T.R.

GEBZE TECHNICAL UNIVERSITY

GRADUATE SCHOOL

MULTIMODAL LEARNING FOR SPEECH ANALYSIS WITH MMWAVE RADAR, ULTRASOUND, AND AUDIO

İLHAN AYTUTULDU

A THESIS OF DOCTORATE

DEPARTMENT OF COMPUTER ENGINEERING

ADVISOR: PROF. DR. YUSUF SİNAN AKGÜL

JUNE 2025

T.R.

GEBZE TECHNICAL UNIVERSITY

GRADUATE SCHOOL

MULTIMODAL LEARNING FOR SPEECH ANALYSIS WITH MMWAVE RADAR, ULTRASOUND, AND AUDIO

İLHAN AYTUTULDU

A THESIS OF DOCTORATE

DEPARTMENT OF COMPUTER ENGINEERING

ADVISOR: PROF. DR. YUSUF SİNAN AKGÜL

JUNE 2025

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

KONUŞMA ANALİZİNDE MMWAVE RADAR, ULTRASON VE SES VERİLERİYLE ÖĞRENME

İLHAN AYTUTULDU

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN: PROF. DR. YUSUF SİNAN AKGÜL

HAZİRAN 2025

DOCTORATE JURY APPROVAL FORM

A thesis submitted by İlhan, Aytutuldu, defended on 11/06/2025 before the jury formed with the 02/06/2025 date and 2025/29 numbered decision of the GTU Graduate Administration Board, has been accepted as a DOCTOR of PHILOSOPHY thesis in the Department of Computer Engineering.

JURY

MEMBER (THESIS ADVISOR)	: Prof. Dr. Yusuf Sinan AKGÜL
MEMBER	: Assoc. Prof. Dr. Yakup GENÇ
MEMBER	: Assist. Prof. Dr. Tarkan AYDIN
MEMBER	: Assoc. Prof. Dr. Habil KALKAN
MEMBER	: Assoc. Prof. Dr. Sabit EKİN

APPROVAL

Gebze Technical University Graduate Administration Board

…..…/…..…/..…… date and ……/…….. numbered decision.

SIGNATURE/SEAL

To my son, ***Tuna*** *—*
For your light, your laughter, and the quiet power you give me every day.

ABSTRACT

Phonetic segment classification is the process of recognizing and categorizing speech sounds (phonemes) from audio or articulation data. This task forms the basis of many speech and language processing applications such as speech recognition, language learning, and clinical speech therapy. This thesis explores novel multimodal approaches based on deep learning that combine mmWave radar, ultrasound tongue imaging (UTI), and voice modalities to achieve robust, practical, and privacy-sensitive phonetic segment classification.

Traditional audio-based methods, such as Gaussian Mixture Model–Hidden Markov Model (GMM-HMM), Deep Neural Network–Hidden Markov Model (DNN-HMM), often yield unreliability in noisy environments and may be insufficient to distinguish phonemes with similar spectral properties. On the other hand, UTI-based approaches provide detailed articulatory information but have limitations in terms of hardware costs and practicality. On the other hand, mmWave radar stands out as a promising alternative that can capture detailed articulatory dynamics in a contactless manner.

In this context, the thesis first introduces a multimodal neural network architecture called MMNet. This architecture is designed to create a common embedding space by taking advantage of the complementary features of audio and UTI modalities. Experimental findings show that this combination leads to significant improvements in classification accuracy and effectively removes ambiguities specific to individual modalities.

Additionally, the architecture named M1 can perform inference using only audio data, thus eliminating the need for UTI, making it easy to use in applications such as speech therapy. Experimental data shows that this approach performs at a level comparable to systems that use both voice and UTI.

In the next phase of the thesis, another multimodal framework named USRadioAI is presented, which combines UTI and voice data with mmWave radar in the training process, but relies only on radar data in inference. This model is able to maintain similar performance to multimodal systems despite working with only radar data, by using techniques such as cross-modal information distillation and embedding space alignment. USRadioAI uses radar's ability to capture articulatory information without contact, reducing both hardware complexity and system load during the inference phase, while maintaining high accuracy rates.

Comprehensive experiments conducted on newly compiled datasets named GTUConsonants and GTUSAudioRadioConsonants demonstrate the effectiveness of the developed multimodal methods. The results show that MMNet provides the highest overall classification performance by effectively using all available modalities, while USRadioAI, despite its simplified inference structure, performs very close to multimodal approaches thanks to optimization techniques.

The findings highlight the potential of multimodal learning in speech and language processing, and suggest that mmWave radar can play an important role as a privacy-preserving, accessible, and user-friendly modality in areas such as clinical speech therapy and interactive applications.

Keywords: Phonetic Segment Classification, Multimodal Learning, mmWave Radar, Ultrasound Tongue Imaging (UTI), Deep Learning, Speech Processing

ÖZET

Fonetik segment sınıflandırması, ses ya da artikülasyon verilerinden konuşma seslerini(fonemleri) tanıma ve kategorize etme sürecidir. Bu görev, konuşma tanıma, dil öğren-imi ve klinik konuşma terapisi gibi pek çok konuşma ve dil işleme uygulamasının temelini oluşturur. Bu tez, sağlam, pratik ve gizliliğe duyarlı bir fonetik segment sınıflandırması elde etmek amacıyla mmWave radarı, ultrason dil görüntüleme (UTI) ve ses modalitelerini birleştiren, derin öğrenmeye dayalı yeni çok modlu yaklaşımları araştırmaktadır.

Geleneksel ses tabanlı yöntemler, gürültülü ortamlarda güvenilirliğini yitirir ve spektral olarak benzer fonemleri ayırt etmekte yetersiz kalabilir. UTI tabanlı sistemler ise ayrıntılı artikülasyon bilgisi sunmalarına rağmen, yüksek donanım maliyetleri ve kullanıcı konforu açısından çeşitli sınırlamalar taşır. Bu noktada mmWave radar, temassız biçimde ayrıntılı artikülasyon dinamiklerini yakalayabilen, umut verici bir alternatif olarak öne çıkmaktadır.

Bu doğrultuda tezde ilk olarak, MMNet adlı çok modlu bir sinir ağı mimarisi sunulmak-tadır. Bu mimari, ses ve UTI modalitelerinin tamamlayıcı özelliklerinden yararlanarak ortak bir yerleştirme (embedding) alanı oluşturmak üzere tasarlanmıştır. Deneysel bul-gular, bu entegrasyonun sınıflandırma doğruluğunu önemli ölçüde artırdığını ve tekil modalitelere özgü belirsizlikleri etkili biçimde ortadan kaldırdığını göstermektedir. Ayrıca, M1 adlı mimari yalnızca ses verilerini kullanarak çıkarım yapabilmekte; böylece UTI kullanımına olan ihtiyacı ortadan kaldırmakta ve konuşma terapisi gibi uygula-malarda pratikliği artırmaktadır. Deneysel sonuçlar, bu yaklaşımın hem ses hem de UTI kullanan sistemlerle karşılaştırılabilir düzeyde performans sunduğunu göstermek-tedir.

Tezin sonraki aşamasında ise, eğitim sürecinde UTI ve ses verilerini mmWave radar ile birleştiren; ancak çıkarım aşamasında yalnızca radar verisine dayanan USRadioAI adlı alternatif bir çok modlu çerçeve tanıtılmaktadır. Çapraz-modal bilgi damıtma ve gömme alanı hizalaması gibi teknikleri kullanan bu model, yalnızca radar verisiyle çalışmasına rağmen çok modlu sistemlerle benzer doğruluk seviyelerini sürdürebilmek-tedir. USRadioAI çerçevesi, radarın eklemsel bilgileri temassız şekilde yakalama yeteneğin-den faydalanarak, çıkarım sırasında donanım karmaşıklığını ve sistem yükünü azaltırken yüksek performans sergilemektedir.

Yeni oluşturulan GTUConsonants ve GTUSAudioRadioConsonants veri kümeleri üzerinde gerçekleştirilen kapsamlı deneyler, geliştirilen çok modlu yöntemlerin etkinliğini ortaya koymaktadır. Sonuçlar, MMNet'in mevcut tüm modaliteleri verimli biçimde kullanarak en yüksek genel sınıflandırma doğruluğunu sağladığını; USRadioAI'nin ise sadeleştirilmiş çıkarım yapısına rağmen, belirli optimizasyon teknikleriyle çok modlu yaklaşımlara oldukça yakın performans elde ettiğini göstermektedir.

Bu bulgular, konuşma ve dil işleme alanında çok modlu öğrenmenin önemini vurgularken, mmWave radarın klinik konuşma terapisi ve etkileşimli uygulamalarda gizliliği koruyan, erişilebilir ve kullanıcı dostu bir modalite olarak önemli bir potansiyele sahip olduğunu ortaya koymaktadır.

Anahtar Kelimeler: Fonetik Segment Sınıflandırması, Çok Modlu Öğrenme, mmWave Radar, Ultrason Dil Görüntüleme (UTI), Derin Öğrenme, Konuşma İşleme

ACKNOWLEDGEMENTS

I would like to sincerely thank my advisor, Prof. Dr. Yusuf Sinan Akgül, for his continuous support, guidance, and encouragement throughout my PhD studies. His knowledge, patience, and thoughtful feedback helped me improve my research and overcome many challenges. I have learned a great deal from him, both academically and professionally, and I truly appreciate the time and effort he has dedicated to this work.

I am also grateful to Assoc. Prof. Yakup Genç, a member of my jury committee, for his valuable time, comments, and suggestions that contributed to the improvement of this thesis.

Special thanks to Asst. Prof. Tarkan Aydın of Bahçeşehir University, who supported me during my master's studies as my advisor and now as a jury member for this thesis. His early guidance played an important role in shaping my academic path.

I am deeply grateful to Assoc. Prof. Habil Kalkan (Gebze Technical University) and Assoc. Prof. Sabit Ekin (Texas A&M University) for accepting to serve as jury members for my Ph.D. thesis defense and for their valuable evaluations and feedback.

I would like to thank my wife, Asst. Prof. Güzin Kaya Aytutuldu of Biruni University, for always being by my side with her endless support, patience, and understanding. Her presence and encouragement have been very important throughout this journey.

To my son, Tuna, thank you for bringing joy and motivation to my life, even during the most difficult times.

This work was supported by the Health Institutes of Türkiye (TUSEB) C-YZ Project (No. 33944) and by Gebze Technical University Scientific Research Projects (No. 2023-A-102-04). I would like to thank both institutions for their financial support.

Finally, I thank all my family members, friends, and colleagues who helped and supported me along the way.

TABLE OF CONTENTS

LIST of ABBREVIATIONS and ACRONYMS

2D	:	Two-dimensional
3D	:	Three-dimensional
A-mode	:	Amplitude Mode
ADC	:	Analog-to-Digital Converter
ASR	:	Automatic Speech Recognition
B-mode	:	Brightness Mode
BiLSTM	:	Bi-directional Long Short-Term Memory
CAS	:	Childhood Apraxia of Speech
CNN	:	Convolutional Neural Network
dB	:	Decibel
dBm	:	Decibel-Milliwatts
EMA	:	Electromagnetic Mid-sagittal Articulography
FLOP	:	Floating Point Operation
FMCW	:	Frequency Modulated Continuous Wave
FFT	:	Fast Fourier Transform
fps	:	Frames Per Second
Fs	:	Sampling Rate
GHz	:	Gigahertz
GTUConsonants	:	GTU Consonants Dataset
GTUSAudioRadioConsonants	:	GTU Ultrasound Audio Radio Consonants Dataset
GTÜ	:	Gebze Teknik Üniversitesi
IF	:	Intermediate Frequency
IPA	:	International Phonetic Alphabet
ksps	:	Kilo Samples Per Second
MHz	:	Megahertz
mmWave	:	Millimeter Wave
MRI	:	Magnetic Resonance Imaging
Msps	:	Mega Samples Per Second
PRI	:	Pulse Repetition Interval
RGB	:	Red Green Blue
RTMRI	:	Real-time MRI
RX	:	Receive Antenna
SLP	:	Speech and Language Processing
STFT	:	Short Time Fourier Transform
t-SNE	:	T-Distributed Stochastic Neighbor Embedding
TI	:	Texas Instruments
TX	:	Transmit Antenna
USRadioAI	:	Ultrasound Radio Audio Integrated model
UXTD	:	Ultrax Typically Developing
UTI	:	Ultrasound Tongue Imaging
ΔR	:	Range Resolution
Δv	:	Velocity Resolution
λ	:	Wavelength
μs	:	Microsecond

LIST of FIGURES

LIST of TABLES

1. INTRODUCTION

1.1. Overview

A fundamental means of human communication, speech helps to transmit complex knowledge by means of coherent articulation, acoustics, and cognitive mechanisms. Recent developments in sensor technology and artificial intelligence are driving speech and language processing outside of traditional audio-based approaches forward. Modern systems are more equipped than only transcription of speech or sound generation to understand and replicate the physical mechanics underpinning speech production. Particularly in fields such education, healthcare, and assistive communication technology, these developments provide fresh opportunities [2–4].

This growing ambition has given rise to a new class of research questions: How can we capture the hidden dynamics of speech production in a way that is accurate, interpretable, and non-invasive? How can rich but complex multimodal data be distilled into practical systems suitable for real-world use? How thus can we balance the simplicity needed for general usability with the technical complexity of multimodal learning?

This thesis is at the center of this inquiry. Using next-generation sensing technologies, the study looks at how actual speech creation may be sensed, modeled, and turned into strong, meaningful representations. Grounding this conversation, the next part explores the fundamental ideas behind this work.

1.2. Motivation

The utilization of speech and language technologies, from voice assistants to language learning tools, have grown indispensable in contemporary human-computer interaction [5]. Models that not only attain high performance but also robust, interpretable, privacy-conscious, and deployable in real-world situations are increasingly sought for as these systems becoming more ubiquitous [6, 7].

The modeling of the physically produced speech to comprehend articulatory information, such as how the tongue, lips, and jaw coordinate to form speech sounds in addition to recognize acoustic patterns is one of the foundational challenges in this field [8, 9].

This level of modeling has implications in a wide range of domains:

- Education, for helping language learners and children master articulation [10].
- Healthcare, for supporting speech therapy and assessment [11, 12].
- Accessibility, by enabling non-invasive communication aids [13]
- Human-computer interaction, through richer and more reliable voice-based interfaces [14]

Thanks to innovations in multimodal data collection, the physical articulatory movements that constitute speech can now be recorded along with traditional audio [15–19]. These technologies provides speech production to be represented in more detail, leading better results in applications such as phoneme recognition, speech generation, and articulation feedback [1, 20].

As multimodal systems require multiple sensors, which can be costly, intrusive, and difficult to integrate into practical environments [21], one of the major challenges in real-world applications is the deployment of these systems. In contrast, unimodal systems—particularly those based solely on audio, radio, or visual data acquisition—are easier to integrate into consumer devices and offer a more viable alternative for real-world applications [22].

To address this trade-off between richness of representation and deployment feasibility, this thesis adopts an indirect cross-modal knowledge distillation approach by using an embedding space [23]. We train deep learning models using data from multiple complementary sources—generated by microphone, radar, and ultrasound devices—to learn detailed articulatory representations (Fig. 1.1), and then transfer this knowledge via the embedding space to models that operate using a single input source at inference time. This strategy enables the system to benefit from the depth and structure provided by richer training data, while remaining practical, efficient, and privacy-conscious in real-world deployment—qualities essential for building next-generation speech and language processing tools.

The basic assumption of the thesis is that detailed articulation information can be collected during the training process and this information may then be converted into lightweight models that are used in real-world scenarios. In support of this approach,

the next section summarizes imaging techniques that are commonly used to capture the dynamics of the speech production process.

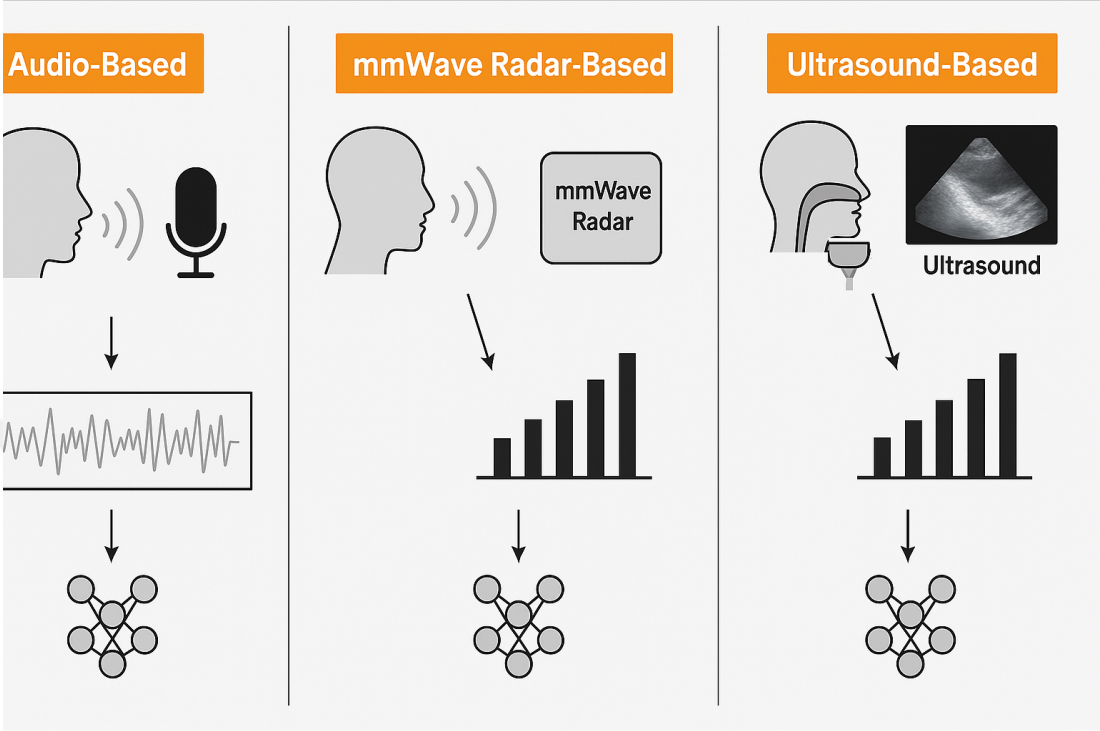


Figure 1.1: Schematic representation of audio-based, mmWave radar-based, and ultrasound-based speech sensing pipelines integrated with deep learning models.

1.3. Speech Production Imaging Techniques

The main ways to spread knowledge and build social ties are speech communications. This complicated process consists in the speaker producing speech sounds and their interpretation by the listener. Tongue movement is an essential component of speech production; it is vital for articulation but stays outside observable in nature [8]. Different imaging approaches are employed to capture the tongue's form as articulatory data [24], so analyzing and modeling articulation inside the vocal tract. Multiple imaging modalities have been used in recent years to investigate speech production including, electromagnetic mid-sagittal articulography (EMA) [25], magnetic resonance imaging (MRI) [9, 26–29], and ultrasound [30, 31].

EMA allows exact tracking of articulatory movements by means of sensor placement

on anatomical features including the lips, jaw, and tongue [32]. EMA data allows one to investigate the kinematics of the vocal tract in both healthy people and Parkinson's disease [33]. It measures tongue and articulator motions directly, however it has limits including small datasets, possible speech alterations due of its invasive character, and reliance on discrete measurement sites that do not fully capture the tongue shape [34, 35].

Furthermore, MRI can help to provide high-resolution images of vocal tract anatomy during speech. MRI also finds uses in research on second language acquisition and speech production [36]. It facilitates the analysis of structural attributes in conjunction with articulation dynamics and acoustics. Not only the tongue but also the articulators such as labial, jaw, velar, pharyngeal, and laryngeal information provide dynamic information with temporal resolution with Real-time MRI (rtMRI) [18, 37]. However, its supine recording orientation is inappropriate for spontaneous speech, and concurrent audio recordings frequently exhibit inferior quality. Moreover, the practical limitations of MRI restrict the dataset sizes in speech processing related research [38].

Ultrasound imaging is based on the measurement of the time it takes for sound waves emitted by a transducer to travel through tissues and reflect back from anatomical structures [40–42]. Different imaging modes exist to interpret these reflections and construct a visual representation of the underlying anatomy. Figure 1.2 (top-left) shows a single ultrasound wave emitted by a transducer directed toward the tongue. Figure 1.2 (middle-left) represents the A-mode (amplitude mode), where the amplitudes of the reflected signals at time tt are plotted, yielding a one-dimensional waveform. In contrast, Figure 1.2 (bottom-left) illustrates the B-mode (brightness mode), in which echo amplitudes are mapped as varying pixel brightness levels. By collecting multiple such reflections from different beams, a two-dimensional (2D) grayscale image is constructed, as shown in Figure 1.2 (right-middle). This study employs B-mode ultrasound imaging to capture and analyze tongue movement.

Ultrasound tongue imaging (UTI), thanks to its ability to provide real-time, high-resolution, and quick 2D imaging, is a useful method for observing the tongue's motions during speech communication. It also provides non-invasive and a risk-free imaging of the articulators during normal and abnormal speech production [43]. Clinical popula-

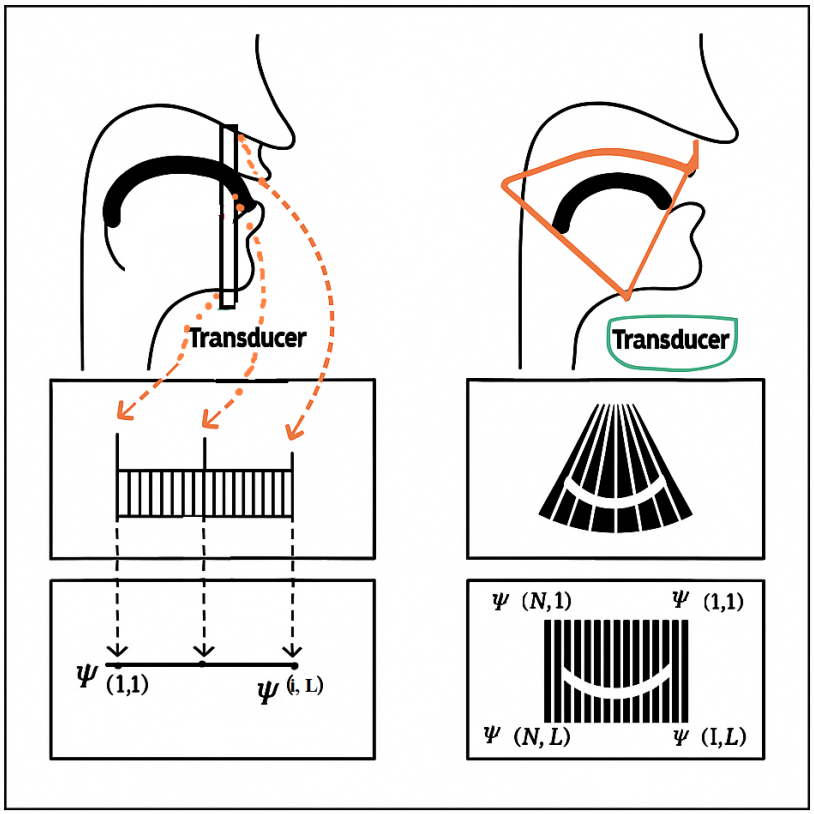


Figure 1.2: Ultrasound imaging modes and the process of ultrasound image acquisition to visualize the vocal tract [39].

tions highly benefit from UTI's visual feedback for speech sound problems assessment and remediation [44]. A number of areas have found uses for the method, such as the diagnosis of obstructive sleep apnea [45, 46] and the assessment of tongue movements in cerebral palsy patients [47]. There has been consistent advancement towards three-dimensional (3D) imaging using ultrasound, which has been a widely used tool in speech production research [48–50]. Platforms for 3D visualization of tongue motion have been developed [51, 52], and researchers have introduced methods to assess tongue shapes and positions using ultrasound imaging. The effectiveness of UTI in analyzing tongue structure and function has been extensively demonstrated in both clinical and scientific settings [53].

1.4. UTI in Speech and Language Processing

UTI is commonly used in clinical settings to capture real-time sagittal or coronal views of the movement of the tongue during speech or swallowing. The transducer is placed

submandibularly behind the chin. It is common practice to use a headset or fixed mount to hold the transducer steady during imaging and then use specialised software to visualise and analyse the ultrasound data [54]. UTI uses conventional medical ultrasound to visualize the surface of the tongue during speech production. This is a noninvasive, clinically safe and progressively affordable technique to visualize the vocal tract [55].

The utilisation of UTI devices in speech therapy research has increased in recent years, due to the enhancement of image quality and accuracy resulting from technological advancements. The decline in costs has led to a greater prevalence of portable ultrasound devices, and the interpretation of ultrasonography data has become more accessible. In addition, the development of various analysis software programs has further contributed to the growth in the use of UTI devices [31, 56].

Research suggests that UTI as a visual biofeedback technology can improve articulatory accuracy in people with acquired apraxia, post-glossectomy speech deficits and children with residual articulation difficulties [11, 57, 58]. Studies have investigated whether ultrasound feedback of the tongue helps some children with childhood apraxia of speech (CAS) to develop more accurate and reliable speech motor plans [59, 60]. The use of UTI feedback, speech therapy delivered in numerous weekly sessions and motor-based speech therapy with intensive schedules have also been found to outperform traditional less frequent service delivery [61, 62].

Therapists may employ the data obtained from these devices to aid the individual in accurately positioning the tongue. The client receives assistance in producing the correct speech sounds and improving their linguistic abilities. These feedback devices enhance the acquisition and practice of precise articulation during treatment by combining auditory input with visual indicators of tongue position [63]. Consequently, the comprehensive imaging of the vocal tract facilitates the evaluation of compliance with the methodologies utilized in speech therapy [64].

However, there are a few issues that still restrict the usefulness of conventional UTI, notwithstanding its significance in comprehending, identifying, and enhancing speech production and therapy. Ultrasound is impractical because it requires a continuous ultrasound probe under the chin, it is still expensive for home use, image quality depends on speaker characteristics such as age and psychology, and synchronization of sound

and image is costly. A need has arisen for a more practical approach to diagnosis, treatment and therapy applications in speech and language disorders.

Moreover, visual feedback alone may not be always enough for quick and accurate interpretation of ultrasound exposures when performing complicated tongue movements [65]. In other words, it is still quite difficult for the UTI to interpret the rapid and complex movements of the tongue and for the therapy to become permanent with visual feedback and for the tongue movements that will provide correct articulation to become automatic for individuals. Also, concerns about image quality and restrictions, interdisplay alternatives, and limited data availability are just a few of the many choices for automated Ultrasound tongue parameters [63]. Due to these challenges, leveraging tongue imaging effectively and accurately in practical scenarios may require an alternative representation or intermediate modeling approach.

1.5. MmWave Radar-based Human Voice Sensing

Millimeter-wave (mmWave) radar has recently emerged as a novel modality for non-contact speech sensing and analysis. Operating in the 30–300 GHz range, mmWave radar captures electromagnetic reflections modulated by the physiological movements of the articulators—particularly the throat, lips, and vocal folds—during speech. Without wearable or contact-based devices [66–68], this sensing mechanism enables robust voice acquisition even under noisy or non-line-of-sight situations.

By using Frequency Modulated Continuous Wave (FMCW) method, mmWave radars sense micro-motions produced by articular dynamics and vocal cord vibrations [69, 70]. These systems offer advantages in terms of resolution, stealth, and robustness against acoustic interference, noise, or spoofing attacks [71, 72].

Radar-based systems utilize the ability of radiofrequency signals to capture subtle articulatory and vocal tract movements during speech production [73]. Earlier approaches, relied on fixed radar antennas attached directly to the face using adhesive methods, often suffered from session-to-session variability and user discomfort [74]. Addressing these limitations, recent research has focused on developing wearable solutions that improve both usability and measurement stability [75]. Specifically, an important step forward in the development of user-friendly and repeatable speech interface sys-

tems has been the incorporation of radar sensors into lightweight, adjustable headsets. These radar-based designs seek to merge the non-invasiveness and portability needed for daily applications, taking inspiration from other biosignal-based speech systems that use techniques like electromyography, magnetic sensing, or ultrasound imaging [76]. The evolution of such systems reflects a broader trend toward silent, privacy-preserving, and contactless speech technologies that can be adapted to a variety of use cases, such as assistive communication, discreet interaction, and speech disorder rehabilitation.

Additionally, during therapy sessions, real-time mmWave feedback allows therapists to observe facial and vocal dynamics and provide immediate corrections. Unlike audio-based systems, which struggle to separate overlapping voices in interactive settings, mmWave radar only captures data from the participant when directed at them, naturally isolating their input during dialogue. This method provides objective, quantitative tracking of therapy progress without requiring any physical attachments to the participant. The portability of mmWave radar systems also makes them viable for home use, increasing the frequency and accessibility of therapy, and enhancing participant outcomes.

Tasks such as speech enhancement, speaker identification, and silent speech recognition can be performed using the analysis of the reflected signals characterized by changes in amplitude, phase, or frequency. This enables to perform applications in healthcare, human-computer interaction, and security [77, 78]. The feasibility of using mmWave radar has been demonstrated in speech signal detection, enhancement, and classification [68, 70, 77]. These studies suggests that the potential of radar-based technology in healthcare, especially in clinics for diagnosis, treatment and therapy open to explore.

1.6. Main Contributions of the Thesis

The studies and experimental setups developed in this thesis are organized into three main parts. The first part presents the preliminary work and foundational setups used to collect synchronized data from UTI, audio, and mmWave radar. This stage focuses on building the multimodal infrastructure required for articulatory speech analysis.

