Towards Radar-Driven Speech Therapy: Multimodal Training with Ultrasound, Audio, and Radar for Unimodal Phonetic Segment Classification

İlhan Aytutuldu[1] Yakup Genç[1] Yusuf Sinan Akgül[1]
[1]Gebze Technical University, Department of Computer Engineering, Türkiye
{iaytutuldu, yakup.genc, akgul}@gtu.edu.tr

Introduction

- Phonetic segment classification (consonants such as /k/, /g/, /t/, /d/, /s/, /z/ or vowels such as /a/, /e/, /i/, /o/) is essential for speech therapy and linguistics.
- Audio-only models fail in noisy conditions and with similar phonemes.
- Ultrasound Tongue Imaging (UTI) provides articulatory detail but is costly and impractical.
- mmWave radar enables contactless, privacy-preserving sensing of articulators and vocal-fold vibrations.
- USRadioAI: trains with audio + UTI + radar, but deploys radar-only at inference.

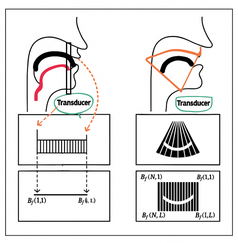
Problem Definition & Objectives

Research Question

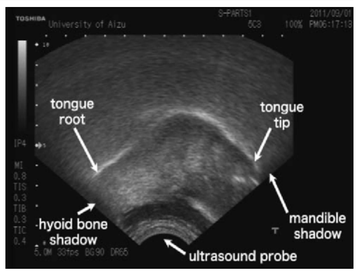
Can radar alone support reliable speech analysis, while leveraging multimodal supervision during training?

- **Assess** UTI and radar classification power independently
- **Train** unimodal vs. multimodal deep learning models
- **Develop** a **radar-only inference** system guided by multimodal knowledge

Preliminary Studies: Ultrasound Tongue Imaging



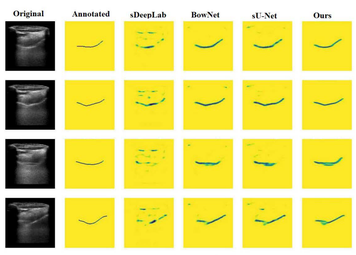
(a) Tongue imaging setup



(b) Example UTI frame

Akgul et al., 2000

Preliminary Studies: UTI in Speech Processing



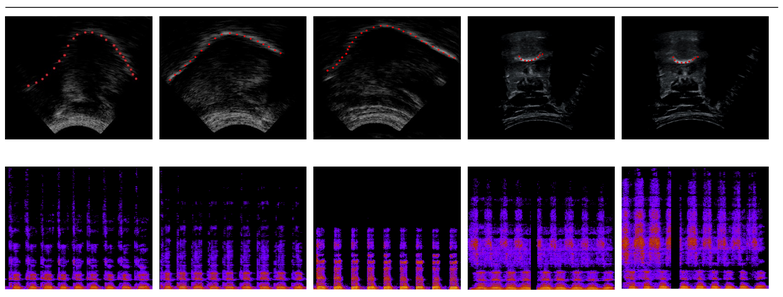
(a) Tongue contour extraction (ours)

(b) Lip + UTI fusion

- **Tongue contours** provide direct articulatory motion cues
- **Lip–UTI fusion** integrates external + internal articulators
- Demonstrates UTI's role in detailed phonetic representation

Ribeiro et al., 2021

Preliminary: Audio + Ultrasound Tongue Imaging



- **Joint UTI + Audio training (MMNet)** achieves 91% accuracy, outperforming unimodal baselines.
- **Audio-only inference (M1)** reaches 81% accuracy, enabling practical, device-free deployment without UTI.

Aytutuldu et al., 2024

References

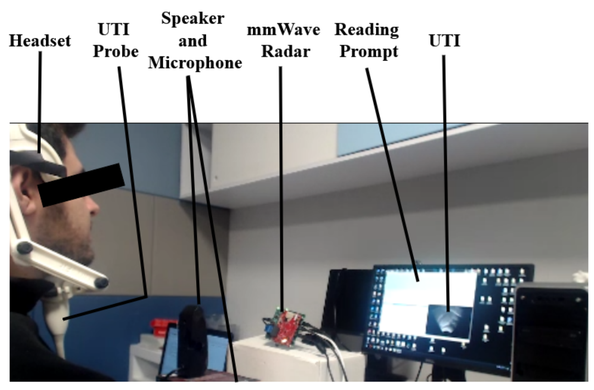
[1] Y.S. Akgul, C. Kambhamettu, and M. Stone. A task-specific contour tracker for ultrasound. In *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. MMBIA-2000 (Cat. No.PR00737)*, pages 135–142, 2000.

