Electronics and Information Technologies, 2026, 34, 181–194
http://publications.lnu.edu.ua/collections/index.php/electronics/index

UDC 004.89:004.93:004.7

ADAPTIVE SPECTRAL NOISE REDUCTION TO IMPROVE SPEECH RECOGNITION

Oleh Osadchuk**, Ihor Olenych
Department of Radioelectronic and Computer Systems,
Ivan Franko National University of Lviv,
50 Drahomanova Str., Lviv, 79005, Ukraine

Osadchuk, O., Olenych, I. (2026). Adaptive Spectral Noise Reduction to Improve Speech Recognition. *Electronics and Information Technologies*, 34, 181–194. https://doi.org/10.30970/eli.34.11

ABSTRACT

Background. Speech recognition technology is an important component of intelligent systems in various fields, in particular for voice control of smart home devices, personal identification in security systems, automation of various technological processes, etc. Increasing requirements for accuracy, speed, and autonomy of systems require improvement of speech recognition methods, especially when affected by background noise. Therefore, the goal of this work was to develop an adaptive noise reduction method to improve the accuracy of audio information recognition.

Materials and methods. The study of the impact of noise environments on the accuracy of automatic speech recognition covers the stages of spectral analysis of audio signals using short-term Fourier transformation, adaptive noise reduction using soft masks and Wiener filter elements, and signal reconstruction based on inverse Fourier transform. The study uses Whisper speech recognition models and Common Voice and LibriSpeech datasets, supplemented with synthesized noisy recordings.

Results and Discussion. It has been established that various types of background noise (specifically low-frequency stationary, percussive, impulse, tonal, and broadband) and signal-to-noise ratios affect the speech recognition accuracy of Whisper models differently. A method to improve speech recognition accuracy based on adaptive noise reduction is proposed. This method provides flexible suppression of unwanted frequencies without significant speech distortion and allows for the compensation of even complex background noises. A reduction in the average consonant confusion level by 50–60% for the Whisper Tiny model and by 15–30% for the Whisper-1 model in real-time has been demonstrated.

Conclusion. As a result of analyzing the speech recognition efficiency of Whisper models based on phonetic distortion metrics, transcription accuracy, consonant confusion, and artifact perception across a wide range of signal-to-noise ratios, it has been established that the application of adaptive noise reduction provides a significant improvement in accuracy. Despite the higher accuracy of the Whisper-1 model, the Whisper Tiny model with a soft mask demonstrates sufficient efficiency for intelligent systems based on real-time IoT platforms.

Keywords: Whisper models, background noise, noise reduction, Wiener filter, spectral masking, IoT.

INTRODUCTION

The current level of development in electronics, computing, and information technology enables significant progress in modeling natural perception processes, such as hearing, sound signal analysis, and human speech processing. Intelligent sound and voice

command recognition systems hold a vital position in automation, security systems [1], smart homes [2], voice-controlled devices, medical and adaptive technologies [3, 4]. As the scale of these applications grows, the requirements for accuracy, speed, and autonomy of systems are increasing, especially in conditions where processing is performed on low-power IoT platforms in real time [5, 6].

Studies have shown that even modern speech recognition models, such as the Whisper family, demonstrate a significant decrease in accuracy under background noise conditions [7, 8]. This is particularly critical for autonomous security systems, where acoustic interference levels change constantly, and some noises possess complex spectral structures comparable to fricative and affricate consonant sounds. Broadband noise from fans and air conditioners, impulse signals when walking and keyboard clicks, low-frequency vibrations from transport, or the rustling of people moving around in a room overlap critical areas of the frequency spectrum and significantly impair intelligibility (s, sh, ch, zh; “c”, “ш”, “ж”, “ч”) of similar sounds in English and Ukrainian. This leads to recognition errors, which mainly affect consonant groups with similar spectral profiles.

A key feature of this field is that IoT-based security systems mostly operate in real-time on hardware platforms with limited computational resources [9, 10]. Therefore, the use of complex recurrent neural networks or deep noise-reduction models is impractical in such scenarios. Instead, there is a relevant need for spectral processing methods that introduce minimal signal delay (no more than 30–50 ms), require low computational power, and significantly enhance automatic recognition accuracy [11].

Classic spectral subtraction approaches do not provide sufficient quality for practical applications, as harsh truncating of energy in frequency bin cells leads to artifacts known as "musical noise" or "metallic timbre" [12, 13]. These features distort the structure of high-frequency consonants, degrade phase coherence, and can reduce the accuracy of Whisper models, especially lightweight versions like Whisper Tiny [14]. This creates a demand for more flexible frequency masking methods that gradually attenuate unwanted components instead of cutting them off entirely [15].

The aim of this work is an in-depth analysis of the impact of various background noise types on automatic speech recognition accuracy and the development of an adaptive noise reduction method based on the Fast Fourier Transform. Unlike traditional algorithms, the proposed approach uses a gradual transition from hard thresholds to soft spectral masks, similar to Wiener filtering [16, 17]. This reduces noise without losing important spectral characteristics of consonants and minimizes anomalies, preserving the natural sound. Particular attention is focused on noise environments most common in household and office settings, which are critical for real-time speech systems: walking, fan or air conditioner operation, keyboard clicking, indoor rustling, and traffic noise.

The paper analyses which consonant groups in English and Ukrainian are most affected by specific noise types, provides spectral examples, collects confusion statistics for Whisper Tiny and Whisper-1 models, and evaluates the effect of the proposed method. A series of experiments demonstrated that adaptive soft masking based on short-term Fourier transform reduces consonant error rates by 50–60% and significantly improves model accuracy in real-world IoT conditions.