The second part investigates a joint embedding space constructed from UTI and au-

dio data. Using this space, audio-only models are trained to benefit from articulatory supervision during training, enabling more robust and interpretable representations of speech sounds while maintaining practicality at inference time.

The third part extends this approach by incorporating mmWave radar into the joint embedding space, resulting in a comprehensive multimodal framework based on UTI, audio, and radar. This setup enables the development of radar-only systems that inherit articulatory knowledge from richer modalities, offering a privacy-preserving and contactless alternative for speech processing applications.

Together, these sections build toward the thesis objective of creating accurate, deployable, and privacy-conscious speech processing models using cross-modal supervision.

This thesis provides five key contributions to the field of multimodal speech processing and human voice sensing:

- Proposes a novel framework, that combines mmWave radar, UTI, and audio for speech analysis.
- Demonstrates that mmWave radar can non-invasively capture articulatory and vocal fold motion, offering a privacy-preserving alternative to traditional sensing methods.
- Develops a unified phonetic embedding space that enables radar- or audio-only inference by leveraging supervision from multimodal data during training.
- Introduces seven complementary methods (M1–M4, RadioUS, RadioAudio, US-RadioAI) that exploit the embedding space through different strategies to enhance speech representation learning.
- Incorporates temporal modeling of articulator dynamics, leading to more consistent and accurate phonetic detection compared to conventional frame-based UTI approaches.

2. PRELIMINARY WORK

This chapter presents the preliminary studies conducted to establish the foundations and technical infrastructure for multimodal speech and human sensing using mmWave radar technology. It begins by outlining the theoretical basis and recent developments in multimodal deep learning, particularly in the domains of healthcare and speech processing. Following this, the chapter details a series of experimental setups and signal processing methods used to evaluate the viability of radar-based sensing. This includes both human motion capture and phonetic articulation experiments, with a focus on micro-Doppler analysis and spectrogram generation. Additionally, a real-time radar configuration and data acquisition framework is introduced, enabling synchronized multimodal data collection and processing. The results presented in this chapter form the experimental backbone for the multimodal AI system proposed in subsequent chapters.

2.1. Multimodal Foundations

In artificial intelligence (AI), multimodal learning becomes ever more important. This method underlines the need of combining many sensing sources that underpin human perception and cognitive decision-making processes: vision, language, and sound. Whereas conventional single-modal systems are constrained by their limited perceptual coverage, multimodal systems seek to make use of complementary and redundant information across many senses. In this regard, innovative work by [79, 80] clarifies in great detail the ideas of representation, alignment, fusion, and joint learning that form the technical basis of multimodal learning, and also highlights basic computational difficulties such the difference between modalities and interaction dynamics.

Multimodal deep learning has emerged that the focus fields of the researchers have been expanded to include many modalities and application domains, such as healthcare [81, 82]. This scenario shows the growing relevance of multimodal solutions, particularly in fields where sophisticated and multi-source data are somewhat widespread [83, 84] increases awareness by exposing the cognitive similarities of multimodal AI with human multisensory learning processes by analyzing how modalities contribute differently to reasoning, emotional involvement, and metacognitive processes.

Still, this development in capacity introduces complication as well. Especially in high-risk sectors like healthcare, where explainability is essential to developing trust [85], the interpretability of multimodal deep neural networks remains a major problem. Real-world applications such as image captioning, video annotation systems [86] , and speech-based systems where acoustic representations are augmented with visual or articulatory data [87, 88] already demonstrate the potential of multimodal integration. As a result, the ability of multimodal AI to combine vision, speech, and reasoning into coherent and context-sensitive systems represents a significant step in the development of more adaptive, reliable, and human-centered AI (Table 2.1 and Table 2.2).

Table 2.1: Detailed modality-wise comparison of multimodal studies with corresponding deep learning techniques

Domain	**Citation**	**Modalities Used**	**DL Techniques Used or Mentioned**	**Key Contribution**	**Main Challenge Addressed**
Healthcare	Tariq et al. (2025) [89]	Radiology + EHR	Transformer-based fusion, CNNs	Enhanced diagnostic accuracy by integrating imaging with clinical history	Data incompleteness, semantic alignment, deployment in clinical workflows
	Schouten et al. (2024) [90]	Radiology + EHR	Attention-based multimodal fusion	Highlighting clinical and system deployment gaps in multimodal radiology AI	Semantic alignment, interpretability, interoperability

Continued on next page

Table 2.1 (continued)

Domain	Citation	Modalities Used	DL Techniques Used	Key Contribution	Main Challenge Addressed
	Milosevic et al. (2024) [91]	MRI + CT + Echo	CNN + Feature-level Fusion	Cardiovascular disease classification with cross-modality fusion	Modality-specific constraints, clinical variability, low real-world adoption
	Acosta et al. (2022) [92]	Clinical text + lab data	Knowledge Graph, Self-attention networks	Broader health monitoring and contextual decision support	Standardization of unstructured data, interoperability
	Li et al. (2023) [93]	Clinical text + imaging	Multimodal embedding learning, Transformers	Survey of fusion techniques in biomedical process acceleration	Cross-source integration and complexity
	Chen et al. (2024) [94]	Multimodal architectures	Situation-aware DL models	Integration of multimodal pipelines in ambient intelligence	Pipeline complexity, context modeling
	Jiang et al. (2021) [88]	Audio + Text	Cross-modal knowledge distillation (BERT to Speech Transformer)	Distillation of linguistic priors into speech-based models	Cross-modal representation mismatch

Table 2.2: Multimodal DL approaches in speech and mmWave applications.

Domain	Citation	Modalities Used	DL Techniques Used	Key Contribution	Main Challenge Addressed
Speech	Xiaoyu (2021) [95]	Audio + Text	Self-supervised distillation	Compact ASR models using large pretrained acoustic-text models	Transfer effectiveness, task-specific tuning
	Morency and Baltrušaitis (2017) [86]	Audio + Video	Multimodal fusion via Transformer/RNN	Robust speech recognition under noise and occlusion	Missing modalities at inference
	Yu et al. (2021) [96]	Audio + Video	Teacher-Student Fusion (AV-TS)	Cross-modal fusion for video classification	Synchronization, cross-modality fusion
	Hori et al. (2019) [97]	Audio + Video	Joint Student-Teacher Learning	Scene-aware dialog systems	Multi-turn coordination, modality dropout
	Gholami et al. (2024) [87]	Audio + Ultrasound	Latent Feature Distillation	Learning latent articulatory representations for compact, robust speech models	Generalization across speech domains
	Kerpicci et al. (2022) [98]	Audio + Ultrasound	Multitask KD with layer loss	Multi-representation distillation in multi-task setup	Efficient knowledge transfer

Continued on next page

Table 2.2 (continued)

Domain	Citation	Modalities Used	DL Techniques Used	Key Contribution	Main Challenge Addressed
	Yoon et al. (2021) [99]	Audio + Ultrasound	Dynamic weighting, Hybrid KD	TutorNet: frame-wise informative feature emphasis	Frame-level attention mismatch
	Chebotar and Waters (2016) [100]	Audio + Ultrasound	Soft target distillation	Single student from ensemble RNNs	Reducing model complexity while preserving performance
mmWave	Xu and Chen (2024) [101]	mmWave + Audio	CNN + RF signal enhancement	Through-wall and privacy-preserving voice sensing	Training/inference mismatch
	Ozturk et al. (2023) [70]	mmWave + Audio	AV fusion transformer	Enhancing noisy speech via radar-guided filtering	Radar-audio calibration, segmentation
	Wang et al. (2022) [22]	mmWave + Audio	Spectrogram Clustering + RF Sensing	Word detection via micro-doppler	Voice activity detection, occlusion
	Aytutuldu et al. (2024) [1]	mmWave + Ultrasound	Joint embedding network (CNN + LSTM)	Audio-only inference for phonetic recognition	Sensor alignment, modality transfer

Continued on next page

Table 2.2 (continued)

Domain	Citation	Modalities Used	DL Techniques Used	Key Contribution	Main Challenge Addressed
	Zhou et al. (2023) [102]	mmWave + Vision	Multi-view Fusion with Transformer/-PointNet	Robust 3D object detection and gesture understanding	View disparity, temporal synchronization
	Wei et al. (2022) [103]	mmWave + Vision	YOLO + mmWave fusion	Object detection in adverse environments	Modality imbalance, false fusion
	Patel and Heath (2024) [104]	mmWave + Comms	Beamformer DNN, Fusion MLP	Scene understanding and user tracking via Comm-Radar fusion	Latency, inference under mobility
	Jain et al. (2024) [105]	mmWave + Comms	Context-aware adaptive DL	CommRad: Efficient channel estimation with radar	Fusion latency, RF synchronization
	Wu et al. (2024) [106]	mmWave + Comms	Diffusion Model for Radar	Scene semantics extraction from radar	High-quality perception from sparse data

2.2. Multimodal Applications in Healthcare

Healthcare is a multimodal domain, where clinicians routinely combine radiological images, laboratory values, clinical notes, and patient histories to make informed decisions intuitively. Multimodal AI aims to replicate this integrative process by combining data across modalities to enhance diagnostic accuracy, contextual understanding, and clinical utility. Foundational surveys highlight the field's potential, with [82] noting that

multimodal deep learning is particularly suited to complex biomedical data landscapes where no single modality suffices.

Radiological images, laboratory data, clinical reports, and patient histories are frequently integrated by medical experts in healthcare to guide decision-making processes. The ability of multimodal AI to support this integrative process is enhanced by the fact that data fusion across modalities increases diagnostic accuracy, contextual understanding, and clinical value. As noted by [82], multimodal deep learning is particularly suited to complex biomedical data environments where a single modality is insufficient.

It has been demonstrated that combining multiple modalities such as imaging with structured or unstructured clinical data enhance the overall performance of the system [2]. The AI-driven process is similar to what human experts perform intuitively, advancing with the integration of image and text data such as psychology, radiology reports. Some studies demonstrated that current AI systems may under-perform when they are lack of integrating image data with EHRs or radiology reports [89]. This fusion improves diagnostic relevance, especially in ambiguous or subtle cases. Similarly, multimodal models achieve a consistent performance improvement—averaging over 6 AUC points—across a wide range of clinical tasks [90], yet emphasize that data incompleteness, misalignment, and cross-departmental silos pose major barriers to deployment.

Furthermore, [91] highlighted that although real-world usage remains limited for ultrasound and biomarkers, the combination of inputs such as MRI, computed-tomography (CT), and echo-cardiography demonstrated promising results in cardiovascular diagnostics. The recent study stated that multimodal AI provides an enhancement for broader monitoring and predictive capabilities in terms of personalized medicine [92].

Despite these advances, practical integration of multimodal AI into clinical workflows remains a significant challenge. Issues such as heterogeneous data formats, lack of interpretability, and insufficient standardization must be addressed before such systems can be safely and widely adopted in medical practice [93, 94]. These limitations underscore the need for models that are not only accurate but also adaptable, interpretable, and aligned with real-world clinical constraints.

2.3. Multimodal Applications in Speech Processing

In speech processing, multimodal learning has become a potent paradigm allowing systems to combine several information sources—audio, visual, articulatory, and linguistic data—such that their resilience, efficiency, and generalizing capacity are raised. Particularly in noisy or data-limited surroundings, when single-modal solutions fall short in offering enough dependability, this paradigm is rather important. Using complimentary information across modalities and frequently mutually reinforcing signals, multimodal systems can better replicate the way people see and interpret speech.

Knowledge distillation has lately become a crucial method enhancing the application of multimodal learning, particularly in situations when only some modalities may be applied in the inference phase. Modern methods see distillation not only as a model compression strategy but also as a basic way of intermodal representation transfer, in which case resource-limited learner models can transfer rich information from high-capacity or multimodal tutor models.

Recent studies reveal a remarkable result: knowledge distillation is more effective when it includes the latent representations in intermediate layers in addition to the final outputs. In particular, temporal and spectral connections, moving intermediate activations, allow learner models to acquire rich representations, including structural and hierarchical aspects of speech. In speech development applications, this method has proven to be very successful, allowing compact learner models with significantly fewer parameters to produce results close to the performance of the tutor model [87]. Distillations of self-supervised speech representation learning models such as wav2vec 2.0 or HuBERT to smaller architectures have achieved model reduction ratios of up to 75% while nearly maintaining performance on downstream tasks [98].

Multimodal distillation, which goes beyond model compression, also has the potential to increase robustness in sequential tasks. For example, in cascade systems combining automatic speech recognition (ASR) and machine translation, distilling knowledge from teacher models trained on clean text to student models working on noisy ASR outputs has been shown to reduce the negative effects of transcription errors and consequently improve translation quality [107]. Such robustness is critical in real-world applications where upstream errors in modular systems are often carried over to the

next stages.

In addition, multimodal distillation enables cross-domain knowledge transfer, where knowledge acquired in one modality or domain drives the learning process in a different domain. For example, transferring syntactic and semantic information from a BERT language model to a speech converter model has been shown to significantly improve intent classification performance. This highlights the importance of utilizing text-based linguistic information in spoken language understanding systems [88].

To enable efficient and usable automatic speech recognition (ASR) systems, many studies have investigated task-oriented knowledge distillation strategies. Among these strategies, the use of specialized loss functions to preserve full output distributions in transformer-based ASR models stands out [95]. Such methods allow compact learning models to preserve not only the performance accuracy but also the structural properties of large-scale self-supervised tutor models.

From a broader perspective, the flexibility of cross-architecture knowledge distillation processes is increased by frameworks such as TutorNet, which provide guidance at both the representation level and the softmax output level, and dynamically weight frames according to the attention distributions of tutor models. This approach allows learning models to focus more effectively on the most informative parts of the data [99]. Early work has shown that transferring knowledge from RNN-based ensemble models to a single model can significantly reduce complexity while maintaining performance [100].

2.4. Multimodal Sensing with mmWave Radar

Because of its unique characteristics—robustness against obstacles, great sensitivity in micro vibration detection, and robustness under strong light or sound conditions—multimodal sensing employing millimeter wave (mmWave) radar has attracted increasing attention recently. Effective integration of mmWave radar with other sensing modalities offers a strong platform that may greatly extend environmental and human perception in sectors including healthcare, automotive systems, and communications, as reported in [17].

As shown in multi-user beamforming applications whereby radar and auxiliary sensor

inputs improve channel and beam area estimation [104], deep learning-based sensor fusion methods have great potential in obtaining structured and meaningful representations from multimodal data. By use of vocal cord vibrations and lip movements, it has been verified that mmWave radar may be employed as a sensor capable of capturing both speech and visual inputs. This functionality helps the radar to give dependable and strong performance even behind walls or in non-line-of-sight environments [101]. Both DiffRadar [106] which uses diffusion-based probabilistic models to produce high-quality semantic scene representations using radar data and CommRad [105], which combines radar detection capabilities with wireless communications to establish highly efficient and directionally adaptive links in obstructed mobile environments, clearly show the detection flexibility of radar.

Advances in distributed sensing, radar-camera fusion, and autonomous driving further exemplify the integration of radar with vision and other modalities for dense, robust 3D object understanding [102, 103]. These efforts collectively demonstrate that mmWave radar is not merely a substitute for other sensors but a potent contributor to multimodal systems, capable of enhancing perception when harmonized with vision, audio, or contextual priors. Ultimately, this line of work exemplifies the transformative potential of multimodal integration, where the fusion of complementary signals yields richer, more adaptable, and human-aligned machine perception.

2.5. Human Sensing using mmWave Radar

In this section, we will describe an automated human sensing system using raw mmWave radar signals which is the first mmWave-based system we developed in this study. Basically, the system observes human in a room occupancy via mmWave radar for indoor counting (we used TI Instruments IWR1843BOOST in our experiments) and collects data for action recognition . The proposed system senses on-site objects by analyzing raw mmWave radar data. The system analyzes raw mmWave radar data to detect objects and determine whether they exhibit displacement (distance) or motion (velocity). The system also analyzes human actions to categorize them in a supervised way to learn common human functional movements.

2.5.1. Introduction

Recent advancements in mmWave radar technology have underscored the efficacy of radar sensors as a complementary modality to visual perception-based methods. The compact structure and advanced imaging capabilities of radar sensors have been demonstrated to be effective solutions, particularly in low-visibility environments and inclement weather conditions. Due to a significant reduction in the size of mmWave radar hardware, radar technology is becoming a powerful sensing platform that improves upon other methods such as cameras and LiDAR. Unlike cameras, radar is resistant to bad weather conditions and is capable of object detection while maintaining individual privacy [108, 109].

Nevertheless, numerous studies have overlooked the sophisticated functionalities of imaging radars, opting for overly simplistic approaches that represent radar data with a single detection point for moving objects [110, 111]. Radars offer significant advantages in distinguishing the speed and micro-level movements of targets thanks to their innate sensitivity to phase shifts. This feature, known as micro-Doppler motion, serves as a distinctive signature for identifying specific objects or types of motion due to its ability to extract precise motion features from phase shifts. The fact that each target has its own unique micro-motion pattern is important because it demonstrates that this information can be integrated into existing target classification and recognition methods and provide valuable contributions [112, 113].

In the realm of radar sensing, the mechanical vibration and rotation of structures within a target can introduce frequency modulation to the returned signals [114, 115]. This phenomenon results in the creation of sidebands around the center frequency of the target's body Doppler frequency . The modulation induced by vibration and rotation, producing lower and higher frequencies relative to the Doppler frequency, is referred to as micro-Doppler. Given that each form of movement is unique in terms of the rotation and vibration of various parts of objects, micro-Doppler serves as a distinctive feature for the classification of different signals [116].

This chapter presents the research conducted to deepen the understanding of mmWave radar technology through the analysis of radar-based motion and velocity representations. Additionally, speech analysis experiments were carried out on the mmWave

radar sensor using those techniques. To identify the optimal operating parameters for speech-related tasks under different configurations, several calibrations of the radar sensor were performed. An experimental setup was established for data collection, revealing the need for careful tuning of numerous data acquisition and processing parameters. The details of the studies conducted throughout this process are outlined in the following sections.

2.5.2. Methods

2.5.2.1. Basic Principles of FMCW Radar

Particularly those using FMCW methods, MmWave radar systems provide a strong basis for motion sensing, range estimate, and velocity analysis. These systems broadcast frequency-modulated signals called chirps, then examine their reflections to define the spatial and dynamic characteristics of objects in the radar field.

A chirp is a single-frequency tone with linear over-time frequency rise. As shown in Fig. 2.1, a frequency synthesizer (1) generates the chirp which is then sent via the transmit (TX) antenna (2). Reflecting off a target, the signal is picked at the receive (RX) antenna (3) and combined with the original broadcast chirp to generate an intermediary frequency (IF) signal (4). The IF signal records the frequency variances between the sent and received chirps. The time delay brought forth by the range of the target determines this discrepancy. Target range may be estimated by first converting this signal into the frequency domain using a Fast Fourier Transform (FFT) applied along the fast time (analog-to- digital converter) sampling dimension. This technique is called the Range-FFT. As Fig. 2.2 shows, the frequency spectrum's peak locations exactly coincide with object range (5). Two chirps sent with a specified time gap can be used to determine the velocity of the object by means of phase changes between related reflections. The Doppler effect lets this happen.

The basic unit of radar data capturing is a succession of chirps forming a frame. The range resolution of each chirp—defined as the rate of change in frequency over time—determines the system's slope. While Fig. 2.3b shows the structure of a radar frame, wherein many chirps are sent consecutively, separated by specified intervals, Fig. 2.3a

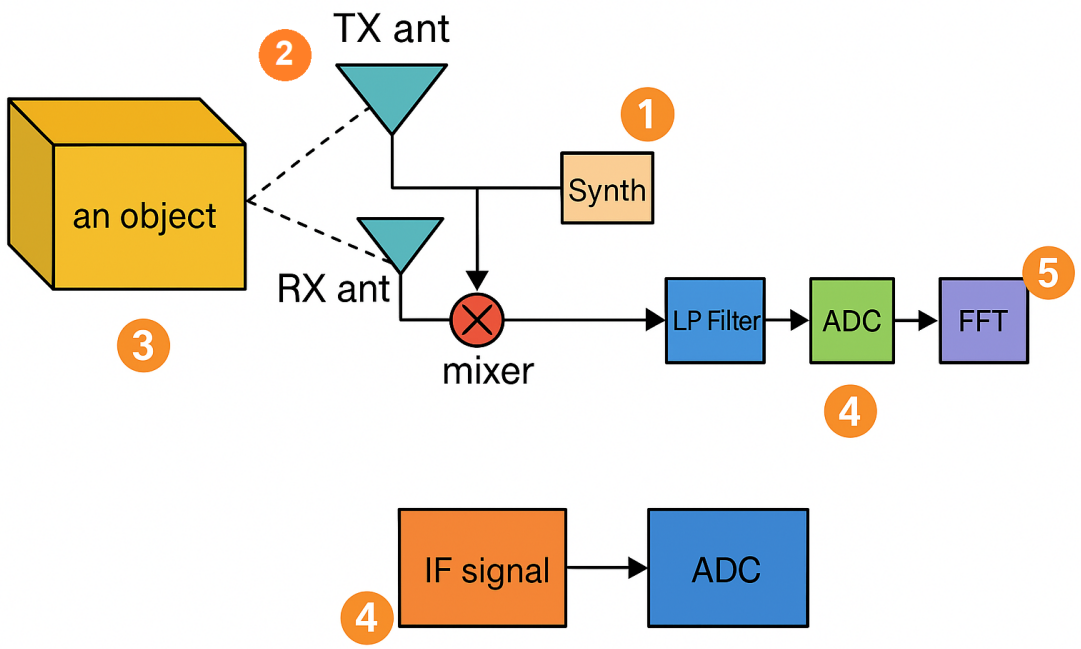


Figure 2.1: Overview of real-time processing.

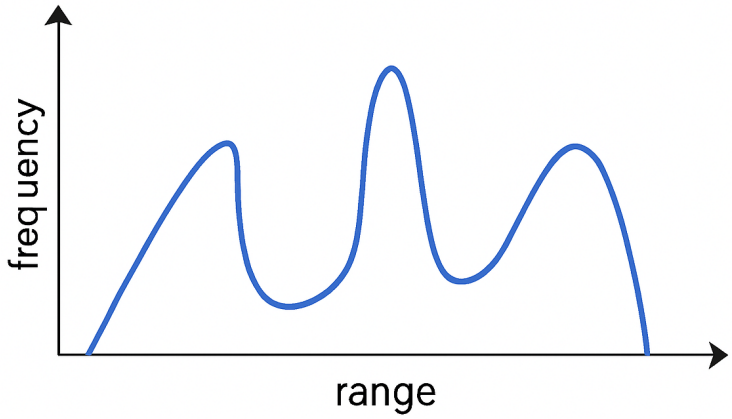
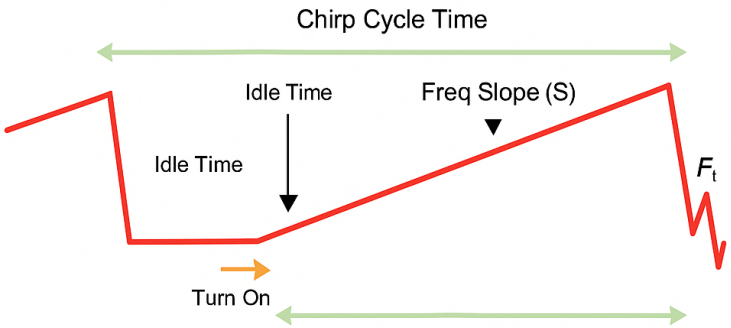


Figure 2.2: Range-frequency peaks.

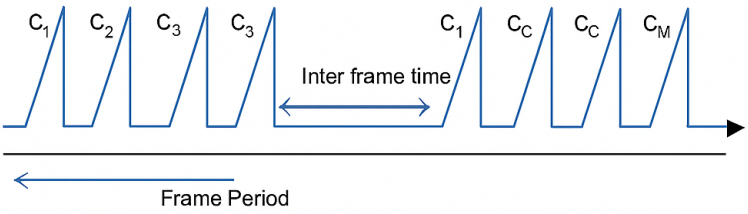
shows the temporal features of a normal chirp.

Signal Processing Pipeline: The processing of mmWave radar data involves multiple steps:

- Chirp and Frame Structuring: Each frame consists of multiple chirps separated by a constant interval.



(a) Typical FMCW Chirp



(b) Typical Frame Structure

Figure 2.3: FMCW radar signal structure: (a) A typical chirp showing frequency sweep over time, (b) A frame composed of multiple chirps with inter-frame timing.

- Range-FFT: Performed on each chirp to extract range information.
- Doppler-FFT (2D-FFT): Applied across chirps within a frame to resolve velocity. This process is particularly useful for detecting multiple objects at the same range but with different speeds.
- Micro-Doppler Analysis: By focusing on specific range bins and aggregating phase data over time, fine-grained movement patterns (micro-Doppler signatures) can be extracted.

The position and velocity of the object directly affect the IF signal phase and frequency; a little movement (e.g., 1 mm) in front of the radar causes a phase change of 180°. Sub-millimeter displacement detection is made possible by the IF signal's changing frequency in direct line with distance travelled. These concepts allow mmWave radars to offer complete motion analysis even in difficult environments with many moving objects.

2.5.2.2. Range-Doppler and Micro-Doppler Analysis

Because of its low visibility, small size, and ability to detect target velocity and micro-Doppler components with great precision, MmWave radars have notable benefits over other technologies [117, 118]. These characteristics help the radar to run consistently even in congested and complicated metropolitan surroundings. In addition, its small size and power-efficient structure make it portable and suitable for use in long-term applications [119, 120].

Range-Doppler is a radar analysis method used to measure both the distance and speed of an object (Fig. 2.4). This method calculates the distance of the object to the radar and the speed of its movement based on the data obtained as a result of the interaction of the radar signals with the target. This technique, which is generally preferred for tracking moving objects, reveals the general dynamics of the object. Micro-Doppler analysis, on the other hand, is more suitable for studying the internal structure of the target or its structural characteristics, such as vibrations. [121, 122].

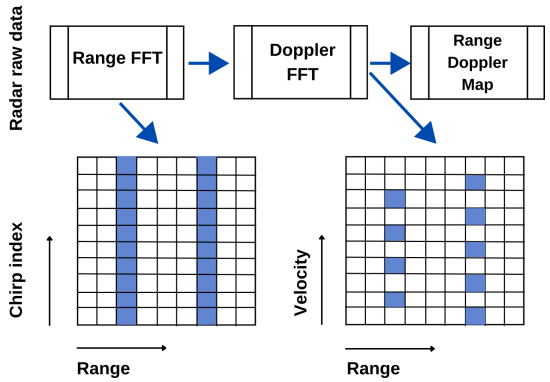


Figure 2.4: Data processing chain: Obtaining distance and velocity for each sequenced radar frame [108].

Using at least two chirps with a T_c spacing, the velocity can be determined by considering total chirp time. Although having different phases, the range-FFTs matching each chirp will show peaks at the same places. The observed phase difference $omega$ matches the motion of an item vT_c. $\omega = \frac{4\pi v T_c}{\lambda}$ which may be rearranged to get the velocity (v) as $v = \frac{\lambda\omega}{4\pi T_c}$. Therefore, the estimate of the object's velocity is made possible by using the phase difference recorded across two successive chirps.

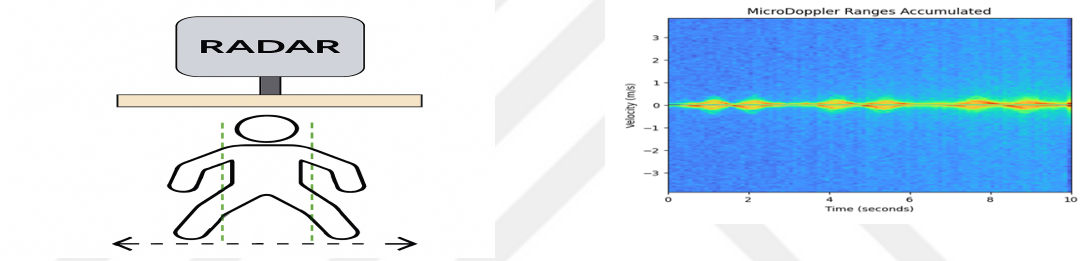
2.5.2.3. Pilot Setup for Human Sensing

Table 2.3: mmWave Radar IWR1843BOOST Configuration Parameters with Typical Values

Parameter	**Units**	**Typical Value / Range**	**Description**
Start Frequency	GHz	76 – 81	Starting frequency of the chirp; determines radar band (e.g., 77 GHz).
Frequency Range	GHz	4	Total bandwidth across the chirp (e.g., 77–81 GHz).
Bandwidth	MHz	250 – 4000	Larger bandwidth gives finer range resolution.
Slope	MHz/μs	20 – 80	Chirp slope; higher slope improves range resolution but increases ADC requirements.
Idle Time	μs	5 – 100	Time between chirps; affects frame timing.
ADC Start Time	μs	2 – 10	Delay after chirp starts before sampling begins.
Ramp End Time	μs	50 – 200	Duration of chirp including frequency ramp.
Sampling Rate (Fs)	Msps	5 – 25	ADC sampling frequency.
Samples per Chirp	-	128 – 1024	Number of ADC samples per chirp.
Number of Chirps per Frame	-	32 – 128	Defines the number of chirps in one radar frame.
Number of Transmit Antennas	-	1 – 3	Impacts angle estimation resolution.
Number of Receive Antennas	-	1 – 4	Used for angle of arrival estimation.
Transmit Power	dBm	10 – 15	Power level of the transmitted signal.
Transmit Gain	dB	0 – 50	Amplifies the outgoing signal.
Receive Gain	dB	0 – 30	Amplifies the incoming signal.
Maximum Detection Range	m	2 – 100+	Depends on chirp design, gain, and object reflectivity.
Range Resolution	m	0.04 – 0.3	Determined by bandwidth: $\Delta R = \frac{c}{2B}$.
Maximum Velocity	m/s or km/h	10 – 300 km/h	Limited by PRI and number of chirps.
Velocity Resolution	m/s or km/h	0.1 – 5 km/h	$\Delta v = \frac{\lambda}{2T_{frame}}$.
Frame Duration	ms	10 – 100	Duration of a full radar frame.

mmWave Radar Configurations: We visited calibration of the sensor to find the optimal parameters of the radar for sensing tasks in a series of configurations. These configurations have been chosen to reflect human sensing tasks such as human counting and human action recognition. We observed that numerous parameters related to data collection and processing required careful adjustment to ensure optimal performance.

