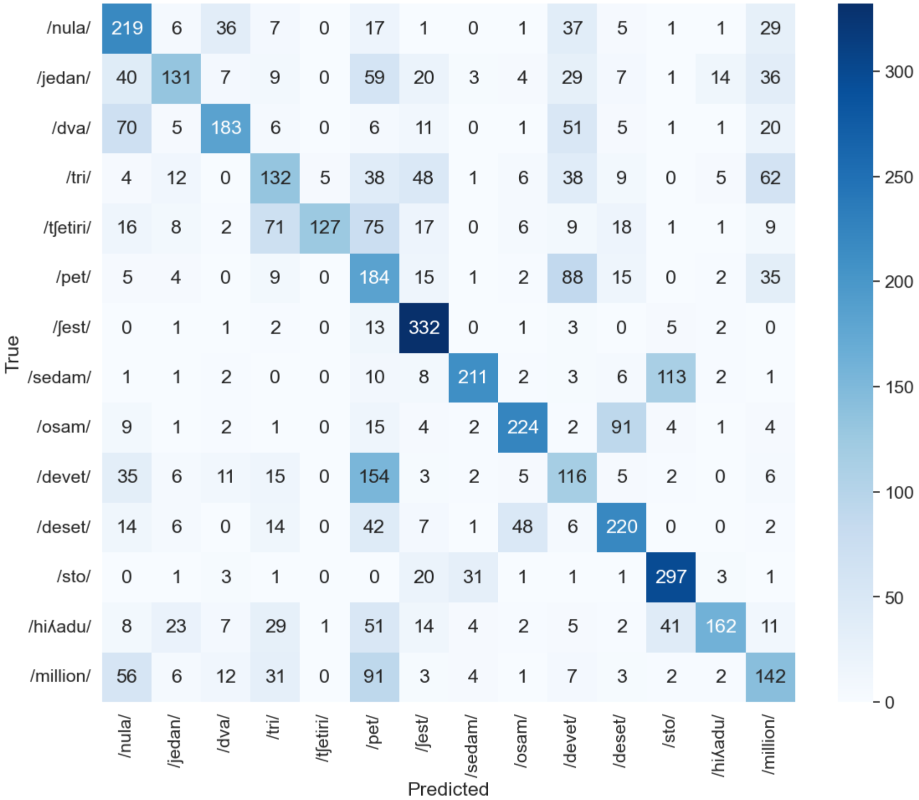
	/nula/	/jedan/	/dva/	/tri/	/tʃetiri/	/pet/	/ʃest/	/sedam/	/osam/	/devet/	/deset/	/sto/	/hiʎadu/	/million/
/nula/	841	15	126	39	0	94	9	1	5	167	9	1	6	145
/jedan/	178	483	30	44	3	236	95	19	13	114	42	2	55	144
/dva/	283	14	708	23	2	38	70	1	5	196	16	6	7	89
/tri/	37	43	1	480	15	166	229	3	22	138	31	0	15	278
/tʃetiri/	77	27	4	327	469	277	84	2	20	53	53	7	12	46
/pet/	60	28	0	40	5	677	77	2	14	374	40	0	6	135
/ʃest/	0	2	3	15	0	59	1306	3	8	13	1	41	7	0
/sedam/	3	17	6	4	0	39	59	803	3	3	18	493	8	2
/osam/	40	7	6	12	1	70	22	5	892	6	367	15	3	12
/devet/	182	30	42	57	0	612	20	6	21	451	15	2	0	20
/deset/	54	26	0	61	0	230	27	5	216	30	797	4	1	7
/sto/	0	11	6	2	0	0	102	120	2	5	5	1184	13	8
/hiʎadu/	36	83	24	132	4	193	61	13	7	18	8	201	642	36
/million/	268	26	50	127	0	377	11	6	3	22	11	5	9	543

Fig. 5 – *Confusion matrix of all classifications combined.*

Reading the matrix row-wise (recall) reveals which labels the model reliably recovers and which it misses. Highest recall is achieved for /ʃest/ (6), which is recovered in 1,306/1,458 ≈ 89.6% of its presentations, followed by /sto/ (100) at 1,184/1,458 ≈ 81.2%, and /osam/ (8) at 892/1,458 ≈ 61.2%. Several numerals in the mid-to-high range exceed 50% recall (e.g. /deset/ ≈ 54.7%, /sedam/ ≈ 55.1%). In contrast, other tokens are systematically under-recognized: /tʃetiri/ (4) is correctly labeled only 469/1,458 ≈ 32.2%, /devet/ (9) 451/1,458 ≈ 30.9%, and /tri/ (3) 480/1,458 ≈ 32.9%. Because true labels are uniform, these disparities point squarely to problems in the model/feature space (rather than to sampling bias).