MATERIALS AND METHODS

Physical Meaning and Theoretical Foundation of Spectral Analysis

To study the impact of the noise environment on automatic speech recognition accuracy, a generalized audio signal preprocessing architecture has been developed, covering the stages of spectral analysis, adaptive noise reduction, and signal reconstruction. This approach allows us to investigate the behavior of consonants in

different spectral conditions and determine how energy distortions in the high-frequency regions of the spectrum affect errors in Whisper models of different scales.

Spectral Estimation and Analysis Tools

Since speech is a non-stationary process, the estimation of spectral changes over time was performed using the Short-Time Fourier Transform (STFT). The signal was divided into overlapping frames of length *N* with a hop size *H*, each of which was multiplied by a window function w[*n*]. For a frame with index m (the temporal position of the frame), the spectral representation was calculated as:

$$S_m[k] = \sum_{n=0}^{N-1} x[n + mH]w[n]e^{-j2\pi\frac{kn}{N}}. \quad (1)$$

This representation allowed for tracking energy changes in frequency bins over time. The concept of a "bin" in this context describes a single discrete frequency band f_k, corresponding to the frequency.

$$f_k = \frac{kf_s}{N} \quad (2)$$

This enables the investigation of local frequency regions where fricative and affricate consonants are formed, which are particularly sensitive to noise overlap and determines which specific speech segments are most distorted under the influence of noise

Adaptive Noise Reduction and Masking

The improvement of the noise reduction system was based on the application of adaptive spectral masks. The base mask was constructed similarly to a Wiener filter:

$$G(m,k) = \frac{|S(m,k)|^\alpha}{|S(m,k)|^\alpha + |N(m,k)|^\alpha} \quad (3)$$

Here, $S(m,k)$ is the clean signal spectrum, $N(m,k)$ is the noise spectrum, and a is the sensitivity parameter (*a* = 1 corresponds to the classic Wiener filter, while *a* = 2 is a modified version). The estimation of the noise spectrum $N(m,k)$ was performed in the initial segment of the recording and was adaptively updated thereafter. During frames with low speech activity, the noise spectrum was adjusted using the formula:

$$N_{t+1} = \beta N_t + (1-\beta)|S_t|, \quad (4)$$

The $\beta \in [0.95,0.995]$ value ensured a slow and stable update of the noise model, which is essential for handling fan, air conditioner, and transport noises.

The next crucial component was the time-frequency smoothing of the mask. The mask underwent exponential smoothing in the time domain to reduce sharp transitions between adjacent frames, as well as a moving average in the frequency domain to prevent point-like fluctuations in the spectrum. Additionally, a minimum mask floor of $\approx 0.01 - 0.05$ was introduced to prevent the complete zeroing of narrow frequency bins, thereby reducing the occurrence of "musical" artifacts typical of classic spectral subtraction.

Signal Reconstruction and Synthesis

The inverse transform for each frame is performed using the Inverse Discrete Fourier Transform (IDFT), described by the expression

$$X[n] = \frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{j2\pi\frac{kn}{N}}, n = 0,1,\dots,N-1 \quad (5)$$

After restoring individual time segments, they are overlaid on each other taking into account the window function, which ensures the correct merging of overlapping segments. The final signal waveform

$$\hat{x}[n] = \frac{\Sigma_m IDFT(S_m)[n-mH]w[n-mH]}{\Sigma_m w^2[n-mH]} \quad (6)$$

This expression guarantees that the signal energy remains constant after summing the overlapping windows. The condition of a non-zero sum of squared windows ensures the absence of artifacts during reconstruction.

Experimental Methodology and Corpus

To conduct experimental studies, a specialized audio corpus was developed based on a combination of high-quality speech standards (Ground Truth) and synthesized noisy samples. The sample was primarily based on recordings from the open datasets Common Voice and LibriSpeech, supplemented by original recordings in a semi-formal style covering security systems, household IoT scenarios, and emergency alerts. Each audio segment, ranging from 3 to 7 seconds in duration, was normalized to a sampling rate of 16 kHz and a 16-bit quantization depth, consistent with the input requirements of the Whisper architecture. Before the additive noise mixing stage, the Root Mean Square (RMS) amplitude of the signals was adjusted to 20 dBFS, providing a consistent baseline for varying the signal-to-noise ratio (SNR) within the range of 0 to 15 dB. The corpus comprised approximately 83% English-language and 17% Ukrainian-language recordings, covering a diverse range of speaker voices across the 850 individual trials. The synthesized noise conditions were produced in two ways: by selecting recordings that already contained real background noise, and by programmatically mixing clean speaker recordings with noise samples sourced from open audio libraries. In **Table 1**, the interval distortion values represent the range observed across SNR levels of 0–15 dB, averaged across all noise instances within each noise category.

The testing methodology involved dividing the corpus into two control groups for a detailed verification of algorithmic robustness. The first group included recordings where target consonants (fricatives, affricates, and plosives) were heavily masked by impulsive, tonal, or low-frequency noise. The second group consisted exclusively of clean reference signals (SNR > 30 dB) passed through an identical processing chain, including spectral subtraction and adaptive soft-mask Wiener filtering. This approach allowed for the verification of the absence of undesirable side effects, such as "musical noise" or spectral degradation, which could trigger new recognition errors in clean speech segments.

