OMÜ Mühendislik Bilimleri ve Teknolojisi Dergisi,
OMU Journal of Engineering Sciences and Technology

e-ISSN: 2791-8858 OMUJEST,Eylül 2025, 5(2): 73-91

Automatic Speech Recognition from Throat Microphone Signals of the NATO Phonetic Alphabet Spoken by Turkish Speakers: Evaluation of Classical and Deep Techniques under Push-to-Talk Conditions

Türk Konuşurlar Tarafından Söylenen NATO Fonetik Alfabesine Ait Gırtlak Mikrofonu Sinyallerinden Otomatik Konuşma Tanıma: Push-to-Talk Koşulları Altında Klasik ve Derin Öğrenme Tekniklerinin Değerlendirilmesi

Julio Cesar VELAZQUEZ GARCIA[1], Selim ARAS[2]

[1]Department of Intelligent Systems Engineering, Graduate School of Natural and Applied Sciences, Ondokuz Mayıs University, Samsun, Türkiye
· **julio.velazquez18@gmail.com** · ORCİD > 0009-0007-2270-8068

[2]Department of Electrical and Electronics Engineering, Faculty of Engineering, Ondokuz Mayıs University, Samsun, Türkiye
· **selim.aras@omu.edu.tr** · ORCİD > 0000-0003-1231-5782

Makale Bilgisi/Article Information

Makale Türü/Article Types: Araştırma Makalesi/Research Article
Geliş Tarihi/Received: 29 Temmuz/July 2025
Kabul Tarihi/Accepted: 25 Eylül/September 2025
Yıl/Year: 2025 | **Cilt–Volume:** 5 | **Sayı–Issue:** 2 | **Sayfa/Pages:** 73-91

Atıf/Cite as: Velazquez Garcia, J.C., Aras, S. "Automatic Speech Recognition from Throat Microphone Signals of the NATO Phonetic Alphabet Spoken by Turkish Speakers: Evaluation of Classical and Deep Techniques under Push-to-Talk Conditions" Ondokuz Mayis University Journal of Engineering Sciences and Technology 5(2), Eylül 2025: 73-91.

Sorumlu Yazar/Corresponding Author: Julio Cesar VELAZQUEZ GARCIA

AUTOMATIC SPEECH RECOGNITION FROM THROAT MICROPHONE SIGNALS OF THE NATO PHONETIC ALPHABET SPOKEN BY TURKISH SPEAKERS: EVALUATION OF CLASSICAL AND DEEP TECHNIQUES UNDER PUSH-TO-TALK CONDITIONS

ABSTRACT

This study evaluates the performance of various Automatic Speech Recognition (ASR) techniques applied exclusively to recordings captured using throat microphones (TM). The objective is to explore their applicability in Push-to-Talk (PTT) operational conditions, where traditional air microphones are limited by environmental noise. A corpus was constructed with 10 native Turkish speakers enunciating the NATO phonetic alphabet. Signals were segmented using Silero VAD, resampled to 16 kHz, and augmented to robustify the models against noise and variations. Two feature extraction approaches were employed: Mel-frequency cepstral coefficients (MFCCs) and Wav2Vec2 embeddings reduced by Principal Component Analysis (PCA). Subsequently, five supervised classifiers were trained and compared: SVM, RF, KNN, MLP, and LightGBM. Evaluation metrics included overall accuracy and Word Error Rate (WER). Results demonstrate the technical feasibility of ASR with laryngeal signals, identifying the combination of LightGBM with MFCCs as the most robust (86.38% accuracy, 0.000 WER) and confirming the potential of RF with MFCCs (84.62% accuracy, 0.000 WER). This work establishes an experimental foundation for the development of robust and low-cost ASR systems in noisy environments. In this context, throat microphones offer a crucial alternative.

Keywords: Automatic Speech Recognition (ASR), Machine Learning, NATO Phonetic Alphabet, Push-to-Talk (PTT), Throat Microphone.

❁❁❁

TÜRK KONUŞURLAR TARAFINDAN SÖYLENEN NATO FONETİK ALFABESİNE AİT GIRTLAK MİKROFONU SİNYALLERİNDEN OTOMATİK KONUŞMA TANIMA: PUSH-TO-TALK KOŞULLARI ALTINDA KLASİK VE DERİN ÖĞRENME TEKNİKLERİNİN DEĞERLENDİRİLMESİ

ÖZ

Bu çalışma, yalnızca gırtlak mikrofonları (TM) kullanılarak kaydedilen ses sinyallerine uygulanan çeşitli Otomatik Konuşma Tanıma (ASR) tekniklerinin performansını değerlendirmektedir. Amaç, geleneksel hava mikrofonlarının çevresel gürültü nedeniyle sınırlı kaldığı Push-to-Talk (PTT) operasyonel koşullarında bu tekniklerin uygulanabilirliğini araştırmaktır. On ana dili Türkçe olan konuşmacının NATO fonetik alfabesini telaffuz ettiği bir veri kümesi oluşturulmuştur. Sinyaller Silero VAD kullanılarak bölümlendirilmiş, 16 kHz'e yeniden örneklenmiş ve modellerin gürültü ve varyasyonlara karşı dayanıklılığını artırmak amacıyla veri artırma teknikleri uygulanmıştır. Özellik çıkarımı için iki farklı yaklaşım kullanılmıştır: Mel frekans kepstrum katsayıları (MFCC) ve Temel bileşen analizi (PCA) ile boyutu azaltılmış Wav2Vec2 gömlemeleri. Daha sonra beş denetimli sınıflandırıcı eğitilmiş ve karşılaştırılmıştır: SVM, RF, KNN, MLP ve LightGBM. Değerlendirme metrikleri arasında genel doğruluk ve Kelime Hata Oranı (WER) yer almaktadır. Sonuçlar, gırtlak sinyalleriyle ASR sistemlerinin teknik olarak uygulanabilir olduğunu göstermekte; LightGBM ile MFCC kombinasyonunun en sağlam yapı olduğunu (%86,38 doğruluk, 0.000 WER) ve MFCC ile RF kullanımının da önemli bir potansiyel sunduğunu (%84,62 doğruluk, 0.000 WER) ortaya koymaktadır. Gürültülü ortamlarda kullanılabilecek dayanıklı ve düşük maliyetli ASR sistemlerinin geliştirilmesi için deneysel bir temel oluşturmaktadır. Bu bağlamda gırtlak mikrofonları önemli bir alternatif sunmaktadır.

Anahtar Kelimeler: Otomatik Konuşma Tanıma (ASR), Makine Öğrenmesi, NATO Fonetik Alfabesi, Push-to-Talk (PTT), Gırtlak Mikrofonu.