To initiate this process, we began with a straightforward configuration in which participants produced various speech scenarios. This approach enabled us to identify optimal calibration parameters and facilitated effective visualization of the raw data. In addition, we conducted simplified tests involving inanimate objects moving along predictable, periodic trajectories. After finalizing the data collection parameters, we proceeded to gather speech-related sound analysis data and evaluated the performance of raw data processing algorithms. We have configured the radar as its samples rate (ksps) 10000, frequency slope (MHz/µs) 29,982, with 2 TX antennas and 3 RX antennas, 10 frames, 200ms periodicity, 128 number of chirp loops, and 128 ADC samples. Table 2.2 summarizes the key configuration parameters of mmWave radar systems along with their typical values used in short-range sensing applications.



(a) Radar setup and subject performing repetitive side-stepping while maintaining original position.

(b) Micro-Doppler signature of repetitive side-stepping. Distinct bursts appear symmetrically around zero velocity due to lateral leg movements.

Figure 2.5: Radar-based analysis of repetitive side-stepping motion. In this movement pattern, the subject remains in place while rhythmically stepping side-to-side.

Experiments with mmWave Radar

Micro-Doppler signatures keep distinct and recognizable patterns in both single and multiple person scenarios, as demonstrated by Fig. 2.5 and Fig. 2.6. The Doppler signals from one or two people walking side by side to the radar stay at a low frequency but show different energy traces. In Fig. 2.5, likewise, symmetric bursts centered at zero velocity are produced by repeated lateral motion of a single individual. Especially, these micro-Doppler patterns are stable across several kinds of motion and repeated trials,

Algorithm 1: 2D Micro-Doppler Spectrogram Generation

Input: ADC data cube (F, C, A, S) where F = frames, C = chirps, A = antennas, S = samples

Output: 2D micro-Doppler image (T, V) where T = time, V = velocity bins

for each frame $i = 1$ to F **do**

  Perform range FFT along samples (S) with Blackman window

  Output shape: (C, A, R) where R = range bins

  Perform Doppler FFT across chirps with Hamming window and clutter removal

  Result: detection matrix $D_i \in \mathbb{R}^{V \times R}$

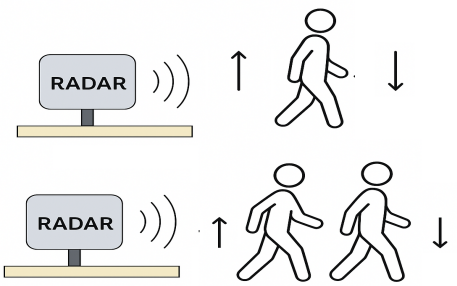
  Apply FFT shift on velocity axis

  Store D_i into micro-Doppler cube $M[i, :, :]$

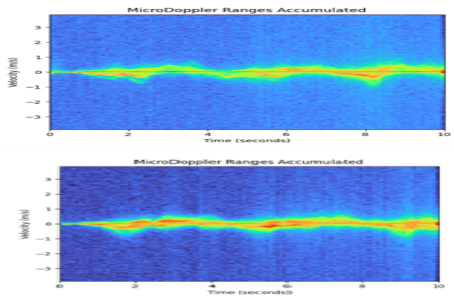
end for

Compute average over range: $M_{2D}[t, v] = \frac{1}{R} \sum_r M[t, v, r]$

Save or visualize M_{2D} as a 2D image (time × velocity)



(a) Illustration: Single-Multiple person walking parallel to the radar.



(b) Micro-Doppler of a Single/Multiple person moving perpendicularly.

Figure 2.6: Movement parallel to the sensor (One and Two People). In both cases — (top) single person and (bottom) two people — the individuals walk parallel to the radar's position.

which helps to identify persons and activities precisely. Whether the motion involves one or more persons, the persistence of these unique micro-Doppler signatures emphasizes the possibilities of radar for human activity detection in practical, multi-person surroundings. Algorithm 1 provide the exact methods for producing 2D micro-Doppler representations.

We first set up to observe basic harmonic motion in raw radar data. We used a lightweight pen fastened to the end of an extended flexible support, like a long thin rod. Fig. 2.7a-b show the simple harmonic motion of a hanging pen where the radar records symmetric, smooth micro-Doppler patterns centered on zero velocity. Figures 2.7c-d show a

human traveling back and forth along the radar's line of sight, producing alternating positive and negative Doppler changes as they approach and retreat. Fig. 2.7e-f shows a person regularly conducting a sit-to-stand action, which produces less symmetric but nonetheless periodic Doppler signatures because of vertical body movement. Fig. 2.7g-h shows two people traveling in different directions—one toward the radar and the other away—producing overlapping but distinctively micro-Doppler patterns with opposite frequency changes. These illustrations show how, depending on their individual micro-Doppler properties, radar can identify and separate distinct human and object movements.

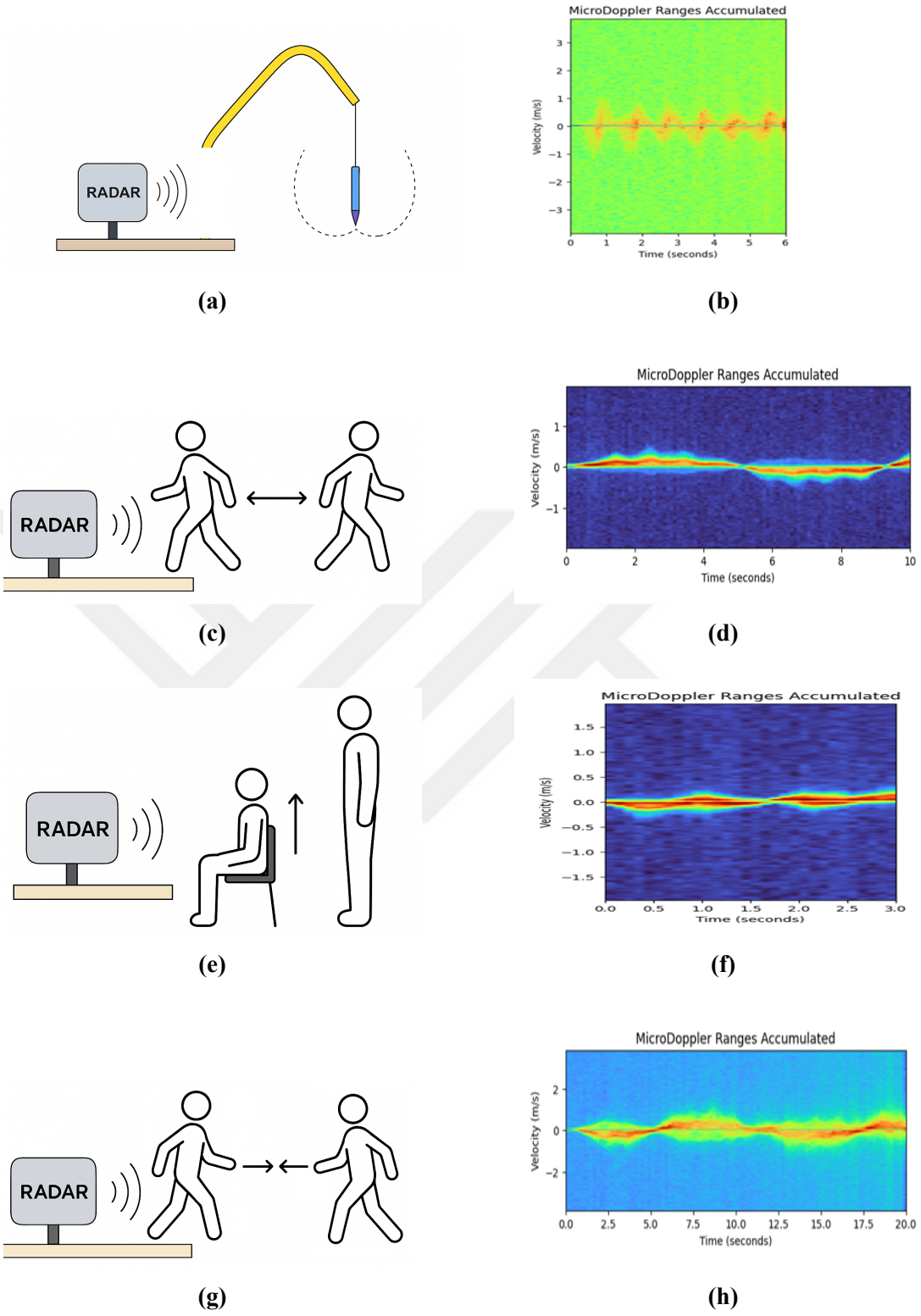


Figure 2.7: Radar-based observation of different types of human and object motion and their corresponding micro-Doppler signatures. (a–b) Simple harmonic motion of a suspended pen observed by the radar. (c–d) A person moving back and forth along the radar's line of sight. (e–f) A subject performing repeated sit-to-stand actions. (g–h) Two individuals walking in opposite directions — one approaching and the other receding from the radar.

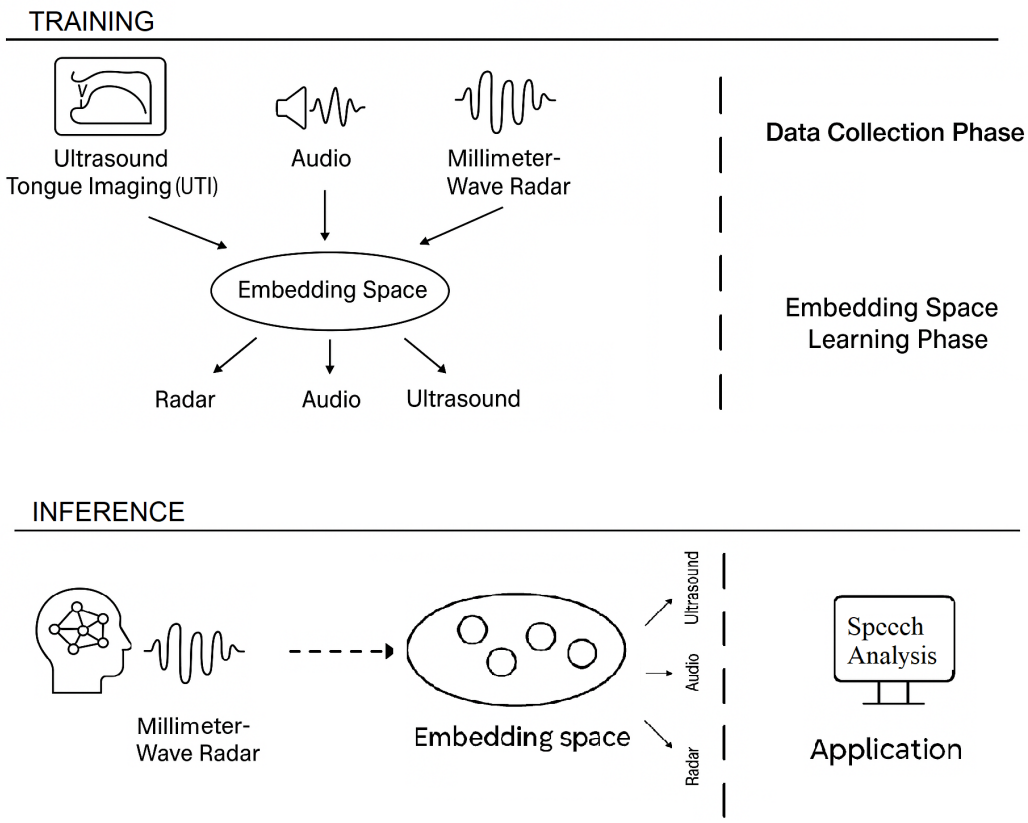


Figure 2.8: Overview of the proposed system comprising three phases: data collection using synchronized UTI, audio, and mmWave radar; embedding space learning with multimodal supervision; and application using radar-only input for downstream tasks such as speech recognition.

2.6. A Multimodal System for Speech Analysis

In this section, we will describe our multimodal speech sensing and analysis system using raw mmWave radar signals which ifs mmWave-based system we developed in this study. Basically, the system observes generated signals in front of mmWave radar for phoneme recognition (we used TI Instruments IWR1843BOOST in our experiments). The proposed system captures both articulatory movements and vocal cord vibrations by analyzing raw mmWave radar, UTI and audio.

2.6.1. The Overall Structure

The overall structure of the proposed system is shown in Fig. 2.8. The system consists of 3 phases; training, embedding space learning, and application phases. In the data

collection phase, we synchronized multiple sources to collect data. During the embedding space learning, the data-driven system trained the collected data, and optimized an embedding space. The radar-only model then utilizes this embedded representation for inference, enabling robust and privacy-preserving predictions using a single modality.

2.6.2. Tongue Contour Extraction as a Foundational Study

The increase in computer hardware, the production of multi-core, powerful graphics cards and their widespread use have significantly increased the speed of computing with computers. As a result of these developments, deep neural networks, which have been known for a long time but cannot be used in practice due to the high cost of computation and the lack of computers to do this, have become popular. To establish a foundation for later experiments, we first investigated the effectiveness of various deep learning models on the task of tongue contour segmentation from real-time ultrasound data. The qualitative performance of the implemented architectures on ultrasound tongue contour segmentation is illustrated in Fig. 2.9, where the proposed model exhibits more accurate and consistent predictions compared to sDeepLab, BowNet, and sU-Net.

sDeepLab is a streamlined adaptation of the DeepLab v3+ architecture, which leverages atrous (dilated) convolutions to increase the receptive field without reducing spatial resolution. In our implementation, we removed the dependency on large-scale pre-trained models (e.g., ImageNet) and instead optimized the model for grayscale ultrasound inputs. BowNet is a custom dilated convolutional network developed specifically for ultrasound tongue contour extraction. Inspired by prior work on peripheral vision and multi-scale context aggregation, BowNet employs a bow-shaped architecture using successive layers of dilated convolutions. This design helps capture both local edge structures and broader anatomical context. Our experiments confirmed BowNet's strength in balancing accuracy and inference speed, particularly under noisy imaging conditions. However, the model occasionally generated artifacts near the tongue root, likely due to echo patterns and shadows misinterpreted as salient features [123].

sUNet is a simplified U-Net variant tailored for efficient training on our modest-sized ultrasound dataset. Unlike standard U-Net implementations, sUNet uses fewer encoder-decoder levels and reduced filter depth to allow faster convergence and lower compu-

tational demands. This architecture proved to be robust across multiple subjects and conditions, making it an effective baseline model. Its skip connections preserved spatial information well, but it occasionally failed to separate close-proximity tongue contours from overlapping structures in the jaw or palate [124].

Figure 2.9: From left to right: original ultrasound image, manually annotated ground truth, and predictions from sDeepLab, BowNet, sU-Net, and our proposed pretrained U-Net based model. While sDeepLab struggles with noise and incomplete contours, BowNet and sU-Net show moderate accuracy. The proposed method demonstrates superior delineation quality, closely matching the ground truth across various subjects and imaging conditions.

2.6.3. Pilot Setup for mmWave Speech Sensing

We set up a radar-camera synchronized data-collecting system by using Texas Instruments' mmWave Studio platform. The customized Lua scripts controlled the radar's settings and automated the acquisition process, including hardware triggering and start/stop instructions. The development of a Python-based multi-threaded video recording module allowed real-time visual alignment. This module ensures that the timing of RGB video frames matches up with radar data, allowing for precise matching of human actions seen on camera and micro-Doppler signatures. This setup provides a solid foundation for combining different types of data and using machine learning, allowing for easy testing of various movements and environmental factors.

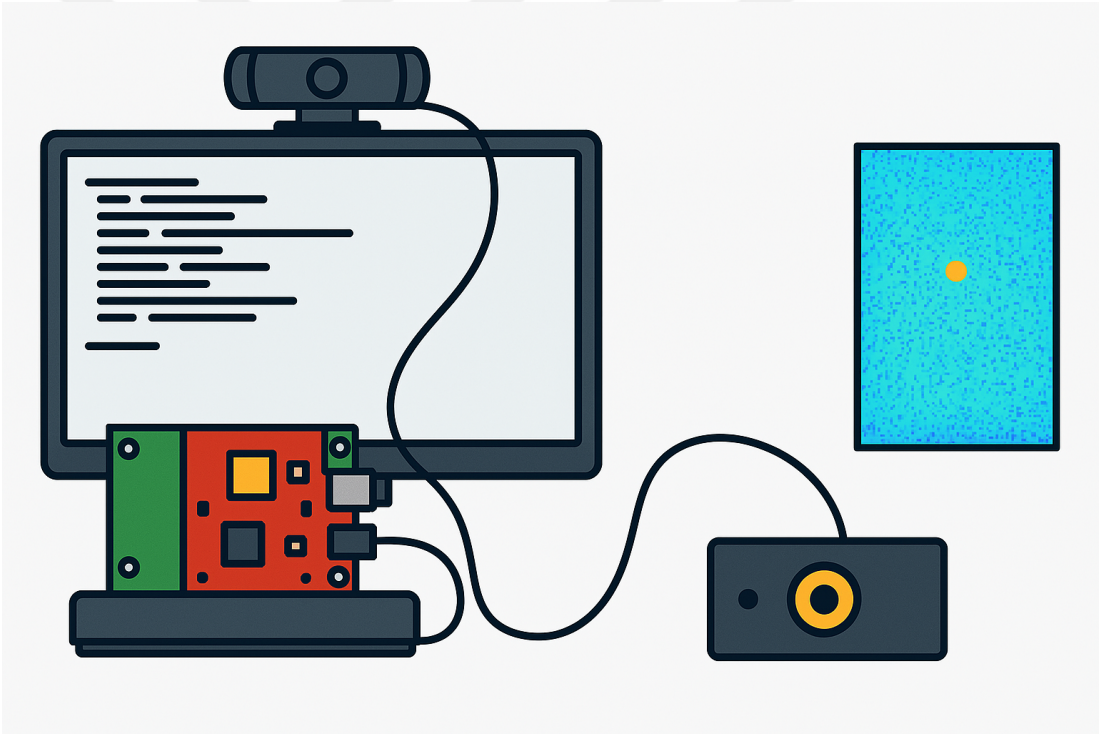


Figure 2.10: The setup for synchronized radar and camera acquisition. The illustration shows a subject positioned in front of a radar and camera system, while a spectrogram is generated in parallel. It emulates a real-time sensing environment using mmWave Studio, custom Lua scripts, and Python-based multi-threaded video capture.

The initial speech sensing investigations employing micro-Doppler signatures demonstrated limited classification performance. The spectrograms derived from radar data

exhibited an absence of discernible, replicable patterns across multiple spoken utterances. Speech, in contrast to physical activities such as walking or waving, generates delicate and fine-grained motions, which are challenging to consistently record using micro-Doppler, particularly at low signal-to-noise ratios and in the absence of exact alignment. The resulting generated spectrograms lacked discernible temporal or spectral structure, hindering the capacity to distinguish phonemes or words. While some studies have demonstrated the feasibility of using micro-Doppler data for speech recognition [78], our results demonstrate that, in this configuration, micro-Doppler signatures alone are inadequate for accurate speech classification (Fig. 2.11). These findings underscore the necessity of incorporating diverse representations or modifying the experimental design to enhance discriminative capacity.

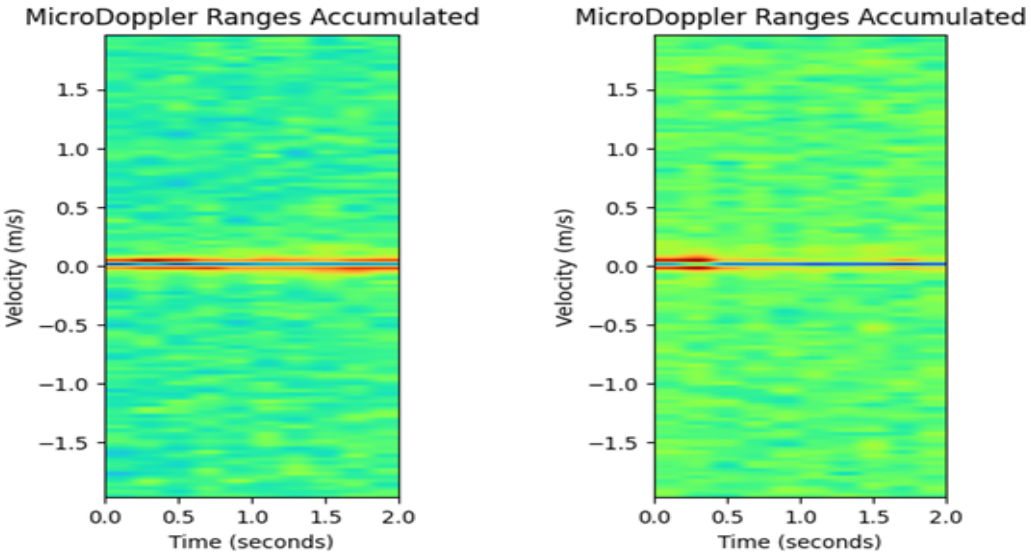


Figure 2.11: Range-accumulated micro-Doppler spectrograms for different speech recordings: the left plot corresponds to the vowel /a/ and the right to the vowel /e/. Each spectrogram shows a 2-second radar capture during speech articulation. The velocity patterns are weak and not clearly distinguishable between the two vowels, supporting the observation that micro-Doppler alone does not provide sufficient discriminatory information for isolated vowel classification.

mmWave Radar Configurations

The mmWave radar setup was modified to enhance system performance and data col-

lection quality. This modification entailed the customization of several pivotal factors, including the number of transmitting and receiving antennas, the frequency slope, chirp periodicity, and the sample rate. Through experimental analysis, it was determined that minor alterations in these parameters can significantly influence downstream signal processing and data collection processes. As indicated in [70], each frame in our configuration runs at a chirp repetition rate of 1 kHz with 256 ADC samples per chirp. This arrangement computes the sampling rate F_s by $F_s = \frac{N}{T}$, where $N = 256$ is the number of samples per chirp and $T = 0.001$ seconds is the chirp duration, hence producing a sampling rate of 256 kHz. This arrangement was found to be appropriate for obtaining the fine motion dynamics required for our use.

A speaker device was positioned 50 cm in front of the mmWave radar and used to emit a sweep sound ranging from 20 Hz to 20,000 Hz (Fig. 2.12). This setup enabled us to evaluate the radar's sensitivity to frequency-varying acoustic signals in air. The known distance and controlled input allowed for consistent testing and spectrogram analysis of the radar's response to airborne sound waves.

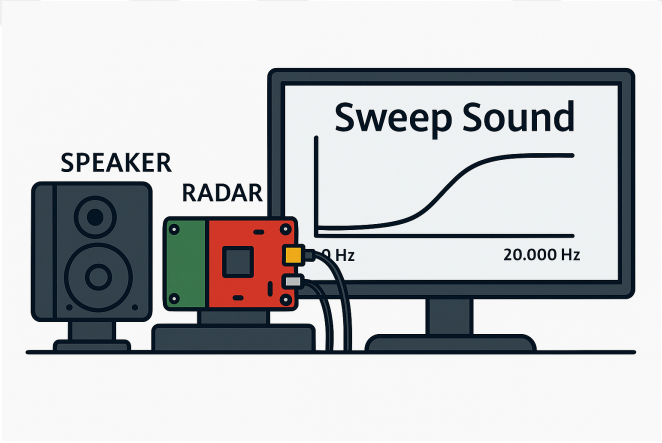
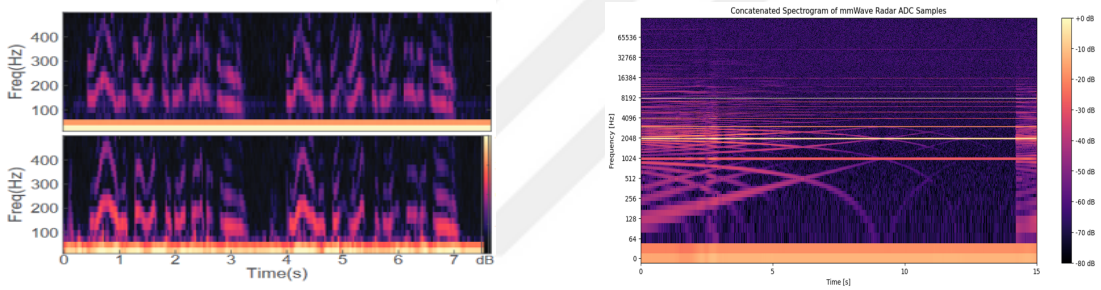


Figure 2.12: Experimental setup illustrating the placement of the speaker device 50 cm in front of the mmWave radar, configured to emit sweep sounds from 20 Hz to 20,000 Hz, facilitating controlled evaluation of the radar's sensitivity to frequency-varying acoustic signals.

In the early phase of our experiments, the radar was configured with moderate values across a range of parameters, including 2 transmit and 3 receive antennas, 10 frames, 128–256 chirp loops, and 128–256 ADC samples per chirp, with a periodicity between

70–200 ms. However, this configuration proved suboptimal for high-resolution temporal tracking, particularly in speech sensing applications. As a result, we revised the settings to significantly improve temporal granularity and control. The updated configuration reduced the system to a single transmitter and four receivers, increased the number of frames to 5000 (can be changed by duration of data capture), and fixed both the number of chirp loops and ADC samples at 10 and 256, respectively. Most notably, the frame periodicity was reduced from a broad 70–200 ms range to a fixed 1 ms, allowing for denser and more temporally precise sampling of micro-movements. This revised setup enhanced the radar's ability to capture fine-grained articulatory motion required for our application, without altering the ADC sample rate (10 Msps) or the frequency slope (29.982 MHz/µs).

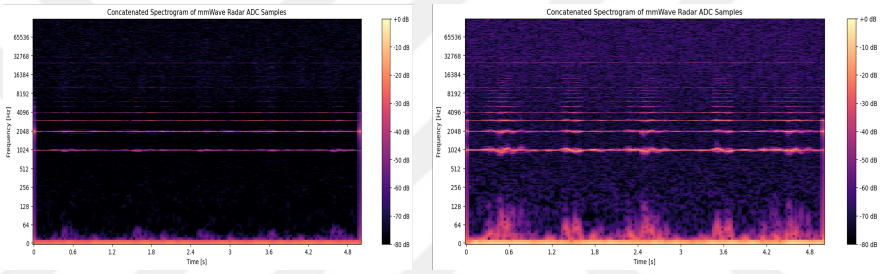


(a) Acoustic baseline [125]. **(b)** Radar-based sweep sound.

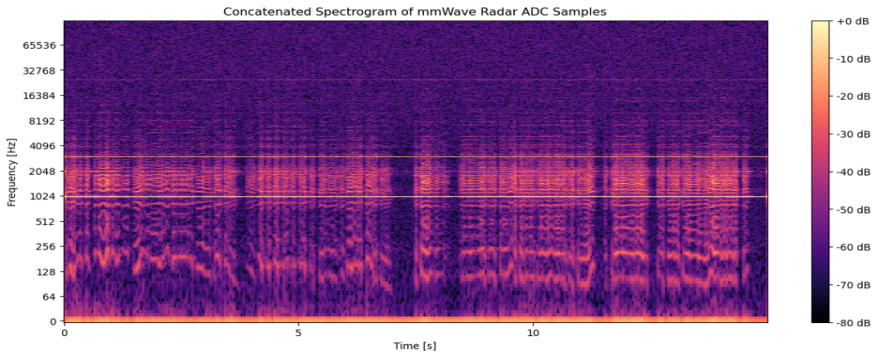
Figure 2.13: (a) Spectrogram of speech recorded from [125]. Distinct harmonic structures are visible across time, enabling speaker device and throat discrimination. (b) Spectrogram of mmWave radar ADC data captured during a 20–20,000 Hz sweep sound test. Harmonic curves reveal how frequency sweeps are reflected and captured across time.

Fig. 2.13 compares spectrograms from different acoustic and radar-based settings. An acoustic spectrogram of speech from [125], where voiced sounds produce well-structured harmonic bands is shown in Fig. 2.13a. Inspired by this, we tested our radar system using a controlled audio sweep from 20 Hz to 20 kHz. As seen in Fig. 2.13b, the radar-captured ADC data reveals curved harmonic patterns in the time-frequency domain, indicating that the radar is responsive to air-coupled audio signals under certain conditions. Finally,these comparisons validate the feasibility of radar-based acoustic signal sensing and guide future configurations for more speech-relevant experiments.

While our work shares conceptual similarities with the MILLIEAR pipeline [126], our spectrogram generation approach differs in execution. Instead of applying Range-FFT and Doppler-FFT to isolate a specific vibrating range bin, we directly concatenate radar frames along the time axis to form a continuous signal. This raw micro-Doppler signal, accumulated across range bins, is then processed using Short-Time Fourier Transform (STFT) to generate the time-frequency domain spectrogram. This approach simplifies the pipeline and provides a global view of speech-induced motion over time. Although less targeted than MILLIEAR's range-selective method, our method is effective for general speech activity sensing and aligns well with the constraints of real-time data acquisition and low-complexity processing [125].