In order to minimize biological deviation and take into account the acoustic characteristics of different voices, announcers aged 20 to 55 with different articulation characteristics were involved in the project. The sample included both female voices with a high fundamental frequency (F0≈200-250 Hz), enabling the analysis of noise overlap on high-register harmonics, and male voices (F0≈85-155 Hz) to study robustness against low-frequency transport interference. This demographic and acoustic balance, combined with

850 iterative trials, ensured the formation of a comprehensive factor space for an objective accuracy assessment of the Whisper Tiny and Whisper-1 models under complex operational conditions.

The experiments used Whisper Tiny (39M parameters, multilingual) deployed locally via the faster-whisper library (v1.1.1) for optimized CPU/GPU inference. Cloud-based transcription was additionally evaluated using the Whisper-1 API endpoint accessed via the OpenAI Python SDK. The preprocessing pipeline was implemented in Python 3.10 using numpy (<2.0.0), scipy (1.15.1), torch (2.3.0), and transformers (4.50.1). Audio capture was performed at a sampling rate of 16 kHz with a buffer size of 512 samples (~32 ms per frame). All local inference runs were performed on Raspberry Pi 4. The measured end-to-end processing latency for the adaptive soft-mask Wiener stage on the target IoT hardware was 3-8 ms per audio frame (window length N=512, hop H=128 at 16 kHz), which is within the declared real-time threshold of 30–50 ms.

Analysis of Consonant Distorion Under Noise

The analysis was conducted under various noise conditions (walking, keyboard, fan, transport) with SNR levels of 0, 3, 6, 9, and 15 dB. The results are shown in **Table 1**.

Evaluation Metrics and Experimental Goals

For all experiments, Word Error Rate (WER), Pair Confusion Rate for specific consonant groups, Perceptual Artifact Index, and latency were measured. The set of experiments allowed for the evaluation of each noise reduction stage's effectiveness and the determination of which algorithm characteristics most significantly influence the accuracy of Whisper models in real-world noise environments.

Table 1. **Probability of consonant distortion in various noise environments**

Background Noise	Most Affected Consonants (Eng/Ukr)	Distorted Pair Examples	Tiny Distortion (%)	Whisper-1 Distortion (%)
Footsteps (Impacts)	short percussions, high-frequency noise — tʃ, dʒ, "ч", "дж"	chip vs cheep	3.2–6.0	0.8–2.5
Fan / AC (Tonal + Broadband)	sibilants — s, ʃ, z; "с", "ш", "ж"; high-frequency fricatives	seal vs steel, sheet vs cheat	2.7–5.2	1.8–3.3
Keyboard Clicks (Impulsive)	intermittent back-grounds — p, k, t, tʃ	tap vs cap (in context)	3.4–4.3	2.2–2.6
Passing Traffic (LF & Harmonics)	low-frequency contours — less impact on sibi-lants, more on vowels	morphological errors, insertions	1.2–1.5	0.5–0.9
Indoor Movement / Rustling	sibilants and fricatives	shaft vs sharpen (in context)	0.8–1.7	0.3–1.1

RESULTS AND DISCUSSION

Experimental Setup and Trial Factors

The experimental study comprised 850 individual trials, forming a full factorial space from combinations of two models (Whisper Tiny, Whisper-1) five noise intensity levels, five noise types (walking, fan/AC, keyboard clicks, transport, broadband rustling. The study also

evaluated the impact of a preprocessing stage, specifically utilizing the adaptive soft-mask Wiener method. Each model received identical speech material with controlled consonant groups, enabling the investigation of noise environment impacts on fricatives, affricates, plosives, and their spectral contours. Evaluation was conducted using four key metrics: Phonetic Distortion (Dist), Word Error Rate (WER), Pair Confusion Rate (PCR), and Perceptual Artifact Index (PAI) which made it possible to complexly evaluate the quality of each stage of the noise reduction system.

The results confirmed that different consonant groups react differently to noise conditions and that the nature of the distortions depends on the spectral structure of the noise. Fricatives and affricates consistently showed the greatest distortion under broadband and tonal noise, while plosives (explosive sounds) proved sensitive to impulsive interference, especially at low SNR. The behavior of the Whisper models depended significantly on the spectral characteristics of the noise reduction method. The hard thresholding (flat mask) of spectral subtraction produced high levels of perceptual distortion, whereas the adaptive soft mask with time-frequency smoothing significantly reduced them.

Metrics

Four key metrics were used to evaluate the accuracy of recognition and the quality of audio signal processing, each of which reflects a specific aspect of the interaction between the speech model and the audio stream. The metrics were calculated for each audio fragment, which made it possible to compare the results in a single experimental context and take into account the variability of background noise, duration, and complexity of the material. The signal-to-noise ratio (SNR) served as the experimental condition parameter rather than an evaluation metric per se.

$$SNR_{dB} = 10 \log_{10} \frac{P_{signal}}{P_{noise}} \tag{7}$$

Here, P_{signal} is the average power of the target signal, and $P_{noise}$is the average power of the noise. The SNR value represents the relative intensity of noise compared to the signal. In this study, the SNR was varied from 0 to 15 dB to evaluate the models' performance across different noise conditions.

The Phonetic Distortion percentage was defined as the proportion of incorrectly reproduced phonemes within a specific consonant group:

$$Dist = \frac{N_d}{N_t} \times 100\% \tag{8}$$

Here, N_d is the number of distorted phonemes in the segment, and N_t is the total number of phonemes in that group. This indicator allows us to quantitatively assess the effect of noise on specific phonetic units and the effectiveness of noise reduction algorithms.

