Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences

Araştırma makalesi / Research article

www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

Acoustic analysis of the NATO phonetic alphabet spoken by native Turkish speakers: A comparison of handheld and throat microphones

Anadili Türkçe olan konuşmacılar tarafından söylenen NATO fonetik alfabesinin akustik analizi: El ve gırtlak mikrofonlarının karşılaştırması

Julio Cesar Velazquez Garcia[1,*], Selim Aras[2]

[1]*Ondokuz Mayıs University, Graduate School of Natural and Applied Sciences, Department of Intelligent Systems Engineering, 55139, Samsun, Türkiye*
[2]*Ondokuz Mayıs University, Faculty of Engineering, Department of Electrical and Electronics Engineering, 55139, Samsun, Türkiye*

Abstract

This pilot study presents a comparative acoustic analysis of speech signals recorded with handheld and throat microphones during the pronunciation of the NATO phonetic alphabet by native Turkish speakers. A total of 2,080 voice samples were collected and preprocessed using voice activity detection (VAD), noise reduction, and silence trimming. Acoustic metrics; duration, root mean square (RMS) energy, zero-crossing rate (ZCR), and zero-crossing density, were extracted. Non-parametric tests revealed significant differences between microphone types: the handheld device yielded higher energy and stability, while the throat microphone showed lower energy but higher spectral complexity. Results also indicated phonatory delay and articulation asymmetries in some speakers. The findings suggest that microphone type affects phonetic structure and signal quality, with implications for automatic speech recognition (ASR) systems in noisy environments. Notably, the throat microphone increased zero-crossing density due to internal vibration sensitivity. This study lays the groundwork for multichannel ASR systems designed for real-world acoustic variability.

Keywords: NATO phonetic alphabet, Throat microphone, Handheld microphone, Speech signal analysis, ASR robustness.

Öz

Bu pilot çalışma, anadili Türkçe olan konuşmacıların NATO fonetik alfabesini telaffuz ederken el tipi ve boğaz mikrofonlarıyla kaydedilen konuşma sinyallerinin karşılaştırmalı akustik analizini sunmaktadır. Toplam 2.080 ses örneği toplanmış ve ses etkinliği tespiti (VAD), gürültü azaltma ve sessizlik kırpma adımlarını içeren bir ön işleme sürecinden geçirilmiştir. Süre, ortalama karekök (RMS) enerjisi, sıfır geçiş oranı (ZCR) ve saniye başına sıfır geçiş yoğunluğu gibi akustik ölçümler çıkarılmıştır. Non-parametrik testler, mikrofon türleri arasında anlamlı farklar olduğunu ortaya koymuştur: El tipi mikrofon daha yüksek enerji ve zamansal kararlılık sağlarken, boğaz mikrofonu daha düşük enerji ancak daha yüksek spektral karmaşıklık göstermiştir. Bazı konuşmacılarda ses üretiminde gecikme ve artikülasyon asimetrileri gözlemlenmiştir. Bulgular, mikrofon türünün fonetik yapı ve sinyal kalitesini etkilediğini ve gürültülü ortamlarda otomatik konuşma tanıma (ASR) sistemleri için önemli çıkarımlar sunduğunu göstermektedir. Bu çalışma, gerçek dünya koşullarına uygun çok kanallı ASR sistemleri için temel oluşturmaktadır.

Anahtar kelimeler: NATO fonetik alfabesi, Gırtlak mikrofonu, El tipi mikrofon, Konuşma sinyali analizi, ASR sağlamlığı.

1 Introduction

In communication-critical domains such as aviation, military operations, and air traffic control systems, ensuring the intelligibility of transmitted speech is paramount. The NATO Phonetic Alphabet, designed to minimize phonetic ambiguity, serves as a global standard to reduce miscommunication during verbal exchanges. According to the International Civil Aviation Organization [1], this phonetic system was developed to maintain clarity even under stressful or noisy operational conditions.

However, despite its international adoption, real-world reports and empirical studies have consistently pointed to the persistence of misunderstandings related to phonetic codes. One study [2], for instance, documented that reduced speech intelligibility in noisy operational environments led to increased communication errors among Navy personnel, emphasizing that factors such as environmental noise and speaker fatigue can critically undermine operational efficiency and safety.

Importantly, the intelligibility of spoken input is not only shaped by human articulation but also by the characteristics of the devices used to capture speech. Handheld microphones, known for their wider frequency response and higher sensitivity, tend to record phonetically rich but noise-susceptible signals [3]. On the other hand, throat microphones, which are in direct contact with the speaker's throat, effectively filter out ambient noise. However, they can distort the spectral and energetic properties of the signal, leading to lower clarity in automatic speech recognition (ASR) systems. This distortion arises from the loss of essential spectral components, resulting in a less natural and more hoarse sound quality [4].