❁❁❁

Highlights

- ASR from throat mics evaluated in PTT conditions
- NATO phonetic alphabet corpus used with Turkish speakers
- LightGBM+MFCCs: highest accuracy (86.38%), WER 0.000
- Establishes basis for robust, low-cost ASR in noise
- TM offer crucial noisy environment solution

1. INTRODUCTION

Automatic Speech Recognition (ASR) has become an essential tool in various applications, ranging from voice control systems to assistance in complex operational environments. However, most traditional ASR studies rely on recordings obtained from air microphones, which are highly sensitive to ambient noise, reverberations, and physical occlusions [1, 2]. This vulnerability limits their effectiveness in situations where speech intelligibility is critical, such as in military, industrial, or healthcare contexts.

In this context, TM (also known as contact microphones or laryngophones) represent a robust alternative. These devices capture vibrations directly from the larynx, allowing an intelligible signal to be maintained even in acoustically adverse conditions [3]. Despite their potential, the use of TM has been relatively underexplored in practical ASR applications, partly due to the attenuated spectral characteristics of their signals [4], which poses additional challenges for designing effective recognition systems.

To contribute to this field, the present study focuses on a comprehensive comparative evaluation of ASR using laryngeal signals. For this purpose, a specific database was constructed where ten native Turkish speakers (five women and five men) enunciated the NATO phonetic alphabet (from Alpha to Zulu). This alphabet is a standardized system of keywords used to represent Latin alphabet letters, designed to avoid ambiguities during oral communication in high-risk or noisy contexts, such as military, aeronautical, or maritime telecommunications [5]. Its controlled structure and universal use make it a phonetically balanced and relevant corpus for speech recognition tests in noisy environments or with bandwidth restrictions. Recordings were exclusively made with a throat microphone, and the data underwent rigorous preprocessing, including voice activity detection using Silero VAD, normalization, and resampling to 16 kHz [6].

Additionally, data augmentation techniques were implemented to simulate real PTT operational conditions. These techniques include the addition of radio-type noise, simulated reverberation, pitch shifting, and electromagnetic distortion [7, 8]. Data augmentation is crucial for improving model robustness and compensating for the limited phonetic variability of small datasets [9].

Regarding feature extraction, two complementary approaches were used: Mel-frequency cepstral coefficients (MFCCs), widely recognized for their effectiveness in phonetic recognition tasks [10]; and embeddings derived from the pre-trained Wav2Vec 2.0 (Wav2Vec2) model, a self-supervised transformer-based framework for speech representation learning that has demonstrated outstanding results even in low-resource conditions [11, 12]. Various supervised classifiers were applied to these feature sets, including Support Vector Machines (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Multilayer Perceptrons (MLP), and Light Gradient Boosting Machine (LightGBM), with the aim of evaluating their comparative performance in realistic PTT scenarios. Model evaluation was carried out using metrics such as overall accuracy and WER, which allow for the analysis of both the quantitative and qualitative performance of the classifiers [13, 14].

The main objective of the present study is to analyze the feasibility of ASR from laryngeal signals in operational contexts, providing experimental evidence on the effectiveness of different combinations of extraction and classification techniques.

This contribution is relevant for the development of robust solutions in noisy environments, where conventional microphones present significant limitations.

The structure of this article is organized as follows: Section 2, “Comparison with Related Work,” provides a comprehensive review of the existing literature, highlighting the unique contributions of our study. Section 3, “Materials and Methods,” details the corpus construction, audio preprocessing, data augmentation techniques, acoustic feature extraction methods, supervised classification algorithms used, and evaluation metrics. Section 4, “Results,” presents our experimental findings, including the overall comparison of model performance, the analysis of explained variance by Principal Component Analysis (PCA), and detailed class-wise evaluation. Section 5, “Discussion,” interprets these results in a broader context, addresses study limitations, and proposes future research directions. Finally, Section 6, “Conclusions,” summarizes the main contributions of this work.

2. COMPARİSON WITH RELATED WORK

The growing interest in ASR with TM in noisy environments (military, industrial, operational) drives research in this field. This section discusses key studies employing TM and similar technologies, analyzing their methodologies, corpora, features, models, and application contexts to contextualize and highlight the contribution of our work.

2.1. Analysis of Relevant Studies

Masuda et al. (2020) proposed an ASR approach using TM in Japanese, focusing on fine-tuning Wav2Vec2 with pseudo-laryngeal data generated through feature mapping from air voice. Unlike our study, they do not employ MFCCs or classical classifiers. Our approach is distinguished by a modular architecture for extracting pre-trained embeddings (reduced with PCA) without costly fine-tuning, and by providing complementary metrics such as class-wise accuracy and WER [15].

The Vibravox corpus (Hauret et al., 2024) is a large-scale French dataset (188 speakers, 45 h/sensor) with multiple body-conducted sensors, including laryngophones. Although its scope is broad (ASR, speech enhancement, speaker verification), it does not focus on command classification or the phonetic alphabet, nor does it perform a detailed acoustic analysis between representations like MFCCs and SSL embeddings. Our work, in contrast, focuses exclusively on TM, comparatively exploring acoustic representation techniques and supervised classifiers [16].

The MoveOn project study (2008) empirically evaluated the superiority of TMs on motorcycles compared to lavalier microphones in real noise environments. This work, although not detailing signal processing techniques or classification models, supports the motivation of our study on the feasibility of TM-based ASR in PTT conditions, providing practical evidence of TM robustness in adverse environments [17].

Schultz and Jou (2004) addressed soft whisper recognition with TMs, significantly improving WER (from 99.3% to 32.9%) through acoustic adaptation (MLLR, FMLLR) on GMM-HMM models. While the speech domain is different, this work highlights the need to adapt models for the laryngeal signal. Our approach, on the other hand, is based on the selection of robust representations (MFCCs vs. embeddings) and classical classifiers, which facilitates practical implementation without extensive retraining [18].

Kim et al. (2025) introduce the TAPS corpus, with paired TM and acoustic microphone recordings in Korean. Their study focuses on TM signal enhancement using deep mapping models, demonstrating that deep learning can restore degraded information in TMs. While not directly addressing ASR, their work complements ours by emphasizing the need for adapted processing to overcome the inherent spectral loss in TMs [19].

