aurum

AURUM MÜHENDİSLİK SİSTEMLERİ VE MİMARLIK DERGİSİ
AURUM JOURNAL OF ENGINEERING SYSTEMS AND ARCHITECTURE

Cilt 9, Sayı 2 | Kış 2025
Volume 9, No 2 | Winter 2025

MAKİNE ÖĞRENME TEKNİKLERİ İLE SES TANIMA: LİTERATÜR TARAMASI

Mutlu Merih AKTUZ[1], Dr. Ayça KURNAZ TÜRKBEN[2], Doç. Dr. Sefer KURNAZ [3]
[1]Elektrik ve Bilgisayar Eğitimi, Altınbaş Üniversitesi, İstanbul, 233720211@ogr.altinbas.edu.tr
(https://orcid.org/0009-0003-9793-1572)
[2]Altınbaş Üniversitesi, İstanbul, ayca.turkben@altinbas.edu.tr
(https://orcid.org/0000-0002-8541-9964)
[3]Altınbaş Üniversitesi, İstanbul, sefer.kurnaz@altinbas.edu.tr
(https://orcid.org/0000-0002-7666-2639)

Received: 22.05.2025 Review Article
Accepted: 01.07.2025 pp.251-270
Published: 31.12.2025
*Corresponding author DOI: 10.53600/ajesa.1704161

Özet

Bu çalışmada, ses tanıma alanında kullanılan makine öğrenme yöntemleriyle ilgili literatür taraması yapılarak, farklı makine öğrenme yöntemlerini karşılaştırmak ve en etkili yöntemi belirlemek amaçlanmıştır. Çalışma kapsamında, Ocak 2023-Mart 2025 tarihleri arasında yayınlanan güncel literatür taranarak 30 çalışma detaylı olarak incelenmiştir. Çalışmada, derin öğrenme tabanlı yaklaşımların, özellikle Transformer mimarileri ve uçtan uca öğrenme modellerinin, geleneksel yöntemlere kıyasla daha yüksek doğruluk ve sağlamlık sergilediği sonucuna varılmıştır. Bununla birlikte, ideal bir ses tanıma sistemi için tek bir "en iyi" yaklaşım yerine, uygulama senaryosuna, mevcut kaynaklara ve hedef kullanıcı grubuna bağlı olarak farklı yaklaşımların kombinasyonunun daha uygun olabileceği değerlendirilmiştir. Ses tanıma sistemlerinin gelişiminde kaydedilen önemli ilerlemelere rağmen, büyük miktarda etiketli veri ihtiyacı, gürültülü ortamlardaki performans düşüşleri gibi çeşitli problemler ve sınırlamalar hala mevcuttur. Gelecekteki araştırmalar için düşük kaynaklı diller için yarı-denetimli ve öz-denetimli öğrenme yaklaşımları, hibrit model mimarileri, çok görevli ve çok modlu öğrenme yaklaşımları, nöromorfolojik hesaplama yaklaşımları ve standartlaştırılmış değerlendirme metrikleri üzerine çalışmalar önerilmiştir.

Anahtar Kelimeler: Ses tanıma, makine öğrenme, derin öğrenme, otomatik konuşma tanıma, özellik çıkarma

BAŞLIK

Abstract

This study aims to compare different machine learning methods and determine the most effective method by reviewing the literature on machine learning methods used in the field of voice recognition. Within the scope of the study, 30 studies were analyzed in detail by reviewing the current literature published between January 2023 and March 2025. The study concluded that deep learning-based approaches, especially Transformer architectures and end-to-end learning models, exhibit higher accuracy and robustness compared to traditional methods. However, instead of a single “best” approach for an ideal voice recognition system, a combination of different approaches may be more appropriate depending on the application scenario, available resources and target user group. Despite the significant progress made in the development of voice recognition systems, several problems and limitations still exist, such as the need for large amounts of labeled data and performance degradation in noisy environments. For future research, semi-supervised and self-supervised learning approaches for low-resource languages, hybrid model architectures, multi-task and multi-modal learning approaches, neuromorphological computational approaches, and standardized evaluation metrics are proposed.

Keywords: Voice recognition, machine learning, deep learning, automatic speech recognition, feature extraction

1. Giriş

Ses tanıma teknolojisi, konuşma sinyallerini analiz ederek dilsel içeriği yazılı metne dönüştürmektedir (Guo, 2023). Makine öğrenme algoritmalarındaki gelişmeler, ses tanıma sistemlerinin doğruluğunu ve kullanılabilirliğini önemli ölçüde iyileştirmiştir (Prabhavalkar vd., 2023). Ses tanıma sistemlerinde kullanılan makine öğrenme teknikleri; (1) Gaussian Karışım Modelleri ve Gizli Markov Modelleri gibi geleneksel yaklaşımlar (Malik vd., 2021), (2) Derin Sinir Ağları ve Transformatör mimarilerini içeren derin öğrenme yaklaşımları (Kheddar vd., 2023) ve (3) geleneksel

ve derin öğrenme tekniklerinin güçlü yönlerini birleştiren hibrit yaklaşımlar (Barhoush vd., 2023) olmak üzere üç ana kategoride incelenebilir.

Ses tanıma sistemlerinin geliştirilmesinde konuşma varyasyonu, gürültülü ortamlar ve farklı aksanlar gibi temel problemler bulunmaktadır (Tandel vd., 2020). Bu problemlerin çözümü için çeşitli teknikler geliştirilmiş olsa da, bunların karşılaştırmalı analizi literatürde yeterince ele alınmamıştır.