Column-wise (precision) analysis complements this picture. High precision indicates that when the model predicts a label it is usually correct; low precision indicates a label is a “sink” that attracts many false positives. The cleanest predictions are /tʃetiri/ (precision ≈ 93.99%) and /hiʎadu/ (1,000) (precision ≈ 81.89%), reflecting that predictions for these labels are conservative and pure (few false positives). By contrast, /pet/ (5) is heavily over-predicted: it is predicted 3,068 times but only 677 of those are true /pet/, giving precision ≈ 22.1%. /devet/ (9) also suffers low precision (≈ 28.4%). Thus the model tends to map many ambiguous examples into a small set of attractor predictions (notably pet), while other labels are predicted infrequently but accurately.

Fig. 6 summarizes the Coif4 subset of 5,040 classification trials (14 labels × 360 presentations per label). Coif4 attains 2,680 correct labels, so overall accuracy is 2,680 / 5,040 ≈ 53.17%, a modest but clear improvement over the 50.34% observed when all four wavelet types are pooled.



	/nula/	/jedan/	/dva/	/tri/	/tʃetiri/	/pet/	/ʃest/	/sedam/	/osam/	/devet/	/deset/	/sto/	/hiʎadu/	/million/
/nula/	219	6	36	7	0	17	1	0	1	37	5	1	1	29
/jedan/	40	131	7	9	0	59	20	3	4	29	7	1	14	36
/dva/	70	5	183	6	0	6	11	0	1	51	5	1	1	20
/tri/	4	12	0	132	5	38	48	1	6	38	9	0	5	62
/tʃetiri/	16	8	2	71	127	75	17	0	6	9	18	1	1	9
/pet/	5	4	0	9	0	184	15	1	2	88	15	0	2	35
/ʃest/	0	1	1	2	0	13	332	0	1	3	0	5	2	0
/sedam/	1	1	2	0	0	10	8	211	2	3	6	113	2	1
/osam/	9	1	2	1	0	15	4	2	224	2	91	4	1	4
/devet/	35	6	11	15	0	154	3	2	5	116	5	2	0	6
/deset/	14	6	0	14	0	42	7	1	48	6	220	0	0	2
/sto/	0	1	3	1	0	0	20	31	1	1	1	297	3	1
/hiʎadu/	8	23	7	29	1	51	14	4	2	5	2	41	162	11
/million/	56	6	12	31	0	91	3	4	1	7	3	2	2	142

Fig. 6 – *Coiflets confusion matrix.*

Because the test sampling is uniform at the true-label side (each true label appears exactly 360 times), the class-by-class differences that follow cannot be ascribed to ground-truth imbalance; instead they reflect how Coif4 features position examples in the model's decision space. Reading the Coif4 matrix row-wise, the best-recovered token is /ʃest/ (6), which Coif4 recovers in 332/360 ≈ 92.2% of presentations; this is followed by /sto/ (100) at 297/360 ≈ 82.5%, and /osam/ (8) at 224/360 ≈ 62.2%. At the opposite end of the spectrum, Coif4 shows the weakest recall for /devet/ (9) at 116/360 ≈ 32.2%, /tʃetiri/ (4) at 127/360 ≈ 35.3%, and /jedan/ (1) at 131/360 ≈ 36.4%. Column-wise (precision), Coif4's predictions are also uneven: the model predicts /pet/ (5) far more than any other label (755 predictions) while predicting /tʃetiri/ (4) the least (133 predictions). This skew produces the familiar "sink" behaviour: /tʃetiri/ is extremely pure when predicted (precision ≈ 95.5%), whereas /pet/ is noisy (precision ≈ 24.4%), meaning Coif4 still funnels many ambiguous examples into a small set of attractor predictions.