* Sorumlu yazar / Corresponding author, e-posta / e-mail: julio.velazquez18@gmail.com (J.C. Velazquez Garcia)
Geliş / Received: 23.05.2025 Kabul / Accepted: 21.10.2025 Yayımlanma / Published: 15.10.2026
doi: 10.28948/ngumuh.1705336

Table 1. Comparative analysis of microphone studies in robust ASR.

Criterion	Present Study	[5] (Acker-Mills et al., 2006)	[6] (Erzin, 2009)	[4] (Tugtekin Turan, 2018 / [7] comparative 2024)
Main Objective	Comparative acoustic analysis (handheld vs. throat) focusing on RMS energy and spectral complexity (ZCR).	Evaluate speech intelligibility for throat vs. acoustic microphones under helicopter/noisy conditions (MRT).	Joint-analysis / multichannel frameworks (throat + acoustic) to improve throat-mic ASR.	Compare enhancement/mapping techniques (GMM, spectral mapping, EMD, LPCC, MFCC) to recover lost high-frequency information in TM recordings.
Microphone Type	Handheld (condenser) & throat (piezo/contact).	Acoustic (boom/helmet) & throat (larynx contact).	Throat + acoustic (paired recordings; joint processing).	Throat (TM) and acoustic/close-talk (AM) for mapping/enhancement experiments.
Key Metrics	RMS energy, ZCR, duration; non-parametric tests; confusion matrices and WER in downstream ASR tests.	Intelligibility percentage (Modified Rhyme Test — MRT), diagnostic consonant tests.	Word Error Rate (WER), feature-mapping gain, objective quality metrics.	Objective speech quality (PESQ/LSD), SNR improvement, subjective A/B preference tests, WER after enhancement.
Relevant Conclusion	TM shows reduced energy and higher ZCR (perceived distortion/complexity), which impacts phonetic realizations and ASR performance.	TM intelligibility is significantly lower than acoustic mics in high noise (no consistent benefit from simple combo)	Multichannel co-analysis (joint TAM approaches) significantly reduces WER for TM-driven ASR.	Mapping and spectral-envelope restoration (phone-dependent GMM, LPCC, EMD) can substantially improve TM intelligibility/quality.

A review of the literature, summarized in Table 1, consistently shows two complementary facts regarding body-coupled microphones: throat microphones (TM) excel at noise robustness because they pick up tissue vibrations rather than airborne noise, but they lose high-frequency spectral content, which degrades phonetic cues critical for ASR. Early experimental work in noisy aviation contexts [5] quantified the severe intelligibility drops using standard testing protocols. Subsequent research has focused on joint-analysis or mapping approaches [6] to reconstruct the missing acoustic information and thus recover ASR performance. Recent enhancement studies [4] further validate that spectral-envelope restoration and mapping techniques can substantially improve TM intelligibility and quality metrics. Our study is positioned within this research area, offering a quantitative acoustic comparison focused on two fundamental metrics: energy (RMS) and spectral complexity (ZCR).

Given the growing integration of ASR in mission-critical environments, these hardware-induced signal differences present a significant concern. The performance of ASR systems (especially those tasked with recognizing isolated terms like phonetic letters) depends heavily on the consistency and clarity of the audio input [8]. As such, understanding how different microphone types affect the acoustic integrity of phonetic speech is essential for designing resilient, adaptive ASR architectures.

The present study builds upon this premise by investigating the acoustic impact of microphone type in a controlled recording setup. Specifically, 2.080 audio samples were recorded from Turkish speakers using both handheld and throat microphones in parallel. The aim was to measure and compare key acoustic features; including duration, RMS energy, and zero-crossing rate (ZCR), under standardized conditions. All signals were preprocessed through a consistent pipeline involving voice activity detection (VAD), noise reduction, silence trimming, and normalization to ensure analytical reliability.

This pilot study sets out to examine whether microphone type introduces significant and systematic acoustic differences in recordings of the NATO Phonetic Alphabet. The broader objective is to provide evidence that can inform the development of ASR systems capable of dynamically adapting to input variability—across microphones, speakers, and linguistic profiles. By exploring how such variability manifests in a multilingual context, this work contributes to ongoing efforts in enhancing the inclusivity, reliability, and practical deployment of ASR technologies in operational environments.