2.2. Comparative Synthesis

Table 1. Comparison of relevant studies in ASR with throat microphones

Study	Microphone	Corpus	Representations	Models	Metrics	Application / Environment
This Work	Throat mic	10 speakers, Turkish, NATO	MFCCs, Wav2Vec2 + PCA	SVM, RF, KNN, MLP, LightGBM	Accuracy, WER	PTT, controlled noise
Masuda et al. (2022)	Throat mic	Japanese (pseudo-TM)	Wav2Vec2 + feature mapping	Wav2Vec2 fine-tuned	CER	High noise, scarce data
Hauret et al. (2024)	Throat mic, others	188 speakers, French	SSL, multiple sensors	ASR + verification + enhancement	PER, various rates	Simulated 3D environments
Move On (2008)	Throat mic	Motorcycle commands	Not specified	Not specified	Empirical	Real noise (traffic)
Schultz & Jou (2004)	Throat mic	Whispers, English	GMM-HMM adaptation	MLLR, FMLLR, retrain	WER	Low-energy voice
Kim et al. (2025)	Throat mic	60 speakers, Korean	Paired signal mapping	Deep learning (mapping)	Signal quality	Factories, subways

2.3. Differentiating Contribution

Compared to the existing literature, this study is distinguished by several key contributions. Firstly, an original corpus in Turkish is introduced, comprising recordings of the NATO phonetic alphabet, captured exclusively with a throat microphone. Furthermore, our work explicitly evaluates and compares two distinct acoustic representations, MFCCs and Wav2Vec2 embeddings, complemented by a dimensionality reduction technique such as PCA. A third differentiating point is the systematic comparison performed among five supervised classifiers, including both traditional models and tree-based approaches. Finally, this study offers a comprehensive insight into classification performance under PTT operational conditions by reporting not only overall accuracy but also WER and a detailed class-wise analysis using confusion matrices.

3. MATERIALS AND METHODS

3.1. Corpus Recording

An audio database was constructed exclusively using a throat microphone to analyze robust voice signals against environmental noise. Recordings were made on a computer using Audacity. Ten native Turkish speakers (5 men, 5 women) participated, each pronouncing the NATO phonetic alphabet (from Alpha to Zulu). Each word was repeated 10 times, generating an initial database of 2,600 original voice samples.

3.1.1. File Naming Convention

Audio files were named following the VV-WW-XX-YY-ZZ format, where:

- **VV:** Language code (“03” for Turkish).
- **WW:** Speaker identifier (01 to 10).
- **XX:** Microphone type (“01” for throat microphone).
- **YY:** NATO phonetic alphabet letter numerically encoded.
- **ZZ:** Repetition number (01 to 10).

This scheme facilitated automated processing and metadata traceability.

3.2. Audio Preprocessing and Segmentation

Audio files were processed using the Silero VAD (Voice Activity Detection) model, integrated via torch.hub, to remove silent fragments and extract only the voice segments. An additional margin of 100 milliseconds was applied before and after each relevant segment. All audio files were resampled to 16 kHz and stored in WAV format. Files without voice detection were discarded.

3.2.1. Data Augmentation

To improve model generalization and simulate real PTT operational conditions, data augmentation techniques were applied to the trimmed samples. Four augmented versions were generated for each original sample:

- Radio-type noise (radio): Addition of white Gaussian noise.
- Simulated reverberation (bunker): Convolution with an artificial impulse response.
- Upward pitch shift (accented): Pitch shifting by +2 semitones using librosa.effects.pitch_shift.
- Electromagnetic interference (electro): Random sample suppression.

Before augmentation, a band-pass filter with cutoff frequencies at 300 Hz and 3,400 Hz was used to simulate typical telephone line conditions. All resulting files were normalized. The augmentation process quintupled the original dataset size, generating a total of 13,000 files.

3.3. Acoustic Feature Extraction

Two main feature extraction approaches were employed:

Mel-frequency Cepstral Coefficients (MFCCs): For each audio sample, 13 static MFCCs were extracted. A 25 ms window with a 10 ms hop size was applied. Subsequently, temporal mean pooling was performed and concatenated with the temporal standard deviation, resulting in a 26-dimensional feature vector for each sample.

Wav2Vec2 Embeddings: Audio embeddings were generated by the pre-trained facebook/wav2vec2-large-xlsr-53 model. This self-supervised model was implemented using Hugging Face's Transformers library. For each audio, high-dimensional representations (1024 dimensions per time frame) were obtained. By applying temporal mean pooling and concatenating the mean and standard deviation of these hidden states, a 2048-dimensional feature vector was obtained for each audio sample. To reduce this dimensionality, PCA was applied. 256 principal components were retained, which explained 99.50% of the total data variance, allowing the most relevant information to be preserved with a lower computational load for subsequent classifiers.

3.4. Supervised Classifiers

Five supervised classification algorithms were evaluated, applied to both MFCC features and PCA-reduced Wav2Vec2 embeddings. Each classifier was trained exclusively with throat microphone data (original and augmented) and evaluated on independent samples to ensure external validity.

3.5. Model Configuration and Hyperparameters

Hyperparameter tuning was deliberately limited to prioritize an equitable comparison. Configurations based on validated practices and default library recommendations were employed, avoiding exhaustive optimization to maintain a uniform experimental framework. Table 2 summarizes the exact configurations of the main hyperparameters used.

Table 2. Hyperparameter configurations used in supervised models

Classifier	Feature Source	Main Hyperparameters	Implementation (Library)
SVM	MFCC / Wav2Vec2+PCA	kernel='rbf', C=10, gamma='scale', random_state=42	scikit-learn
RF	MFCC / Wav2Vec2+PCA	n_estimators=200, max_depth=25, random_state=42	scikit-learn
KNN	MFCC / Wav2Vec2+PCA	n_neighbors=5, metric='euclidean'	scikit-learn
MLP	MFCC / Wav2Vec2+PCA	hidden_layer_sizes=(100,), max_iter=500, activation='relu', solver='adam', random_state=42	scikit-learn
LightGBM	MFCC	n_estimators=300, max_depth=25, learning_rate=0.1, objective='multiclass', num_class=26, random_state=42	lightgbm
LightGBM	Wav2Vec2 + PCA	n_estimators=500, learning_rate=0.05, objective='multiclass', num_class=26, random_state=42	lightgbm

3.6. Evaluation Metrics

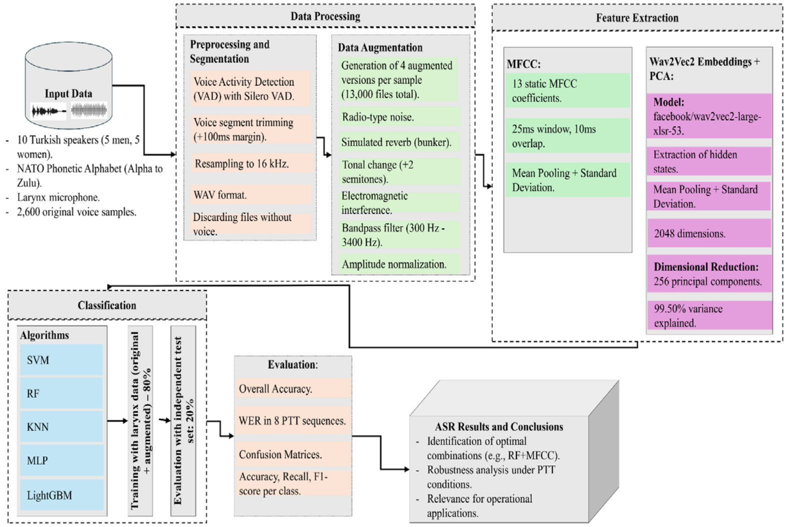
Performance evaluation was carried out using quantitative and qualitative metrics. Overall accuracy was defined as the proportion of correct predictions. For a more detailed class-wise analysis, precision, recall, and f1-score were also employed.