[2] Ilhan Aytutuldu, Yakup Genc, and Yusuf Sinan Akgul. Audio-only phonetic segment classification using embeddings learned from audio and ultrasound tongue imaging data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[3] Ribeiro M., Eshky Aciel, Richmond Korin, and Renals S. Speaker-independent Classification of Phonetic Segments from Raw Ultrasound in Child Speech. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2019.

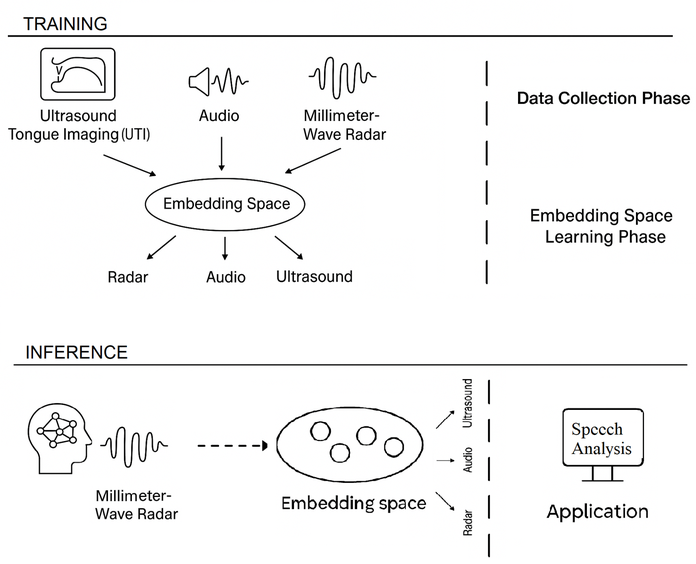
Dataset & Acquisition

- **Participants:** Six native speakers (2F / 4M), ages 18–65
- **Speech Material:** 2,400 utterances of four consonants (/d/, /g/, /k/, /t/)
- **Sensing Modalities:**
  - TI IWR1843 **mmWave radar**
  - Shure Beta 58A **microphone** (48 kHz)
  - **Ultrasound Tongue Imaging (UTI)** device
- **Synchronization:** Achieved via audible ping



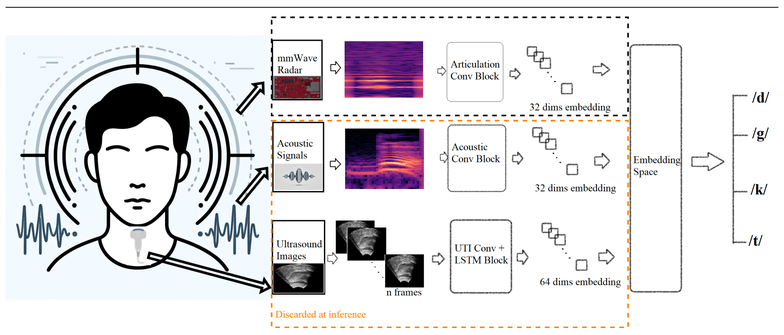
Experimental data collection setup

System Schema



- **Input:** Audio, UTI, and radar signals
- **Training:** Multimodal learning across all three
- **Inference:** Radar-only operation for practicality

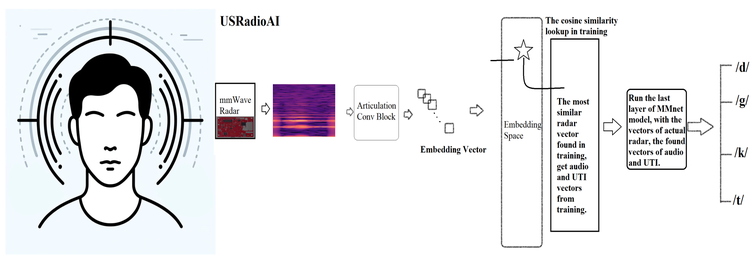
Multimodal Architecture Overview



- Parallel networks process UTI, audio, and radar
- Shared embedding aligns articulatory and acoustic cues
- Enables robust phonetic classification