(a) Human speech with varying radar distances: too close (left), further (right)



(b) Radar-based speech recording with speaker device results in high frequencies.

Figure 2.14: Radar-based speech data collection setups with human and speaker device.

Fig. 2.14 illustrates two radar-based speech sensing setups. In Fig. 2.14a, a speaker repeats the phrase "Adana Mersin Adana Mersin Adana" over five seconds while an mmWave radar captures motion dynamics. Variations in distance between the speaker

and the radar affect the resulting spectrogram. In Fig. 2.14b, a speaker device plays a 15-second common speech sourced from YouTube, recorded using a 1 ms chirp interval, resulting in 15,000 frames. These frames are concatenated to generate the final spectrogram.

2.6.3.1. Experiments with mmWave Radar

Ultrasuite Repository Captured by mmWave Radar

In the preliminary process, we focused on capturing articulatory patterns of selected consonants—/k/, /p/, /r/, and /s/—based on their place and manner of articulation. These consonants were chosen to represent diverse articulatory mechanisms, including bilabial (/p/), alveolar (/s/, /r/), velar (/k/), and retroflex (/r/), as well as fricative sounds (/s/). The total number of training and testing samples collected across all modalities is summarized in Table 2.3.

To ensure consistency and reproducibility in the radar reflections, a loudspeaker setup (Fig. 2.12) was used to play pre-recorded consonant utterances. The utterances were sourced from the UltraSuite repository, which provides high-quality audio recordings aligned with articulatory targets. These '.wav' files were played through the speaker in a controlled environment while the radar captured the resulting signal reflections.

Table 2.4: Distribution of training and testing data across modalities for selected consonants.

Consonants (Articulation-based)	Number of Train Data (All Modalities)	Number of Test Data (All Modalities)
/k/	619	166
/p/	1278	343
/r/	1199	329
/s/	583	164

The concatenated spectrograms of ADC samples acquired from the mmWave radar for various speech utterances played over a speaker demonstrated in Fig. 2.15. The spectro-

grams in the bottom row show more consistent and periodic spectrum activity, implying maybe stronger radar reflections from louder or more ordered sound segments. The top spectrograms, on the other hand, show somewhat less and more scattered patterns that indicate smaller articulatory motions. These time-frequency representations are essential input elements for deep learning-based classification tasks as they allow the study of speech-sensing micro-movements recorded by the radar.

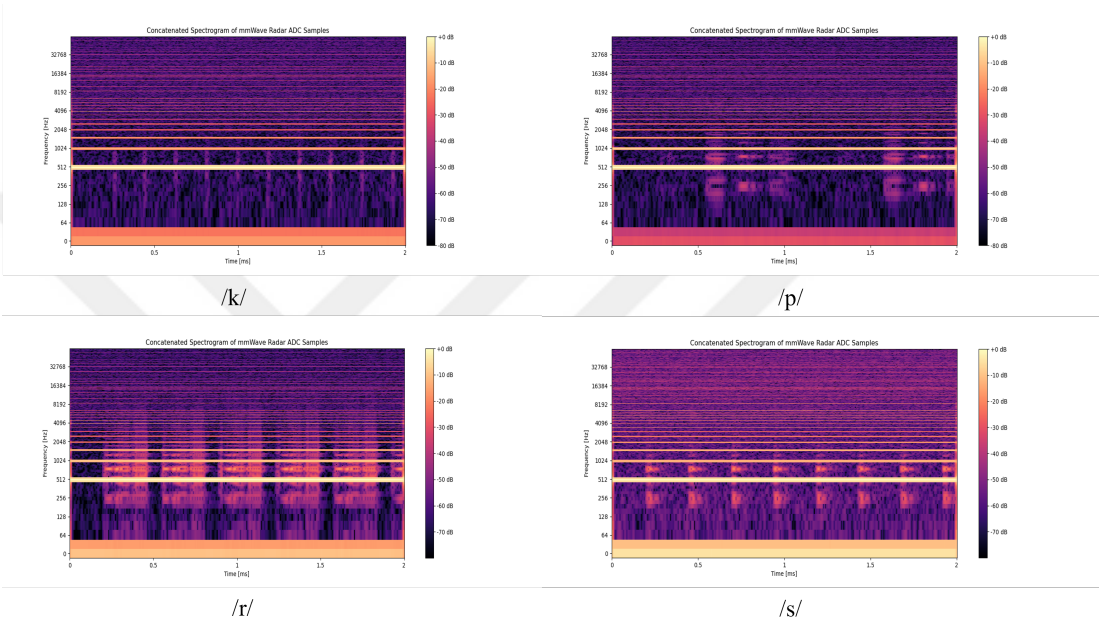


Figure 2.15: Concatenated spectrograms illustrating radar reflections from various speech utterances.

In this study, mmWave radar spectrograms were initially produced by subjecting voice recordings to a process of signal transmission through a speaker, followed by the acquisition of reflected signals over a designated 2-second interval. To enhance signal clarity and ensure adequate representation of radar reflections, each recording incorporated numerous repetitions of the same consonant sound. However, this approach gave rise to a potential challenge for deep learning models: the presence of recurrent patterns within a constrained time frame could lead to temporal overfitting. The potential outcome of this process is the development of models that are capable of linking specific attributes of each phoneme with predetermined locations within the spectrogram, as opposed to the more nuanced and complex process of learning the underlying articulatory traits of each individual letter. To address this challenge, we employed a

data augmentation approach by dividing the continuous two-second spectrograms into smaller, repeat-aware segments, which were then matched with individual utterances. The enhanced precision in labeling, attributable to this segmentation, along with the reduction in temporal bias, enabled the model to attain greater generalizability across a spectrum of speaker contexts and temporal patterns.

Table 2.5: Performance comparison of three models in terms of precision, recall, F1-score, and support across consonant classes and overall. **Overall Accuracies:** Model 1: **92%**, Model 2: **89%**, Model 3: **77%**

Model	**Metric**	**/k-g/**	**/p-b-v-f/**	**/r/**	**/t-d-z-s/**
Model 1	Precision	0.76	0.98	0.98	1.00
	Recall	1.00	0.93	1.00	0.77
	F1-Score	0.86	0.95	0.99	0.87
	Support	166	343	329	164
Model 2	Precision	1.00	1.00	0.92	0.69
	Recall	0.84	0.79	1.00	1.00
	F1-Score	0.91	0.88	0.96	0.82
	Support	166	343	329	164
Model 3	Precision	0.96	1.00	0.58	0.90
	Recall	0.98	0.34	1.00	0.85
	F1-Score	0.97	0.50	0.74	0.88
	Support	166	343	329	164

As illustrated in Table 2.4, the performance of three models trained on various modality combinations (radar, ultrasound, and audio) is shown for consonant classification. The model that combines all three modalities (Model 1) achieved the best overall accuracy of 92%, as indicated by its exceptionally good F1-scores across all consonant groups. Model 2, which utilizes radar and ultrasonic data exclusively, exhibited an overall accuracy of 89%, demonstrating competitive performance, though with slightly diminished recall for the /p-b-v-f/ group. Conversely, Model 3, designed exclusively on radar data, exhibited a substantial decline in performance, with an overall accuracy of 77%.

The findings underscore the necessity of multimodal information for enhancing categorization accuracy, particularly in cases where speech are acoustically and articulatory comparable.

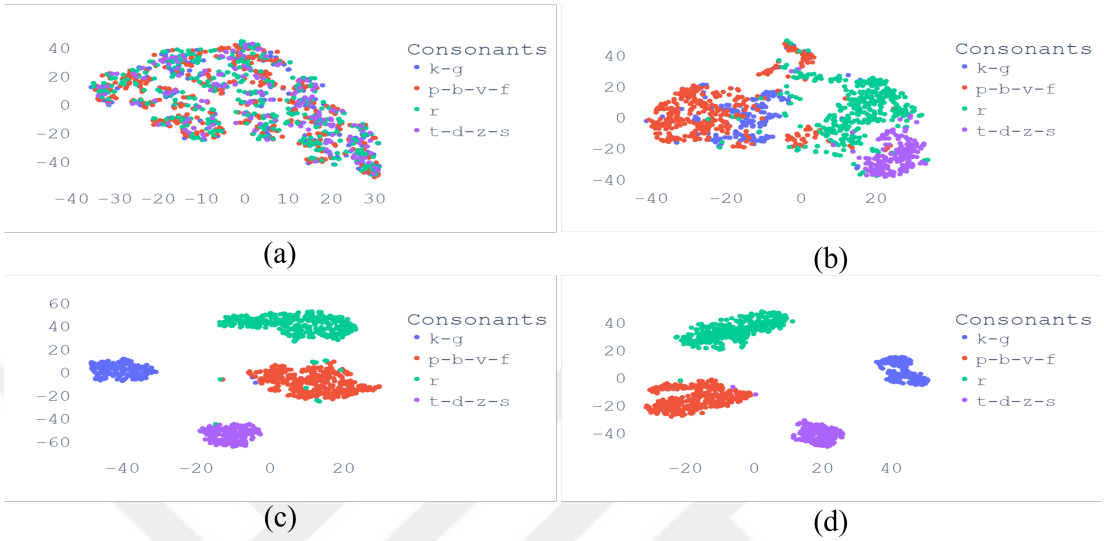


Figure 2.16: t-SNE visualizations under different modality configurations during inference.

Fig. 2.16 presents t-SNE visualizations of the learned embeddings from a model trained on all three modalities—UTI, mmWave radar, and audio—under four different inference conditions. Fig. 2.16a shows the embedding space when noise is added to all data sources, reflecting a baseline for separation under degraded input.

In Fig. 2.16b, the model operates solely on mmWave radar data, with other modalities replaced by noise. Although the embedding structure remains partially separable, the absence of UTI and audio leads to overlapping clusters, highlighting the importance of multimodal inputs. Fig. 2.16c includes radar and UTI data (audio replaced by noise), yielding improved class separation compared to radar-only, confirming that UTI data contributes significantly to articulatory representation. Finally, Fig. 2.16d shows the embedding when all modalities are available at inference, producing the most distinct and well-separated clusters. These results demonstrate the complementary nature of the modalities and the robustness of the learned representation even when some modalities are missing during inference.

2.7. Conclusion

This chapter explored the use of the mmWave radar for human and speech sensing, presenting a detailed overview of its operational principles, signal processing pipeline, experimental setups, and integration with multimodal systems. The main findings and takeaways can be summarized as follows:

- mmWave radar systems offer robust and privacy-preserving alternatives to camera-based sensing, particularly effective in capturing motion and velocity through micro-Doppler analysis.
- We demonstrated the radar's ability to detect and differentiate various human actions, including harmonic motion, walking patterns, and sit-to-stand transitions, using both 2D and 3D micro-Doppler representations.
- Initial speech sensing experiments revealed the limitations of micro-Doppler-based radar for isolated vowel and consonant classification due to the subtlety of articulatory movements and low signal-to-noise conditions.
- To address this, we concatenated raw radar frames along the time axis to form a continuous micro-Doppler signal and applied STFT to generate speech-relevant spectrograms. This approach provided a global view of motion over time and improved the representation of articulatory events.
- We also employed a speaker device to play articulatory diverse consonants from the UltraSuite repository, enabling clean and repeatable radar captures for phoneme-level analysis.
- Segmenting the 2-second spectrograms into utterance-aligned chunks helped reduce temporal overfitting and improved generalization in machine learning models.
- Multimodal learning—combining radar, ultrasound, and audio data—led to significantly higher classification performance compared to radar-only models, as validated by F1-scores and t-SNE visualizations.

In conclusion, mmWave radar shows great promise for speech-related sensing tasks when integrated within a multimodal framework. The methods presented in this chapter—based on speaker-emitted signals—lay a strong foundation for subsequent experiments

using radar data directly captured from human speech. The next chapter extends this work by analyzing naturally produced speech and further investigating the radar's capacity for capturing fine-grained articulatory dynamics in real-world scenarios.

3. TOWARDS ROBUST PHONETIC SEGMENT CLASSIFICATION VIA ARTICULATORY AND AUDIO MODALITIES

3.1. Introduction

The classification of phonetic segments underlies many applications in speech and language processing (SLP), such as automatic speech recognition (ASR), pronunciation modeling, and articulation analysis. Additionally, advancements in speech recognition are intrinsically linked to accurate consonant classification, as the precise identification of phonetic segments is crucial for developing effective speech recognition systems and therapeutic interventions. Research supports that the phonetic placement method has been effective in enhancing articulation skills, specifically for bilabial consonants in children with physical disabilities [127] and individuals with Down syndrome[128].

Although traditional approaches are largely based on audio signals, these methods face several limitations due to factors such as inter-speaker differences, environmental noise, and articulatory ambiguities [129–131]. The phonetic segment classification based solely on audio may fail to distinguish spectrally similar but articulatory different segments, especially when segments vary depending on context or when the differences are subtle.

To address these challenges, articulation data provided by UTI has been introduced as a complementary modality to speech processing. UTI offers the visualization of the tongue shape during the production of each phoneme in a non-invasive and real-time manner (Figure 3.1). This technique stands out for its ability to capture fine-grained articulation information that is usually difficult to extract from audio signals alone, and is therefore an important tool for understanding both individual and inter-individual variations [132]. Thanks to advances in ultrasound technology, UTI systems are becoming more portable, more affordable, and more amenable to integration with machine learning models for phonetic analysis [31, 56].

Despite its advantages, UTI is has a couple of limitations. First, the setup requires precise probe placement under the chin, which reduces its feasibility for unsupervised or

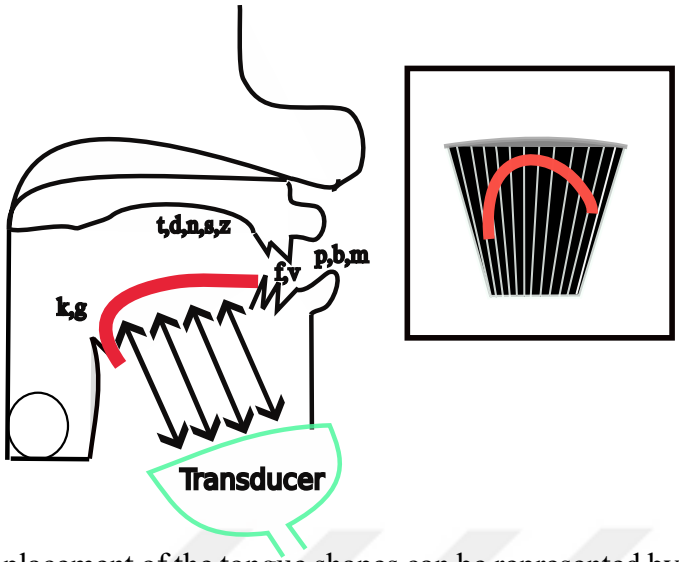
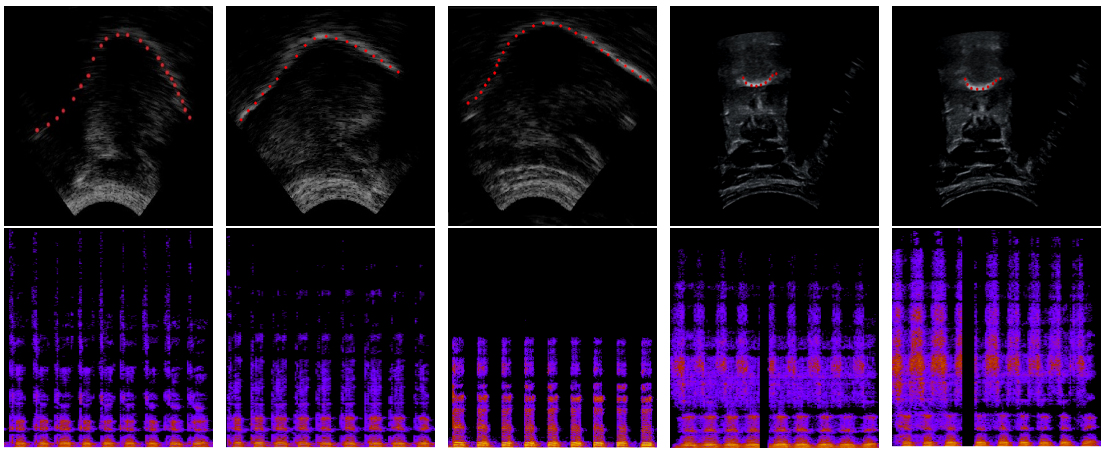


Figure 3.1: Final placement of the tongue shapes can be represented by phonemes [1].

remote use. Second, interpreting UTI images is inherently challenging due to tongue movement complexity, image quality fluctuations, and the absence of visible information about other articulators such as the lips or glottis [65]. These constraints introduce ambiguity in classification when ultrasound images alone are used.



(a) /k/: ”cat” **(b)** /t/: ”time” **(c)** /d/: ”dog” **(d)** /z/: ”zebra” **(e)** /s/: ”sun”

Figure 3.2: Sample phonetics segments of ultrasound tongue (top) and visible spectrum (bottom) imaging [1].

Therefore, the viability of ultrasound and the accuracy of audio-only applications have both come under scrutiny. For example, speech segments that differ primarily in vocal

fold vibration—such as the voiceless /s/ and its voiced counterpart /z/—produce similar tongue shapes in ultrasound frames, making them nearly indistinguishable in UTI images alone (Fig. 3.2d and Fig. 3.2e). Yet, these segments are easily separable in the acoustic domain due to the presence or absence of voicing cues. Similarly, some vowels, such as /i/ and /y/, are similar in terms of tongue shape but differ in lip rounding. However, since standard UTI systems cannot display such lip movements, they may lead to misclassifications unless supported by auido features. Such examples clearly demonstrate the need for a multimodal approach that can capture both articulatory and audio features of speech.

In light of these concerns, this chapter introduces a novel framework (MMNET) that uses both ultrasound tongue images and audio during the training phase to create a shared embedding space to address these issues. This embedding space integrates articulatory patterns with spectral and temporal features of speech, enabling more robust phonetic classification.

At inference time, we proposed several methods methods (M1-M4) operate using only audio input. However, the influence of articulatory information—learned during training—remains embedded in the model's representations. Our design (M1) allows to maintain high classification performance in realistic, audio-only environments, while benefiting from the structured articulatory grounding that UTI provides during training.

M1 offers several key advantages. Firstly, it enhances the classification performance that is vulnerable to acoustic noise and speaker variability through the utilization of joint representation learning. Second, it mitigates ambiguities arising from unimodal imaging or audio inputs by enabling cross-modal supervision [133]. Finally, it enables practical deployment in low-resource or real-time settings where articulatory imaging is unavailable, making it suitable for broader applications in speech modeling and assistive technologies.

3.2. Methods

We propose a shared embedding space and classification methods based solely on audio data in order to address the phonetic segment classification problem. The proposed approach consists of three main stages: data collection, learning the multimodal em-

bedding space, and classification based only on audio.

In the first stage, a multimodal dataset (GTUConsonants) was compiled by collecting both audio and simultaneous UTI data from the speakers. The data collection process was carefully designed to obtain high-quality and synchronized recordings and was supported by reading materials covering various phonetic contents. This stage plays a critical role in the generalizability of the model by being planned to include both individual and inter-individual articulatory variations.

In the second stage, MMNet, a multimodal neural network architecture trained on the task of classifying phonetic segments, learns the common embedding space. This embedding space is defined by the vector representations extracted from the last hidden layer of the model (just before the classification layer). Thus, these vectors learned from audio and UTI data provide a meaningful and discriminative representation for classification, carrying both articulatory and audio information together. Since this representation space is optimized to directly increase the classification performance, it allows the model to focus on the discriminative information.

In the last stage, this embedding space is estimated using only audio data and phonetic segment classification is performed on this representation. Thus, while the models (M1-M4) are enabled to benefit from rich articulatory information such as UTI in the training process, a more practical and portable system that can work only with audio data is obtained in the application phase. This approach allows both to increase the classification accuracy and to improve the usability in real-world speech technology applications.

3.2.1. Datasets

3.2.1.1. GTUConsonants

To have a data-driven model for phonetic segment classification, we compiled a sufficient and balanced dataset dataset by using [134]. For this purpose, the data collection process was carried out within the scope of the dataset we named GTUConsonants [1].

During the data collection process, participants were selected from Turkish-origin adults

[1] A sample recording of our dataset: `https://youtu.be/golisLoMSxY?si=TpmKGr1Gs9vu3YQR`

between the ages of 20 and 40, and were sampled to reflect phonetic diversity as much as possible. Participants were asked to read full sentences in a manner close to the natural flow of speech, and during these readings, both audio data and synchronized UTI recordings were taken. The obtained dataset includes recordings of three participants vocalizing six different consonants, which also determines the number of classes in the target classification task (Table 3.1).

The International Phonetic Alphabet (IPA) [135] chart is internationally recognized and standardized for phonetic notation. It was created by the International Phonetic Association as a way to represent the sounds of spoken language. The IPA provides a consistent system to transcribe the phonetic sounds of any language, allowing linguists, language learners, and professionals in related fields to accurately describe the sounds they hear and produce. The phonetic segments for this study are selected based on IPA standards. We ensured cross-lingual generalization by choosing phonetic segments based on IPA standards, which provides consistency and accuracy across different languages.

Table 3.2 presents a summary of the GTUConsonants dataset developed for phonetic segment classification in Turkish. The dataset comprises 67 video clips, totaling over one hour of synchronized UTI and audio recordings, captured at 25 frames per second with a resolution of 640×480 pixels. UTI image clips recorded in parallel with the audio data were also included in the analysis process. These images were rescaled from their original 640×480 resolution to 96×96 size in order to keep the model size at a manageable level and optimize the training time. During this process, care was taken to preserve the meaningful articulatory information in the ultrasound images.

On the other hand, the processing of audio data was carried out by following a workflow based on spectrogram extraction over labeled segment intervals. Each audio segment was obtained from audio files with a sampling rate of 48 kHz, matching millisecond-level timestamps. In order to analyze the frequency content of the segments over time, the Short Time Fourier Transform (STFT) was applied. This transformation was calculated using a window size of 1024 samples and a shift length of 512 samples, providing 50% overlap. The resulting spectrograms were represented in a form that visualizes the intensity of the signal in decibels, with logarithmic scaling applied to the frequency axis. This approach provides a strong basis for audio-based classification by capturing detailed frequency information over a given time period [136].

Table 3.1: The number of ultrasound frames of train and test sets on GTUConsonants for different phonetics and subjects.

Consonant	Subject 1		Subject 2		Subject 3	
	Train	Test	Train	Test	Train	Test
/t/	642	198	609	324	522	312
/k/	531	336	813	258	756	246
/z/	438	213	459	186	462	156
/d/	522	120	660	306	885	387
/g/	387	159	459	180	411	123
/s/	546	172	687	279	648	245

3.2.1.2. UltraSuite Repository

To assess model performance more precisely and to examine general validity, the Ultrax Typically Developing (UXTD) dataset was incorporated into the analysis as an additional data source. UXTD is a subset of the UltraSuite database, featuring rich annotations including speaker variability, word-level transcriptions, and detailed phonetic content. This dataset has been labeled by expert language therapists within the scope of speech therapy applications and is a widely used and reliable source in the literature [63, 137, 138].

One of the most successful results in the literature was carried out by Ribeiro et al. on the dataset [63]. In order to classify the speech sounds, four different classes can be organized based on the places of articulation (place of articulation).

The defined classes are grouped according to their articulation features as follows: The first class includes sounds with bilabial and labiodental articulations such as /p, b, v, f/. The second class includes dental, alveolar and post-alveolar sounds such as /th, d, t, z, s, sh/. The third class represents phonemes produced in the velar region, such as /k, g/. Finally, the fourth class consists of the alveolar approximant sound represented by the /r/ sound alone. This four-class structure allows the evaluation of models capable of articulatory-based generalization. The total number of ultrasound frames used for each class within the scope of the study is presented in Table 3.3.

Table 3.2: GTUConsonants Dataset Summary.

Video Properties	
Longest video duration	00:01:14 sec
Shortest video duration	00:00:18 sec
Number of video clips	67
Total video duration	01:05:15 sec
Frame rate	25 frame/sec
Frame resolution	640×480
Phonetic Position Labels	
`w_in`	Word initial
`w_fin`	Word finish
`med_in`	Word middle, syllable initial
`med_fin`	Word middle, syllable finish
Sample Annotated Segment	
File	esra_coronal_z1
Segment	z_w_in
Consonant	z
Position	w_in
Start time (s)	3.170
End time (s)	3.245
Duration (s)	0.074

Table 3.3: Total number of ultrasound frames.

Consonants	US frames count
/k-g/	3,255
/p-b-v-f/	8,952
/t-d-z-s/	6,945
/r/	9,551

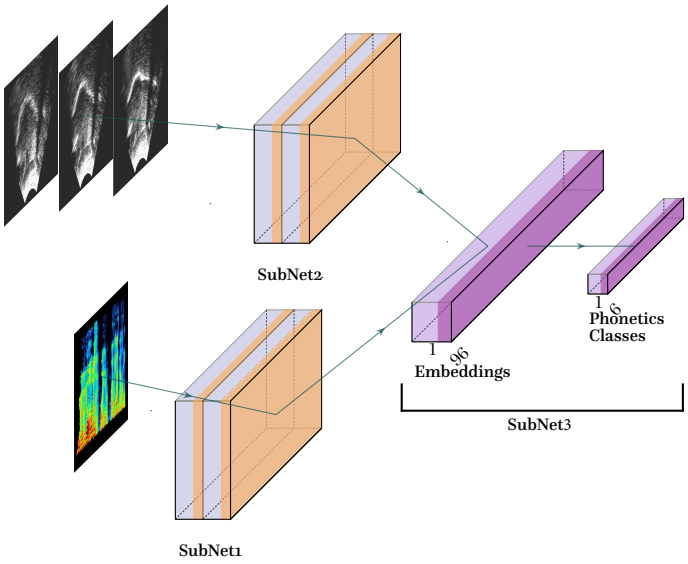
3.2.2. MMNet: Audio and UTI Joint Embedding Space

Our multimodal architecture, MMNet, plays a major role in this framework. It consists of two separate sub-networks (SubNet1 and SubNet2) as shown in Fig. 4.4a. One of these sub-networks uses spectral sound waves (Fig. 4.4b), the other uses ultrasound images (Fig. 4.4c). Both networks reproduce embedding vectors specific to their modality and add these vectors to the phonetic embedding memory. In the last layer, the MMNet, which combines these two different representations, creates a common embedding vector with a total dimension of 96 and performs productions based on this representation. These sub-architectures serve as the basic components of the systems to be created later and provide the direction that will improve the model as a whole.

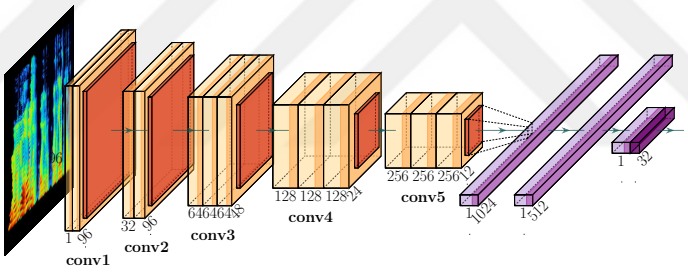
Each of the sub-architectures is built using 2D Convolutional Neural Network (CNN) and Bi-directional Long Short-Term Memory (BiLSTM) layers as the basic building blocks to efficiently capture temporal and spatial information. Time-distributed layers are also integrated into the architecture to support the structure in which sequential images (image sequences that progress in time) are processed through CNN layers. This structure ensures that the spatial features at each time step are preserved and transferred to the temporal enhancement.

Our multimodal architecture combines structures running on audio-only or UTI-only data by integrating both types of input - a spectrogram image and a simultaneous UTI frame sequence - into a single network. This combination makes it possible to synthesise the augmentative common representation offered by each modality.

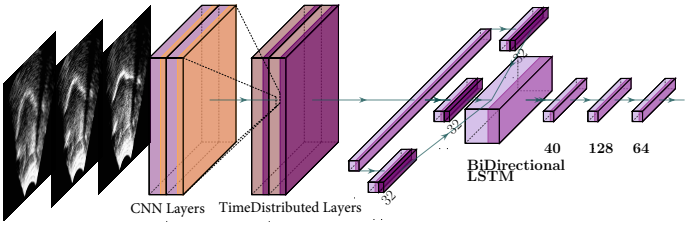
To improve the performance of the MMNET, various training systems such as hyperparameter variation, on-the-fly augmentation, dropout and early stopping have been prepared. In addition, the joint use of multiple UTI frames allows segments to be obtained based not only on the data received at one time, but also on the totality of continuous articulatory movements over time. In this way, our models show satisfactory performance in key feature metrics such as accuracy, flexibility (recall), specificity (specificity) and precision (precision).



(a) MMNet: Overview of the multimodal network architecture that generates a final 96-dimensional embedding by concatenating the output vectors from SubNet1 and SubNet2.



(b) SubNet1: Generates a 32-dimensional embedding from spectral sound wave inputs.



(c) SubNet2: Utilizes CNN layers followed by TimeDistributed and BiDirectional LSTM layers to process each frame (n). The network ultimately produces a 64-dimensional embedding from the input data, including UTI.

Figure 3.3: The sub-components of the phonetics embedding space [1].

3.2.3. Phonetics Embedding-based Classification

The embedding vectors obtained from the last hidden layer of MMNet (Fig. 4.4a) are based on multimodal (audio + UTI) data during training, but allow inference to be made only with audio data during the test phase. This architectural feature offers a great advantage in terms of real-world applications of speech technologies. Since the use of imaging systems such as UTI is costly, has limited portability and requires expertise, the development of systems that work only with audio data both increases accessibility and makes them practical for use in applications such as speech therapy.