To evaluate performance in realistic contexts, the WER was used, calculated as:

$$WER = \frac{S + D + I}{N} \qquad (1)$$

Where S represents the number of substitutions, D deletions, I insertions, and N the total number of units (letters) in the reference sequence. WER was adapted to the recognition of complete sequences of NATO phonetic alphabet letters under PTT conditions. Eight arbitrary sequences of four letters each were defined, selected from individual throat microphone recordings. Predictions were concatenated and compared with the expected sequence using the wer() function from the jiwer library. A WER of 0.0 indicates a completely correct transcription.

Finally, confusion matrices were generated for each experimental configuration, allowing visualization of error patterns and systematic confusions between phonetically similar letters. Figure 1 illustrates the evaluation methodology flowchart.



Input Data
- 10 Turkish speakers (5 men, 5 women).
- NATO Phonetic Alphabet (Alpha to Zulu).
- Larynx microphone.
- 2,600 original voice samples.
Data Processing
Preprocessing and Segmentation
Voice Activity Detection (VAD) with Silero VAD.
Voice segment trimming (+100ms margin).
Resampling to 16 kHz.
WAV format.
Discarding files without voice.
Data Augmentation
Generation of 4 augmented versions per sample (13,000 files total).
Radio-type noise.
Simulated reverb (bunker).
Tonal change (+2 semitones).
Electromagnetic interference.
Bandpass filter (300 Hz - 3400 Hz).
Amplitude normalization.
Feature Extraction
MFCC:
13 static MFCC coefficients.
25ms window, 10ms overlap.
Mean Pooling + Standard Deviation.
Wav2Vec2 Embeddings + PCA:
Model: facebook/wav2vec2-large-xlsr-53.
Extraction of hidden states.
Mean Pooling + Standard Deviation.
2048 dimensions.
Dimensional Reduction: 256 principal components.
99.50% variance explained.
Classification
Algorithms
SVM
RF
KNN
MLP
LightGBM
Training with larynx data (original + augmented) – 80%
Evaluation with independent test set: 20%
Evaluation:
Overall Accuracy.
WER in 8 PTT sequences.
Confusion Matrices.
Accuracy, Recall, F1-score per class.
ASR Results and Conclusions
- Identification of optimal combinations (e.g., RF+MFCC).
- Robustness analysis under PTT conditions.
- Relevance for operational applications.

Figure 1. Flowchart of the Methodology for Automatic Speech Recognition from Throat Microphone Signals.

4. RESULTS

4.1. Experimental Configurations and Evaluation Framework

The evaluation dataset comprised 13,000 audio files, derived from 2,600 original samples of the NATO phonetic alphabet spoken by Turkish speakers, along with four augmented variants for each sample. The augmentation techniques (white noise, reverberation, pitch shifting, electromagnetic distortion) emulated real-world PTT conditions in noisy or degraded environments.

A common test set, equivalent to 20% of the total data, including both original and augmented samples, was used. Recognition was performed across 26 classes, corresponding to the letters of the NATO phonetic alphabet. Overall accuracy and average WER were used for evaluation. It is important to note that exhaustive hyperparameter optimization (e.g., GridSearch) was not applied; standard configurations were maintained to ensure comparative homogeneity.

4.2. Overall Performance Comparison

Table 3 presents an overall comparison of the performance of the classification models trained with throat microphone signals.

Table 3. Accuracy and WER per model with throat microphone signals.

Model	Features	Accuracy (%)	WER	Model	Features	Accuracy (%)	WER
LightGBM	MFCC	86.38	0.000	MLP	MFCC	67.19	0.188
RF	MFCC	84.62	0.000	KNN	MFCC	58.5	0.281
LightGBM	Wav2Vec2 + PCA	75.54	0.063	SVM	MFCC	52.65	0.375

The LightGBM model with MFCC features achieved the best overall performance, with an accuracy of 86.38% and a WER of zero. This result highlights its generalization capability and robustness against significant acoustic variations, positioning it as the most effective alternative for real-world applications. The RF model, also with MFCCs, closely followed (84.62% accuracy, WER = 0.000), confirming that tree-based ensemble methods offer high noise tolerance in laryngeal signals.

Figure 2 illustrates the comparative performance of each model through key metrics such as overall accuracy, macro precision, macro recall, and macro F1-score, complementing the information presented in Table 3.

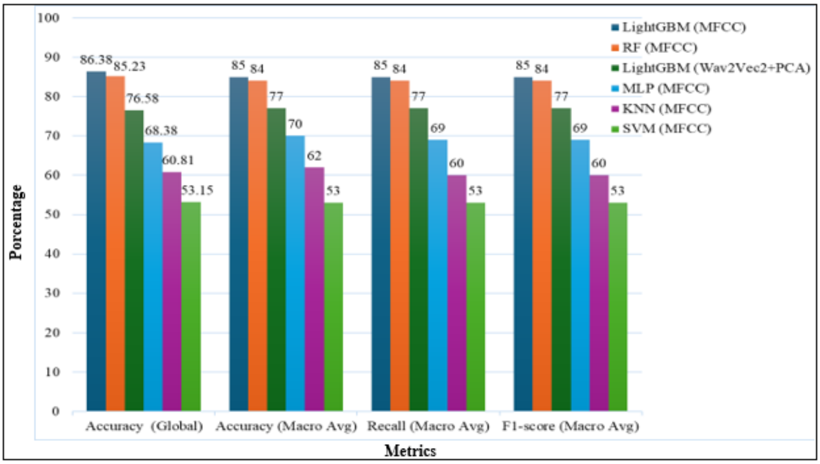


Figure 2. Comparative Performance of Models by Metric (Overall Accuracy, Macro Precision, Macro Recall, Macro F1-score) for Throat Microphone Signals.

The LightGBM configuration using Wav2Vec2 embeddings and PCA reduction achieved an accuracy of 75.54% and a WER of 0.063. Although slightly lower than the models with MFCCs, this result demonstrates the utility of self-supervised embeddings in degraded signal contexts.