2 Materials and methods.

2.1. Database and recording configuration

The audio database used in this study consists of 2.080 speech recordings in WAV format, produced through a synchronous dual-channel recording setup. Two types of microphones were employed: a handheld condenser microphone (RODE AI-1) and a piezoelectric throat microphone, which was attached directly to the neck area. This configuration allowed the simultaneous capture of each utterance under phonetically equivalent conditions, ensuring reliable comparisons [9-11]. The corpus includes four native Turkish speakers (two female and two male), each of whom pronounced the 26 NATO phonetic alphabet letters 10 times per microphone. Recordings were carried out in a quiet domestic environment, using the software VoiceMeeter Banana for routing and Audacity for audio capture. All files were stored in uncompressed WAV format to preserve signal fidelity and maintain compatibility with tools used for acoustic analysis and automatic speech recognition, such as librosa, torchaudio, Kaldi, and ESPnet [12, 13]. Each file was named using a structured five-part identifier that encoded the language, speaker, microphone type, NATO letter, and sample number. This systematic naming approach [14] enhanced corpus scalability and traceability, allowing for seamless integration into multilingual speech analysis pipelines.

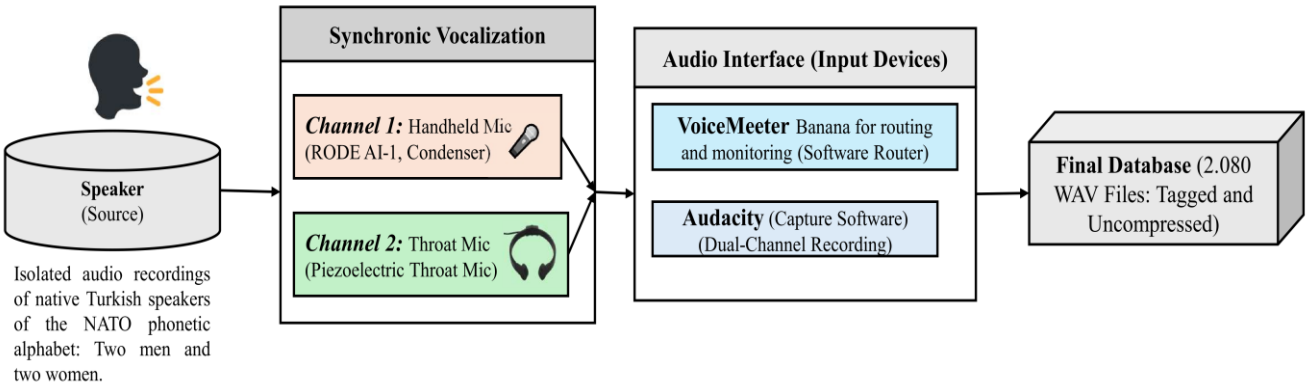

Figure 1. Synchronous dual-channel speech recording setup

Figure 1 visually represents the synchronous recording architecture used in this study. This dual-channel configuration was essential for capturing simultaneous signals from the Handheld condenser microphone and the Piezoelectric throat microphone for each utterance. The signals were routed and time-aligned via VoiceMeeter Banana and captured by Audacity to ensure phono-temporal equivalence across both recording channels.

2.2. Preprocessing of speech samples

Prior to acoustic analysis, all recordings underwent a structured preprocessing pipeline designed to isolate relevant speech segments and reduce signal variability. This included four main stages: voice activity detection, noise reduction, trimming of residual silences, and normalization. Voice activity detection (VAD) was conducted using the Silero VAD model, which is based on convolutional neural networks trained on multilingual datasets. A safety margin of 100 milliseconds was applied before and after each detected segment to preserve transition integrity [15]. All audio files were resampled to 16 kHz prior to applying VAD to ensure temporal precision. Once the active speech was isolated, a spectral subtraction-based noise reduction technique was applied using the noisereduce library [16, 17]. This method has proven effective in domestic environments, particularly when background interference is non-stationary. Its application was especially relevant for samples recorded with the handheld microphone. Residual silences not captured by the VAD model were trimmed using the librosa.effects.trim() function with a threshold of 40 dB. This ensured the removal of low-energy sections while preserving vocalic structure. Samples shorter than 0.2 seconds after trimming were excluded [18]. This preprocessing procedure produced a clean and consistent dataset with minimal loss of phonetic content. As previously noted in similar work, such normalization is essential to balance intelligibility and analytical tractability [19, 20]

2.3. Acoustic feature extraction

Three acoustic descriptors were computed from the cleaned dataset: duration, root mean square (RMS) energy, and zero-crossing rate (ZCR). These features are commonly used in phonetic and speech recognition research due to their low computational cost and strong discriminative power [21]. The duration of each recording was calculated in seconds using the librosa.get_duration() function, which accounts for non-silence regions only [18]. RMS energy was computed to quantify the average acoustic intensity level. It was calculated using the following expression:

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\chi_i^2} \tag{1}$$

where χ_i, denotes the i-th sample in the signal and N, is the total number of samples. Higher RMS values reflect stronger amplitude dynamics, which are particularly relevant when comparing signals captured by devices with different sensitivity profiles [22].