Comparing Coif4 to the pooled four-wavelet matrix shows two consistent effects. First, Coif4 improves both recall and precision on most tokens: the overall accuracy rises from 50.34% to 53.17%, and every class's recall increases (for example, /deset/ (10) improves by ≈ 6.45 percentage points and /pet/ (5) by ≈ 4.68 points). Precision likewise rises for most labels: /ʃest/ (6) gains ≈ 5.9 percentage points in precision under Coif4, /tri/ (3) gains ≈ 5.15 points, and /nula/ (0) gains ≈ 5.07 points. Second, the gross shape of the error surface is preserved: the same label pairs remain the dominant confusions, but with substantially reduced magnitude under Coif4. In both matrices, the single largest off-diagonal error is /devet/→/pet/ (612 counts in the pooled matrix versus 154 in Coif4), followed by /sedam/→/sto/ (493 vs. 113) and /osam/→/deset/ (367 vs. 91). In short, Coif4 reduces the absolute number of mistakes across the board but does not alter which pairs of labels are most frequently confused; the model still tends to map difficult examples into the same attractor classes, merely less often when Coif4 features are used.

This comparison has practical implications. The Coif4 representation is the best single-wavelet performer here: it both raises true positive rates for most classes and dampens the extreme over-prediction of certain sink labels (/pet/ predictions fall from 3,068 in the pooled matrix to 755 under Coif4). Nevertheless, the persistent error modes (e.g., 9→5, 7→100, 8→10, 4→3) suggest that the separability problem is structural in the current feature set rather than an idiosyncrasy of a particular wavelet. To capitalize on Coif4's advantage one might upweight Coif4-derived features in an ensemble or use Coif4 as the primary transform for targeted feature engineering, while also pursuing class-specific remedies (data augmentation or discriminative features) to address the handful of stubborn pairwise confusions that survive across representations.

The concentration of error is striking. The top ten off-diagonal substitutions together account for 3,656 mislabels, or ≈36.1% of all errors (3,656 / 10,136 total errors). The single largest substitution is /devet/ → /pet/ (612 occurrences, as shown on Figure 7), followed by /sedam/ → /sto/ (493) and /million/ → /pet/ (377); other high-count pairs include /pet/ → /devet/ (374) and /osam/ → /deset/ (367). These pairs reveal two qualitatively different failure modes: (a) bidirectional confusions between acoustically or temporally similar numerals (for example /pet/ ↔ /devet/ appears on both sides of the top list), and (b) directional sinks where a true class is systematically mapped to another label much more often than the reverse (e.g., /sedam/ → /sto/ is far more common than /sto/ → /sedam/). Several confusions also span different magnitude types (e.g. /million/ → /pet/ and /million/ → /nula/), suggesting that duration/energy/segment boundary cues that normally separate short numerals from magnitude words are sometimes lost or misinterpreted.

The main reasons for the frequent mismatches shown in the confusion matrix (e.g., /pet/ → /devet/, /sedam/ → /sto/) are the phonetic similarity and the similar energy distribution across frequencies. In the case of /pet/ → /devet/, both patterns share almost the same syllable (/pet/~/vet/), which is dominant in the words. In other cases, such as (/sedam/ → /sto/), the initial consonant /s/ is prolonged and carries most of the energy in the speech patterns.

The confusion matrix reveals that the system is not uniformly confused: errors are concentrated in a small set of systematic substitutions and are amplified by a skewed predicted distribution. Fixing the decoder/post-processing bias and applying targeted data/feature interventions for the handful of top confusions, especially the /devet/ ↔ /pet/ pair and the /sedam/ → /sto/ direction, is the most efficient route to materially increase overall recognition performance.