The MLP model with MFCCs showed intermediate performance (67.19% accuracy, average WER = 0.188), offering a balance between simplicity and generalization capability. Conversely, distance-based (KNN) and margin-based (SVM) models showed significantly inferior performance, with accuracies of 58.50% and 52.65%, respectively. Both exhibited high Word Error Rates (average WER of 0.281 for KNN and 0.375 for SVM), indicating greater sensitivity to spectral overlap and noise in the laryngeal channel.

Figure 3 shows the cumulative percentage of variance explained by the principal components extracted using PCA. It is observed that the first few components explain a substantial proportion of the total variance. In particular, the first 256 components reached a threshold of 99.50% cumulative variance, which justifies the dimensionality reduction from 2048 to 256 without significant information loss for training the subsequent classifiers.

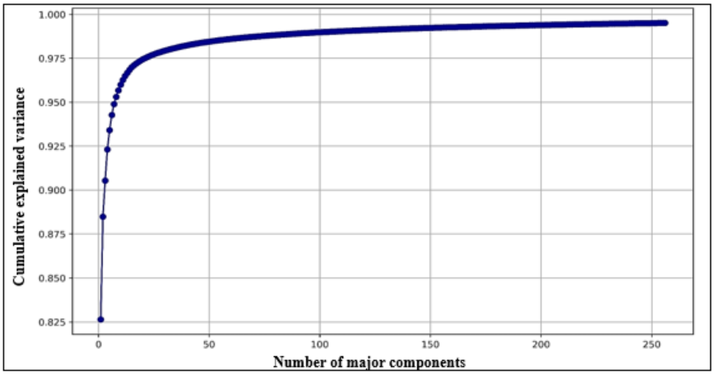
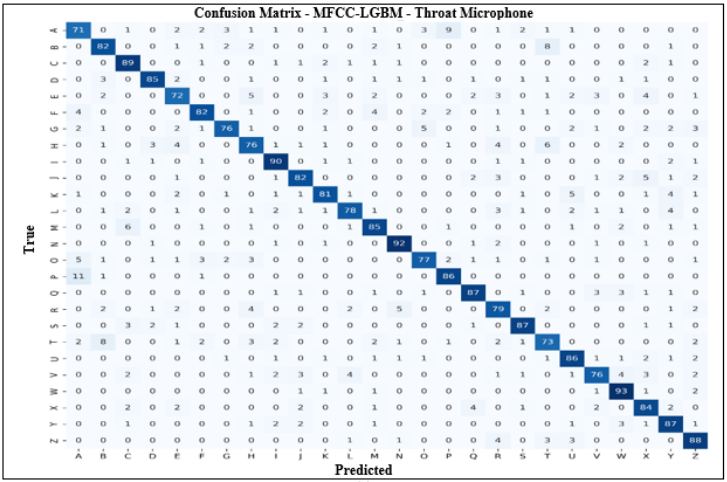


Figure 3. Cumulative Explained Variance by PCA.

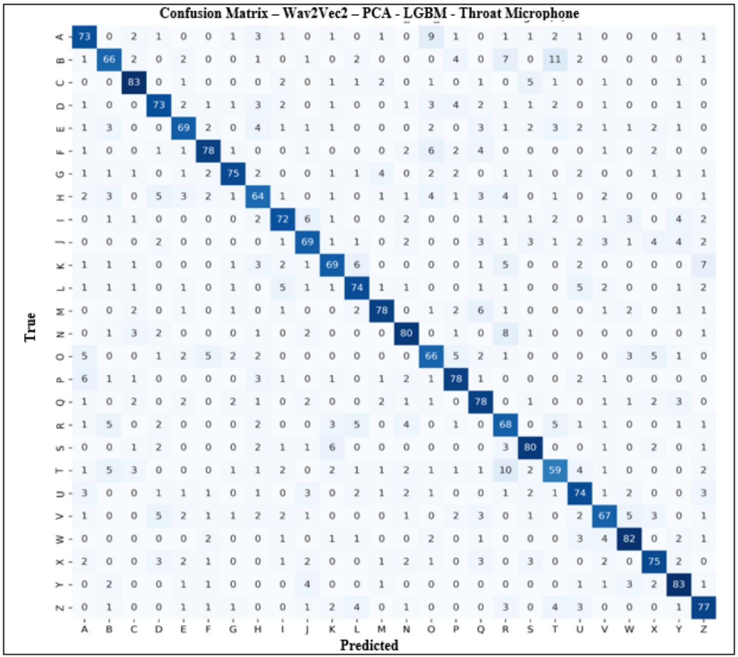
4.3. Class-wise Classification Evaluation (Matrices and Metrics)

For a deeper analysis of classifier behavior, the confusion matrices obtained with LightGBM trained with MFCC features (Figure 4a) and with Wav2Vec2 embeddings (Figure 4b) are representative. Both configurations were evaluated on the 13,000-sample test set under PTT conditions. Visual analysis of the matrices reveals a predominance of correct classifications along the main diagonal, indicating adequate discriminatory capability by both models. However, recurring error patterns were also identified.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	71	0	1	0	2	2	3	1	1	0	1	0	1	0	3	9	0	1	2	1	1	0	0	0	0	0
B	0	82	0	0	1	1	2	2	0	0	0	0	2	1	0	0	0	0	0	8	0	0	0	0	1	0
C	0	0	89	0	0	1	0	0	1	1	2	1	1	1	0	0	0	0	0	0	0	0	0	2	1	0
D	0	3	0	85	2	0	0	1	0	0	1	0	1	1	1	0	1	0	1	1	0	0	1	1	0	0
E	0	2	0	0	72	0	0	5	0	0	3	0	2	0	0	0	2	3	0	1	2	3	0	4	0	1
F	4	0	0	0	0	82	0	1	0	0	2	0	4	0	2	2	0	1	1	1	0	0	0	0	0	0
G	2	1	0	0	2	1	76	1	0	0	1	0	0	0	5	0	0	1	0	0	2	1	0	2	2	3
H	0	1	0	3	4	0	0	76	1	1	1	0	0	0	0	1	0	4	0	6	0	0	2	0	0	0
I	0	0	1	1	0	1	0	0	90	0	1	1	0	0	0	0	0	1	1	0	0	0	0	0	2	1
J	0	0	0	0	1	0	0	0	1	82	0	0	0	0	0	0	2	3	0	0	0	1	2	5	1	2
K	1	0	0	0	2	0	1	0	1	1	81	1	0	0	0	0	0	0	1	0	5	0	0	1	4	1
L	0	1	2	0	1	0	0	1	2	1	1	78	1	0	0	0	0	3	1	0	2	1	1	0	4	0
M	0	0	6	0	0	1	0	1	0	0	0	1	85	0	0	1	0	0	0	0	1	0	2	0	1	1
N	0	0	0	1	0	0	0	0	1	0	1	0	0	92	0	0	1	2	0	0	0	1	0	1	0	0
O	5	1	0	1	1	3	2	3	0	0	0	0	0	0	77	2	1	1	0	1	0	1	0	0	0	1
P	11	1	0	0	0	1	0	0	0	0	0	0	0	0	0	86	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	1	1	0	0	1	0	1	0	87	0	1	0	0	3	3	1	1	0
R	0	2	0	1	2	0	0	4	0	0	0	2	0	5	0	0	0	79	0	2	0	0	0	0	1	2
S	0	0	3	2	1	0	0	0	2	2	0	0	0	0	0	0	1	0	87	0	0	0	0	1	1	0
T	2	8	0	0	1	2	0	3	2	0	0	0	2	1	0	1	0	2	1	73	0	0	0	0	0	2
U	0	0	0	0	0	0	1	0	1	0	1	0	1	1	1	0	0	0	0	1	86	1	1	2	1	2
V	0	0	2	0	0	0	0	1	2	3	0	4	0	0	0	0	0	1	1	0	1	76	4	3	0	2
W	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	1	93	1	0	2
X	0	0	2	0	2	0	0	0	0	2	0	0	1	0	0	0	4	0	1	0	0	2	0	84	2	0
Y	0	0	1	0	0	0	0	1	2	2	0	0	1	0	0	0	0	0	0	0	1	0	3	1	87	1
Z	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	4	0	3	3	0	0	0	0	88