The ZCR was calculated using librosa.feature.zero_crossing_rate() and expressed in two forms: average value and normalized density per second. The latter, referred to as ZCR_sec, was obtained with the following relation:

$$ZCR_{seg} = \frac{C}{T} \tag{2}$$

where C, represents the number of zero crossings and T, the duration of the sample in seconds. This normalization removes temporal bias and is useful when comparing recordings of variable lengths [21, 23].

2.4. Statistical analysis

To validate the significance of the observed acoustic differences between microphone types, statistical analysis was performed using SPSS Statistics v25. Initial tests for normality were applied to the four measured variables: duration, RMS energy, ZCR, and ZCR_sec. Both Kolmogorov–Smirnov and Shapiro–Wilk tests were used. In all cases, the results were significant ($p < 0.05$), indicating non-normal data distributions, a pattern frequently encountered in speech datasets recorded outside professional studios [24]. Given the non-normality, the Mann–Whitney U test was selected to compare the two microphone groups.

This test is appropriate for detecting differences in central tendency between two independent samples without assuming a Gaussian distribution [25]. The confidence level was set at 95 percent for all comparisons. Alongside the inferential analysis, descriptive statistics (mean and standard deviation) were also computed for each variable across both microphone conditions. These values support a clearer interpretation of the patterns observed and offer insight into the magnitude of acoustic variation [9, 10, 21].

3. Results and discussion

3.1. Comparative analysis of acoustic features by microphone type

Figure 2 shows the mean RMS energy values of the 26 NATO phonetic alphabet letters, grouped by microphone type. Handheld microphones consistently registered higher acoustic intensity levels across nearly all phonemes, with the largest differences observed in FOXTROT, MIKE, and X-RAY. This confirms that handheld microphones capture broader amplitude variations in open environments, whereas throat microphones tend to attenuate these dynamics due to their internal coupling mechanism [9].

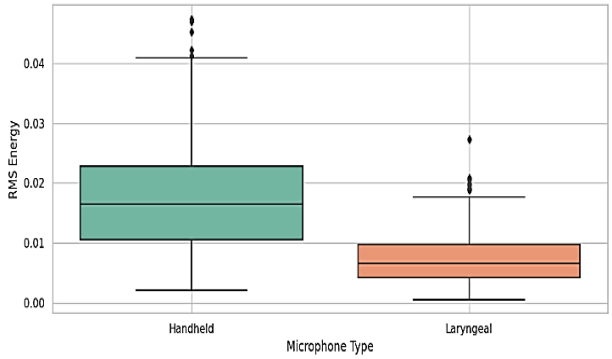

Figure 2. Mean RMS energy per letter by microphone type.

Table 2 illustrates the mean zero-crossing rate per letter. Throat microphones presented higher values, particularly in letters characterized by fricatives or transient consonants, such as SIERRA, WHISKEY, and VICTOR. These patterns suggest a higher spectral sensitivity of the throat device to rapid phonatory events, which may also introduce mechanical artifacts [21]. Specifically, the throat microphone is in direct physical contact with the skin, making it highly sensitive to high-frequency mechanical vibrations, which are typically associated with turbulent air flow during the production of fricatives (e.g., /s/ in SIERRA, /f/ in FOXTROT). These non-speech, high-frequency signals significantly increase the zero-crossing count, leading to the observed higher ZCR values compared to the air-conducted signal captured by the handheld microphone.

To visually aid the interpretation of the spectral differences presented in Table 2, the mean ZCR values for each NATO phonetic letter, grouped by microphone type, are visualized in Figure 3. This comparative bar chart highlights the systematic and pronounced increase in zero-crossing rate captured by the throat microphone across the entire alphabet, a pattern which is especially evident in letters containing highly turbulent phonemes.

Table 2. Zero-crossing rate (ZCR) per letter for each microphone.

LETTER	HANDHELD	THROAT
ALFA	973.4	1630.45
BRAVO	645.9	954.25
CHARLIE	936.45	1749.12
DELTA	767.58	1345.1
ECHO	952.35	1253.42
FOXTROT	1909.08	2787.32
GOLF	613.25	1152.83
HOTEL	867.35	1246.8
INDIA	704.38	1514.62
JULIETT	611.7	2815.25
KILO	508.25	1579.65
LIMA	604.33	1405.35
MIKE	848.45	1059.42
NOVEMBER	1069.55	1505.74
OSCAR	1654	2650.1
PAPA	559.38	742.4
QUEBEC	583.25	1573.55
ROMEO	689.8	1217.4
SIERRA	1665	2700.35
TANGO	656.2	992.9
UNIFORM	729.85	1644.92
VICTOR	642.9	2200.8
WHISKEY	1603.98	2534.68
X-RAY	2039.28	2785.5
YANKEE	790.15	1503.15
ZULU	447.52	1922.68