Based on this motivation, the four different architectural structures we propose (M1–M4) are designed to eliminate the need for additional hardware while increasing the accuracy of systems based only on acoustic data. These structures are important steps in a research line that aims to make phonetic segment classification more accessible, economical and faster.

3.2.3.1. M1:Embedding Extraction by Similarity

This method implicitly uses the audio ($\vec{a}' \in \mathbb{R}^{32}$) and UTI ($\vec{u}' \in \mathbb{R}^{64}$) embedding vector pairs obtained in the training phase to create a system that works only with audio data in the inference phase. In the inference phase of the model, the acoustic embedding vector $\vec{a}$ produced by SubNet1 for the test sample is compared with all the audio embedding vectors ($\vec{a}'$'s) in the training set. This comparison is based on the cosine similarity between the vectors, and the $\vec{a}'_k$ vector closest to the test sample is selected.

Since each vector $\vec{a}'$ is paired with a corresponding UTI embedding $\vec{u}'$ during the training process, the vector $\vec{u}'_k$ corresponding to the nearest $\vec{a}'_k$ is easily accessible. This vector is treated as a real UTI input during the inference process of the model and is used as input for the phonetic classification. This approach offers the possibility of indirectly maintaining UTI-based classification performance even when the UTI data is not actually present at test time.

In addition, to make the UTI representations more robust and generalizable, the matching with the audio data is enhanced by adding artificial Gaussian noise to the UTI data during the training process. All these steps are detailed in Algorithm 2, which summa-

Algorithm 2: M1 Test Phase

Data: $\vec{a}'$, $\vec{u}'$ as training audio and UTI data embedding.

Result: $< c_i$ class $label >$

foreach $\vec{a}$ *in test space produced by SubNet1* **do**

Find the most similar acoustic embedding vector $\vec{a}'_k$ using cosine similarity:

$$\vec{a}'_k = \arg\max_{\vec{a}'_k} \left\{ \cos(\theta_k) = \frac{\vec{a} \cdot \vec{a}'_k}{\|\vec{a}\|_2 \, \|\vec{a}'_k\|_2} \right\} \tag{3.1}$$

Retrieve the corresponding UTI embedding $\vec{u}'_k$;

Use $\vec{u}'_k$ as test-time input for further phonetics classification ;

Run SubNet3 with $[\vec{a}\ \vec{u}'_k]$ concatenated ;

rizes the algorithmic structure of our proposed method. This algorithm systematically demonstrates how articulatory knowledge can be used indirectly by using only audio data at test time.

3.2.3.2. M2:Embedding Extraction by Encoders

This approach aims to estimate articulatory information indirectly by creating an embedding space based solely on audio data. The model is based on an encoder-decoder architecture. Thanks to this structure, the audio-based representation space is learned by mapping it to the articulatory, UTI, embedding space during training. This mapping allows the system to approximate articulatory representations during the test phase (Fig. 3.4) using only audio data.

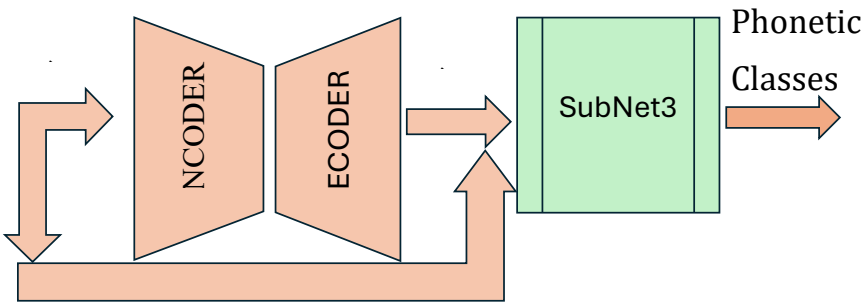


Figure 3.4: During the test phase, the M2 architecture takes the input vector $\vec{a}$ and generates an output vector $\vec{u}$, which is then concatenated with $\vec{a}$ and passed through SubNet3 for final processing [1].

The encoder part of the model consists of three Conv2D layers that process spectral sound data, and each layer is followed by a 2×2 dimensional max-pooling operation to reduce the size of the feature maps. This structure allows abstract but meaningful representations to be extracted from the input data. The decoder part has three Conv2D layers in a symmetrical manner, and each layer is followed by a 2×2 dimensional up-sampling operation. In this way, the acoustic information compressed by the encoder is projected back into the articulatory space. The output layer is a Conv2D layer with a 3×3 dimensional filter and a sigmoid activation function, which allows the reconstruction of the UTI embedding vectors.

In the test phase, only acoustic embedding vectors $\vec{a}$ are used, and the decoder transforms these acoustic representations into vectors corresponding to the articulatory space $[\vec{a}; \vec{u}]$. The combined vectors, which are the output of this decoder, are then used as inputs by SubNet3 for classification purposes. Thus, although the system only uses audio data at inference, it may allow improving phonetic classification accuracy by producing embedding vectors that are similar to the articulatory representations. This structure allows the effective reconstruction of articulatory representations learned from acoustic signals without the direct need for UTI data.

3.2.3.3. M3:Embedding Extraction by Auto-Encoders

In this architecture, instead of using the embedding vectors obtained from SubNet3, the raw audio and the UTI data are fed directly into the system. The model is based on two separate autoencoder structures that process both acoustic and articulatory modalities simultaneously. These autoencoders learn their own modality-specific representations (latent embedding), which are then combined and transferred to the final classification layer (Fig. 3.5).

A key feature of the model is its end-to-end training architecture. The entire system—from input layers to the final classifier—is optimized using a unified loss function, ensuring seamless information flow across modalities and consistent parameter updates throughout the network. This integrated training strategy allows the direct observation of how features extracted from both data types influence overall system performance.

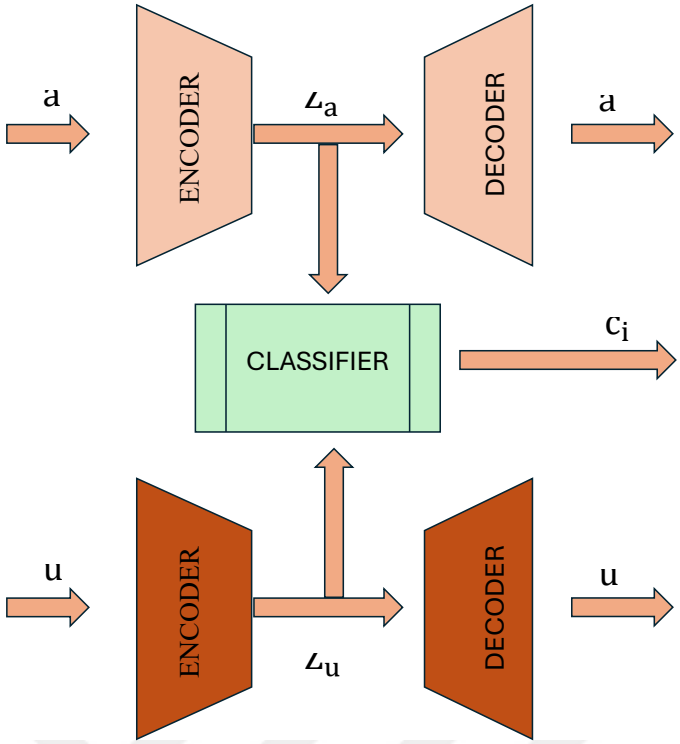


Figure 3.5: The M3 architecture employs three distinct input configurations during training: (a_i, u_i, c_i), (a_i, None, c_i), and (None, u_i, c_i), where a and u represent audio and UTI data, respectively, and c denotes the corresponding class label [1].

Encoder-decoder architectures trained separately for each modality extract low-dimensional but meaningful representations of the data. These latent representations (latent embeddings) are fed into the classifier network as a combined vector. This combination process aims to provide a richer representation by exploiting the complementary structure of audio and articulatory information. This structure is expected to enhance classification accuracy, particularly in cases where phonetic segment distinctions are subtle or ambiguous.

The model allows in the test phase to have more robust and stable architectures compared to structures using only acoustic data or only UTI data. This lead us to processing the two modalities together provides a superior generalization and classification capacity compared to single modalities. In addition, the end-to-end joint training of the system enabled overlapping features between the modalities to be learned and effectively transferred to the classifier.

3.2.3.4. M4: UTI-Free Inference via Noise-Augmented Training

This model employs an architecture that directly utilizes multiple data modalities—specifically audio and UTI—as inputs, rather than relying on precomputed embedding vectors. To support this approach, a special dataset has been created containing simultaneously collected audio and UTI data. This dataset is fed directly into the model during the training process, allowing it to process both audio and image-based samples together (Fig. 3.6).

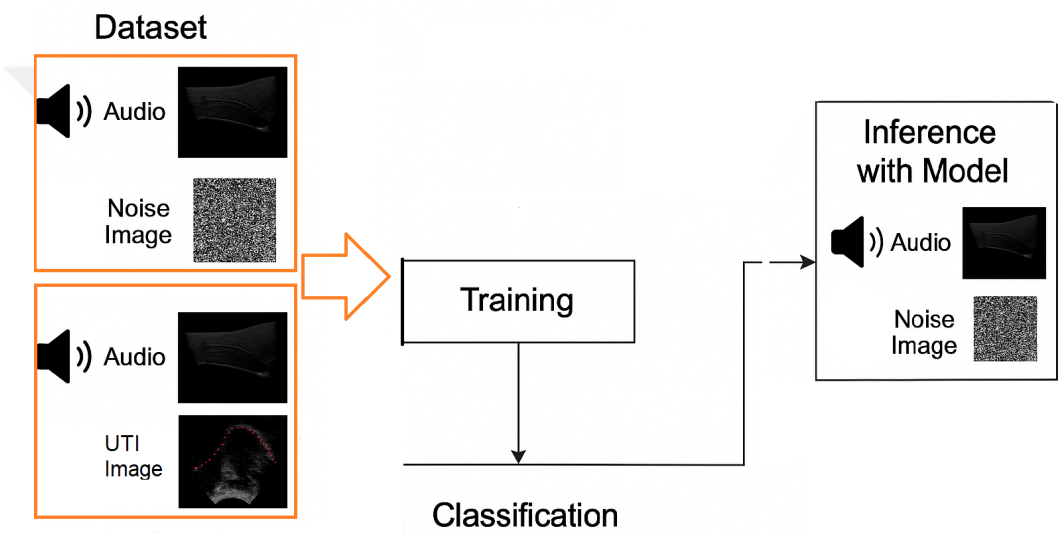


Figure 3.6: Training with both audio–UTI and audio–noise pairs enables robust inference using only audio and noise images, reducing dependency on UTI data.

However, as UTI images are often not available during the inference phase, the applicability of this model in the real world remains limited. To address this issue, only real UTI data are used in part of the training data, while the remaining part is presented with random noise images along with audio data. Thus, the model is trained with both complete modality pairs (audio + UTI) and incomplete modality conditions (audio + noise), which increases the model's ability to classify with audio alone and to learn relationships between UTI and audio.

This strategy allows the generalizability and robustness of the model to noise. It ensures that the model maintains its performance, especially in the absence of real data, and

provides more flexible solutions for field use.

In the inference, the model is configured to run only with audio data and accompanying noise images. The use of these images, which do not contain structural information, removes the reliance on real UTI data and makes the technology more accessible in the field. Especially in environments where ultrasound imaging systems are expensive, difficult to utilize, or inapplicable, this model allows the reasonably robust architectures.

3.3. Experimental Results and Analysis

This section presents the architectural designs of the proposed models—ResNet50, SubNet1, SubNet2, and MMNet—along with their intermediate evaluation results, thereby providing a detailed overview of how each model contributes to phonetic segment classification within the proposed system. In addition to the implemented architectures, comparative analyses are made with the basic approaches that are widely used in the literature. These comparisons facilitate a comprehensive evaluation of the system's overall success level, its flexibility across different data modalities, and the distribution of the application.

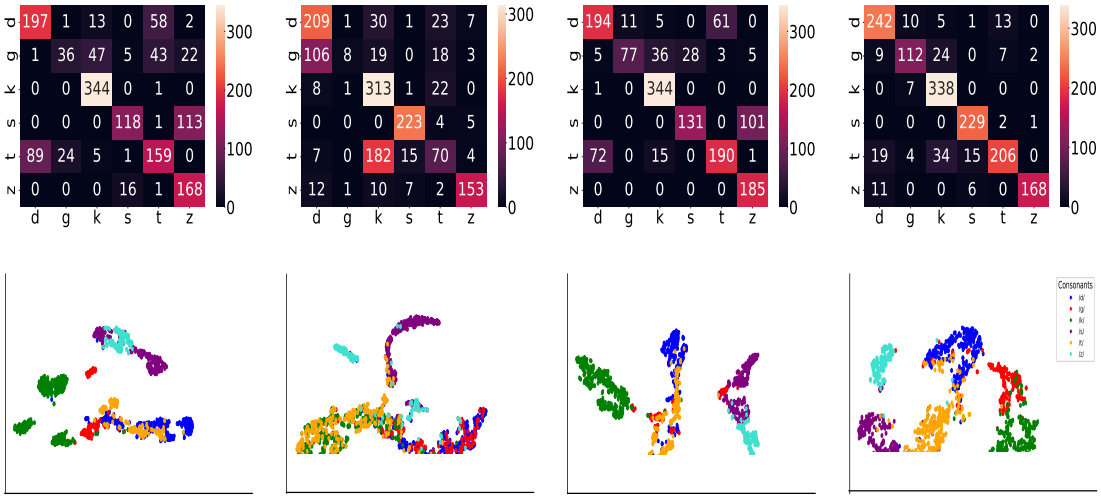
3.3.1. Baseline Models

ResNet50: This architecture is notable for its strong performance in image classification tasks [139]. One of the most prominent innovations of the ResNet architecture is the use of residual connections, which facilitate the network to directly learn a series of transformations that map the input to the target output. These connections structure the learning process at each layer by adding a "difference" function directly to the input. Thus, instead of learning the target output directly, the network learns to model the difference between the input and the output. This approach provides an effective solution to the vanishing gradient problem, one of the main problems that make it difficult to train deep networks.

We trained this architecture using just a single UTI input and is regarded as one of the baseline models. It aims to classify phonetic segments using spatial patterns marked from ultrasound images. This structure provides an important starting point in terms of the limitations of systems based solely on visual data and potential evaluation.

SubNet2: While the ResNet50 model utilizes a single static UTI frame to perform phonetic classification, the SubNet2 model (Fig. 4.4c) adopts a temporal approach by incorporating multiple consecutive UTI frames. This design enables the model to learn dynamic articulatory representations by capturing tongue movement over time, rather than relying solely on static spatial information.

SubNet1: This model (Fig. 4.4b) is designed to operate exclusively on audio data, relying on spectral representations of speech signals to perform phonetic classification. Unlike models such as ResNet50 and SubNet2, which utilize articulatory information from UTI images, SubNet1 extracts phonetic cues solely from the acoustic modality. This approach leverages the rich temporal and frequency-based patterns present in speech signals, allowing the model to capture distinctions that are rooted in spectral variation. It is particularly well-suited for phoneme pairs that, while articulatorily similar, exhibit contrasting acoustic properties—such as the difference between plosives and fricatives. In such cases, SubNet1 processes the dynamic structure of the sound wave to derive class-relevant features purely from the audio stream.



(a) ResNet50 **(b)** SubNet1 **(c)** SubNet2 **(d)** MMNET

Figure 3.7: The confusion matrices (top) and 2D t-SNE projections (bottom) of consonant embeddings from ResNet50, SubNet1, SubNet2, and MMNET models on the GTUConsonants test data [1].

3.3.1.1. The Analysis of the Baseline Models

The ResNet50 model, which has been trained exclusively on UTI data and works with single-frame images, has been shown to successfully perform the classification task between phonetic segments that are visually distinct in terms of articulatory representation. This is particularly evident in the case of phoneme pairs such as /t/ and /z/ or /d/ and /s/, which exhibit clear differences in terms of tongue position and shape in UTI images. The model demonstrates an acceptable level of accuracy, enabling it to successfully distinguish between classes (Fig. 3.7a). This finding indicates that the spatial variations evident in the articulation region of the tongue in UTI images are conducive to the classification learning of a CNN-based structure.

However, the efficacy of the model is constrained to instances where the visual contrast between classes is substantial. This is particularly evident in the case of phoneme pairs such as /d/ and /t/, or /s/ and /z/, where the model's ability to discriminate between the sounds is significantly diminished due to the similarity of their articulatory structures in UTI images. The visual patterns of these phonemes are largely overlapping in UTI images due to the fact that they are produced in the same articulatory region (e.g., alveolar) and with similar language form. Consequently, a UTI representation based on a single frame may prove inadequate in capturing these subtle variations.

For instance, the ResNet50 model exhibits a capacity to accurately classify only 118 out of 232 instances of the /s/ phoneme, while the SubNet2 model demonstrates a superior performance with 131 accurate classifications within the same phoneme category (Fig. 3.7a and Fig. 3.7c) . For the /s/ vs. /z/ pair, ResNet50 shows substantial confusion: out of 232 /s/ instances, only 118 are correctly classified, while 113 are misclassified as /z/. Similarly, 16 /z/ instances are misclassified as /s/, indicating limited boundary resolution between the two classes. In contrast, SubNet2 improves correct /s/ classification to 131 and eliminates confusions between /z/ and /s/ entirely (0 instances), while reducing /s/ to /z/ confusion to 101.

A similar trend is observed in the /t/ vs. /d/ pair. ResNet50 correctly classifies 159 /t/ instances and 197 /d/ instances, but 58 /d/ samples are incorrectly labeled as /t/ — a significant overlap. SubNet2 improves correct /t/ classification to 190 and /d/ to 194, while slightly increasing /d/→/t/ misclassifications to 61. Although the confusion re-

mains notable, the overall classification performance improves, suggesting that tongue motion captured across multiple frames helps refine predictions for stop consonants with closely overlapping articulatory gestures.

For the /k/ vs. /g/, ResNet50 achieves only 36 correct classifications for /g/, while 47 /g/ samples are confused with /k/, showing that the model struggles to separate these velars. SubNet2, however, raises correct /g/ classifications to 77 and reduces /g/→/k/ misclassifications to 36. The correct classification for /k/ remains consistent (344) in both models, confirming that temporal features particularly help disambiguate voiced/unvoiced pairs that share articulation zones but differ subtly in timing and movement.

On the positive side, the SubNet1 model, based on audio data alone, shows remarkable performance in distinguishing certain phonetic classes. This is particularly evident in cases of phoneme pairs that are articulatory similar but possess distinctive acoustic features, such as /s/ and /z/, or /t/ and /d/ (Fig. 3.2). In such instances, the SubNet1 model demonstrates a higher level of classification accuracy in comparison to models such as ResNet50 and SubNet2, which employ solely UTI data (Fig. 3.7a - Fig. 3.7c). These results demonstrate that the spectral-based acoustic representation possesses significant discriminatory potential for specific phonetic classes and that the model can effectively learn these distinctions. This is particularly evident in phoneme pairs exhibiting contrasting acoustic structures, such as plosives with discontinuous plosives and fricatives with continuous plosives. In such cases, the model demonstrates an effective capacity to discern the dynamic structure embedded within the sound waves.

However, it should be noted that the efficacy of this model varies across different phonetic classes. The SubNet1 model, for instance, demonstrates limited efficacy in differentiating sounds such as /k/ and /t/. This is attributable to the strikingly similar patterns in the spectrogram images of these two phonemes (Figs. 3.2a and 3.2b). The similarity in high frequency components and short-term energy densities of these sounds leads to a substantial reduction in acoustic signal-based discrimination. This similarity hinders the model's capacity to discern discriminative features, consequently leading to a decline in classification accuracy.

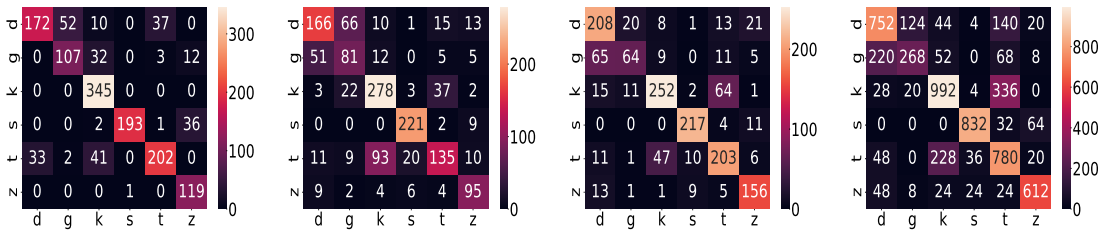
To facilitate the visualization of our intermediate results, the classification vectors obtained in the final layer of each classification architecture were projected into a 2D

space using the t-Distributed Stochastic Neighbor Embedding t-SNE method (Fig. 3.7 (bottom)). These offer a qualitative insight into the extent to which the model differentiates between classes in the embedding space.

These results are considered as a qualitative complementary evidence to the confusion matrices presented in Fig. 3.7 (top); supports the effects of methodological differences between using UTI-only data (Fig. 3.7a), audio data-only (Fig. 3.7b), multi-frame ultrasound images (Fig. 3.7c) and multimodal data (Fig. 3.7d).

3.3.2. Multimodal (MMNET) vs. Audio-only (M1-M4) models

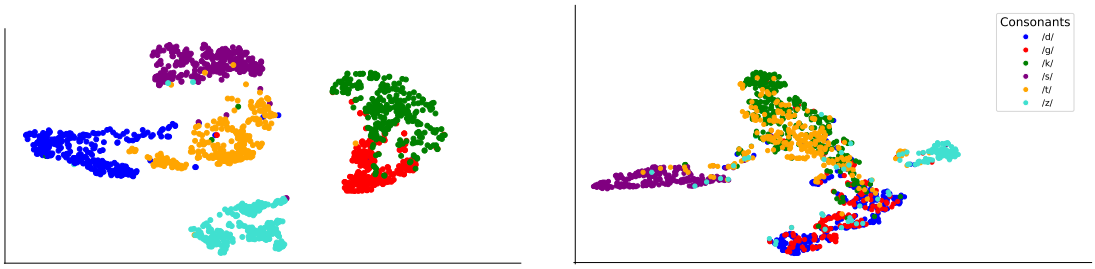
As illustrated in Table 3.4, the accuracy outcomes of our proposed neural network architectures are summarized, along with the data modalities utilized in consonant classification experiments conducted on the GTUConsonants dataset. Among the architectural variations ranging from M1 to M4, only the M1 model demonstrated an accuracy increase of approximately 5% compared to the SubNet2 model using UTI data in both training and testing stages. This finding underscores the significance of multimodality integration for enhancing classification performance (Fig. 3.8).



(a) M1 **(b)** M2 **(c)** M3 **(d)** M4

Figure 3.8: Confusion matrices of the proposed architectures (M1–M4) on the test dataset. The M1 model demonstrates the most effective phonetic segment classification among all.

To support the experimental findings, two-dimensional visualizations were performed with the t-SNE method to qualitatively examine the embedding spaces (Fig. 3.9). These reveal the proximity relationships in the embedding space of the samples obtained with only sound-based methods on the GTUConsonants dataset. The embeddings of the M1 model displays a distinctive clustering behaviour among phonetic pairs with articula-



(a) M1 method **(b)** M2 method

Figure 3.9: 2D visualizations of consonant embeddings generated by t-distributed stochastic neighbor embedding (t-SNE) for the proposed M1 and M2 models on the GTUConsonants dataset.

Table 3.4: Accuracy and modality configurations for consonant classification using the proposed architectures.

Test Accuracy	Model	Training		Test	
		UTI	Audio	UTI	Audio
0.67	SubNet1		+		+
0.69	ResNet50	+		+	
0.69	M2	+	+		+
0.72	M4	+	+		+
0.75	M3	+	+		+
0.77	SubNet2	+		+	
0.81	M1	+	+		+
0.91	MMNet	+	+	+	+

tory similarities, such as /d/–/t/ and /k/–/g/. This behaviour reflects the model's capacity to distinguish between such classes, thereby establishing its representation. We assessed model performance using paired t-tests and 95% bootstrap confidence intervals over five runs. MMNET (88.40%) and M1 (81.45%) showed statistically significant gains over all other models ($p < 0.05$). MMNET also significantly outperformed M1. Results confirm the robustness and superiority of the proposed multimodal architecture on the GTUConsonants dataset. However, further investigation is required to determine the generalization capabilities of the pro-posed methods on different datasets.

3.3.3. Performance on UltraSuite Repository

The present experiment was conducted with the objective of comparing the results of the system with those of other successful methods as documented in the literature. The most successful study similar to our work, as cited in reference [63], utilities four classes based on different places of articulation.

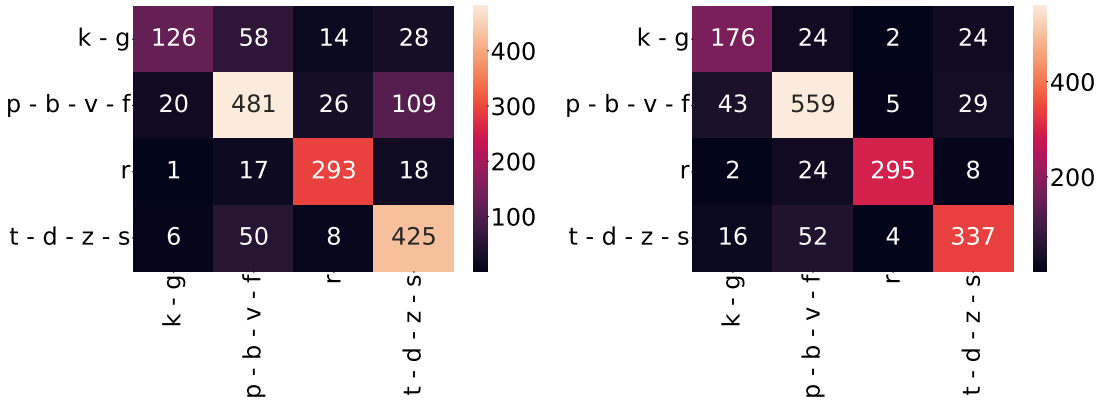
A series of tests were conducted under a speaker-independent scenario using the UTDX dataset in the UltraSuite Repository to evaluate the generalizability of the models and their suitability for real-world conditions. This experimental configuration enables the analysis of the model's performance on unfamiliar speakers, with the assurance that the speakers utilized in the test phase are distinct from those presented to the model during the training process. Consequently, the evaluation process aimed not only to detect signs of rote memorization, but also to assess the model's ability to generalize phonetic representations effectively across unseen speakers.

Table 3.5: UltraSuite Repository classification test results.

Model	Accuracy
Independent Cnn-Raw [2]	0.59
SubNet2 (ours)	0.63
Independent with Speaker Mean [2]	0.67
Adapted Cnn-Raw [2]	0.72
Adapted with Speaker Mean [2]	0.71
MMNet (ours)	0.78
M1 (ours)	0.85

The findings obtained from this study are summarized in Table 3.5 for direct comparison with similar studies in the literature. In order to analyze the behavior of the model in more detail, the confusion matrices related to the classification performance are presented in Fig. 3.10, and the 2D representations of the embedding spaces learned by the model are presented in Fig. 3.11. These visual elements constitute an important refer-

ence point, especially in terms of qualitatively evaluating the capacity of the model to produce distinctive representations between different phonetic classes.



(a) MMNet **(b)** M1

Figure 3.10: Confusion matrices from testing the MMNet (a) and M1 (b) architectures in a multimodal setting.

In our evaluation of various models based on accuracy, the SubNet2 model achieves a score of 0.63, while MMNet surpasses it with an impressive accuracy of 0.78. Particularly noteworthy is the M1 model, demonstrating the accuracy at 0.85.

The findings obtained from our study indicate a potential for reducing the dependency on ultrasound tongue imaging data in the testing phase. The visualization of the classification vectors obtained in the last layer of each classification architecture by reducing them to 2D space with the t-SNE method is shown in Fig. 3.11. This visual representation clearly reveals that the embedding space of the M1 model in particular has a high segregation ability and a regular clustering structure.

When having comparison between Table 3.4 and Table 3.5 shows that as the diversity in the test dataset increases, M1 performs better. However, M1 is dependent on the training dataset and requires similarity comparison during inference. This suggests that it may be heavily reliant on the training dataset and perhaps not suitable for handling larger scale data due to the time cost of comparing similarities.

To address the concerning points, we present Fig. 3.12 , which shows the accuracy of two models, MMNET and M1, as a function of the number of subjects in training.