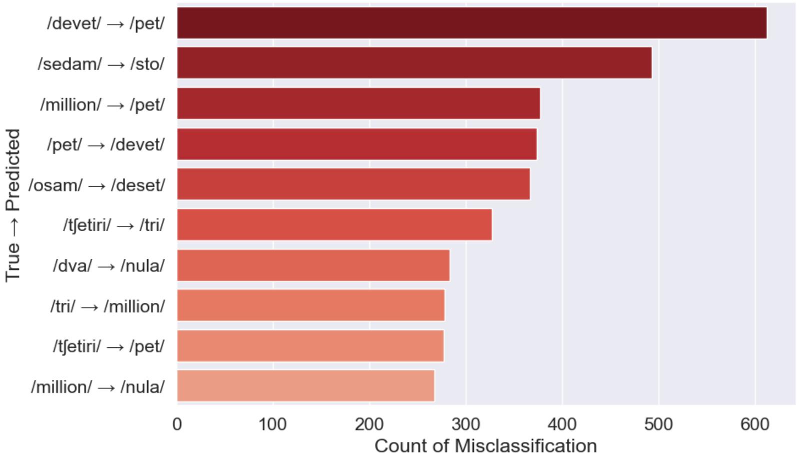


Fig. 7 – *Top* 10 *most common misclassifications.*

Fig. 8 exposes a consistent and meaningful gender-dependent performance gap that aligns with and helps explain the concentrated error modes evident in the confusion matrix.

Averaged at the word level, female speakers are consistently better recognized than male speakers for most labels, as expected [23, 24]. For example, /ʃest/ is identified almost perfectly for females (97.4%) but substantially worse for males (81.5%), and /sto/ is recognized at 90.9% vs. 71.2%. Several numerals show large female advantages (e.g. /nula/ 67.5% vs. 47.6%, /pet/ 54.3% vs. 38.3%), while a few words buck the trend: /sedam/ and /devet/ show higher male accuracy (57.8% and 34.0%, respectively) than female (52.4% and 27.9%). Because every true label is sampled uniformly in the dataset, these per-word gender differences cannot be attributed to class-frequency imbalance; rather, they point to systematic differences in how the learned features and classifier treat male and female voices. In practical terms, the female advantage on many high-support classes means that a large portion of the system's overall accuracy (≈50.3% overall) is being driven by stronger female recognition on a handful of well-recognized words, while persistent failures on other words, especially those where male accuracy is low, contribute disproportionately to total error.

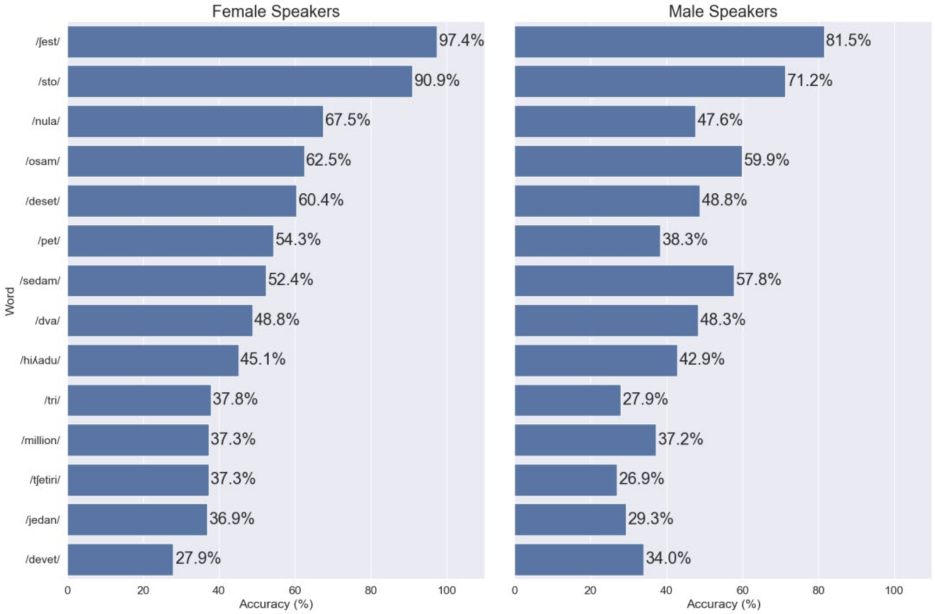


Fig. 8 – *Word accuracy (%) per speaker gender.*

Fig. 9 isolates the Coiflets representation and shows that the gender gap is most pronounced when the phonation modes match (N/N and W/W) and essentially disappears when modes mismatch.

Under Coif4, Normal/Normal accuracy is 89.4% for females versus 81.0% for males, and Whisper/Whisper is 78.7% versus 61.3%, which is a very large 17.4 percentage-point gap in the whisper condition. By contrast, cross-mode conditions (N/W and W/N) produce poor performance for both genders (≈31–32% and ≈25–26% respectively) and only trivial gender differences.