(a)



Confusion Matrix – Wav2Vec2 – PCA - LGBM - Throat Microphone

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	73	0	2	1	0	0	1	3	1	0	1	0	1	0	9	1	0	1	1	2	1	0	0	0	1	1
B	1	66	2	0	2	0	0	1	0	1	0	2	0	0	0	4	0	7	0	11	2	0	0	0	0	1
C	0	0	83	0	1	0	0	0	2	0	1	1	2	0	1	0	1	0	5	1	0	1	0	0	1	0
D	1	0	0	73	2	1	1	3	2	0	1	0	0	1	3	4	2	1	1	2	0	1	0	0	1	0
E	1	3	0	0	69	2	0	4	1	1	1	0	0	0	2	0	3	1	2	3	2	1	1	2	1	0
F	1	0	0	1	1	78	1	0	0	1	0	0	0	2	6	2	4	0	0	0	0	1	0	2	0	0
G	1	1	1	0	1	2	75	2	0	0	1	1	4	0	2	2	0	1	1	0	2	0	0	1	1	1
H	2	3	0	5	3	2	1	64	1	0	1	0	1	1	4	1	3	4	0	1	0	2	0	0	0	1
I	0	1	1	0	0	0	0	2	72	6	1	0	0	2	0	0	1	1	1	2	0	1	3	0	4	2
J	0	0	0	2	0	0	0	0	1	69	1	1	0	2	0	0	3	1	3	1	2	3	1	4	4	2
K	1	1	1	0	0	0	1	3	2	1	69	6	0	0	0	0	1	5	0	0	2	0	0	0	0	7
L	1	1	1	0	1	0	1	0	5	1	1	74	1	1	0	0	1	1	0	0	5	2	0	0	1	2
M	0	0	2	0	1	0	1	0	1	0	0	2	78	0	1	2	6	1	0	0	0	1	2	0	1	1
N	0	1	3	2	0	0	0	1	0	2	0	0	0	80	0	1	0	8	1	0	0	0	0	0	0	1
O	5	0	0	1	2	5	2	2	0	0	0	0	0	0	66	5	2	1	0	0	0	0	3	5	1	0
P	6	1	1	0	0	0	0	3	1	0	1	0	1	2	1	78	1	0	0	0	2	1	0	0	0	0
Q	1	0	2	0	2	0	2	1	0	2	0	0	2	1	1	0	78	0	1	0	0	1	1	2	3	0
R	1	5	0	2	0	0	0	2	0	0	3	5	0	4	0	1	0	68	0	5	1	1	0	0	1	1
S	0	0	1	2	0	0	0	2	1	1	6	0	0	0	0	0	0	3	80	0	0	1	0	2	0	1
T	1	5	3	0	0	0	1	1	2	0	2	1	1	2	1	1	1	10	2	59	4	1	0	0	0	2
U	3	0	0	1	1	1	0	1	0	3	0	2	1	2	1	0	0	1	2	1	74	1	2	0	0	3
V	1	0	0	5	2	1	1	2	2	1	0	0	0	1	0	2	3	0	1	0	2	67	5	3	0	1
W	0	0	0	0	0	2	0	0	1	0	1	1	0	0	2	0	1	0	0	0	3	4	82	0	2	1
X	2	0	0	3	2	1	0	0	1	2	0	0	1	2	1	0	3	0	3	0	0	2	0	75	2	0
Y	0	2	0	0	1	1	0	0	0	4	0	0	1	0	0	0	0	0	0	0	1	1	3	2	83	1
Z	0	1	0	0	1	1	1	0	0	1	2	4	0	1	0	0	0	3	0	4	3	0	0	0	1	77

(b)

Figure 4. Confusion Matrices for LightGBM (MFCC) and LightGBM (Wav2Vec2 + PCA).

In the case of the MFCC-based model (Figure 4a), when analyzing the numerical data, the most notable confusions include:

- Letter 'A' (Real) → 'P' (Predicted): 9 cases, with a diagonal value for 'A' of 71.
- Letter 'E' (Real) → 'H' (Predicted): 5 cases, with a diagonal value for 'O' of 72.
- Letter 'T' (Real) → 'B' (Predicted): 8 cases, with a diagonal value for 'B' of 72.
- Letter 'G' (Real) → 'O' (Predicted): 5 cases, with a diagonal value for 'G' of 76.
- Letter 'H' (Real) → 'T' (Predicted): 6 cases, with a diagonal value for 'H' of 76.

The model employing Wav2Vec2 embeddings (Figure 4b) showed different confusion patterns:

- Letter 'T' (Real) → 'Z' (Predicted): 10 cases, with a diagonal value for 'T' of 59.
- Letter 'H' (Real) → 'D' (Predicted): 5 cases, with a diagonal value for 'H' of 64.

- Letter ‘B’ (Actual) → ‘T’ (Predicted): 11 cases, with a diagonal value for ‘V’ of 66.
- Letter ‘O’ (Actual) → ‘A’ (Predicted): 9 cases, with a diagonal value for ‘O’ of 66.
- Letter ‘U’ (Actual) → ‘W’ (Predicted): 5 cases, with a diagonal value for ‘Z’ of 67.