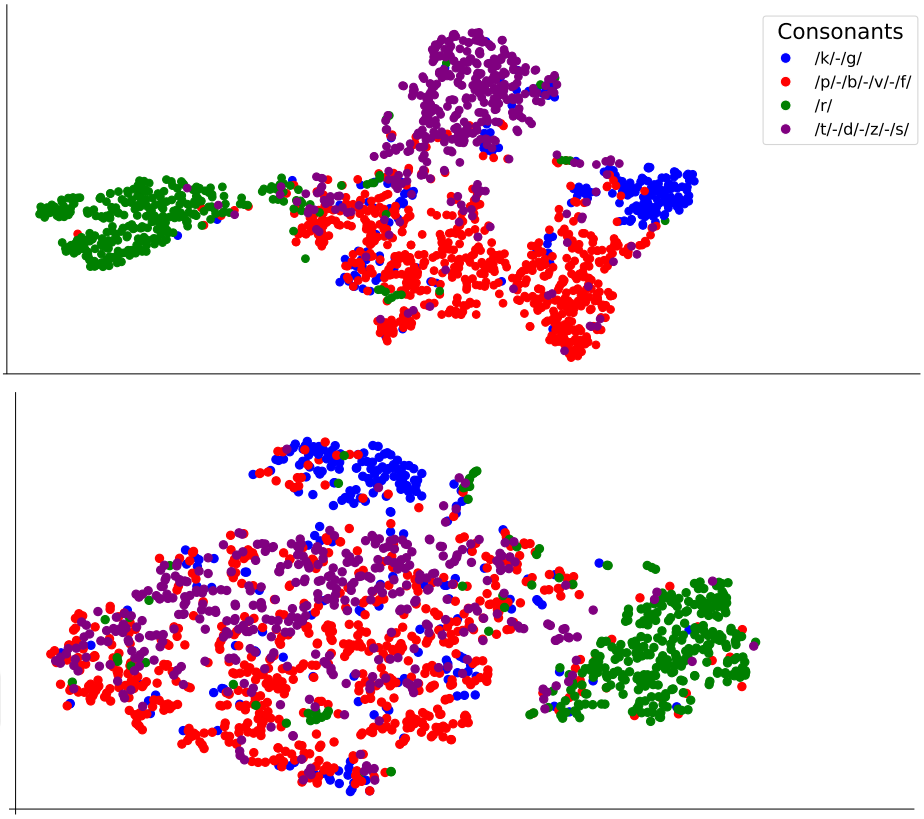


Figure 3.11: The findings provide supporting evidence for the effectiveness of the M1 and M2 methods on the GTUConsonants dataset. Among them, M1 demonstrates the highest accuracy, which is further illustrated through the 2D t-SNE visualization of consonant embeddings.

While we utilized the same set of test data, we used different subsets of subjects during training. The plot demonstrates that the accuracy of both MMNET and M1 increases with the number of subjects in training.

MMNET starts with an accuracy of around 0.30 (actually indicates randomness) when training on 5 subjects, reaching its peak accuracy of about 0.85 at 25 subjects. After this, it slightly fluctuates but remains relatively stable around 0.80. M1 starts with a lower accuracy of around 0.10 (actually indicates randomness) at 5 subjects, and then steadily increases, and by 25 subjects, it reaches 0.80, maintaining stability around this value with slight improvements, eventually reaching 0.85. As the number of subjects increases beyond 25, M1 slightly outperforms MMNET, reaching a higher accuracy of 0.85 compared to MMNET's 0.79. Our experiments showed that both models show a performance plateau starting from 25 subjects, indicating that additional subjects do not significantly improve accuracy beyond this point (Fig. 2).

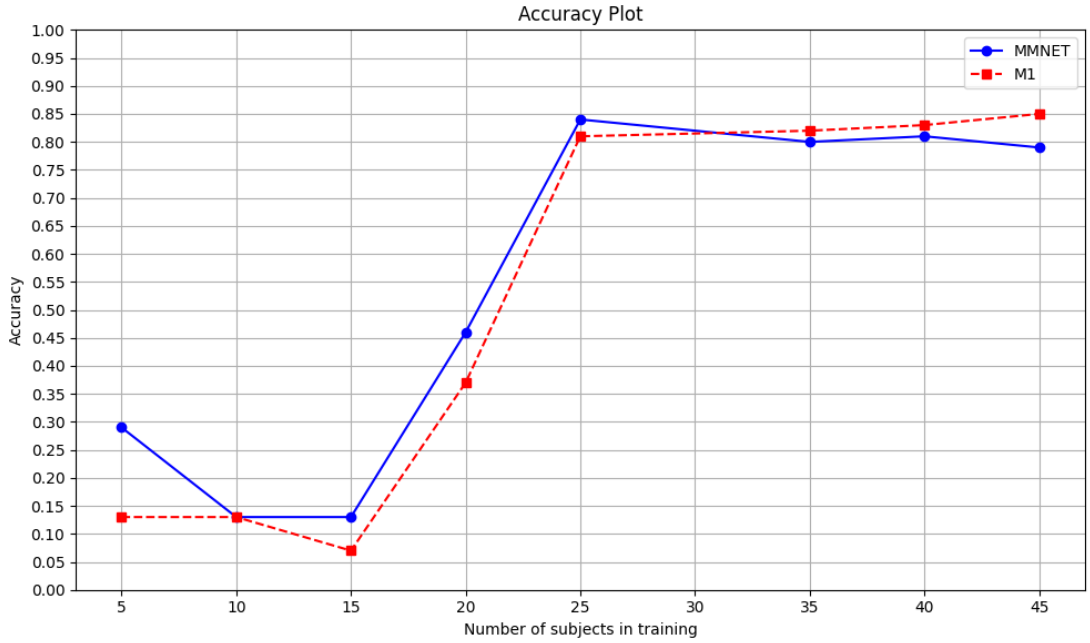


Figure 3.12: Test results of M1 and MMNET models trained by different number of subjects.

Additionally, while M1 currently relies on direct similarity comparisons during inference, we acknowledge the potential challenges posed by scalability with larger datasets. We understand the need for optimizations, such as using efficient data structures for embeddings and employing advanced vector search algorithms like FAISS [140] or ScaNN [141]. These approaches could enable us to conduct similarity comparisons on datasets of billions of records, addressing these challenges effectively. These optimizations could reduce the time complexity of the similarity search, making M1 computational tractable for handling larger scale data without compromising performance.

3.3.4. Computational Efficiency and Real-time Feasibility

A thorough investigation was carried out to evaluate the runtime performance of the suggested models under reasonable and practical situations; this is reported in Table 3.6. Raw audio input entered via a recording interface starts the assessment process. The step of audio segmentation comes next, in which the continuous audio stream is split into smaller, reasonable sections. After that, every segment is turned into a spectrogram, providing the signal's time-frequency domain picture. Following their projection into the audio embedding space, these spectrograms are stored as fixed-size vector

representations.

Table 3.6: Training and Inference Characteristics.

Model	Training Time (per epoch)	Audio Only Inference Time (per sample)	Trainable Parameter Size	FLOPs
SubNet1	21 seconds	103 milliseconds	655,140	9.9×10^9
SubNet2	28 seconds	-	810,532	2.8×10^{10}
MMNET	44 seconds	-	1,267,108	3.3×10^{10}
M1	47 seconds	133 milliseconds	-	-
M2	8 seconds	193 milliseconds	5,315,904	1×10^7
M3	14 seconds	71 milliseconds	156,324	2.8×10^7
M4	53 seconds	141 milliseconds	1,267,108	3.3×10^{10}

Important processing components such embedding search—which matches the recovered embeddings with entries in a reference database—and audio-to–UTI embedding transformation—which uses cross-modal information to improve retrieval precision—are included in later phases of the pipeline. These sequential processes are fundamental elements in assessing the feasibility and efficiency of model deployment in real-world situations as each of them adds to the whole computing cost.

Executed on a machine fitted with an RTX 490 GPU and a 6-core Intel Xeon CPU, the M1 architecture—which combines all six stages of the pipeline—achieves a throughput of 7.5 frames per second (fps). Particularly in high-performance hardware settings, the pipeline might theoretically be further enhanced with multi-threaded implementation and parallel processing improvements, hence enabling real-time applications.

Distinct performance trade-offs revealed by a comparison of many model architectures concerning training time, inference time, and computational complexity guide their fit for distinct application environments. Especially highlighting its fit for latency-sensitive real-time applications, the M3 model has minimal floating point operation (FLOP) needs and fast inference capability. Although more computationally intensive,

the M2 and M4 models produce superior classification accuracy, hence they are better suitable for applications requiring fine-grained phonetic distinction.

There exists a notably harmonious trade-off between the SubNet1 and M3 models. Presenting a well-rounded performance profile, SubNet1 offers a reasonable training time of 21 seconds per epoch and an inference time of 103 milliseconds per sample. By contrast, M3 gets a shorter training time of 14 seconds per epoch and the quickest inference time of 71 milliseconds per sample. Fascinatingly, the M2 model runs with 10,624,320 FLOPs and shows the smallest training time of 8 seconds each epoch although incurs a greater inference time of 193 milliseconds per sample. These properties imply that M2 is especially useful in situations when quick model training comes first over inference efficiency.

Additionally, we accept that the real-time application can be implemented as a pipelined multi-threaded application to achieve better running time performance. Fig. 3.13 shows how the pipelined model can be implemented for the real time application. Depending on the system characteristics, we theoretically achieve around 20 fps, excluding the overhead of synchronization and pipelining.

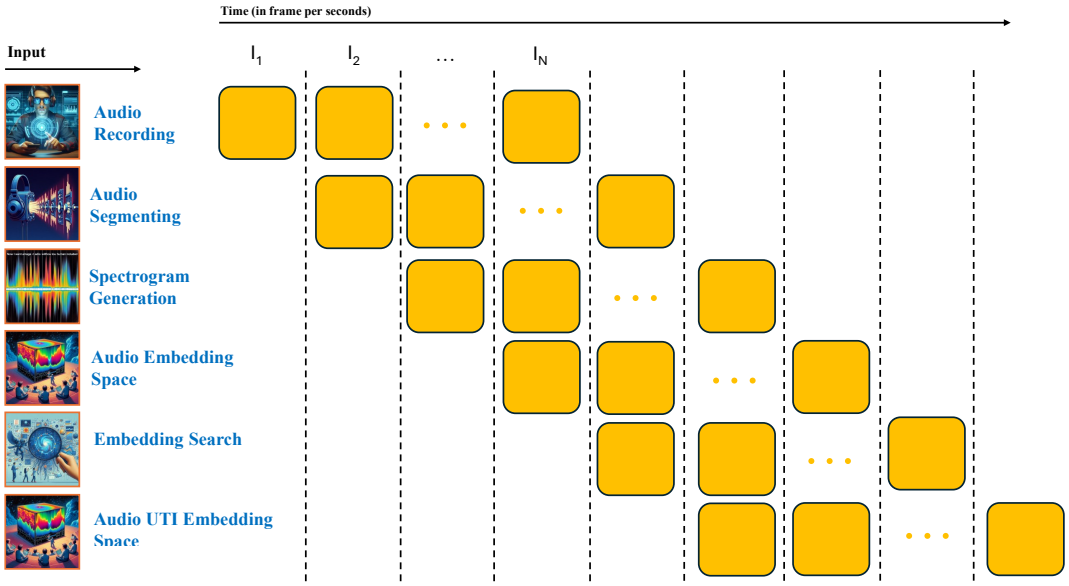


Figure 3.13: Overview of real-time processing.

3.3.5. Key Findings and Insights

It has been suggested that multimodal models may demonstrate superior performance in comparison to single-modal models, due to their capacity to integrate and utilize a variety of data types, thereby facilitating a more comprehensive representation of complex tasks [66, 142]. This performance advantage becomes even more evident in scenarios where interaction between modalities is decisive [143]. Our experiments have shown that superior results are achieved on the GTUConsonants dataset, especially with the joint use of the MMNet architecture (Table 3.4).

Furthermore, Fig. 3.2 shows that some phonetics appear visually similar in UTI data, but have spectrally distinguishable features. The multimodal MMNet architecture we developed combines the various modalities under the same roof by processing the audio and UTI data together to capture these differences. This integration has the potential to enhance performance in tasks such as phonetic segment classification (Fig. 3.7).

The findings appear to align with the growing body of literature highlighting the merits of multimodal models. It is also worth noting that our models trained and tested on audio data alone also yielded significant results. In particular, the findings reveal that the M1 model yielded similar results to MMNet in terms of accuracy (see Table 3.4), and outperformed MMNet on some datasets (see Table 3.5). These results suggest that our audio-only modelling approach is effective and could be a strong alternative in practical applications where only audio data is available.

It seems that the integration of multiple data streams allows multimodal models to identify correlations and patterns that unimodal approaches may not capture, which could produce more robust and generalizable results [144]. The M1 model appears to demonstrate a superior performance in comparison to current methods that utilize solely UTI data, a phenomenon that may be attributed to the inherent advantages of multimodality phonetic embedding spaces. Furthermore, while existing high-level models mostly focus on spatial data, they suggest important lines of research for future studies, such as using all frames corresponding to a phoneme and adopting iterative architectures to capture temporal dynamics. These models underscore the significance of integrating audio data in applications such as speech therapy, emphasizing that the combination of ultrasound and audio data can offer complementary insights. The M1 model, which

uses joint embedding spaces for both audio and UTI data and is trained with a spatiotemporal approach, is an example of this approach and shows how the integration of dynamic temporal information with multimodal data leads to higher performance.

Another important finding reveals the spatiotemporal improvements provided by our models. When the accuracy and t-SNE results are evaluated together, it appears that the SubNet2 architecture designed to process multi-frame UTI data exhibits superior performance compared to both ResNet (Fig. 3.7, Table 3.4) and the Independent CNN Raw method in [63] (Table 3.5).

The M1 method achieved approximately 15% higher accuracy during testing compared to the SubNet1 model that uses only spectral sound waves. It is interesting to note that both models work with only audio data in the testing phase, but the M1 model uses the joint placement space obtained from audio and UTI data in the training phase, which seems to provide a significant advantage over SubNet1 (Table 3.4).

However, upon the testing of the models on different datasets, a number of discrepancies in performance were observed between MMNet and M1. MMNet demonstrated higher performance on the GTUConsonants dataset, which comprised only three participants (Fig. 3.7 d, Fig. 3.8a, and Table 3.4). Conversely, the phonetic embedding space of the M1 model exhibited high robustness on a larger and more diverse dataset consisting of 58 speakers (Fig. 3.10, Table 3.5).Although the M1 and SubNet2 models produced similar outcomes on GTUConsonants, a significant performance difference favoring M1 was seen on the UDTX dataset (Fig. 3.11, Table 3.5). Nevertheless, owing to the modest sample size of the GTUConsonants dataset, these outcomes should be regarded with caution. Nevertheless, the consistent results obtained from the UDTX data set support the generalizability and reliability of the M1 model (Fig. 3.12).

3.4. Conclusion

This study demonstrates the potential of deep neural networks to play a transformative role in speech therapy, particularly through the fusion of audio and UTI data. The proposed systems offer a substantial reduction in workload by automating the intensive and time-consuming manual analysis processes frequently encountered by speech therapists. This is due to the deep learning architectures that can process both spatial and

temporal information. The development of spatiotemporal models of this kind paves the way for intelligent assistant systems that will support human experts in clinical settings.

A comparative analysis of the various sub-architectures proposed has revealed that the highest classification accuracy was achieved in the architecture where spectral representations of audio and ultrasound tongue imaging data were used in combination. This finding was further reinforced by the creation of a dataset. The dataset comprised UTI images and audio data collected from three volunteer speakers, covering six different phonetic classes. The meticulous and balanced nature of the data collection process enabled the models to learn more robust and generalizable representations.

Furthermore, the analysis of the embedding vectors extracted from the models has yielded two notable insights. Firstly, it has provided a comprehensive understanding of the internal structure of the models. Secondly, it has demonstrated the potential to reduce the need for ultrasound imaging. This transformation renders the system more accessible and cost-effective, signifying an important stride towards its practical implementation as a tool for speech therapists, particularly in clinical settings.

One of the primary contributions of our research is the formulation of a model architecture that is constructed using the placement areas of models trained on audio and UTI data and operates solely with audio input during inference. The efficacy of this model has been demonstrated by its superior classification performance in comparison to conventional models based solely on ultrasound data or audio alone. This superiority is evident in both accuracy and inter-class separation metrics. Consequently, we advocate the utilization of this audio-only model, particularly for remote speech therapy applications. The system's reliance on microphone-based hardware not only reduces installation costs but also enhances the user experience.

The lightweight structure of the proposed model (with a small number of model parameters) and the minimization of video processing load reveal its applicability not only to clinical applications but also to interactive platforms such as real-time in-game feedback systems. The study demonstrates the practical potential of multimodal signal processing methods in both health and technology fields and provides a solid foundation for future research.

4. RADAR-DRIVEN PHONETIC SEGMENT CLASSIFICATION

4.1. Introduction

Phonetic segment classification is widely regarded as a foundational problem in the domains of speech therapy and linguistics. This process entails the allocation of each phonetic segment to a predefined classes based on audio or imaging-derived data [63]. Traditional unimodal approaches – methods based solely on audio or articulatory data – have yielded effective results in many SLP applications [145–147]. Nevertheless, audio-based models are limited in their ability to discriminate acoustically similar phonemes and are also less robust to noise. Currently, the use of articulatory data, such as UTI, significantly enhances classification accuracy by providing additional and important insights into the complexities of speech production [1].

Despite their advantages, UTI-based approaches face several practical limitations [148, 149]. These limitations include high hardware costs, the need for specialized equipment, and physical discomfort that reduces user comfort during long-term use [65]. As a solution to such challenges, multimodal methods that combine articulatory data with acoustic information have become a focus of research [150].

Specifically, multimodal systems that integrate UTI data or video-based lip movement analysis with audio aim to improve phonetic classification performance with their ability to represent both internal and external articulatory processes [151, 152]. While UTI is a valuable modality, especially in terms of its ability to provide detailed internal articulatory dynamics, [153, 154], the practical limitations outlined above limit its broad applicability [155]. Similarly, although video-based lip movement analysis is effective in tracking external articulators, its inability to provide information about internal speech dynamics is a fundamental limitation that prevents holistic and accurate modeling of speech production.

In recent years, mmWave radar has emerged as a promising approach in the field of SLP. This new generation technology enables the acquisition of both acoustic and articulatory data in a contactless manner [66, 68, 125]. This technology offers the opportunity

to track the movements of structural elements that play a role in speech production, such as the jaw, lips, teeth, tongue and vocal cords, without requiring any physical contact and while preserving user privacy. Despite the obvious benefits of mmWave radar for phonetic segment classification, a systematic and comprehensive investigation has yet to be conducted.

Although speech signals contain phonetically critical information, they may not be preferred in situations where privacy and security are at the forefront; they may be affected by environmental noise or may not be suitable for use. In such cases, mmWave radar emerges as a compelling alternative. Given their sensitivity to vocal cord vibrations, both speech spectrograms and radar can record speech patterns. But by showing precise joint motions beyond acoustic signals [156], radar offers complimentary information. The spectrograms produced from the radar (left) and the microphone (right) show comparable temporal and frequency patterns, as shown in Figure 4.1, thereby indicating the potential of radar as a non-invasive and privacy-preserving technique to speech processing.

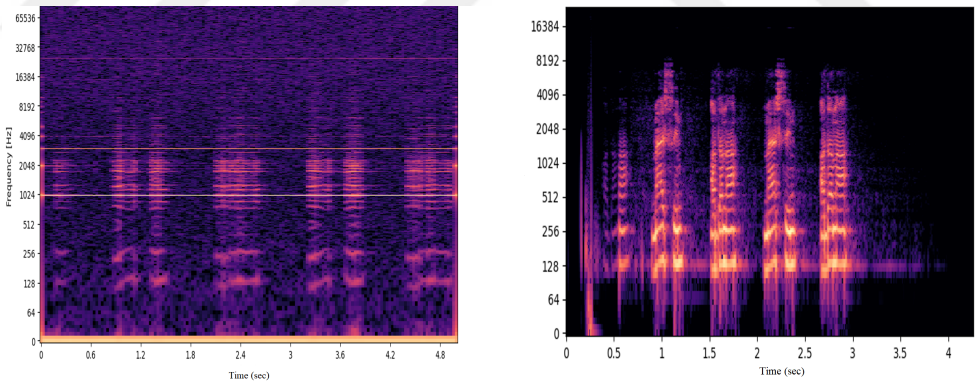


Figure 4.1: Radar-captured (left) and microphone-captured (right) spectrograms for the phrase ”Adana Mersin Adana Mersin.” Both reveal similar speech patterns, with radar also capturing subtle articulatory movements beyond the audio signal.

In order to improve the performance, robustness, and practical application of phonetic segment classification, a multimodal learning framework is provided in this chapter integrating mmWave radar, UTi, and audio data. USRadioAI uses the strengths of every modality during the training process. The proposed method, USRadioAI, capitalizes

on the strengths of each modality during the training process. However, it is configured to operate solely with mmWave radar data in the inference phase, thereby eliminating the requirement for UTI or audio input by the model. To achieve this objective, a joint embedding space is constructed that can capture articulatory and acoustic features together, thus enhancing the model's ability to generalize phonetic representations.

4.2. Methods

4.2.1. Data Collection

Hardware: As shown in Figure 4.2, a data acquisition setup was created to collect multimodal data for training, validation and evaluation of the USRadioAI system. Synchronized audio data was collected using a Shure Beta 58A microphone at a sampling rate of 48 kHz; UTI data was collected using [134]; and radar data was collected using a Texas Instruments (TI) IWR1843 mmWave radar. The radar was configured to operate at a bandwidth of 3,89 GHz and a sampling rate of 25,600 Hz. For synchronization purposes, the ultrasound device operated continuously throughout the session (approximately 2 minutes) in sync with the audio. The mmWave radar was triggered separately with a ping tone played at the beginning of each recording to provide off-line synchronisation with a known signal. A ping tone was played at the beginning of each recording to alert the narrator and align the ultrasound device with the radar recording.

A video showing this process can be viewed in the sample dataset provided in the supplementary material [1]. It should be noted that the experiments were conducted in a real speech therapy room that is actively used for clinical purposes.

The processing pipeline begins by converting raw mmWave radar ADC data into spectrograms. Once the data is loaded into the system, missing samples are filled in and the signals are structured into frames, followed by alignment using the DCA1000 processing method. After averaging across antenna-channels, the signal is flattened and processed using the STFT, resulting in a decibel-scale spectrogram that allows time-frequency analysis.

[1] For a sample view of our dataset: CollectionSetup.mp4.

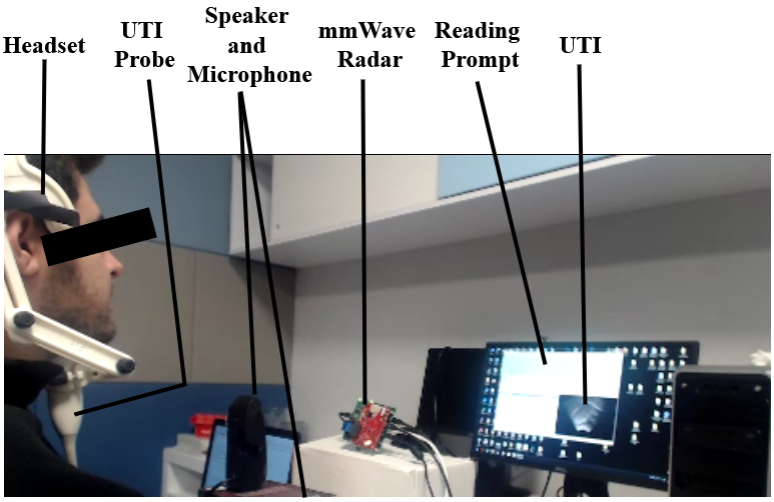


Figure 4.2: Data collection setup.

Improved settings in the radar system help to guarantee high-resolution signal collecting. The ADC is set to catch signals promptly by having a 10,000 ksps sampling rate. The system uses four RX antennas for spatial variation and improved signal processing and one TX antenna. The 200-frame, one chirp cycle arrangement ensures continuity and uniformity in data collection. Each chirp comprises 256 ADC samples that establish the resolution of the resultant signal patterns. Moreover, the 10-millisecond cyclic repeat ensures a consistent and continuous radar data stream, facilitating real-time speech and articulation processing.

Our audio data processing pipeline is centered on producing spectrograms from segmented audio intervals. A spectrogram was generated from a 48kHz audio segment according to the specified durations in milliseconds. The STFT was executed with a window size of 1024 samples and a hop length of 512 samples, resulting in a 50% overlap. The spectrogram employed logarithmic scaling on the frequency axis, illustrating signal strength in decibels relative to the maximum magnitude. The UTI clips have been reduced from their original dimensions of 640 × 480 to 96 × 96 to save model size and training duration

Real-Time Radar Configuration and Data Acquisition Pipeline: To enable real-time signal analysis and multimodal data collection, we implemented a Python-based radar driver for automated configuration and streaming using the Texas Instruments IWR1843BOOST radar and DCA1000EVM data capture board. This setup allows for low-latency data acquisition, repeatable experiment control, and live visualization

support — crucial for deploying radar-based sensing systems in practical settings.

The radar configuration and data acquisition pipeline consist of three primary stages: (1) radar configuration over UART, (2) FPGA setup using UDP-based command transmission, and (3) real-time data streaming over a high-speed UDP listener interface.

Step 1: Radar Configuration via UART: Using a configuration file (e.g., IWR1843.cfg), radar parameters are transmitted via UART through a wrapper class called SerialConfig. This module communicates over the COM port (in our case, COM4) at 115200 baud, issuing stop/start/reset commands as well as sending the full radar chirp configuration string-by-string to initialize the sensor.

Listing 4.1: Serial configuration using Python

```
radar = SerialConfig("ConnectRadar", "COM4", 115200)
radar.StopRadar()
radar.SendConfig("../config/IWR1843.cfg")
radar.StartRadar()
```

Step 2: FPGA Configuration over UDP: The DCA1000EVM must be initialized and configured separately from the radar sensor to stream raw ADC samples. This is achieved using UDP socket programming. Each configuration command is constructed using low-level hexadecimal packets, consisting of specific command codes (e.g., connect, configure FPGA, set packet parameters, and start recording). Commands are sent from the host PC (e.g., 192.168.33.30:4096) to the FPGA address (192.168.33.180:4096), and acknowledgments are parsed for validation.

Listing 4.2: Sending FPGA command packets

```
for cmd in ['9', 'E', '3', 'B', '5']:
    sockConfig.sendto(send_cmd(cmd), FPGA_address_cfg)
    response, _ = sockConfig.recvfrom(2048)
    print(f"Sentcommand{cmd},response:{response.hex()}")
```

Step 3: Real-Time UDP Data Listener: After successful radar and FPGA initialization, the radar streams raw binary ADC data to a designated UDP port. A custom listener class UdpListener is launched asynchronously using Python's threading and queue modules. The binary data are parsed into fixed-length frames based on the following configuration:

- ADC samples per chirp: 64
- Chirps per frame: 32
- TX antennas: 1
- RX antennas: 4

The resulting frame size is calculated as:

$$\text{Frame Length} = \text{ADC} \times \text{Chirps} \times \text{TX} \times \text{RX} \times 2 = 64 \times 32 \times 1 \times 4 \times 2 = 16384 \text{ Bytes}$$

Listing 4.3: Launching the UDP listener

```
BinData = Queue()
udp_listener = UdpListener("Listener", BinData, frame_length,
                           ('192.168.33.30', 4098), 2097152)
udp_listener.start()
```

Step 4: Real-Time Data Check: To ensure functionality, the main thread monitors whether the BinData queue receives any data within a timeout period. If data is successfully received, the radar is confirmed to be operational.

Listing 4.4: Data verification

```
timeout = time.time() + 10
while time.time() < timeout:
    if not BinData.empty():
        data = BinData.get()
        print(f"Received␣{len(data)}␣data␣points␣from␣radar!")
        break
```

This real-time configuration and streaming pipeline provides:

- Automated and reproducible radar setup using software-defined configuration.
- Low-latency data acquisition suitable for live monitoring and AI inference.
- Modular expansion capability, allowing synchronized recording with cameras or microphones.
- Parameter flexibility for testing different radar sensing setups.

This implementation served as the backbone of our data acquisition protocol in both human motion sensing and speech classification experiments, enabling rapid iteration and accurate alignment across sensing modalities.

Real-Time Radar Configuration and Data Collection Framework: To support real-time multimodal sensing and facilitate continuous experimentation, a complete radar configuration and data collection framework was developed using Python. This system enables automated control and streaming of raw ADC data from the Texas Instruments IWR1843BOOST radar and the DCA1000EVM FPGA data capture board. The architecture is modular, supporting seamless integration with GUI-based visualization, signal processing modules, and machine learning pipelines.

The framework is composed of several coordinated components:

- **Radar Configuration Module:** Implements a serial interface to load '.cfg' profiles and initialize the radar system (e.g., number of chirps, sampling rate, frequency slope, etc.).
- **FPGA Command Handler:** Sends low-level UDP commands to the DCA1000 FPGA to configure packet structure and initiate recording sessions.
- **UDP Listener Thread:** A dedicated thread captures the streamed binary data packets over UDP and segments them into fixed-size frames based on radar configuration parameters.
- **Data Processing Thread:** This thread consumes raw frames, reshapes the signal structure, and performs spectrogram transformations using STFT operations. The processed spectrograms are then fed to downstream queues.
- **Graphical Interface:** A PyQt-based interface allows real-time visualization of spectrogram outputs and provides basic start/stop control over the radar pipeline.

The system is designed to support multiple experimental setups. Configuration parameters, such as ADC sample count, number of chirps per frame, and antenna configurations, can be flexibly updated depending on the sensing task. Threading and queue-based design ensures that data streaming, processing, and display are decoupled, allowing for robust operation even in high-throughput settings. This framework was used extensively in both human activity and speech-related experiments. The processed spectrograms were saved to disk in real time and labeled using a pseudo-model or user input for downstream model training. Data collection sessions are automatically timestamped and structured into subject-specific directories, improving traceability and dataset management.

The GUI provides dynamic rendering of raw mmWave signal images with adaptive colormaps and fixed viewports, allowing researchers to inspect radar returns in real time and evaluate signal integrity during live capture. This real-time framework was critical to the multimodal system described in this study, enabling synchronized, high-resolution radar capture and facilitating the training of data-driven models. It forms the backbone for future extensions into live speech recognition and interactive radar-based applications.