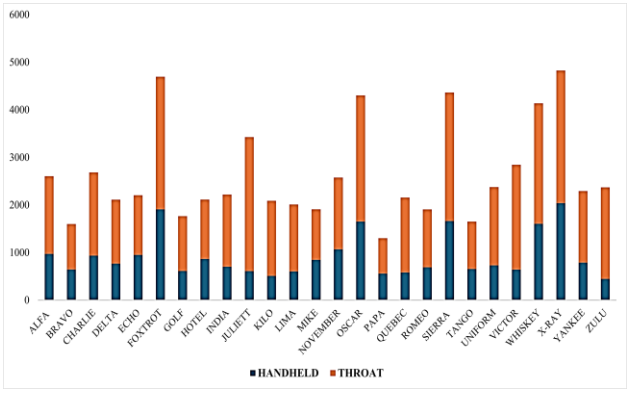

Figure 3. Comparative bar chart of mean zero-crossing rate (ZCR) per letter by microphone type.

The average duration of the recorded signals is shown in Figure 4. While differences were subtle, throat microphone samples tended to be slightly shorter, suggesting a more aggressive trimming of low-energy segments during preprocessing, especially in recordings of JULIETT and YANKEE. This effect has been previously described in studies involving voice activity detection and contact microphones [10].

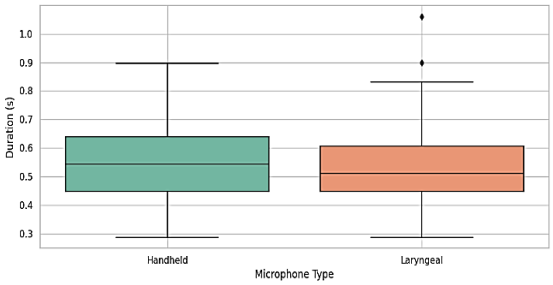

Figure 4. Average signal duration (s) for each NATO letter by microphone type.

3.2. Microphone-induced acoustic patterns per speaker

A cross-reference analysis of all four speakers revealed systematic patterns between microphone type and phoneme behavior. Table 3 summarizes the main phenomena observed for each individual.

Table 3. Dominant microphone-induced acoustic henomena observed per speaker.

Speaker	Microphone Effects	Most Affected Letter
Man1	High ZCR, high RMS, asymmetry	FOXTROT, WHISKEY, X-RAY.
Man2	Moderate variability, energy drop	SIERRA, VICTOR, JULIETT.
Woman1	High ZCR, compressed signals	ECHO, MIKE, ZULU, CHARLIE.
Woman2	Low energy, high spectral density	BRAVO, YANKEE, SIERRA, DELTA.

The throat microphone was particularly sensitive to inter-speaker variability, reflecting not only vocal tract characteristics but also differences in articulatory dynamics and microphone contact pressure [26, 27]. This individual variability is critical, as phonetic cue weighting and individual acoustic realizations directly influence ASR performance [28]. This supports the hypothesis that channel-specific and speaker-specific adaptations are necessary for robust ASR performance [29].

3.3. Statistical validation of microphone-induced differences

To support the acoustic observations with empirical evidence, a series of inferential statistical tests were conducted using SPSS v25, in accordance with best practices for experimental phonetics [30].

Initially, normality tests were performed on the main acoustic variables: duration, RMS energy, zero-crossing rate (ZCR), and zero crossings per second. As shown in Table 4, both Kolmogorov–Smirnov and Shapiro–Wilk tests yielded significant results ($p < 0.05$) for all variables, confirming the non-normal distribution of the data, a pattern typical of recordings obtained in semi-controlled environments [25].

Table 4. Normality test results for the analyzed variables.

Variable	Kolmogorov–Smirnov p	Shapiro–Wilk p
Sample Duration (s)	.000	.000
Signal Energy (RMS)	.000	.000
Zero Crossing Rate (ZCR)	.000	.000
ZCR Density (crossings/sec)	.000	.000

Given this outcome, the Mann–Whitney U test was selected to compare the two independent groups (handheld vs. throat). Table 5 presents the statistical outcomes, revealing highly significant differences ($p < 0.001$) across all features. These findings are consistent with earlier waveform and spectrogram analyses, reinforcing the robustness of the patterns detected [20, 25].

Table 5. Mann–Whitney U test results comparing microphone types.