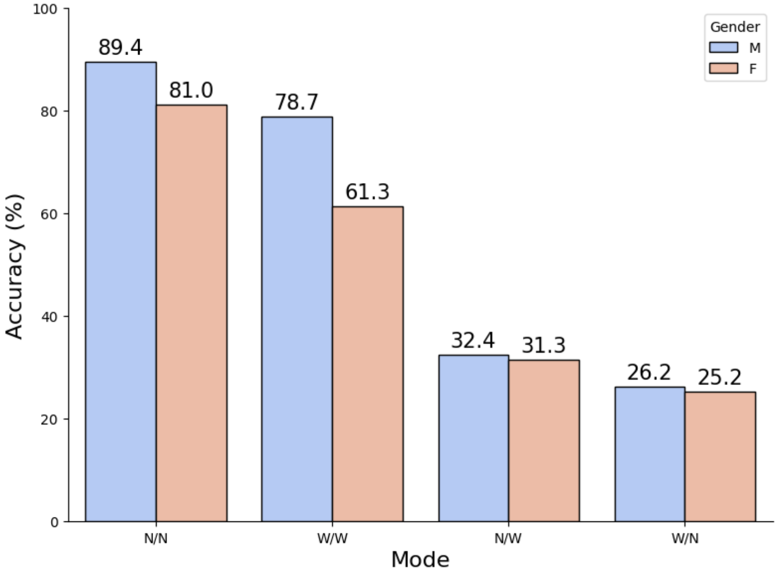


Fig. 9 – *Coiflets (coif4) accuracy by speaker gender.*

This pattern implies two distinct phenomena: first, Coiflets capture speech characteristics that allow strong, gender-dependent discrimination when training and testing modes match; second, mode mismatch is such a dominant source of error that it overwhelms gender-related effects, yielding equally poor results for both genders. Thus, improvements that address mode mismatch (data augmentation, domain adaptation, or mode-aware modeling) are likely to lift both genders simultaneously, whereas gains from gender-specific adaptation would be most effective for within-mode conditions.

When these gender-specific results are combined with the confusion-matrix findings discussed earlier, a clearer causal picture emerges. The confusion matrix shows that errors are highly concentrated in a small set of pairs (notably the /pet/↔/devet/ pair and confusions involving /sto/, /deset/, and /tri/, /tʃetiri/), and Fig. 5 shows that the model's competency on the ground-truth tokens involved in those pairs differs by gender. For instance, female recognition of /pet/ (54.3%) exceeds male (38.3%), yet females have particularly low accuracy on /devet/ (27.9%), which may create an asymmetric flow of mass from one token to another dependent on gender-specific spectral characteristics. Likewise, tokens like /ʃest/

and /sto/, which the system predicts with high purity and high recall for females, act as stabilizers for female performance but less so for males, explaining why improving a few tokens for male speakers could yield a meaningful rise in aggregate accuracy.

The combined evidence suggests concrete, targeted remediation paths. Because Coiflets already perform best and show a large within-mode gender gap, a two-step strategy is indicated: first, reduce mode mismatch with mode-aware training or robust feature transforms so that cross-mode accuracy (currently ~25–32%) no longer dominates error budgets for either gender; second, apply gender-aware calibration or speaker-adaptive fine-tuning for within-mode models, focusing especially on words with large gender discrepancies (e.g. /pet/, /sto/, /nula/) and on the top confusion pairs highlighted by the confusion matrix. In addition, the striking female superiority in Whisper/Whisper (78.7% vs. 61.3%) implies that the feature set and Coiflets in particular capture aspects of whispered female speech (perhaps higher relative spectral tilt or clearer transient cues) better than male whispered speech, so augmentation that simulates male whisper characteristics or explicit feature normalization by estimated fundamental frequency could be especially beneficial.