Participants: To enable a radar-based implementation of phonetic segment classification, a sufficiently large and balanced dataset was compiled. Accordingly, the data collection process was conducted under the framework of the GTUSAudioRadioConsonants dataset [1]. For phonetic readings, we enlist the help of six people who are all native speakers. Our sample of native speakers includes two females and four males, with ages ranging from twenty-five to forty-five. Without any speaking issues, all individuals have typical speech ability. We trained the model using data from five speakers and then tested it on data from a sixth speaker that was not used during training in order to examine how well the model generalized.

The chosen phonetic segments for this investigation were selected in line with IPA criteria; following this criterion has been demonstrated to ensure consistency and accuracy across languages. Two,400 utterances in all were generated, with four distinct phonetic consonants—shown by /d/, /g/, /k/, /t/ in the International Phonetic Alphabet). An auditory signal (a ping sound) was emitted for each data collection, prompting the subject to articulate the sentence immediately following the cessation of the sound. Participants were situated roughly 50 cm from the radio equipment for data collection. They articulated each syllable at a moderate level while limiting extraneous movement to maintain data integrity.

4.2.2. Radar-Driven Embedding Space and Classification

In this study, we introduce the USRadioAI, a radar-based multimodal approach developed for phonetic segment classification and utilizing a common embedding space. The proposed deep neural network architecture MMNet combines the mmWave radar,

[1] For a sample recording from our dataset: `CollectionSetup.mp4`

audio data, and UTI modalities (Fig. 4.3).

The MMNet model necessitates the concurrent operation of components such as a mmWave radar sensor, a microphone, and a UTI device employed for feature extraction. The integration of these components guarantees that features derived from various data sources are effectively aligned inside a unified embedding space.

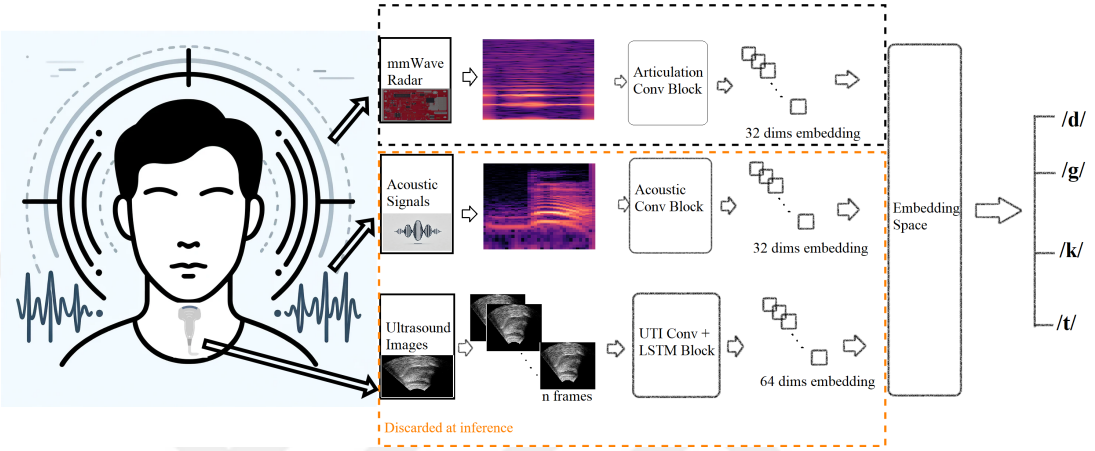
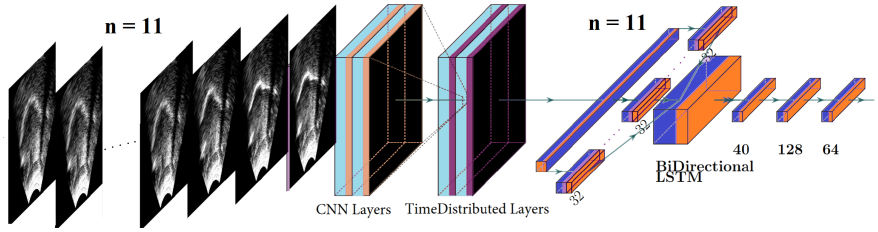


Figure 4.3: System overview of a multimodal speech processing framework combining mmWave radar, audio, and UTI.

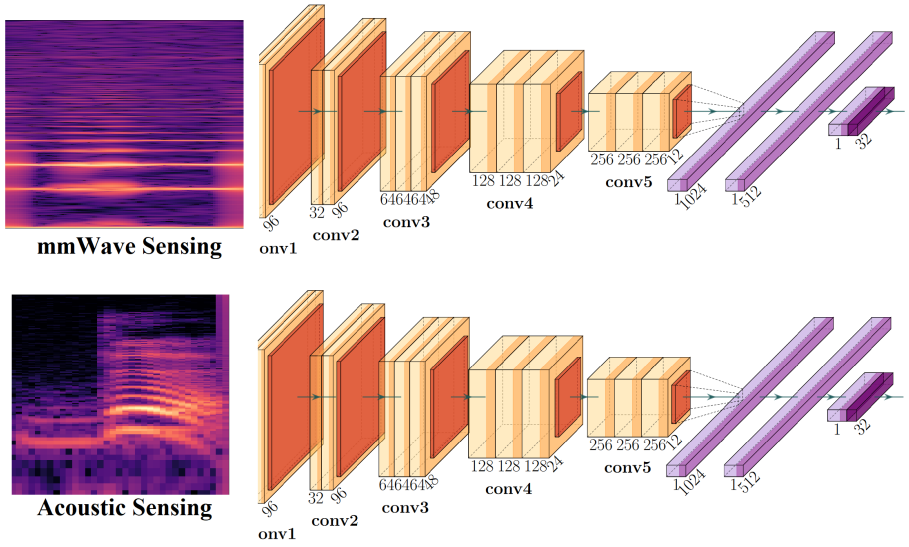
The development of neural network architectures has enabled the creation of robust and generalizable phonetic segment classification systems, with each architecture designed to extract meaningful representations from different perception modalities. Structural details of these architectures are presented in Fig. 4.4, where the two main architectures represent the main structures contributing to the embedding space.

The architecture, (Fig. 4.4a), first extracts visual features with CNN, and then integrates the spatio-temporal features by modelling the sequential dependencies with Time Distributed layers and BiLSTM networks. This architecture processes UTI data to generate a 64-dimensional embedding vector representing joint movements. The second architecture, Fig. 4.4b, employs stacked CNN layers to learn spectro-temporal patterns by processing mmWave radar (upper path) and acoustic signal (lower path) data, and generates a 32-dimensional embedding vector for each input.

In the training phase, a shared embedding space is created that can capture common representations among radar, audio, and UTI data. This embedding space integrates the unique features of each modality, thereby enabling the model to learn cross-modal



(a) This network employs dense layers to produce a 64-dimensional embedding from the UTI data.



(b) A 32-dimensional embedding is generated for mmWave and acoustic sensing.

Figure 4.4: The phonetics embedding space (MMNet) with its sub-architectures.

relationships. However, in the inference phase, the system only extracts embeddings from the radar data and performs classification through a special subnetwork that can independently process mmWave detection.

The USRadioAI aims to estimate the missing audio and UTI representations by comparing the radar embeddings extracted during training with the vectors in the common embedding space created during training. The selection of closest matches is achieved by calculating the cosine similarity between radar outputs and previously stored embedding vectors [1]. These most similar embedding vectors provide estimated audio and UTI representations corresponding to the missing modalities (Fig. 4.5). These representations are then integrated into the feature extraction process by combining them

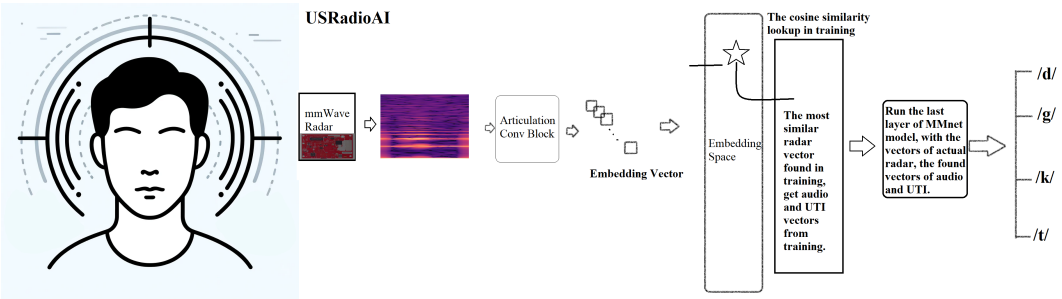


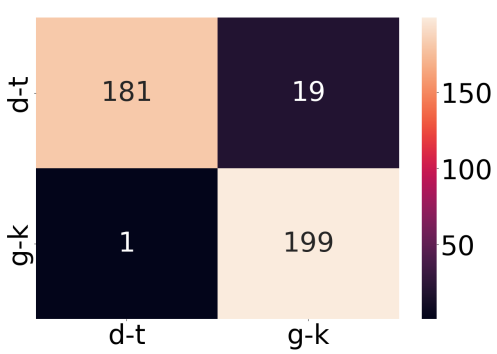
Figure 4.5: USRadioAI infers missing audio and UTI embeddings by matching radar features to training data via cosine similarity, enabling radar-only phonetic classification.

with real-time radar inputs. This process improves the classification performance of the model and increases the robustness of the system under different conditions.

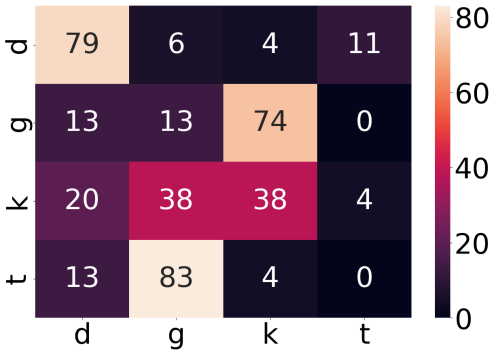
4.3. Experimental Results and Analysis

4.3.1. UTI-only Classification

After the data collection process was completed, we first examined the phonetic segment classification using only UTI data. It is important to check whether some phonemes, such as /k/ and /g/ or /t/ and /d/, which are difficult to distinguish in UTI images because they have similar articulatory patterns, can be distinguished from each other.



(a) Binary classification (velar vs. alveolar). UTI distinguishes broader articulatory categories more reliably.



(b) Four-class classification. UTI shows inconsistent performance due to articulatory similarity.

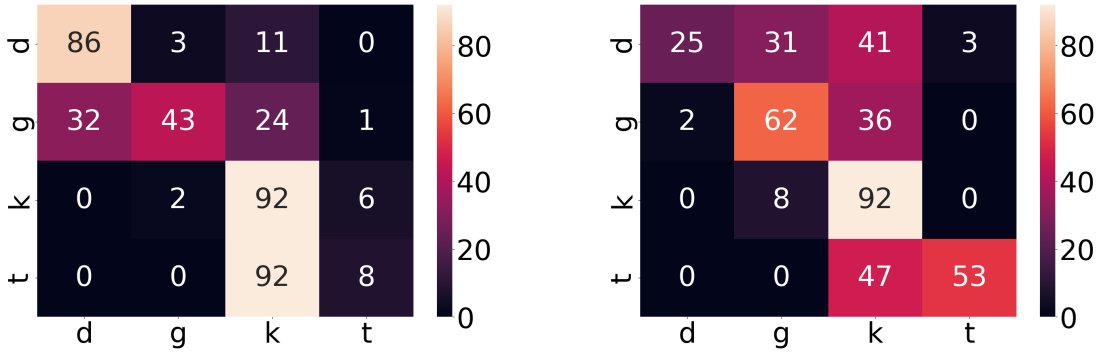
Figure 4.6: Confusion matrices for phoneme classification using UTI data.

The initial stage of the four-class classification task (/k/, /g/, /t/, /d/) produced inconsistent outcomes (Fig. 4.6(a)), indicating that UTI-based models, as previously noted [1], struggle to differentiate between voiced and unvoiced phoneme pairs. Transforming the task into a binary classification (velar versus alveolar) led to a notable increase in performance (Fig. 4.6(b)). While it is useful for differentiating more general articulatory types, UTI data may be insufficient for phoneme classification. Emphasizing the need of multimodal approaches using additional information sources, including audio or radar data, this experimental investigation highlights the limitations of phoneme recognition limited to UTI alone.

4.3.2. Unimodal vs. Multimodal Classification

Emphasizing the benefits of our phonetic segment classification approach over current techniques, this part offers model implementations and performance comparisons.

The classification findings show different benefits of only audio and only radar phoneme recognition. Since both /k/ and /t/ have no vocal cord vibrations while articulation (voiceless), it difficult to identify their articulatory motions using mmWave radar (Fig. 4.7(a). But since mmWave captures articulatory variations that UTI finds difficult, it does better in differentiating /k/ and /g/ that have very similar tongue shapes.

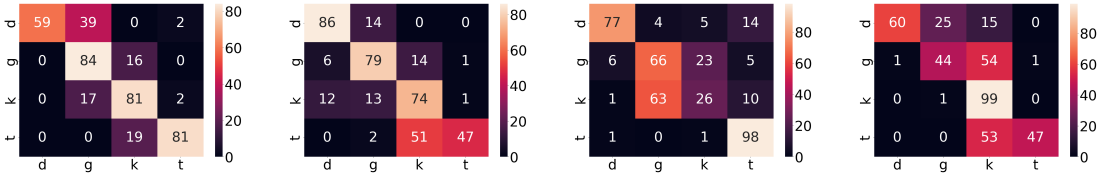


(a) Radar **(b)** Audio

Figure 4.7: Comparison of phoneme classification using (a) mmWave radar and (b) audio.

Conversely, The Audio model outperforms the Radar model for /k/ and /t/ classifica-

tion as it recognizes aspiration and high-frequency bursts, hence separating voiceless phonemes. Figures 4.7(b). But since radar monitors articulatory motion more precisely than audio, it detects voiced phonemes like /d/ and /g/, which often confounds audio because of their comparable spectral characteristics.



(a) MMnet **(b)** UsRadioAI **(c)** AudioUS **(d)** RadarUS

Figure 4.8: Comparison of phoneme classification performance across proposed architectures. The MMnet model demonstrates superior accuracy on the test set, outperforming other multimodal configurations.

Impact of integrating UTI with audio: The inclusion of UTI alongside audio significantly influences phoneme classification outcomes, offering both enhancements and new complexities (Fig. 4.7b vs. Fig. 4.8c). A notable improvement is observed in the classification of /t/ and /k/, where the integration of UTI reduces misclassification errors—likely due to its capacity to capture tongue movement more effectively, which complements audio limitations.

Conversely, the addition of UTI introduces increased confusion between /k/ and /g/, as UTI struggles to differentiate these phonemes given their similar dorsal articulatory patterns. While audio alone better distinguished /k/ and /g/, the fusion with UTI appears to blur this distinction. These results suggest that multimodal fusion can strengthen phoneme discrimination in some cases, yet also introduce new sources of ambiguity, underscoring the importance of thoughtful modality selection and feature-level integration in multimodal systems.

Impact of integration UTI with radar: Integrating UTI with radar data yields a notable enhancement in phoneme classification performance, with an approximate 5% increase in accuracy—most evident in the improved discrimination of /k/ and /t/, which radar alone struggled to distinguish (Table 4.1). This improvement suggests that UTI effectively captures articulatory dynamics, particularly tongue movements, thus mitigating radar's limitations in detecting voiceless phonemes. The integration of UTI reduces

Table 4.1: Accuracy and modalities for phoneme classification with our proposed architectures.

Test Accuracy	Model	Training			Inference		
		UTI	Audio	mmWave	UTI	Audio	mmWave
0.33	US	+			+		
0.57	Radar			+			+
0.58	Audio		+			+	
0.60	RadioAudio		+	+		+	+
0.62	RadioUS	+		+	+		+
0.67	AudioUS	+	+		+	+	
0.71	USRadioAI	+	+	+			+
0.76	MMnet	+	+	+	+	+	+

confusion between /d/ and /g/, which were frequently misclassified in the radar-only model, indicating that articulatory information contributes positively to the classification of voiced phonemes as well (Fig. 4.7(a) vs. Fig. 4.8(d)).

Nonetheless, the incorporation of UTI also introduces new challenges, such as increased misclassification of /d/ with other segments. These findings underscore the importance of careful modality fusion, as the benefits of enhanced articulatory resolution must be balanced against the risk of introducing modality-specific noise or ambiguity.

Figure 4.9: We used t-SNE to project the final classification layer's output vectors into 2D space for visualization. MMnet (top) and USRadioAI (bottom).

Comparing MMNet and USRadioAI: Accuracy vs. Practicality MMNet achieves the highest classification accuracy at 76%, leveraging the full combination of UTI, audio, and radar during both training and inference (Table 4.1, Fig. 4.9). However, this performance gain comes at the cost of increased complexity, as it requires the simultaneous operation of multiple sensing devices, limiting its practicality for real-world deployment. In contrast, USRadioAI is trained with all three modalities but operates solely on radar data at inference. Despite this simplification, it attains a competitive accuracy of 71%, representing a substantial improvement over the radar-only baseline, which achieves just 57%. This highlights the effectiveness of multimodal training in enhancing performance, even when inference relies on a single modality. While USRadioAI falls slightly short of MMNet in terms of accuracy, its operational simplicity, reduced hardware requirements, and cost-efficiency make it a highly viable solution for real-world scenarios, balancing performance with deployability. Table 4.2 summarizes the per-class F1-scores across selected models, highlighting USRadioAI gains.

Table 4.2: F1-scores for each model.

Model	/d/	/g/	/k/	/t/
Audio-only	0.39	0.62	0.58	0.68
Radar-only	0.79	0.58	0.58	0.14
USRadioAI	0.84	0.76	0.62	0.63
MMNet	0.74	0.70	0.75	0.88

4.3.3. Key Findings and Insights

This study investigates phonetic segment classification through a multimodal framework that fuses mmWave radar, UTI, and audio modalities. By analyzing confusion matrices and t-SNE visualizations, we extract critical insights about modality-specific contributions, the robustness of multimodal learning, and practical implications for real-world deployment.

Each modality reveals distinct strengths and weaknesses in phoneme classification:

- The radar-only model performs well for voiced phonemes like /d/ and /g/ but

struggles with voiceless ones, especially /t/, which is misclassified as /d/ in 14 instances. This supports earlier findings that radar sensors are more sensitive to articulatory motion and voicing cues than to aspiration bursts [157].

- The audio-only model excels at classifying voiceless stop consonants due to their acoustic bursts, but shows 39 instances of /d/ being confused with /g/, likely due to spectral similarities—echoing classic results from speech acoustics literature [158].
- The UTI-only (4-class) classifier struggles with /k/ vs. /g/ due to their shared dorsal articulation; 63 samples of /k/ are misclassified as /g/, confirming earlier observations that tongue shape alone is insufficient for voicing contrast [1].

Multimodal integration substantially improves classification performance:

- The MMNet model achieves the highest overall accuracy with clean diagonal structure and minimal cross-phoneme confusion. Its embedding visualization shows clear separation among classes, validating the effectiveness of multimodal learning.
- USRadioAI improves upon the radar-only model by leveraging multimodal training but operates with radar-only input. It achieves 71% accuracy, outperforming radar-only by 14%. The corresponding t-SNE plot reveals structured but overlapping clusters, demonstrating how shared embeddings encode multimodal information.
- AudioUS improves /t/ and /k/ classification compared to audio-only, but introduces confusion between /k/ and /g/. This demonstrates that combining modalities can amplify beneficial cues but also propagate conflicting signals if not integrated carefully.
- RadarUS shows a 5% increase in accuracy over radar-only, especially for /t/ and /d/, indicating that UTI complements radar in resolving tongue-placement ambiguities.
- The UTI-based binary classifier achieves 90%+ accuracy when classifying broad articulatory categories (velar vs. alveolar). However, its four-class performance drops significantly, emphasizing UTI's limitations in resolving fine phonemic

contrasts—particularly voiced/voiceless pairs, consistent with Badin et al. [159].

4.4. Conclusion

While the proposed system has demonstrated encouraging results in the context of radar-assisted phonetic segment classification, it is evident that the system's accuracy, generalizability, and robustness can be further enhanced, particularly under diverse speech conditions, speaker variations, and environmental factors. While the current structure demonstrates competitive performance in classification tasks, the implementation of more advanced models and learning strategies is necessary to adapt to the diversity that can be encountered in real-world scenarios.

Furthermore, the incorporation of the system with radar-based gamification components holds considerable promise in enhancing interaction and enriching the user experience in speech therapy applications. The integration of radar-based gamification components has the potential to enhance the system's interactivity, engagement, and accessibility, particularly in therapeutic processes involving children and individuals with special needs. The contactless and privacy-protecting nature of the radar offers additional advantages in this context.

Subsequent studies will prioritize the enhancement of multimodal training methodologies, the exploration of data synthesis and augmentation techniques, and the restructuring of the system, particularly for real-time speech analysis, with the objective of further extending the boundaries of the proposed system. However, the development of applications that maximize the practical benefits of radar-based speech processing technology is critical to increasing the impact of the system in both clinical and daily use areas.

5. CONCLUSIONS

In this thesis, the problem of phonetic segment classification addressed via deep learning-based multimodal approaches by combining mmWave radar, ultrasound tongue imaging (UTI) and audio data. This multimodal strategy aimed to capture both acoustic and articulatory features, enhancing the model's ability to differentiate subtle phonetic variations. This research addresses the limitations of audio-only systems, such as sensitivity to noise and inability to distinguish similar phonemes, as well as the practical challenges of UTI-based systems, such as high cost and user discomfort.

In this context, the architecture of MMNet, a multimodal neural network designed to take advantage of the complementary aspects of audio and UTI data proposed firstly. The experimental results showed that this integration significantly increases the classification accuracy and reduces the uncertainties encountered in single-mode systems. By combining temporal and spatial articulatory patterns, MMNet demonstrated improved phoneme boundary resolution and reduced confusion among visually similar segments.

The M1 framework was introduced next, which can perform inference with audio data only. This structure revealed how multimodal training can guide single-mode inference processes. Experimental findings showed that the performance of M1 is comparable to approaches that use audio and UTI together, emphasizing its ease of use, especially in applications such as clinical speech therapy. This approach also serves as a practical example of knowledge distillation in multimodal systems, where richer training supervision improves a lightweight inference model.

Considering the practical limitations of fully multimodal systems, the USRadioAI has been developped as an architecture that uses radar, UTI, and audio data during the training phase, but only works with mmWave radar data during inference. By using cross-modal information distillation and joint embedding space techniques, USRadioAI achieved similar accuracy to fully multimodal systems while significantly reducing the implementation complexity. This design also brings strong advantages in terms of privacy, as mmWave signals do not capture identifiable audio or visual features, making the system highly suitable for sensitive or crowded environments.

Extensive experiments on newly created GTUConsonants and GTUSAudioRadioCon-

sonants datasets have shown that the proposed methods are strong in terms of robustness and generalizability. The results showed that multimodal training and embedding space alignment achieve high accuracy despite modality constraints during inference phase, and offer significant advantages in terms of system cost and user experience. The effectiveness of these approaches across datasets collected under varying conditions also underscores their potential for transferability to unseen speakers and acoustic settings.

In conclusion, this thesis demonstrated the importance of integrating mmWave radar with other modalities in speech processing, and provides robust, practical, and privacy-aware solutions. The proposed methods have promising application potential in many areas such as clinical speech therapy, remote diagnosis, assistive technologies and human-computer interaction, and provide a solid ground for future research. Altogether, this work contributes not only a set of practical architectures but also a framework for future multimodal learning systems that aim to balance performance, usability, and privacy in real-world speech applications.

As potential directions for future work, real-time deployment on embedded platforms such as NVIDIA Jetson or other mobile AI hardware could be explored to bring the system closer to in-field or wearable use cases. Another extension involves applying the proposed architectures to speaker-dependent or pathological speech datasets, which would validate their effectiveness in more diverse and clinically relevant conditions. Additionally, exploring alternative modality fusion strategies, such as attention-based or transformer-driven architectures, could further improve model interpretability and performance. The robustness of mmWave radar sensing can also be evaluated under varying environmental conditions, such as clothing interference or device movement, to assess its practical reliability. Lastly, the framework may be expanded for paralinguistic tasks such as emotion recognition, prosody estimation, or speech fluency monitoring, where articulatory features could provide valuable complementary cues.

REFERENCES

[1] Aytutuldu I., Genc Y., Akgul Y. S., (2024), "Audio-only phonetic segment classification using embeddings learned from audio and ultrasound tongue imaging data", IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[2] Wang Y., Yin C., Zhang P., (2024), "Multimodal risk prediction with physiological signals, medical images and clinical notes", Heliyon.

[3] Narayanan S., (2023), Bridging speech science and technology—now and into the future, in "Proc. Interspeech 2023", 1–1.

[4] Grzyb B. J., Vigliocco G., (2022), "Beyond robotic speech: Mutual benefits to cognitive psychology and artificial intelligence from the study of multimodal communication.".

[5] Sezgin E., Militello L. K., Huang Y., Lin S., (2020), "A scoping review of patient-facing, behavioral health interventions with voice assistant technology targeting self-management and healthy lifestyle behaviors", Translational Behavioral Medicine, 10(3), 606–628.

[6] Anwar U., Dong Y., Arslan T., Dashtipour K., Gogate M., Hussain A., Abbasi Q.H., Imran M. A., Russ T. C., Lomax P., (2025), "Privacy-preserving facial emotion classification with visual micro-doppler signatures for hearing aid appli-cations", IEEE Transactions on Instrumentation and Measurement.

[7] Aloufi R., Haddadi H., Boyle D., (2021), "Configurable privacy-preserving auto-matic speech recognition", arXiv preprint arXiv:2104.00766.

[8] Stone M., (2005), "A guide to analysing tongue motion from ultrasound images", Clinical linguistics & phonetics, 19(6-7), 455–501.

[9] Narayanan S., Nayak K., Lee S., Sethy A., Byrd D., (2004), "An approach to real-time magnetic resonance imaging for speech production", The Journal of the Acoustical Society of America, 115(4), 1771–1776.

[10] Chang Y.-h. S., (2023), "Effects of production training with ultrasound biofeedback on production and perception of second-language english tense–lax vowel contrasts", Journal of Speech, Language, and Hearing Research, 66(5), 1479–1495.

[11] Preston J. L., Molfese P. J., Mencl W. E., Frost S. J., Hoeft F., Fulbright R. K., Pugh K. R., (2014), "Structural brain differences in school-age children with resid-ual speech sound errors", Brain and Language, 128(1), 25–33.

[12] Mehta S., (2020), "Effects of Visual Feedback on the Production and Perception of Second Language Speech Sounds: A Comparison of Articulatory and Auditory Instruction", The University of Texas at Dallas.

[13] Denby B., Schultz T., Honda K., Hueber T., Gilbert J. M., Brumberg J. S., (2010), "Silent speech interfaces", Speech Communication, 52(4), 270–287.

[14] Shariff K. K. M., Yusni A. N., Ali M. A. M., Ali M. S. A. M., Tajuddin M. Z. M., Younis M., (2022), Cw radar based silent speech interface using cnn, in "2022 IEEE Symposium on Wireless Technology & Applications (ISWTA)", IEEE, 76–81.

[15] Hueber T., Aversano G., Chollet G., Denby B., Dreyfus G., Oussar Y., Roussel P., Stone M., (2007), Eigentongue feature extraction for an ultrasound-based silent speech interface, in "2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07", Vol. 1, IEEE, I-1245.

[16] Hu S., Xie X., Geng M., Cui M., Deng J., Li G., Wang T., Liu X., Meng H., (2022), "Exploiting cross-domain and cross-lingual ultrasound tongue imaging features for elderly and dysarthric speech recognition", arXiv preprint arXiv:2206.07327.

[17] Wang S., Mei L., Liu R., Jiang W., Yin Z., Deng X., He T., (2024), "Multi-modal fusion sensing: A comprehensive review of millimeter-wave radar and its integra-tion with other modalities", IEEE Communications Surveys & Tutorials.

[18] Laprie Y., Aron M., Berger M.-O., Wrobel-Dautcourt B., (2014), Studying mri acquisition protocols of sustained sounds with a multimodal acquisition system, in "10th International Seminar on Speech Production (ISSP)".

[19] Aron M., Berger M.-O., Kerrien E., Wrobel-Dautcourt B., Potard B., Laprie Y.,(2016), "Multimodal acquisition of articulatory data: Geometrical and temporal registration", The Journal of the Acoustical Society of America, 139(2), 636-648.

[20] Gaddy D., Klein D., (2021), "An improved model for voicing silent speech", arXiv preprint arXiv:2106.01933.

[21] Baljekar P., Sitaram S., Muthukumar P. K., Black A. W., (2015), Using articula-tory features and inferred phonological segments in zero resource speech process-ing., in "INTERSPEECH", 3194-3198.

[22] Wang C., Lin F., Ba Z., Zhang F., Xu W., Ren K., (2022), "Wavesdropper: Through-wall word detection of human speech via commercial mmwave devices", Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Tech-nologies, 6(2), 1-26.

[23] Gou J., Yu B., Maybank S. J., Tao D., (2021), "Knowledge distillation: A survey", International Journal of Computer Vision, 129(6), 1789-1819.

[24] Stone M., (1997), "Laboratory techniques for investigating speech articulation", The handbook of phonetic sciences, 1, 1-32.

[25] Ying J., Shaw J. A., Carignan C., Proctor M., Derrick D., Best C. T., (2021), "Ev-idence for active control of tongue lateralization in australian english/1", Journal of Phonetics, 86, 101039.