Transcription accuracy (Word Error Rate, WER) was calculated according to the standard formula:

$$WER = \frac{S + D + I}{N} \times 100\%. \tag{9}$$

Here, *S* is the number of substitutions, *D* is deletions, *I* is insertions, and *N* is the total number of words in the reference text. WER reflects the overall recognition quality and allows models to be compared under different noise conditions.

The Pair Confusion Rate (PCR) was defined as

$$PCR = \frac{N_{cp}}{N_{ttp}} \times 100\%. \quad (10)$$

Here, N_{cp} is the number of confused phoneme pairs, and N_{ttp} is the total number of tested pairs. This metric evaluates the model's ability to accurately distinguish between similar sounds.

The Perceptual Artifact Index (PAI) evaluates the presence of spectral artifacts following noise reduction, such as resonant "musical noise" and is calculated as a dimensionless coefficient in arbitrary units or in a normalized format from 0 to 1. Lower PAI values correspond to a more natural sound of the processed audio and indicate a smaller amount of undesirable distortion.

Percussive Walking Noise (Short Impacts)

Percussive walking noise functions as high-amplitude, low-frequency, non-stationary impulsive interference, which nevertheless triggers a broadband response within the recording path. Although its primary energy may be low-frequency, the spectral contour of the burst reaching the decoder includes high-frequency harmonics that directly mask the fricative components of the affricates /tʃ/ and /dʒ/.

Table 2. **Effectiveness of the adaptive noise reduction method for short percussions (walking noise)**

Model	SNR, dB	Without noise reduction				With noise reduction			
		Dist, %	WER, %	PCR, %	PAI, a. u.	Dist, %	WER, %	PCR, %	PAI, a. u.
Whisper Tiny	0	6.00	11.2	3.10	0.440	3.90	6.4	1.70	0.352
	3	5.30	10.3	2.90	0.389	2.65	5.7	1.50	0.229
	6	4.60	9.4	2.60	0.356	2.30	5.1	1.30	0.218
	9	3.90	8.3	2.30	0.314	2.54	4.8	1.20	0.233
	15	3.20	6.8	1.90	0.272	2.88	4.1	1.10	0.256
Whisper-1	0	2.50	2.4	1.05	0.230	1.63	1.3	0.48	0.168
	3	2.08	2.1	0.93	0.198	1.04	1.1	0.41	0.125
	6	1.65	1.9	0.82	0.177	0.83	1.0	0.36	0.128
	9	1.23	1.7	0.63	0.153	0.80	0.9	0.31	0.127
	15	0.80	1.4	0.55	0.128	0.72	0.8	0.27	0.120

The Tiny model demonstrates a high "WER no noise reduction (NNR)%" (11.2% at 0 dB, see **Table 2**), indicating its increased error variance in the presence of impulsive noise. This is a consequence of its acoustic space's inability to effectively discriminate the spectral attack of the noise from the acoustic transients of speech.

The use of the Wiener Mask reduces error dispersion, as illustrated by the decrease in PCR with noise reduction (WNR) % from 3.1% to 1.1%. The Wiener Mask utilizes short-term power estimation for each frequency bin. When a walking impulse causes a sudden and localized power surge, the gain coefficient in that specific bin decreases sharply. This prevents the masking of critical transition formants (which indicate the onset or offset of an affricate), allowing the system to restore the correct temporal localization of the phoneme.

Tonal and Broadband Noise (Fan / AC)

Fan noise is a quasi-stationary source that combines tonal components (rotational harmonics) and broadband noise, covering frequencies critical for sibilant fricatives /s/ and

/ʃ/ (primarily 4–8 kHz). This noise causes consistent energetic masking, reducing the local SNR across the entire temporal range of speech.

Table 3. **Effectiveness of the adaptive noise reduction method for tonal and broadband noise (fan)**

Model	SNR, dB	Without noise reduction				With noise reduction			
		Dist, %	WER, %	PCR, %	PAI, a. u.	Dist, %	WER, %	PCR, %	PAI, a. u.
Whisper Tiny	0	5.20	11.3	1.82	0.392	3.380	6.9	0.95	0.288
	3	4.35	10.8	1.52	0.342	2.830	6.1	0.76	0.256
	6	3.95	10.1	1.38	0.318	1.980	5.8	0.63	0.197
	9	3.33	8.9	1.17	0.280	2.160	5.0	0.55	0.148
	15	2.70	7.9	0.97	0.244	2.430	4.6	0.50	0.223
Whisper-1	0	3.30	2.6	0.90	0.278	2.145	1.4	0.45	0.197
	3	2.73	2.4	0.75	0.236	1.774	1.3	0.39	0.163
	6	2.25	2.3	0.63	0.215	1.125	1.2	0.31	0.148
	9	1.98	2.1	0.54	0.197	1.287	1.2	0.29	0.166
	15	1.80	1.9	0.48	0.188	1.620	1.1	0.26	0.168

The "PCR NNR %" in **Table 3** for the Tiny model (1.82% at 0 dB) reflects direct masking, as the noise energy consistently exceeds the fricative discrimination threshold. Since the noise is stationary, the Wiener Mask demonstrates high predictability and effectiveness, achieving steady suppression in constant noise-dominant regions. The reduction in "PCR WNR %" from 0.95% to 0.50% indicates successful spectral separation. The method preserves the internal amplitude modulation of the speech fricative noise above the background, which is key for phoneme identification by the decoder whereas the subtractive method could introduce "musical noise," destroying these vital acoustic features.