Variable	U	Z	p (2-t)	Interp.
Duration (s)	490.521.00	-3.328	.001	Significant difference
Signal Energy (RMS)	135.710.00	-29.45	.000	Highly significant difference
Zero Crossing Rate (ZCR)	151.293.00	-28.27	.000	Highly significant difference
ZCR Density (crossings/s)	151.300.50	-28.26	.000	Highly significant difference

Finally, a descriptive summary of the acoustic values, presented in Table 6, confirms that handheld microphones produced signals with higher RMS energy and lower zero-crossing values, while throat microphones yielded denser ZCR patterns and slightly shorter durations. These trends are in line with previous findings on the interaction between microphone mechanics and speech signal properties [9, 21].

Table 6. Comparison of means and standard deviations of acoustic variables between microphone types.

Variable	Handheld Microphone	Throat Microphone
Duration (s) (mean ± SD)	0.55 ± 0.11	0.53 ± 0.11
RMS Energy (mean ± SD)	0.02 ± 0.01	0.01 ± 0.00
ZCR (mean ± SD)	0.10 ± 0.05	0.20 ± 0.07
ZCR Density (mean ± SD)	1660.09 ± 855.55	3138.15 ± 1102.94

3.4. Impact of Turkish accent on NATO phonetic alphabet articulation

Beyond hardware effects, phonetic deviations resulting from the Turkish accent were systematically identified, even

under standardized pronunciation conditions. The most notable modifications were found in JULIETT, YANKEE, and WHISKEY, where prosodic deviations and altered intonation patterns affected signal structure and energy profiles. These accent-related features, particularly temporal features like duration and prosodic timing, are known predictors of L2 proficiency and intelligibility [31] These observations align with previous findings [32], which emphasized how L1 (First Language) phonology influences L2 (Second Language) articulation at the suprasegmental level. Such accent-related variability poses challenges for ASR systems trained predominantly on native English phonetic patterns. As proposed in [33], the integration of multi-stream self-attention mechanisms can enhance the robustness of speech recognition systems, particularly in multilingual or operational contexts like military communication. It is acknowledged that the claims regarding 'prosodic deviations and altered intonation patterns' are primarily observational within the scope of this pilot study. Future dedicated research, employing quantitative measures such as fundamental frequency (F0) contours and spectrographic analysis, is necessary to fully validate and detail these L1 interference effects.

3.5. Implications for ASR development and microphone selection

The comparative results indicate that the choice of microphone has a direct influence on the spectral and temporal features relevant to phoneme recognition. While handheld microphones provide greater energy stability and broader amplitude resolution, their susceptibility to noise limits their effectiveness in uncontrolled environments. Throat microphones, on the other hand, preserve phoneme transitions and high-frequency detail but introduce distortion through physical coupling.

This duality must be considered in the design of ASR systems for tactical or noisy settings. Preprocessing techniques that compensate for amplitude compression or fricative amplification, as well as model calibration for each channel type, will be essential for maximizing recognition accuracy [20]. Research confirms that techniques like spectral mapping and enhancement algorithms (e.g., GMM or LPCC-based methods) can successfully recover lost high-frequency information and substantially improve the objective quality of TM recordings [4].

4. Conclusions

This study examined the influence of microphone type and speaker profile on the acoustic properties of speech signals, with a particular focus on their implications for automatic speech recognition (ASR) in multilingual and operational environments. The findings confirmed that these variables are not minor technical considerations, but central elements that directly impact the fidelity and usability of phonetic data in computational models.

Handheld microphones demonstrated superior performance in preserving key acoustic features. They consistently captured signals with higher energy levels, temporal stability, and phonetic richness, making them the more suitable option for generating training data in ASR models that rely on detailed acoustic input. Throat microphones, while advantageous in environments with high ambient noise or physical constraints (such as military, medical, or aeronautical contexts), introduced notable limitations. These include systematic attenuation of amplitude and irregular spectral behavior, which necessitate careful preprocessing and adaptation strategies to integrate them effectively into high-performance ASR systems [9, 10].

It is important to acknowledge that, as a pilot study, the use of only four native Turkish speakers limits the generalizability of the findings to the broader Turkish-speaking population. Future work should prioritize validating these results with a wider, more diverse corpus of speakers to fully confirm the systematic acoustic patterns observed.

The results also highlighted significant inter-speaker variability. Despite the use of a controlled reading protocol, differences were observed in articulation dynamics, energy patterns, and signal consistency. These differences underscore the need to implement speaker normalization and channel-specific calibration mechanisms, particularly in systems intended for broad deployment across diverse populations [29, 30].

Moreover, phonetic deviations influenced by the Turkish accent were clearly observed, especially in terms of duration, intensity, and prosodic structure. While no direct comparison was made with native English speakers, these deviations are consistent with previous findings on how L1 phonology interferes with L2 articulation patterns [32]. This reinforces the importance of incorporating accent adaptation techniques in ASR models, particularly in multilingual or operational contexts where reliance on monolingual training data may hinder recognition accuracy [34].