5 Conclusion

The experiments show three clear results. Recognition strongly depends on phonation mode, with within-mode matching far outperforming cross-mode evaluation: Normal/Normal averaged 80.4% while Whisper/Whisper averaged 67.1%, and cross-mode conditions collapse to the mid 20s to low 30s. Coif4 is the best single representation, raising pooled accuracy from 50.34% to 53.17%. Errors are highly concentrated in a small set of substitutions such as /devet/ to /pet/, /sedam/ to /sto/, and /osam/ to /deset/, and gender mediates performance so that female speakers are generally recognized more reliably within-mode while mode mismatch eliminates the gender advantage. In general, these results are lower than those obtained with feature vectors based on Linear Frequency Cepstral Coefficients (LFCC), Mel Frequency Cepstral Coefficients (MFCC), Gammatone Frequency Cepstral Coefficients (GFCC), etc. [25]. Obviously, the wavelet transformations, as implemented here, cannot capture acoustic features as effectively as LFCC, MFCC, or GFCC.

To improve results, we recommend three immediate modeling paths. First, reduce mode mismatch through mode-aware training strategies, for example by adding an explicit phonation indicator, using domain-adversarial objectives to encourage mode-invariant embeddings, or routing with a mode detector to mode-specific classifiers. Second, expand and diversify training data via targeted augmentations that simulate whisper effects and male whisper characteristics, and add discriminative acoustic features such as duration, sub-band energy ratios, and voicing probability to separate the most frequent confusion pairs. Third, apply confusion-aware training and decoding, for example through weighted losses or focal loss for high-impact substitutions, calibration of output probabilities, and a small rescoring step to reduce sink-label attraction. Normalization was not applied in these experiments; however, implementing Cepstral Mean Subtraction [26] or similar techniques in future work may improve accuracy.

Complementary steps will make progress measurable and robust. Build a Coif4-weighted ensemble or a learned meta-learner to combine wavelet representations while preserving Coif4's advantage. Explore speaker and gender adaptation by fine-tuning within-mode models by gender or by learning small per-speaker adaptation vectors. Finally, strengthen evaluation with stratified cross-validation, per-class and per-gender confidence intervals, paired statistical tests across folds, and a prioritized error-repair metric that reports reductions in the top-10 misclassification mass.

The next experimental phase should incorporate these interventions and provide cross-validated results along with targeted reductions in error mass, ensuring that the improvements are both statistically robust and practically meaningful. Furthermore, applying state-of-the-art Deep Neural Network architectures such as Convolutional Neural Networks, Recurrent Neural Networks, or Transformers, at the ASR back-end is expected to further enhance speech recognition performance.

6 Acknowledgments

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. 451-03-136/2025-03/200132 with the University of Kragujevac – Faculty of Technical Sciences Čačak.

7 References

[1] C. Zhang, J. H. L. Hansen: Analysis and Classification of Speech Mode: Whisper through Shouted, Proceedings of the Interspeech, Antwerp, Belgium, August 2007, pp. 2289 – 2292.

[2] S. T. Jovičić: Formant Feature Differences Between Whispered and Voiced Sustained Vowels, ACUSTICA - Acta Acustica, Vol. 84, No. 4, 1998, pp. 739 – 743.

[3] S. T. Jovičić, Z. M. Šarić: Acoustic Analysis of Consonants in Whispered Speech, Journal of Voice, Vol. 22, No. 3, May 2008, pp. 263 – 274.

[4] M. Matsuda, H. Kasuya: Acoustic Nature of the Whisper, Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH), Budapest, Hungary, September 1999, pp. 133 – 136.

[5] S. Mallat: A Wavelet Tour of Signal Processing, 3rd Edition, Academic Press, Amsterdam, Boston, 2008.

[6] O. Rioul, M. Vetterli: Wavelets and Signal Processing, IEEE Signal Processing Magazine, Vol. 8, No. 4, October 1991, pp. 14 – 38.

[7] M. van Berkel, G. Witvoet, P. W. J. M. Nuij, M. Steinbuch: Wavelets for Feature Detection: Theoretical Background, Eindhoven University of Technology, 2010.

[8] R. Sarikaya, B. L. Pellom, J. H. L. Hansen: Wavelet Packet Transform Features with Application to Speaker Identification, Proceedings of the IEEE Nordic Signal Processing Symposium, Vigsø, Denmark, June 1998, pp. 81 – 84.

[9] B. Marković, Đ. Damnjanović: Experiments in Whispered Speech Recognition Based on Wavelet Transformation, Proceedings of the 11th International Conference on Electrical, Electronics and Computer Engineering (IcETRAN), Niš, Serbia, June 2024, pp. 9 – 13.