[26] Scott A. D., Wylezinska M., Birch M. J., Miquel M. E., (2014), "Speech mri: morphology and function", Physica Medica, 30(6), 604--618.

[27] Toutios A., Byrd D., Goldstein L., Narayanan S., (2019), Advances in vocal tract imaging and analysis, in "The Routledge handbook of phonetics", Routledge, 34--50.

[28] Lingala S. G., Sutton B. P., Miquel M. E., Nayak K. S., (2016), "Recommendations for real-time speech mri", Journal of Magnetic Rezonantça Imaging, 43(1), 28--44.

[29] Zhao Z., Lim Y., Byrd D., Narayanan S., Nayak K. S., (2021), “Improved 3d real-time mri of speech production”, Magnetic Resonance in Medicine, 85(6), 3182–3195.

[30] Akgul Y. S., Kambhamettu C., Stone M., (1998), Extraction and tracking of the tongue surface from ultrasound image sequences, in “Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No. 98CB36231)”, IEEE, 298–303.

[31] Wilson I., Gick B., O’Brien M., Shea C., Archibald J., (2006), Ultrasound technol-ogy and second language acquisition research, in “Proceedings of the 8th Gener-ative Approaches to Second Language Acquisition Conference (GASLA 2006)”, Cascadilla Proceedings Project Somerville, MA, 148–152.

[32] Pattem A. K., Illa A., Afshan A., Ghosh P. K., (2018), “Optimal sensor place-ment in electromagnetic articulography recording for speech production study”, Computer speech & language, 47, 157–174.

[33] Wong M. N., Murdoch B. E., Whelan B.-m., (2010), “Kinematic analysis of lin-gual function in dysarthric speakers with parkinson’s disease: An electromag-netic articulograph study”, International Journal of Speech-Language Pathology, 12(5), 414–425.

[34] Stella M., Bernardini P., Sigona F., Stella A., Grimaldi M., Gili Fivela B.,

(2012), “Numerical instabilities and three-dimensional electromagnetic articulography”, The Journal of the Acoustical Society of America, 132(6), 3941–3949.

[35] Marinelli F., Venegas C., Alarcón J., Navarro P., Fuentes R., (2023), “Chewing analysis by means of electromagnetic articulography: Current developments and new possibilities”, Sensors, 23(23), 9511.

[36] Carey D., McGettigan C., (2017), “Magnetic resonance imaging of the brain and vocal tract: applications to the study of speech production and language learning”, Neuropsychologia, 98, 201–211.

[37] Ramanarayanan V., Tilsen S., Proctor M., Töger J., Goldstein L., Nayak K. S., Narayanan S., (2018), “Analysis of speech production real-time mri”, Computer Speech & Language, 52, 1–22.

[38] Ball M., (2021), “Manual of clinical phonetics”, Routledge.

[39] Aslan E., (2019), Ultrason görüntülerinden uzam-zamansal derin ağlara dayalı dil konturu tespiti, Master’s thesis, Fen Bilimleri Enstitüsü.

[40] McDicken W., Anderson T., (2011), “Basic physics of medical ultrasound”, U: Allan PL, Baxter GM, Weston MJ. Clinical Ultrasound, Volume One, 3, 3–16.

[41] Saini K., Dewal M., Rohit M., (2010), “Ultrasound imaging and image segmenta-tion in the area of ultrasound: a review”, International Journal of advanced science and technology.

[42] Martin K., (2011), “Basic equipment, components and image production”, Clini-cal ultrasound, 1, 16–30.

[43] J. C., (2021), “Ultrasound Tongue Imaging”, Manual of Clinical Phonetics.

[44] J. P., Tara M. B., S. B., S. H., M. T., Emily P., Ahmed R.-C., D. W., (2017), “Ultrasound Images of the Tongue: A Tutorial for Assessment and Remediation of Speech Sound Errors.”, Journal of Visualized Experiments.

[45] Chih-Yen C., Jeng-Wen C., Chun-Hsiang C., Chih-Chung H., (2017), “Tracking Dynamic Tongue Motion in Ultrasound Images for Obstructive Sleep Apnea.”, Ultrasound in Medicine and Biology.

[46] P. B., Stanley Y. L., Peiyu C., A. C., C. K., (2024), “Using standardized ultrasound imaging to correlate OSA severity with tongue morphology.”, Sleep Medicine.

[47] Lianne R., G. W., M. N. D. N.-v. d. S., B. G., C. D. d. K., (2015), “Ultrasound Imaging for Analyzing Lateral Tongue Movements during Mastication in Adults with Cerebral Palsy Compared with Adults without Oral Motor Disabilities.”, Ultrasound in Medicine and Biology.

[48] Huang Q., Zeng Z., (2017), “A review on real-time 3d ultrasound imaging tech-nology”, BioMed research international, 2017(1), 6027029.

[49] Lulich S. M., Pearson Jr W. G., (2019), Three-/four-dimensional ultrasound technology in speech research, Technical report, American Speech-Language-Hearing Association.

[50] Naga Karthik E., Karimi E., Lulich S. M., Laporte C., (2020), “Automatic tongue surface extraction from three-dimensional ultrasound vocal tract images”, The Journal of the Acoustical Society of America, 147(3), 1623–1633.

[51] Ménard L., Aubin J., Thibeault M., Richard G., (2012), “Measuring tongue shapes and positions with ultrasound imaging: A validation experiment using an articulatory model”, Folia Phoniatrica et Logopaedica, 64(2), 64–72.

[52] Xu K., Yang Y., Jaumard-Hakoun A., Leboullenger C., Dreyfus G., Rous-sel P., Stone M., Denby B., (2016), “Development of a 3d tongue motion visualization platform based on ultrasound image sequences”, arXiv preprint arXiv:1605.06106.

[53] Abbott S., Zelesco M., Gibson D., (2020), “Tongue sonography: An introduc-tion”, Sonography, 7(3), 118–125.

[54] Shawker T. H., Sonies Phd B. C., (1984), “Tongue movement during speech: a real-time ultrasound evaluation”, Journal of Clinical Ultrasound, 12(3), 125–133.

[55] Stone M., Davis E. P., (1995), “A head and transducer support system for mak-ing ultrasound images of tongue/jaw movement”, The Journal of The Acoustical Society of America, 98(6), 3107–3112.

[56] Gick B., Bernhardt B., Bacsfalvi P., Wilson I., Zampini M., et al. (2008), “Ul-trasound imaging applications in second language acquisition”, Phonology and second language acquisition, 36(6), 309–322.

[57] Preston J. L., Byun T. M., Boyce S. E., Hamilton S., Tiede M., Phillips E., Whalen D. H., (2017), "Ultrasound images of the tongue: A tutorial for assessment and remediation of speech sound errors", JoVE (Journal of Visualized Experiments), 119, e55123.

[58] Blyth K. M., Mccabe P., Madill C., Ballard K. J., (2016), "Ultrasound visual feedback in articulation therapy following partial glossectomy", Journal of Communication Disorders, 61, 1–15.

[59] Byun T. M., Hitchcock E. R., Swartz M. T., (2014), "Retroflex versus bunched in treatment for rhotic misarticulation: Evidence from ultrasound biofeedback intervention", Journal of Speech, Language, and Hearing Research, 57(6), 2116–2130.

[60] Preston J. L., Hull M., Edwards M. L., (2013), "Preschool speech error patterns predict articulation and phonological awareness outcomes in children with histories of speech sound disorders", Journal of Speech, Language, and Hearing Re-search, 56(5), 1511–1524.

[61] Kaipa R., Peterson A. M., (2016), "A systematic review of treatment intensity in speech disorders", International Journal of Speech-Language Pathology, 18(6), 507–520.

[62] Murray E., McCabe P., Ballard K. J., (2014), "A systematic review of treatment outcomes for children with childhood apraxia of speech", American Journal of Speech-Language Pathology, 23(3), 486–504.

[63] Ribeiro M. S., Eshky A., Richmond K., Renals S., (2019), Speaker-independent classification of phonetic segments from raw ultrasound in child speech, in "ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", IEEE, 1328–1332.

[64] Gibbon F., Wolters M., (2005), A new application of ultrasound to image tongue behaviour in cleft palate speech, in "Poster presentation at the Craniofacial Society of Great Britain and Ireland Annual Scientific Conference", Swansea UK.

[65] Annand C. T., Lamb M., Dugan S., Li S. R., Woeste H. M., Mast T. D., Riley M. A., Masterson J. A., Mahalingam N., Eary K. J., et al. (2019), Using ultrasound imaging to create augmented visual biofeedback for articulatory practice., in "INTERSPEECH", 974–975.

[66] Chen F., Li S., Zhang Y., Wang J., (2017), "Detection of the vibration signal from human vocal folds using a 94-ghz millimeter-wave radar", Sensors, 17(3), 543.

[67] Dong Y., Yao Y.-D., (2020), "Secure mmwave-radar-based speaker verification for iot smart home", IEEE Internet of Things Journal, 8(5), 3500–3511.

[68] Ozturk M. Z., Wu C., Wang B., Liu K. R., (2022), "Radiomic: Sound sensing via radio signals", IEEE Internet of Things Journal, 10(5), 4431–4448.

[69] Li J., Zhang J., (2021), A study of voice print recognition technology, in "2021 International Wireless Communications and Mobile Computing (IWCMC)", IEEE, 1802–1808.

[70] Ozturk M. Z., Wu C., Wang B., Wu M., Liu K. R., (2023), "Radio ses: mmwave-based audioradio speech enhancement and separation system", IEEE/ACM Trans-actions on Audio, Speech, and Language Processing, 31, 1333–1347.

[71] Papastratis I., (2021), “Speech recognition: a review of the different deep learning approaches”, Accessed on, 2, 2021.

[72] Zhu J., Chen L., Xu D., Zhao W., (2022), “Backdoor defence for voice print recognition model based on speech enhancement and weight pruning”, IEEE ac-cess, 10, 114016–114023.

[73] Holzrichter J., Burnett G., Ng L., Lea W., (1998), “Speech articulator measure-ments using low power em-wave sensors”, The Journal of the Acoustical Society of America, 103(1), 622–625.

[74] Birkholz P., Stone S., Wolf K., Plettemeier D., (2018), “Non-invasive silent phoneme recognition using microwave signals”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(12), 2404–2411.

[75] Digehsara P. A., de Menezes J. V. P., Wagner C., Bärhold M., Schaffer P., Plet-temeier D., Birkholz P., (2022), A user-friendly headset for radar-based silent speech recognition., in “INTERSPEECH”, 4835–4839.

[76] Spreafico L., Pucher M., Matosova A., (2018), Ultrafit: A speaker-friendly head-set for ultrasound recordings in speech science, in “Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018”, International Speech Communication Associ-ation, 1517–1520.

[77] Xu C., Li Z., Zhang H., Rathore A. S., Li H., Song C., Wang K., Xu W., (2019), Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface, in “Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services”, 14–26.

[78] Fan L., Xie L., Lu X., Li Y., Wang C., Lu S., (2023), mmmic: Multi-modal speech recognition based on mmwave radar, in “IEEE INFOCOM 2023-IEEE Confer-ence on Computer Communications”, IEEE, 1–10.

[79] Baltrusaitis T., Ahuja C., Morency L.-P., (2019), “Multimodal Machine Learning: A Survey and Taxonomy”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

[80] Liang P. P., Zadeh A., Morency L.-P., (2022), “Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions”, arXiv.org.

[81] Jabeen S., Xi L., Amin M. S., Songyuan L., Abdul J., (2021), “Recent Advances and Trends in Multimodal Deep Learning: A Review”, arXiv.org.

[82] Jabeen S., Li X., Amin M. S., Bourahla O., Li S., Jabbar A., (2023), “A Review on Methods and Applications in Multimodal Deep Learning”, ACM Transactions on Multimedia Computing, Communications, and Applications, 19(2s), 1–41.

[83] Dimitri G. M., (2022), “A Short Survey on Deep Learning for Multimodal Integra-tion: Applications, Future Perspectives and Challenges”, Computers, 11(11), 163.

[84] Thodupunori S. R., (2017), “Analysis of Multimodal Data Using Deep Learning and Machine Learning”, Asian Journal of Humanity, Art and Literature, 4(2), 171–176.

[85] Joshi G., Walambe R., Kotecha K., (2021), “A Review on Explainability in Mul-timodal Deep Neural Nets”, IEEE Access, 9, 59800–59821.

[86] Morency L.-P., Baltrušaitis T., (2017), Multimodal Machine Learning: Integrat-ing Language, Vision and Speech, in “Proceedings of ACL 2017, Tutorial Ab-stracts”, Association for Computational Linguistics, 3–5.

[87] Gholami B., El-Khamy M., Song K., (2024), Knowledge Distillation for Tiny Speech Enhancement with Latent Feature Augmentation, in “Interspeech 2024”, ISCA, 652–656.

[88] Jiang Y., Sharma B., Madhavi M., Li H., (2021), Knowledge Distillation from BERT Transformer to Speech Transformer for Intent Classification, in “Interspeech 2021”, ISCA, 4713–4717.

[89] Tariq A., Banerjee I., Trivedi H., Gichoya J., (2025), “Multimodal artificial intelligence models for radiology”, BJR|Artificial Intelligence.

[90] Schouten D., Nicoletti G., Dille B., Chia C., Vendittelli P., Schuurmans M., Litjens G., Khalili N., (2024), “Navigating the landscape of multimodal AI in medicine: a scoping review on technical challenges and clinical applications”, arXiv.org.

[91] Milosevic M., Jin Q., Singh A., Amal S., (2024), “Applications of AI in multi-modal imaging for cardiovascular disease”, Frontiers in Radiology.

[92] Acosta J. N., Falcone G. J., Rajpurkar P., Topol E. J., (2022), “Multimodal biomedical AI”, Nature Medicine, 28(9), 1773–1784.

[93] Li J., Han X., Qin Y., Tan F., Chen Y., Wang Z., Song H., Zhou X., Zhang Y., Hu L., Hu P., (2023), “Artificial intelligence accelerates multi-modal biomedical process: A Survey”, Neurocomputing, 558, 126720.

[94] Chen J., Seng K. P., Smith J., Ang L.-M., (2024), “Situation Awareness in AI-Based Technologies and Multimodal Systems: Architectures, Challenges and Ap-plications”, IEEE Access, 12, 88779–88818.

[95] Xiaoyu Y., (2021), “Knowledge Distillation for End-to-End Automatic Speech Recognition”.

[96] Yu B. X., Liu Y., Chan K. C., (2021), “Multimodal Fusion via Teacher-Student Network for Indoor Action Recognition”, Proceedings of the AAAI Conference on Artificial Intelligence, 35(4), 3199–3207.

[97] Hori C., Cherian A., Marks T. K., Hori T., (2019), Joint Student-Teacher Learning for Audio-Visual Scene-Aware Dialog, in “Interspeech 2019”, ISCA, 1886–1890.

[98] Kerpicci M., Nguyen V., Zhang S., Visser E., (2022), “Application of Knowledge Distillation to Multi-task Speech Representation Learning”, Interspeech.

[99] Yoon J. W., Lee H., Kim H. Y., Cho W. I., Kim N. S., (2021), “Tutornet: Towards Flexible Knowledge Distillation for End-to-End Speech Recognition”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1626-1638.

[100] Chebotar Y., Waters A., (2016), Distilling Knowledge from Ensembles of Neural Networks for Speech Recognition, in “Interspeech 2016”, ISCA.

[101] Xu Y., Chen Q., (2024), “Radar Can See and Hear as Well: A New Multi-modal Benchmark Based on Radar Sensing”, IEEE Internet of Things Journal, 11(15), 26459–26469.

[102] Zhou T., Chen J., Shi Y., Jiang K., Yang M., Yang D., (2023), “Bridging the View Disparity Between Radar and Camera Features for Multi-Modal Fusion 3d Object Detection”, IEEE Transactions on Intelligent Vehicles, 8(2), 1523–1535.

[103] Wei Z., Zhang F., Chang S., Liu Y., Wu H., Feng Z., (2022), “Mmwave Radar and Vision Fusion for Object Detection in Autonomous Driving: A Review”, Sensors, 22(7), 2542.

[104] Patel K., Heath R. W., (2024), “Harnessing Multimodal Sensing for Multi-User Beamforming in mmWave Systems”, IEEE Transactions on Wireless Communi-cations, 23(12), 18725–18739.

[105] Jain I. K., MM S., Bharadia D., (2024), “Commrad: Context-Aware Sensing-Driven Millimeter-Wave Networks”, ACM International Conference on Embed-ded Networked Sensor Systems.

[106] Wu J., Geng R., Li Y., Zhang D., Lu Z., Hu Y., Chen Y., (2024),

Diffradar:high-Quality Mmwave Radar Perception With Diffusion Probabilistic Model, in “ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)”, IEEE, 8291–8295.

[107] Fukuda R., Sudoh K., Nakamura S., (2021), On Knowledge Distillation for Translating Erroneous Speech Transcriptions, in “Proceedings of the 18th Interna-tional Conference on Spoken Language Translation (IWSLT 2021)”, Association for Computational Linguistics, 198–205.

[108] Aytutuldu I., Bilici Z., Genc Y., Akgul Y. S., (2024), Indoor human counting based on range-doppler spectrum in the presence of material penetration and oc-clusion, in “2024 6th International Conference on Communications, Signal Pro-cessing, and their Applications (ICCSPA)”, IEEE, 1–5.

[109] Bilici Z., Aytutuldu I., Genc Y., Akgul Y. S., (2024), mmwave frequency mod-ulated continuous wave radar-based human action recognition, in “2024 32nd Signal Processing and Communications Applications Conference (SIU)”, IEEE, 1–4.

[110] Pu W., Wu J., Huang Y., Yang J., Yang H., (2019), “Fast factorized backprojec-tion imaging algorithm integrated with motion trajectory estimation for bistatic forward-looking sar”, IEEE Journal of Selected Topics in Applied Earth Observa-tions and Remote Sensing, 12(10), 3949–3965.

[111] Palmer P., Krüger M., Altendorfer R., Bertram T., (2024), Multi-object tracking based on imaging radar 3d object detection, in “International ATZ Conference”, Springer, 93–108.

[112] Zabalza J., Clemente C., Di Caterina G., Ren J., Soraghan J. J., Marshall S.,

(2014), “Robust pca micro-doppler classification using svm on embedded sys-tems”, IEEE Transactions on Aerospace and Electronic Systems, 50(3), 2304–2310.

[113] Ding C., Hong H., Zou Y., Chu H., Zhu X., Fioranelli F., Le Kernec J., Li C.,(2019), “Continuous human motion recognition with a dynamic range-doppler trajectory method based on fmcw radar”, IEEE Transactions on Geoscience and Remote Sensing, 57(9), 6821–6831.

[114] Chen V. C., (2019), “The micro-Doppler effect in radar”, Artech house.

[115] Luo F., Poslad S., Bodanese E., (2019), “Human activity detection and coarse localization outdoors using micro-doppler signatures”, IEEE Sensors Journal, 19(18), 8079–8094.

[116] Alhadhrami E., Al-Mufti M., Taha B., Werghi N., (2019), “Learned micro-doppler representations for targets classification based on spectrogram images”, IEEE Access, 7, 139377–139387.

[117] Schumann O., Hahn M., Dickmann J., Wöhler C., (2018), Semantic segmentation on radar point clouds, in “2018 21st International Conference on Information Fusion (FUSION)”, IEEE, 2179–2186.

[118] Zhang Z., Meng W., Song M., Liu Y., Zhao Y., Feng X., Li F., (2022), “Application of multi-angle millimeter-wave radar detection in human motion be-havior and micro-action recognition”, Measurement Science and Technology, 33(10), 105107.

[119] Karabacak C., Gürbüz S. Z., Gürbüz A. C., (2013), Radar simulation of human micro-doppler signature from video motion capture data, in “2013 21st Signal Processing and Communications Applications Conference (SIU)”, IEEE, 1–4.

[120] Singh V., Bhattacharyya S., Jain P. K., (2020), “Micro-doppler classification of human movements using spectrogram spatial features and support vector ma-chine”, International Journal of RF and Microwave Computer-Aided Engineering, 30(8), e22264.

[121] Shen W., Wang S., Wang Y., Li Y., Lin Y., Zhou Y., Xu X., (2023), “Moving target detection algorithm for millimeter wave radar based on keystone-2dfft”, Electronics, 12(23), 4776.

[122] Wen Q., (2023), Exploring the capabilities of millimeter-wave radar for sensing and imaging: Range-doppler flow based point cloud clustering and effects of motion errors in sar, Master’s thesis, The University of Arizona.

[123] Mozaffari M. H., Lee W.-S., (2019), “Bownet: Dilated convolution neural network for ultrasound tongue contour extraction”, arXiv preprint arXiv:1906.04232.

[124] Wen S., (2018), Automatic tongue contour segmentation using deep learning, PhD thesis, Université d’Ottawa/University of Ottawa.

[125] Ozturk M. Z., Wu C., Wang B., Liu K. R., (2021), Sound recovery from radio signals, in “ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)”, IEEE, 8022–8026.

[126] Hu P., Ma Y., Santhalingam P. S., Pathak P. H., Cheng X., (2022), Mil-liear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary, in “IEEE INFOCOM 2022-IEEE Conference on Computer Communications”, IEEE, 11–20.

[127] Lusiana L., Abadi R. F., Utami Y. T., (2021), “Penerapan metode phonetic placment untuk meningkatkan kemampuan pengucapan konsonan bilabial”, Jurnal UNIK: Pendidikan Luar Biasa.

[128] Lestari N. L. N., Jamal I. M. J. I. M., Maesyabani A. T. T., Bintari H., (2023), “Implementation of phonetic placement method on articulation disorder of down syndrome client”, Jurnal Terapi Wicara.

[129] Guenther F. H., Espy-Wilson C. Y., Boyce S. E., Matthies M. L., Zandipour M., Perkell J. S., (1999), “Articulatory tradeoffs reduce acoustic variability during american english/r/production”, The Journal of the Acoustical Society of America, 105(5), 2854–2865.

[130] Kjellström H., Engwall O., Abdou S. M., Bälter O., (2007), Audio-visual phoneme classification for pronunciation training applications., in “Interspeech”, 702–705.

[131] Evans R., Mullins S., Noel G., (2022), Challenges of vehicle classification using acoustics, in “The International FLAIRS Conference Proceedings”, Vol. 35.

[132] Johnson K., (2023), Individual differences in speech production: What is “phonetic substance?”, in “Proceedings of the International Congress of Phonetic Sciences”, 1102–1106.

[133] Ma Z., Fokoué E., (2014), “Accent recognition for noisy audio signals”, Serdica Journal of Computing, 8(2), 169–182.

[134] Instruments A., (2012), “Articulate assistant advanced user guide: Version 2.16”, Articulate Instruments Ltd.

[135] International Phonetic Association (2020), “International Phonetic Alpha-bet Chart”,
https://www.internationalphoneticassociation.org/IPAcharts/IPA_chart_orig/pdfs/IPA_Kiel_2020_full.pdf, Accessed: March 29, 2025.

[136] Zhang Y., Li B., Fang H., Meng Q., (2022), Spectrogram transformers for audio classification, in “2022 IEEE International Conference on Imaging Systems and Techniques (IST)”, IEEE, 1–6.

[137] Eshky A., Ribeiro M. S., Cleland J., Richmond K., Roxburgh Z., Scobbie J., Wrench A., (2019), “Ultrasuite: A repository of ultrasound and acoustic data from child speech therapy sessions”, arXiv preprint arXiv:1907.00835.

[138] Ani S. A., Cleland J., Zoha A., (2024), “Automated classification of phonetic segments in child speech using raw ultrasound imaging”, arXiv preprint arXiv:2402.17482.

[139] He K., Zhang X., Ren S., Sun J., (2016), Deep residual learning for image recognition, in “Proceedings of the IEEE conference on computer vision and pattern recognition”, 770–778.

[140] Douze M., Guzhva A., Deng C., Johnson J., Szilvasy G., Mazaré P.-E., Lomeli M., Hosseini L., Jégou H., (2024), “The faiss library.(2024)”, arXiv preprint

arXiv:2401.08281.

[141] Sun P., Simcha D., Dopson D., Guo R., Kumar S., (2024), “Soar: improved indexing for approximate nearest neighbor search”, Advances in Neural Informa-tion Processing Systems.

[142] Poria S., Cambria E., Hazarika D., Mazumder N., Zadeh A., Morency L.-P.,(2017), Multi-level multiple attentions for contextual multimodal sentiment analysis, in “2017 IEEE international conference on data mining (ICDM)”, IEEE, 1033–1038.

[143] Baltrušaitis T., Ahuja C., Morency L.-P., (2018), “Multimodal machine learning: A survey and taxonomy”, IEEE transactions on pattern analysis and machine in-telligence, 41(2), 423–443.

[144] Ngiam J., Khosla A., Kim M., Nam J., Lee H., Ng A. Y., et al. (2011), Multimodal deep learning., in “ICML”, Vol. 11, 689–696.

[145] Loizou P. C., (2007), “Speech enhancement: theory and practice”, CRC press.

[146] Wang D., Brown G. J., (2006), “Computational auditory scene analysis: Principles, algorithms, and applications”, Wiley-IEEE press.

[147] Xu Y., Du J., Dai L.-R., Lee C.-H., (2014), “A regression approach to speech enhancement based on deep neural networks”, IEEE/ACM transactions on audio, speech, and language processing, 23(1), 7–19.

[148] Chen T., Shi Y., Zheng Z., Yan B., Hu J., Zhu X. X., Mou L., (2024), Ultrasound Image-to-Video Synthesis via Latent Dynamic Diffusion Models , in

“proceedings of Medical Image Computing and Computer Assisted Intervention–MICCAI 2024”, Vol. LNCS 15004, Springer Nature Switzerland.

[149] Wu H., Fu J., Ye H., Zhong Y., Zou X., Zhou J., Wang Y., (2024), Towards Multi-modality Fusion and Prototype-based Feature Refinement for Clinically Significant Prostate Cancer Classification in Transrectal Ultrasound , in “pro-ceedings of Medical Image Computing and Computer Assisted Intervention –MICCAI 2024”, Vol. LNCS 15005, Springer Nature Switzerland.

[150] Zheng R.-C., Ai Y., Ling Z.-H., (2024), “Incorporating ultrasound tongue images for audio-visual speech enhancement”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32, 1430–1444.

[151] Beeson R., Richmond K., (2023), Silent speech recognition with articulator positions estimated from tongue ultrasound and lip video, in “Interspeech 2023”, ISCA, 1149–1153.

[152] Eshky A., Cleland J., Ribeiro M. S., Sugden E., Richmond K., Renals S.,

(2021), “Automatic audiovisual synchronisation for ultrasound tongue imaging”, Speech Communication, 132, 83–95.

[153] Xu K., Roussel P., Csapó T. G., Denby B., (2017), “Convolutional neural network-based automatic classification of midsagittal tongue gestural targets us-ing b-

mode ultrasound images", The Journal of the Acoustical Society of America, 141(6), EL531–EL537.

[154] Tatulli E., Hueber T., (2017), Feature extraction using multimodal convolutional neural networks for visual speech recognition, in "2017 IEEE international con-ference on acoustics, speech and signal processing (ICASSP)", IEEE, 2971–2975.

[155] Cleland J., (2023), "Ultrasound tongue imaging in research and practice with people with cleft palate±cleft lip", The Cleft Palate Craniofacial Journal, 10556656231202448.

[156] Ozturk M. Z., Wu C., Wang B., Wu M., Liu K. J. R., (2023), "Ra-dioses: mmwave-based audioradio speech enhancement and separation system", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 1333–1347.

[157] Ozturk M. Z., Wu C., Wang B., Wu M., Liu K. R., (2024), "Radiovad: mmwave-based noise and interference-resilient voice activity detection", IEEE Internet of Things Journal.

[158] Stevens K. N., (2000), "Acoustic phonetics", Vol. 30, MIT press.

[159] Badin P., Bailly G., Reveret L., Baciu M., Segebarth C., Savariaux C., (2002), "Three-dimensional linear articulatory modeling of tongue, lips and face, based on mri and video images", Journal of Phonetics, 30(3), 533–553.

BIOGRAPHY

Ilhan AYTUTULDU received the B.S. degree in information systems engineering from Istanbul Technical University, Istanbul, Turkey, in 2015, and the M.S. degree in com-puter engineering from Bahcesehir University, Istanbul, Türkiye, in 2019. From 2015 to 2018, he was a Software Engineer with INTERTECH, Istanbul. Since 2020, he has been a Research Assistant with the Department of Computer Engineering, Gebze Technical University, Gebze, Türkiye. His research interests include the application of computer vision in medical image analysis, speech and language processing, and signal processing.

PUBLICATIONS WITHIN the SCOPE of THESIS STUDY

Paper titled ”Towards Radar-Driven Speech Therapy: Multimodal Training with Ultrasound, Audio, and Radar for Unimodal Phonetic Segment Classification” has been submitted and is currently under review for MICCAI 2025.

Aytutuldu, I., Genc, Y., and Akgul, Y. S. (2024). Audio-Only Phonetic Segment Classification using embeddings Learned from Audio and Ultrasound Tongue Imaging Data. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Aytutuldu, Z. Bilici, Y. Genc and Y. S. Akgul, ”Indoor Human Counting Based on Range-Doppler Spectrum in the Presence of Material Penetration and Occlusion,” 2024 6th International Conference on Communications, Signal Processing, and their Applications (ICCSPA), Istanbul, Turkiye, 2024, pp. 1-5, doi: 10.1109/ICCSPA61559.2024.10794279.

Bilici, Z., Aytutuldu, I., Genc, and Y., Akgul, Y. S. (2024, May). mmWave Frequency Modulated Continuous Wave Radar-based Human Action Recognition. In 2024 32nd Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.