Looking forward, several research avenues are proposed. One involves applying state-of-the-art transcription models such as Whisper or Wav2Vec2 to the multichannel dataset to evaluate their performance under varying acoustic conditions using metrics like Word Error Rate (WER) and Character Error Rate (CER) [35]. Another promising direction is the development of microphone-adaptive models that treat the input channel as a controllable variable, allowing dynamic adjustments to architecture or preprocessing pipelines based on the recording source.

Future studies should also expand the analysis to include additional languages and accents already available in the current corpus (such as Arabic, Russian, French, and English), which may reveal deeper interactions between linguistic structure and microphone characteristics. Lastly, assessing these systems in real-world conditions (such as aircraft cockpits, emergency response scenarios, or field military operations) would provide valuable insights into their robustness under environmental and cognitive stressors. In summary, this study contributes to the understanding of how technical and linguistic variability influences phonetic signal quality and ASR performance. It highlights the need for more inclusive and adaptive design in speech recognition technologies, where variability is not treated as a limitation but as a fundamental component in building systems that are accurate, human-centered, and operationally reliable.

Acknowledgments

The author would like to express sincere gratitude to Dr. Öğr. Üyesi Selim Aras for his valuable academic guidance and support throughout the development of this study. His insightful suggestions and supervision played a crucial role in shaping the direction and quality of the research.

Conflict of Interest

The authors declared that there is no conflict of interest.

Ethical consideration

This study was conducted with the approval of the Ethics Committee for Social and Human Sciences Research of Ondokuz Mayıs University. The approval was granted on 29 November 2024 under the decision number 2024-1118. The research involved voice recordings and interviews carried out as part of a master's thesis titled "Machine Learning-Based Analysis and Resolution of Multilingual Pronunciation Issues in the NATO Phonetic Alphabet", under the supervision of Dr. Öğr. Üyesi Selim Aras.

Similarity index (Intihal.net): 6%

References

[1] I.C.A.O., Manual on the ICAO phonetic alphabet. ICAO Publishing, 2007.

[2] M.D. Keller, J.M. Ziriax, W. Barns, B. Sheffield, D. Brungart, T. Thomas, B. Jaeger, and K. Yankaskas, Performance in noise: impact of reduced speech intelligibility on sailor performance in a Navy command and control environment. Hearing Research, 349, 55–66, 2017. https://doi.org/10.1016/j.heares.2016.10.007.

[3] P. Boersma and D. Weenink, Praat: doing phonetics by computer. University of Amsterdam, 2023.

[4] M.A.T. Turan, Enhancement of throat microphone recordings using Gaussian mixture model probabilistic estimator. arXiv:1804.05937, 2018. https://doi.org/10.48550/arXiv.1804.05937

[5] B.E. Acker-Mills, A.J. Houtsma, and W.A. Ahroon, Speech intelligibility in noise using throat and acoustic microphones. Aviation, Space and Environmental Medicine, 77 (1), 26–31, 2006.

[6] E. Erzin, Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings. IEEE Transactions on Audio, Speech and Language Processing, 17 (7), 1316–1324, 2009.

[7] M.E. Arafat, I. Misra, and M.E. Hamid, A comparative study for throat microphone speech enhancement with different approaches. International Journal of Science and Research Archive, 13 (1), 850–859, 2024. https://doi.org/10.30574/ijsra.2024.13.1.1631

[8] X. Huang, A. Acero, and H.W. Hon, Spoken language processing: a guide to theory, algorithm, and system development. Pearson Prentice Hall, Upper Saddle River, 2001.

[9] Y. Kobayashi, K. Watanabe, and M. Akagi, Analysis and synthesis of speech captured by a contact microphone using neural vocoders. IEEE/ACM Transactions on Audio, Speech and Language Processing, 28, 2619–2632, 2020.

[10] T. Nguyen and S. Kim, An investigation of throat microphone signals for speech enhancement in noisy environments. Sensors, 21 (3), 987, 2021.

[11] R. Microphones, RØDE AI-1 Audio Interface – Specifications. 2022. Accessed 21 April 2025. https://www.rode.com/interfaces/ai-1

[12] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.E. Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, ESPnet: end-to-end speech processing toolkit. 2018.