Impulsive Noise (Keyboard Clicks)

Keyboard noise is a high-intensity impulsive noise with broad spectral coverage that affects plosive consonants /p/, /k/, /t/, which are characterized by an energy burst and Voice Onset Time (VOT). A click impulse can be erroneously identified by the ASR decoder as a false acoustic transient (i.e., a false burst) or can completely mask the genuine burst, leading to consonant omission or substitution errors.

The high "WER NNR %" (13.1% at 0 dB, see **Table 4**) for the Tiny model indicates that false keyboard transients destabilize the decoder's temporal synchronization. The reduction in "PCR WNR %" from 0.95% to 0.58% results from the successful operation of the Wiener Mask with time-window adaptation. The mask, reacting to the impulse, rapidly reduces its amplitude, minimizing the probability of its false identification as a speech burst. This reduces the energetic correlation between noise and speech in the time domain, restoring the integrity of the VOT for genuine consonants.

Low-Frequency Harmonics (Transport)

Transport noise is stationary and predominantly low-frequency, concentrating on the first and second formants (F1, F2) of vowels. Its primary impact lies in the deformation of vowel frequency contours, which complicates their identification and leads to global decoding errors.

The low "PCR NNR %" (0.45% at 0 dB) in **Table 5** confirms that this noise has a lesser impact on the acoustic accuracy of consonants while the high "WER NNR %" (13.9% at 0 dB) highlights morphological and lexical errors caused by vowel deformation.

The effectiveness of the Wiener Mask against stationary, low-frequency noise is high. The reduction in "WER WNR %" from 11.2% to 4.3% demonstrates the mask's ability to

Table 4. **Effectiveness of the adaptive noise reduction method for impulsive noise (keyboard clicks)**

Model	SNR, dB	Without noise reduction				With noise reduction			
		Dist, %	WER, %	PCR, %	PAI, a. u.	Dist, %	WER, %	PCR, %	PAI, a. u.
Whisper Tiny	0	4.300	13.10	1.46	0.335	2.795	8.40	0.95	0.215
	3	3.950	12.40	1.34	0.323	2.370	6.00	0.73	0.195
	6	3.850	11.80	1.30	0.317	1.925	5.60	0.67	0.155
	9	3.625	10.70	1.26	0.306	2.356	5.40	0.63	0.141
	15	3.400	9.80	1.19	0.292	3.060	5.10	0.58	0.268
Whisper-1	0	2.600	2.60	0.62	0.236	1.690	1.55	0.31	0.140
	3	2.450	2.35	0.57	0.225	1.590	1.30	0.28	0.125
	6	2.400	2.25	0.55	0.221	1.200	1.20	0.26	0.108
	9	2.310	2.10	0.52	0.216	1.500	1.15	0.22	0.125
	15	2.200	1.95	0.50	0.215	1.980	1.05	0.10	0.197

Table 5. **Effectiveness of the adaptive noise reduction method for low-frequency harmonics (transport)**

Model	SNR, dB	Without noise reduction				With noise reduction			
		Dist, %	WER, %	PCR, %	PAI, a. u.	Dist, %	WER, %	PCR, %	PAI, a. u.
Whisper Tiny	0	1.500	13.9	0.45	0.170	0.975	11.2	0.32	0.138
	3	1.425	11.3	0.42	0.162	0.712	6.6	0.25	0.123
	6	1.350	9.4	0.40	0.162	0.675	5.4	0.23	0.118
	9	1.275	8.6	0.38	0.155	0.829	4.7	0.19	0.119
	15	1.200	7.8	0.36	0.152	1.080	4.3	0.18	0.168
Whisper-1	0	0.900	7.1	0.24	0.134	0.585	6.3	0.15	0.125
	3	0.825	5.2	0.22	0.129	0.412	3.4	0.14	0.099
	6	0.750	4.4	0.21	0.125	0.375	2.4	0.12	0.095
	9	0.675	3.7	0.19	0.121	0.439	2.2	0.10	0.099
	15	0.560	3.2	0.17	0.116	0.540	1.9	0.09	0.141

restore the clarity of formant peaks by suppressing constant energy noise in low-frequency bins. This improves the internal acoustic model of vowels within the decoder, significantly increasing the accuracy of word searches in the lexicon and minimizing global errors.

Broadband Non-Stationary Noise (Rustling)

Rustling (indoor movement) is the most challenging type of noise, as it is both non-stationary and broadband.

Its spectral characteristics change rapidly, resulting in the worst "WER NNR %" (17.8% at 0 dB) for the Tiny model, as detailed in **Table 6**. This high error rate reflects the model's inability to effectively process rapidly shifting spectral contours that mask fricatives and other consonants. The effectiveness of the Wiener mask here is noticeable but not as significant as for stationary noise; its performance is markedly lower (with "WER WNR %" decreasing only to 15.9% at 0 dB). This is explained by the fact that noise estimation in an

Table 6. **Effectiveness of the adaptive noise reduction method for broadband non-stationary noise (rustling)**