[10] B. Marković, J. Galić, Đ. Grozdić, S. T. Jovičić: Application of DTW Method for Whispered Speech Recognition, Proceedings of the 4th International Conference on Fundamental and Applied Aspects of Speech and Language, Belgrade, Serbia, October 2013, pp. 308 – 315.

[11] J. Galić, S. T. Jovičić, Đ. Grozdić, B. Marković: HTK-Based Recognition of Whispered Speech, Proceedings of the 16th International Conference Speech and Computer (SPECOM), Novi Sad, Serbia, October 2014, pp. 251 – 258.

[12] J. Galić, B. Popović, D. Šumarac Pavlović: Whispered Speech Recognition Using Hidden Markov Models and Support Vector Machines, Acta Polytechnica Hungarica, Vol. 15, No. 5, 2018, pp. 11 – 29.

[13] Đ. T. Grozdić, S. T. Jovičić, M. Subotić: Whispered Speech Recognition Using Deep Denoising Autoencoder, Engineering Applications of Artificial Intelligence, Vol. 59, March 2017, pp. 15 – 22.

[14] A. Alsobhani, H. M. A. ALabboodi, H. Mahdi: Speech Recognition Using Convolution Deep Neural Networks, Journal of Physics: Conference Series, Vol. 1973, No. 1, August 2021, p. 0121166.

[15] B. Marković, S. T. Jovičić, J. Galić, Đ. Grozdić: Whispered Speech Database: Design, Processing and Application, Proceedings of the 16th International Conference Text, Speech and Dialogue (TSD), Pilsen, Czech Republic, September 2013, pp. 591 – 598.

[16] B. R. Marković, V. Lončarević, J. Galić: A Comparative Analysis of Different Wavelet Transformations on Normal and Whispered Speech Recognition, Proceedings of the 12th International Conference on Electrical, Electronic and Computing Engineering (IcETRAN), Čačak, Serbia, June 2025, pp. 1 – 5.

[17] I. Daubechies: Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[18] K. K. Shukla, A. K. Tiwari: Efficient Algorithms for Discrete Wavelet Transform: With Applications to Denoising and Fuzzy Inference Systems, Springer, London, New York, 2013.

[19] O. N. Pavlova, G. A. Guyo, A. N. Pavlov: Multiresolution Wavelet Analysis of Noisy Datasets with Different Measures for Decomposition Coefficients, Physica A: Statistical Mechanics and its Applications, Vol. 585, January 2022, p. 126406.

[20] Y. V. Parkale, S. L. Nalbalwar: Application of 1-D Discrete Wavelet Transform Based Compressed Sensing Matrices for Speech Compression, SpringerPlus, Vol. 5, November 2016, p. 2048.

[21] A. Dixit, S. Majumdar: Comparative Analysis of Coiflet and Daubechies Wavelets Using Global Threshold for Image De-Noising, International Journal of Advances in Engineering and Technology, Vol. 6, No. 5, November 2013, pp. 2247 – 2252.

[22] S. Zhang, S. Zhang, Y. Wang: Biorthogonal Wavelets in Image Compression, Proceedings of the 3rd International Conference on Intelligent Control and Information Processing, Dalian, China, July 2012, pp. 590 – 593.

[23] I. Eklund, H. Traunmüller: Comparative Study of Male and Female Whispered and Phonated Versions of the Long Vowels of Swedish, Phonetica, Vol. 54, No. 1, 1997, pp. 1 – 21.

[24] D. R. R. Smith: Speaker-Sex Discrimination for Voiced and Whispered Vowels at Short Duration, i-Perception, Vol. 7, No. 5, October 2016, pp.1 – 13.

[25] B. R. Marković: Analiza obeležja u govornom signalu za potrebe prepoznavanja multimodalnog govora, Ph.D. Thesis, University of Belgrade, Belgrade, 2018.

[26] Đ. Grozdić, S. Jovičić, D. Šumarac Pavlović, J. Galić, B. Marković: Comparison of Cepstral Normalization Techniques in Whispered Speech Recognition, Advances in Electrical and Computer Engineering, Vol. 17, No. 1, February 2017, pp. 21 – 27.