Overall, the Wav2Vec2 model demonstrates a distributed error structure, without collapses into specific classes, although its overall accuracy is slightly lower. Figure 5 shows the F1-score per class for both LightGBM configurations. It is observed that the model with MFCC features generally achieved higher F1-scores for most letters compared to the Wav2Vec2 + PCA-based model. Specifically, letters such as ‘N’ (0.91), ‘S’ (0.88), ‘I’ (0.90), ‘W’ (0.88), ‘D’ (0.88), ‘I’ (0.87), and ‘Q’ (0.87) showed notably strong performance with MFCCs. In contrast, letters ‘H’ (MFCC: 0.79, Wav2Vec2: 0.65), ‘B’ (MFCC: 0.81, Wav2Vec2: 0.69), ‘A’ (MFCC: 0.72, Wav2Vec2: 0.72), ‘E’ (MFCC: 0.74, Wav2Vec2: 0.72), and ‘T’ (MFCC: 0.74, Wav2Vec2: 0.61 presented some of the lowest F1-score values across both configurations, which is consistent with the confusions observed in the matrices. The utilization of Wav2Vec2 embeddings, although with slightly lower overall F1-score performance per class, demonstrated the ability to handle phonetic variability in unconventional recording environments, and for some specific letters, its performance was comparable or even superior to that of MFCCs.

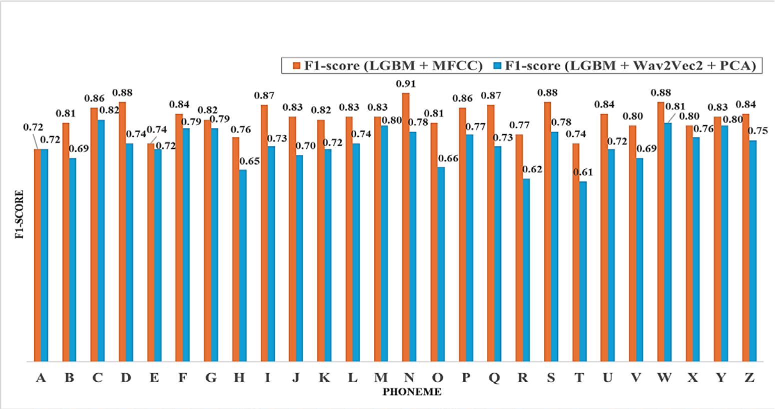


Figure 5. Class-wise F1-score Comparison for LightGBM with MFCC and with Wav2Vec2 + PCA (Throat Microphone).

4.4. Robustness Considerations under Realistic Conditions

One of the central objectives was to evaluate the capability of ASR approaches to operate under PTT conditions and with TM. The WER results obtained from the eight evaluation sequences reveal notable differences. The LightGBM with MFCC and RF with MFCC configurations showed exceptionally robust performance, achieving a WER of 0.000 across all test sequences for both models. This implies near-perfect transcription, reaffirming their robustness not only at the phoneme classification level but also in the recognition of complete sequences. The LightGBM configuration with Wav2Vec2 + PCA obtained a WER of 0.063, indicating very good performance, albeit with some isolated errors in the PTT sequences.

In contrast, models such as SVM, KNN, and MLP with MFCC showed less consistent performance. The SVM classifier obtained an average WER of 0.375, while KNN and MLP recorded WERs of 0.281 and 0.188, respectively. These results suggest that approaches based on distances or linear hyperplanes are more affected by the spectral distortions inherent to the laryngeal channel and by the acoustic variability induced by augmentation techniques.

The outstanding performance of methods like LightGBM and RF is attributed to their decision tree ensemble structure, which captures complex non-linear relationships and adapts to degraded input patterns. LightGBM, in particular, optimizes training by prioritizing the correction of residual errors, which explains its robustness. This advantage holds even when using deep representations like Wav2Vec2.

From a practical perspective, these findings indicate that ASR systems based on LightGBM and RF, combined with appropriate acoustic representations, are promising for applications in environments where TM are used. The ability to maintain a null or near-zero WER in real-world test situations is a key indicator of the system's robustness and usability.

5. DISCUSSION

The results of this study confirm the technical feasibility of ASR using signals captured exclusively with a throat microphone, even under simulated adverse acoustic conditions. The outstanding performance of decision tree-based classifiers (LightGBM and RF) utilizing MFCC representations suggests that these traditional features remain a robust and efficient option for phonemic classification tasks. The near-zero WER for these models is highly relevant for critical applications requiring high intelligibility.

However, this performance should not be interpreted as a generalized superiority of traditional representations. Embeddings derived from self-supervised models, such as Wav2Vec2, offered significant qualitative advantages, particularly in reducing systematic errors associated with phonemes exhibiting high acoustic ambiguity. Although their overall accuracy was slightly lower in some experiments with LightGBM, their ability to abstract phonological information from spectrally degraded signals indicates high potential for environments with greater input variability or structural noise. PCA analysis revealed the efficiency of dimensionality reduction for Wav2Vec2 embeddings (from 2048 to 256 components), preserving over 99.50% of the variance. This finding opens opportunities for pipeline compression and optimization, reducing computational cost without compromising predictive performance.

From an architectural standpoint, simpler models like MLP achieved competitive performance, positioning them as a viable alternative for embedded or low-power applications. In contrast, SVM and KNN exhibited significant limitations in accuracy and stability. Their low performance can be attributed to a combination of factors, including sensitivity to spectral compression and difficulties in modeling complex class distributions in high-dimensional spaces. Their inclusion in the study served as a methodological reference to gauge the real impact of algorithmic improvements.

Qualitatively, the analysis of confusion matrices and class-wise F1-score identified recurring error patterns, which is consistent with the inherent limitations of laryngeal signals, such as the loss of harmonics and the attenuation of formant transitions—characteristic phenomena of the laryngeal signal. Notable confusions in the MFCC-LGBM model, such as 'T' and 'H' predicted as 'B', and in the Wav2Vec2-LGBM model, such as 'I' with 'L' or 'T' with 'R', underscore the specific phonetic challenges that persist. Despite these acoustic limitations, Wav2Vec2-based models demonstrated greater robustness against some types of phonetic ambiguity, reinforcing their relevance in restricted communication scenarios.

A key limitation of the present study is the absence of fine-tuning of the self-supervised models on a specific corpus of laryngeal speech. Although the use of pre-trained embeddings in a zero-shot modality allowed for a practical approximation, additional training in the target domain could significantly enhance performance. Similarly, the non-implementation of speaker adaptation strategies might have affected the models' generalization to inter-individual variability. Furthermore, the decision not to perform systematic hyperparameter optimization, while maintaining experimental control, probably impacted the models most sensitive to this configuration (such as SVM and KNN) more negatively. These aspects should be considered in future studies to evaluate the true performance ceiling of these architectures.

Additionally, the quantity and diversity of the available data, although mitigated by augmentation, represent a limitation. A larger and more representative dataset would allow evaluating model robustness against a greater variety of operational or dialectal conditions.