Model	SNR, dB	Without noise reduction				With noise reduction			
		Dist, %	WER, %	PCR, %	PAI, a. u.	Dist, %	WER, %	PCR, %	PAI, a. u.
Whisper Tiny	0	1.700	17.8	0.38	0.182	1.105	15.9	0.26	0.146
	3	1.530	14.3	0.36	0.172	0.765	8.8	0.23	0.122
	6	1.250	12.1	0.34	0.158	0.625	6.4	0.22	0.116
	9	1.050	10.7	0.32	0.143	0.682	6.0	0.18	0.126
	15	0.800	9.2	0.29	0.128	0.720	5.5	0.17	0.169
Whisper-1	0	1.100	8.9	0.26	0.146	0.715	8.2	0.16	0.121
	3	0.950	6.2	0.24	0.134	0.570	4.3	0.15	0.105
	6	0.800	5.0	0.22	0.128	0.400	2.6	0.14	0.092
	9	0.675	4.1	0.20	0.121	0.439	2.4	0.12	0.099
	15	0.300	3.5	0.18	0.098	0.270	2.2	0.11	0.094

adaptive system requires a certain accumulation time (noise estimation latency). The rate of change in the spectral characteristics of rustling exceeds the mask's adaptation time constant, leading to inaccurate local SNR estimates. Therefore, signal recovery is partial, although the consistent reduction in Dist and PCR still confirms that the Wiener mask provides an overall reduction in energy pollution across the entire spectrum.

Qualitative and Quantitative Assessment of ASR Performance

In most scenarios, such as transport (low-frequency stationary noise), fan (tonal), walking (percussive), and keyboard (impulsive), the proposed noise reduction method ensures full restoration of phonematic accuracy. A qualitative comparison of these effects across different noise types is presented in **Table 7**. This indicates that when noise spectral profiles are relatively stable or predictable (as with tonal or impulsive interference), the adaptive approach effectively reconstructs critical speech features, specifically Voice Onset Time (VOT) and formant structure.

The adaptive soft-mask Wiener filter demonstrates a particular advantage in complex acoustic environments, showing high robustness even with variability in the spectral profile

Table 7. **Qualitative Illustration of Noise Impact and Noise Reduction Method on ASR Decoding**

Ground Truth	Noise (SNR)	Phonematic Substitution Type	ASR before Denoising	ASR after Adaptive Soft-Mask
A large van is speeding	Transport (3 dB)	Vowels /æ/, /ʌ/ (Low)	A large fun is speeding	A large van is speeding
The file must be closed	Fan (7 dB)	Fricatives /f/, /θ/, /v/ (High)	The thigh must be closed	The file must be closed

She will grasp the concept	Walking (5 dB)	Fricatives /s/, /z/ (Aspiration)	She will grass the concept	She will grasp the concept
The dock is far away	Keyboard (8 dB)	Plosives /d/, /g/ (VOT/Burst)	The gock is far away	The dock is far away
He used a winch to pull	Rustling (4 dB)	Affricates /tʃ/, /ʃ/ (Spectral)	He used a witch to pull	He used a winch to pull

of the interference. A characteristic example is rustling (broadband non-stationary noise): due to its capability for precise dynamic modelling, the soft-mask successfully recovers (e.g., witch). This confirms that the mechanism of soft suppression of non-stationary energy is critically important for preserving the integrity of acoustic speech features and ensuring high recognition accuracy.

Based on the aggregated data (**Table 8**), which reflect the average performance of the Whisper Tiny and Whisper-1 models under mean acoustic noise conditions, a consistent conclusion regarding system robustness is established. It is fundamentally established that the Whisper-1 model demonstrates significantly higher intrinsic robustness, as its average "WER NNR %" at 0 dB SNR (4.24%) is 2.65 times lower than that of the Tiny model (11.22%). This difference indicates that the larger capacity of the acoustic model more effectively forms invariant spectral representations capable of reliably segregating speech from noise.

Further analysis confirms a monotonic decrease in all error metrics (Dist, WER, PCR) as the SNR increases, consistent with the normative behavior of ASR systems. The average "PCR NNR %" for Tiny (0.822% at 0 dB) is twice as high as that of Whisper-1 (0.404%), highlighting the smaller model's critical vulnerability to the masking of high-frequency transients and fricatives. This indicator directly correlates with Tiny's inability to accurately localize energy emissions and VOT consonants in the time domain, amplified by background noise.

Table 8. **General assessment of background noise impact using Whisper Tiny and Whisper-1 models speech recognition metrics and the effectiveness of noise reduction method**

Model	SNR, dB	Without noise reduction				With noise reduction			
		Dist, %	WER, %	PCR, %	PAI, a. u.	Dist, %	WER, %	PCR, %	PAI, a. u.
Whisper Tiny	0	2,540	11,22	0,822	0,215	1,651	8,48	0,496	0,157
	3	2,251	9,76	0,728	0,199	1,335	5,50	0,394	0,139
	6	3,000	8,68	0,684	0,191	1,041	4,64	0,350	0,160
	9	2,636	7,78	0,626	0,176	1,205	4,22	0,310	0,153
	15	2,020	5,38	0,490	0,132	1,242	3,04	0,250	0,183
Whisper-1	0	1,580	4,24	0,404	0,158	1,027	3,49	0,214	0,116
	3	1,391	3,23	0,356	0,144	0,869	2,06	0,192	0,098
	6	1,570	3,17	0,486	0,173	0,786	1,68	0,238	0,114
	9	1,374	2,74	0,416	0,161	0,893	1,57	0,208	0,123
	15	0,972	2,11	0,266	0,123	0,882	1,25	0,112	0,120

Evaluation of noise reduction effectiveness through PAI metrics reveals the greatest reduction in objective errors due to the adaptive soft-mask Wiener filter, particularly in the 6 dB SNR range. This fact confirms the principle that ASR optimization should be directed toward preserving acoustic speech features rather than maximizing subjective noise suppression. The Wiener mask minimizes the introduction of spectral artifacts (such as "musical noise") and ensures the restoration of the time-frequency integrity of the speech signal. At high SNR (15 dB), the effectiveness of noise reduction decays asymptotically, as

further processing of a clean signal has limited improvement potential and may even slightly degrade performance due to subtle spectral distortion effects.