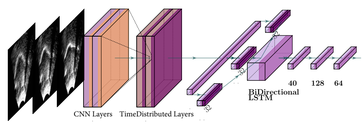
Radar-Only Inference

- **Input:** Radar signals processed into embeddings
- **Comparison:** Radar embeddings are matched against **multimodal reference embeddings** using cosine similarity
- **Advantages:** Lightweight, device-free, and privacy-preserving speech analysis

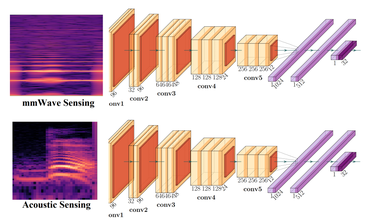


Joint Embedding (MMnet)

- **Two sub-networks** project into a shared embedding space:
  - (a) **UTI:** Spatio-temporal CNN + BiLSTM → 64-D embedding
  - (b) **Radar/Audio:** CNN stacks on spectrograms → 32-D embedding
- **Training:** All modalities are aligned in a common embedding space
- **Inference:** Radar-only, guided by multimodal supervision

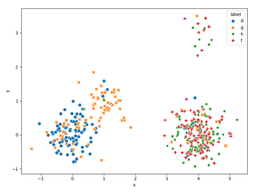


(a) UTI subnetwork

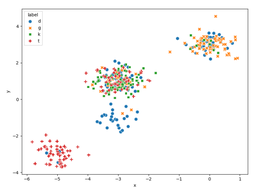


(b) Radar/Audio subnetwork

Unimodal Results: Radar vs. Audio



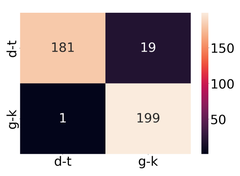
Radar: Strong in **voiced** sounds (/d/, /g/)

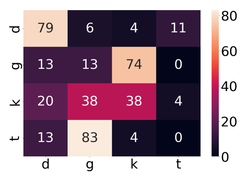


Audio: Strong in **voiceless** sounds (/t/, /k/)

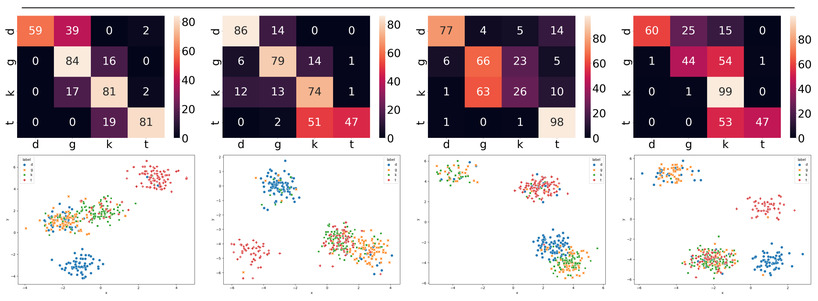
Unimodal Results: UTI

- UTI-only classification
- **2-class:** Velar (/k/, /g/) vs. Alveolar (/t/, /d/) → clear separation
- **4-class:** Harder to distinguish similar tongue positions





Multimodal Fusion Performance



MMNet USRadioAI Audio+UTI Radar+UTI

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	+	+	+	+	+	+

Conclusion and Future Work

MMNet vs. USRadioAI

- **MMNet:** 76% accuracy (UTI + Audio + Radar) → highest performance, but low practicality
- **USRadioAI:** 71% accuracy (radar-only inference) → practical, cost-effective, easy deployment
- Tradeoff: small accuracy drop ↔ major usability gain

Future Work

- Clinical validation on speaker-dependent & pathological speech
- Extend to paralinguistic tasks (emotion, prosody, fluency)
- **Dual-radar setup:** track face + vocal folds simultaneously

Acknowledgment

This work was supported by the **Health Institutes of Türkiye (TÜSEB), C-YZ Project (No. 33944)** and by **Gebze Technical University Scientific Research Projects (No. 2023-A-102-04)**. The authors thank both institutions for their financial support. Special thanks are extended to the Department of Speech and Language Therapy and the MEDKOM Research Group at Medipol University for their valuable support in data collection.