Regarding the most relevant future directions, we propose:

1. Exploring the fine-tuning of self-supervised models (e.g., Wav2Vec2, HuBERT) specifically on laryngeal speech corpora to adapt their representations to the peculiarities of this signal.

2. Investigating the use of hybrid representations (e.g., combinations of MFCCs with deep embeddings) to leverage the advantages of both approaches.

3. Implementing speaker adaptation mechanisms to improve model generalization to inter-individual variability.

4. Evaluating lightweight models or advanced architectures of lightweight Transformers and convolutional networks to obtain more efficient embeddings and optimize computational cost in embedded applications.

Collectively, this work reinforces the idea that optimal ASR performance depends not only on model sophistication or the novelty of representations but on the overall coherence and compatibility of the system, including architecture, signal nature, computational complexity, and specific application conditions.

6. CONCLUSIONS

This study has demonstrated the technical feasibility of ASR using signals captured exclusively with TM under PTT conditions and simulated acoustically adverse scenarios. It was confirmed that robust phonemic classification systems can be implemented even with a degraded signal spectrum.

Decision tree-based classifiers, LightGBM and RF with MFCC features, showed outstanding effectiveness, achieving high accuracies and exceptionally low WERs (0.000), which makes them highly promising for critical applications. Similarly, Wav2Vec2 embeddings, efficiently reduced with PCA, showed high potential by improving the representation of ambiguous phonemes and offering greater resilience to systematic errors, achieving a WER of 0.063, despite not having undergone specific fine-tuning for the laryngeal domain.

Clear differences in model adaptability to the laryngeal domain were observed: while MLP offered a favorable balance, SVM and KNN exhibited significant limitations, underscoring the importance of careful model selection based on signal characteristics. This work emphasizes that optimal ASR performance is achieved through the strategic alignment between model architecture, signal nature, data volume, and application purpose. Future research should focus on fine-tuning self-supervised models and exploring hybrid systems to advance towards more generalizable and operationally viable solutions in environments with TM.

Sources of Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Credit Authorship Contribution Statement

Julio Velazquez: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization.

Selim Aras: Writing – review & editing, Visualization, Validation, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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.

Author Contribution Rates

Design of Study: JCVG(%50), SA(%50)

Data Acquisition: JCVG(%50), SA(%50)

Data Analysis: JCVG(%50), SA(%50)

Writing Up: JCVG(%50), SA(%50)

Submission and Revision: JCVG(%50), SA(%50)

REFERENCES

[1] Y. Görmez, “Customized deep learning based Turkish automatic speech recognition system supported by language model,” *PeerJ. Computer Science*, vol. 10, p. e1981, 2024.

[2] Z. Y. Ren, Nurmement, H. Wang, and W. Slamu, “Exploring Turkish speech recognition via hybrid CTC/attention architecture and multi-feature fusion network,” arXiv preprint arXiv:2308.10654 [cs.SD], 2023. (For arXiv preprints, it’s good practice to include the arXiv ID).

[3] E. T. Erzin and T. Tugtekin, “Improving phoneme recognition of throat microphone speech recordings using transfer learning,” Speech Communication, vol. 129, p. 104764, 2021. (Added article number/page if available, assuming “10” was part of it. If it was just the number of the article, no. 10 is fine).

[4] J. L. K. E. T. Fendji, D. C. M. Diane, B. O. Yenke, and M. Atemkeng, “Automatic speech recognition using limited vocabulary: A survey,” Applied Artificial Intelligence: AAI, vol. 36, no. 1, 2022.

[5] NATO, Allied Communications Publication ACP 125(G) Communication Instructions Radiotelephone Procedures. 2008. (This is a report/publication from an organization).

[6] A. Veysov, “silero-vad: Silero VAD: pre-trained enterprise-grade Voice Activity Detector,” GitHub, 2022. [Online]. Available: https://github.com/snakers4/silero-vad (For software/code, it’s best to treat it like a technical report or use the most formal citation possible, often including the repository host and year. The original example didn’t have a year or formal publisher).

[7] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio augmentation for speech recognition,” in Proc. Interspeech 2015, Dresden, Germany, 2015, pp. 3586-3589. (For conference papers, it’s good to include the location of the conference if known).

[8] D. S. C. Park, W. William, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition,” arXiv preprint arXiv:1904.08779 [eess.AS], 2019. (Added arXiv ID).

[9] C. Shorten and T. M. K. Taghi, “A survey on image data augmentation for deep learning,” Journal of Big Data, vol. 6, no. 1, 2019.

[10] S. M. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, 1980. (Corrected author ‘P’ to ‘P. Mermelstein’ as per common authorship in this field, assuming this is the full name, and added volume, number, and pages).

[11] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” arXiv preprint arXiv:2006.11477 [cs.CL], 2020. (Added arXiv ID).

[12] O. B. Mercan, S. Sercan, D. E. Tasar, and S. Ozan, “Performance comparison of pre-trained models for speech-to-text in Turkish: Whisper-small and Wav2Vec2-XLS-R-300M,” arXiv preprint arXiv:2310.04652 [cs.CL], 2023. (Added arXiv ID).

[13] S. Goldwater, D. Jurafsky, and C. D. Manning, “Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates,” Speech Communication, vol. 52, no. 3, pp. 181-200, 2010.

[14] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, “Hybrid CTC/attention architecture for end-to-end speech recognition,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240-1253, 2017.

[15] K. Masuda, M. Nishida, J. Ogata, and M. Nishimura, “Throat microphone speech recognition using wav2vec 2.0 and feature mapping,” in Proceedings of the 2022 IEEE 11th Global Conference on Consumer Electronics (GCCE), Osaka, Japan, 2022. (Added likely location).

[16] J. Hauret, M. O. Malo, T. Joubaud, C. Langrenne, S. Poir’ee, V. Zimpfer, and É. Bavu, “Vibravox: A dataset of french speech captured with body-conduction audio sensors,” Speech Communication, 2024. (This is an “abs” entry, so it means it’s an accepted paper but maybe not yet in a specific volume/issue).

[17] T. Winkler, S. Pronkine, R. Bardeli, and J. Köhler, “A study of throat microphone performance in automatic speech recognition on motorcycles,” in Proceedings of the Sixth International Language Resources and Evaluation (LREC’08), Marrakech, Morocco, 2008. (Added location for conference).

[18] S.-C. Jou, T. Schultz, and A. Waibel, “Adaptation for soft whisper recognition using a throat microphone,” in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, QC, Canada, 2004, pp. I-457-I-460. (Added likely location and full page numbers where applicable).

[19] Y. Kim, Y. Song, and Y. Chung, “TAPS: Throat and Acoustic Paired Speech Dataset for Deep Learning-Based Speech Enhancement,” arXiv preprint arXiv:2502.11478v2, 2025. (Added arXiv ID).