CONCLUSION

The work conducted involved an in-depth study of the impact of typical background noise on the accuracy of automatic speech recognition (ASR) based on Whisper models and an assessment of the effectiveness of the proposed adaptive spectral noise reduction method. The evaluated models, Whisper Tiny and Whisper-1, were assessed using Word Error Rate (WER), Pair Confusion Rate (PCR), and Perceptual Artifact Index (PAI) metrics across a wide range of Signal-to-Noise Ratios (SNR).

It was established that the Whisper-1 model demonstrates systematically higher intrinsic robustness, as evidenced by a significantly lower average WER NNR % at low SNR (4.24% at 0 dB), which is 2.65 times lower than that of the Tiny model (11.22%). Its larger capacity enables more effective speech-noise segregation and maintains a lower PCR error level. In contrast, the Whisper Tiny model showed the highest vulnerability to consonant masking and the greatest error dispersion, limiting its application in high noise environments.

The proposed adaptive spectral noise reduction method based on the Wiener soft-mask provided an accuracy boost, reducing the average PCR WNR level by 50–60% for Whisper Tiny and by 15–30% for Whisper-1. The Soft-Mask method is optimal for ASR because it provides low PAI WNR (less distortion) and preserves the integrity of acoustic speech features. An analysis of the metrics leads to the conclusion that the choice of the optimal model depends on accuracy requirements and hardware resources. Although Whisper-1 provides the highest accuracy, its computational cost is excessive for real-time IoT platforms. Considering the need for minimal latency and limited resources, Whisper Tiny — thanks to a significant boost in accuracy (reducing WER by 50% or more) after applying effective Soft-Mask noise reduction — becomes the optimal choice for IoT-based intelligent security systems.

ACKNOWLEDGMENTS AND FUNDING SOURCES

The authors received no financial support for the research, writing, and/or publication of this article.

COMPLIANCE WITH ETHICAL STANDARDS

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

Conceptualization, [O.O., I.O.]; methodology, [O.O.]; investigation, [O.O.]; writing – original draft preparation, [O.O.]; writing – review and editing, [O.O., I.O.]; visualization, [O.O.].

All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] Smart Secure Homes: A Survey of Smart Home Technologies that Sense, Assess, and Respond to Security Threats / J. Dahmen, D. J. Cook, X. Wang, W. Honglei. *Journal of Reliable Intelligent Environments*. 2017. Vol. 3, no. 2. P. 83–98. https://doi.org/10.1007/s40860-017-0035-0

[2] Knowledge-based Cyber Physical Security at Smart Home: A Review / A. Alsufyani, O. Rana, C. Perera. *ACM Computing Surveys*. 2024. Vol. 57, iss. 3. Art. 53. P. 1–36. https://doi.org/10.1145/3698768

[3] Noise Reduction in Industry Based on Virtual Instrumentation / R. Martinek, R. Jaros, J. Baros [et al.]. *Computers, Materials and Continua*. 2021. Vol. 69, iss. 1. P. 1073–1096. https://doi.org/10.32604/cmc.2021.017568
[4] Real-Time Speech Recognition for IoT Purpose using a Delta Recurrent Neural Network Accelerator / C. Gao, S. Braun, I. Kiselev, J. Anumula. *2019 IEEE International Symposium on Circuits and Systems (ISCAS)* : Conference Paper. 2019. https://doi.org/10.1109/ISCAS.2019.8702290
[5] Application of speech recognition technology in IoT smart home / P. Wang, X. Lu, H. Sun, W. Lv. *2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)* : Conference Paper. 2019. https://doi.org/10.1109/IMCEC46724.2019.8984175
[6] Towards Energy-Efficient and Low-Latency Voice-Controlled Smart Homes: A Proposal for Offline Speech Recognition and IoT Integration / P. Huang, I. Ullah, X. Wei, T. A. Ahanger. *arXiv*. 2025. https://doi.org/10.48550/arXiv.2506.07494
[7] Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also *Strong General Audio Event Taggers / Y. Gong, S. Khurana, L. Karlinsky, J. Glass. Proceedings of the 24th Annual Conference of the International Speech Communication Association (INTERSPEECH 2023)*. 2023. P. 2798–2802. https://doi.org/10.21437/Interspeech.2023-2193
[8] Improving Speech Recognition Performance in Noisy Environments by Enhancing Lip Reading Accuracy / D. Li, Y. Gao, C. Zhu [et al.]. *Sensors*. 2023. Vol. 23, iss. 4. Art. 2053. https://doi.org/10.3390/s23042053
[9] AI-Driven Real-Time Distress Detection Through Speech Recognition for Emergency Response Systems / M. Halilaj, E. Myrto, A. F. Dragoni. CEUR Workshop Proceedings (CEUR-WS.org). 2025. Vol. 4044, paper 03. https://ceur-ws.org/Vol-4044/paper03.pdf
[10] IoT based speech recognition system / K. K. Sethy, L. M. Varalakshmi, R. E. [et al.]. *AIP Conference Proceedings*. 2022. Vol. 2393, iss. 1. Art. 020096. https://doi.org/10.1063/5.0074140
[11] Unsupervised Spectral Subtraction for Noise-Robust ASR / G. Lathoud, M. Magimai-Doss, B. Mesot, H. Bourlard. *IDIAP Research Institute*. 2006. https://publications.idiap.ch/downloads/papers/2005/lathoud05e.pdf
[12] Musical Noise Reduction By Processing Spectrogram of Spectral Subtraction Output / H. R. Abutalebi, B. Dashtbozorg. *Proceedings of the 15th International Congress on Sound and Vibration (ICSV15)*. 2008. P. 2979–2986. https://www.researchgate.net/publication/228916452_Signal_Processing_Musical_Noise_Reduction_By_Processing_Spectrogram_of_Spectral_Subtraction_Output
[13] Vaseghi S. V. Advanced Digital Signal Processing and Noise Reduction. 2nd ed. John Wiley & Sons Ltd, 2000. 488 p. ISBN 0-471-62692-9. https://scispace.com/pdf/advanced-digital-signal-processing-and-noise-reduction-375hxslkau.pdf
[14] Wiener Filter and Deep Neural Networks: A Well-Balanced Pair for Speech Enhancement / D. Ribas, A. Miguel, A. Ortega, E. Lleida. *Applied Sciences*. 2022. Vol. 12, iss. 18. Art. 9000. https://doi.org/10.3390/app12189000
[15] Reddy A., Raj B. Soft Mask Methods for Single-Channel Speaker Separation. *IEEE Transactions on Audio, Speech, and Language Processing*. 2007. Vol. 15, iss. 6. P. 1766–1776. https://doi.org/10.1109/TASL.2007.901310
[16] Hout J. van, Alwan A. A novel approach to soft-mask estimation and Log-Spectral enhancement for robust speech recognition. *2012 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP). 2012. P. 4105–4108. https://doi.org/10.1109/ICASSP.2012.6288821