[13] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N.E. Yalta Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, A comparative study on Transformer vs RNN in speech applications. arXiv:1909.06317, 2019. https://doi.org/10.48550/arXiv.1909.06317

[14] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D.W. Oard, and M. Post, The Multilingual TEDx Corpus for speech recognition and translation. arXiv:2102.01757, 2021. https://doi.org/10.48550/arXiv.2102.01757

[15] N. Wilkinson and T. Niesler, A hybrid CNN-BiLSTM voice activity detector. arXiv:2103.03529, 2021. https://doi.org/10.48550/arXiv.2103.03529

[16] J. Thomas, noisereduce: a Python package for real-time noise removal from speech signals. 2022. Accessed 2 May 2025. https://pypi.org/project/noisereduce/

[17] G. Ioannides and V. Rallis, Real-time speech enhancement using spectral subtraction with minimum statistics and spectral floor. arXiv:2302.10313, 2023. https://doi.org/10.48550/arXiv.2302.10313

[18] B. McFee, C. Raffel, D. Liang, D.P.W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, librosa: audio and music signal analysis in Python. Proceedings of the 14th Python in Science Conference, 18–25, 2015. https://doi.org/10.25080/Majora-7b98e3ed-003

[19] K.A. Abdalmalak and A. Gallardo-Antolín, Enhancement of a text-independent speaker verification system by using feature combination and parallel-structure classifiers. arXiv:2401.15018, 2024. https://doi.org/10.48550/arXiv.2401.15018

[20] S. Verma, R. Banerjee, and A. Singh, A comprehensive review of preprocessing strategies in speech recognition systems. ACM Transactions on Speech and Language Processing, 19 (1), 1–26, 2022.

[21] G. Degottex, P. Lanchantin, M.J. Gales, and S. King, A survey on acoustic representations for voice analysis and synthesis. IEEE/ACM Transactions on Audio, Speech and Language Processing, 29, 116–137, 2021.

[22] J.M. Valin and J. Skoglund, A real-time wideband neural vocoder at 1.6 kb/s using LPCNet. Proceedings of Interspeech, 3406–3410, 2019. https://doi.org/10.21437/Interspeech.2019-1255

[23] A. Saeed, D. Grangier, and N. Zeghidour, Contrastive learning of general-purpose audio representations. arXiv:2010.10915, 2021. https://doi.org/10.48550/arXiv.2010.10915

[24] W. Wang, J. Yi, M. Wu, and X. Lei, Improving ASR robustness via uncertainty modeling and consistency training. Proceedings of Interspeech, 2022.
[25] Laerd Statistics, Mann–Whitney U Test using SPSS Statistics. 2021. Accessed 2 May 2025. https://statistics.laerd.com/spss-tutorials/mann-whitne y-u-test-using-spss-statistics.php
[26] Y. Zhao, C. Ni, C.C. Leung, S. Joty, E.S. Chng, and B. Ma, A unified speaker adaptation approach for ASR. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 9398–9410, 2021. https://doi.org/10.18653/v1/2021.emnlp-main.737
[27] Y. Zhang, C. Wu, and Y. Wang, Acoustic environment and recording conditions for voice applications: a comparative study. IEEE Access, 8, 213187–213200, 2020.
[28] J.X. Ou, X. Ming, and A.C.L. Yu, Individual variability in subcortical neural encoding shapes phonetic cue weighting. Scientific Reports, 13 (1), 9991, 2023. https://doi.org/10.1038/s41598-023-37212-y
[29] D. Prabhu, P. Jyothi, S. Ganapathy, and V. Unni, Accented speech recognition with accent-specific codebooks. arXiv:2310.15970, 2023. https://doi.org/10.48550/arXiv.2310.15970
[30] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, Montreal Forced Aligner: trainable text–speech alignment using Kaldi. Proceedings of Interspeech 2017, 498–502, 2017. https://doi.org/10.21437/Interspeech.2017-1386
[31] H. Kallio, S. Antti, and J. Šimko, Fluency-related temporal features and syllable prominence as prosodic proficiency predictors for learners of English with different language backgrounds. Language and Speech, 65 (3), 571–597, 2022. https://doi.org/10.1177/00238309211040175
[32] O. Türk and H. Açıkgoz, Phonological influence of Turkish on English pronunciation: a comparative spectrographic study. Journal of Phonetics and Speech Sciences, 14, 157–170, 2022.
[33] K.J. Han, R. Prieto, K. Wu, and T. Ma, State-of-the-art speech recognition using multi-stream self-attention with dilated 1D convolutions. arXiv:1910.00716, 2019. https://doi.org/10.48550/arXiv.1910.00716
[34] K. Tomanek, V. Zayats, D. Padfield, K. Vaillancourt, and F. Biadsy, Residual adapters for parameter-efficient ASR adaptation to atypical and accented speech. arXiv:2109.06952, 2021. https://doi.org/10.48550/arXiv.2109.06952
[35] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, Robust speech recognition via large-scale weak supervision. arXiv:2212.04356, 2023. https://doi.org/10.48550/arXiv.2212.04356