[17]] Radfar M. M., Dansereau R. M. Single-Channel Speech Separation Using Soft Mask Filtering. *IEEE Transactions on Audio, Speech, and Language Processing*. 2007. Vol. 15, iss. 8. P. 2299–2310. https://doi.org/10.1109/TASL.2007.904233

АДАПТИВНЕ СПЕКТРАЛЬНЕ ШУМОЗНИЖЕННЯ ДЛЯ ПІДВИЩЕННЯ ТОЧНОСТІ РОЗПІЗНАВАННЯ МОВЛЕННЯ

Олег Осадчук*, Ігор Оленич
Кафедра радіоелектронних і комп'ютерних систем,
Львівський національний університет імені Івана Франка,
вул. Драгоманова 50, Львів, 79005, Україна

АНОТАЦІЯ

Вступ. Технологія розпізнавання мовлення є важливою складовою інтелектуальних систем у різних галузях, зокрема, для голосового керування пристроями розумного будинку, ідентифікації особи у системах безпеки, автоматизації різноманітних технологічних процесів тощо. Підвищення вимог до точності, швидкодії та автономності систем потребує удосконалення методів розпізнавання мовлення, особливо за впливу фонових шумів. Тому мета роботи полягала у розробленні адаптивного методу шумозниження для підвищення точності розпізнавання аудіоінформації.

Матеріали та методи. Дослідження впливу шумового середовища на точність автоматичного розпізнавання мовлення охоплює етапи спектрального аналізу аудіосигналів за допомогою короткочасного перетворення Фур'є, адаптивного шумозниження з використанням м'яких масок і елементів Вінер-фільтрації та реконструкції сигналу на основі зворотного перетворення Фур'є. Для дослідження використано моделі розпізнавання мовлення Whisper і набори даних Common Voice та LibriSpeech, доповнені синтезованими зашумленими записами.

Результати. Встановлено, що різні типи фонових шумів (зокрема, низькочастотні стаціонарні, перкусійні, імпульсні, тональні, широкосмугові) і співвідношення сигналу до шуму по різному впливають на точність розпізнавання мовлення моделями Whisper. Запропоновано метод підвищення точності розпізнавання мовлення на основі адаптивного шумозниження, який забезпечує гнучке придушення небажаних частот без суттєвого спотворення мовлення і дає змогу компенсувати вплив навіть складних фонових шумів. Продемонстровано зменшення середнього рівня плутанини приголосних на 50–60% для моделі Whisper Tiny та 15–30% для моделі Whisper-1 у реальному часі.

Висновки. У результаті аналізу ефективності розпізнавання мовлення моделями Whisper за метриками фонетичних спотворень, точності транскрипції, плутанини приголосних та сприйняття артефактів у широкому діапазоні співвідношенням сигналу до шуму встановлено, що застосування адаптивного шумозниження забезпечує суттєве підвищення точності моделей Whisper. Попри вищу точність моделі Whisper-1, модель Whisper Tiny з м'якою маскою демонструє достатню ефективність для інтелектуальних систем на основі IoT-платформ реального часу.

Ключові слова: моделі Whisper, фоновий шум, шумоподавлення, Wiener-фільтр, спектральне маскування, IoT.

Received / Одержано	Revised / Доопрацьовано	Accepted / Прийнято	Published / Опубліковано
15 April, 2026	09 June, 2026	10 June, 2026	29 June, 2026