Bu çalışmanın amacı, ses tanıma alanındaki mevcut literatürü inceleyerek farklı makine öğrenme yöntemlerini karşılaştırmak ve en etkili yöntemleri belirlemektir. Araştırma kapsamında, IEEE Xplore, ACM Digital Library ve diğer önemli veri tabanlarında 2023-2025 yılları arasında yayınlanan çalışmalar incelenmiştir. İlk taramada tespit edilen 51 çalışmadan, belirlenen kriterlere uygun 30 çalışma detaylı analize tabi tutulmuştur.

Çalışmada, farklı makine öğrenme tekniklerinin performans, hesaplama karmaşıklığı, veri gereksinimi, gürültüye dayanıklılık ve uygulanabilirlik açısından karşılaştırılması yapılmıştır. Çalışma sonucunda, ses tanıma alanında en etkili makine öğrenme teknikleri belirlenmiş ve gelecekteki araştırmalar için öneriler sunulmuştur.

2. Kavramsal Çerçeve

2.1. Ses Tanımanın Temelleri

Ses tanıma sistemlerinin temeli, konuşma sinyallerinin dijital temsiline dayanır. Akustik dalga formundaki konuşma, mikrofon aracılığıyla elektrik sinyallerine, ardından sayısal verilere dönüştürülür. Sayısallaştırma sürecinde, sürekli ses dalgası belirli aralıklarla örneklenerek (genellikle saniyede 16,000 örnek) işlenebilir hale getirilir (Mishra vd., 2024). Sistemlerin işlevsel mimarisi ön işleme, özellik çıkarma, akustik modelleme, dil modelleme ve kod çözme aşamalarından oluşur.

Ön işleme aşamasında gürültü azaltma ve sinyal normalizasyonu uygulanarak ses kalitesi artırılır (Wang, 2023). Temel zorluklardan biri konuşma varyasyonudur. Aksanlar, konuşma hızı ve duygusal durum gibi faktörlerden kaynaklanan farklılıkları ele alabilmek için sistemlerin geniş örneklem üzerinde eğitilmesi gerekir (Tandel vd., 2020). Konuşma sinyallerinden anlamlı özelliklerin çıkarılması büyük önem taşımaktadır. Yaygın kullanılan yöntemlerden biri olan Mel Frekansı Kepstral Katsayıları (MFCC), insan işitme sisteminin frekans algısını taklit ederek spektral özellikleri çıkarır (Ahmed vd., 2023). Akustik modelleme, ses tanımanın merkezinde yer alır ve konuşma sinyallerindeki özelliklerin dilin temel birimlerindeki karşılıklarını modeller. Geleneksel sistemlerde Gizli Markov Modelleri (HMM) ve Gauss Karışım Modelleri (GMM) kullanılırken, günümüzde derin öğrenme tabanlı modeller, özellikle Tekrarlayan Sinir Ağları (RNN) ve Evrişimli Sinir Ağları (CNN) üstün performans göstermektedir (Gençyılmaz ve Karaoğlan, 2024). Dil modelleme, akustik model çıktılarını daha doğru metinlere dönüştürmeye yardımcı olur. Geleneksel N-gram modellerinin yerini günümüzde Transformer mimarisine dayalı derin öğrenme modelleri almıştır (Liu vd., 2023). Son aşamada, kod çözme veya arama adımıyla akustik ve dil modellerinin çıktıları birleştirilerek en olası metin dizisi Viterbi algoritması veya ışın arama gibi yöntemlerle belirlenir (Yao vd., 2024).

2.2. Makine Öğrenme Yaklaşımları

Ses tanıma sistemlerinde kullanılan makine öğrenme yaklaşımları geleneksel, derin öğrenme ve hibrit yöntemler olarak sınıflandırılabilir. Geleneksel yaklaşımlar arasında Gaussian Karışım Modelleri (GMM), Gizli Markov Modelleri (HMM), Destek Vektör Makineleri (SVM) ve K-En Yakın Komşu (KNN) algoritmaları öne çıkmaktadır (Malik vd., 2021). Bu yöntemler özellikle sınırlı veri ve hesaplama kaynakları ile çalışırken etkili sonuçlar üretebilmektedir. Örneğin, KNN algoritması kullanılarak Türkçe izole kelime tanımada gürültülü ortamlarda %86,51'e varan tanıma oranları elde edilmiştir (Keser, 2023).

Derin öğrenme yaklaşımları, son on yılda geleneksel yöntemlere kıyasla kelime hata oranında %50'den fazla azalma sağlayarak ses tanımada önemli bir dönüşüm yaratmıştır (Prabhavalkar vd., 2023). Derin Sinir Ağları (DNN), Evrişimli Sinir Ağları (CNN) ve Tekrarlayan Sinir Ağları (RNN) gibi mimariler, ham ses verilerinden anlamlı özellikleri otomatik olarak çıkarabilme yetenekleriyle manuel özellik mühendisliği ihtiyacını azaltmıştır. Son yıllarda Transformatör mimarileri, öz-dikkat mekanizmaları sayesinde ASR sistemlerinde devrim yaratmıştır (Kheddar vd., 2023). Conformer mimarisi, Transformatör ve CNN'nin güçlü yönlerini birleştirerek hem yerel hem de global bağımlılıkları öğrenmek için popüler bir kodlayıcı modeli haline gelmiştir (Yao vd., 2024).

Uçtan uca ASR sistemleri, geleneksel sistemlerin ayrı bileşenlerini tek bir sinir ağı modeline entegre ederek Connectionist Temporal Classification (CTC), Attention-based Encoder-Decoder (AED) ve RNN-Transducer (RNN-T) gibi çeşitli yöntemler kullanmaktadır. Derin transfer öğrenme (DTL), özellikle düşük kaynaklı dillerde veya özel alanlarda önemli performans iyileştirmeleri sağlamıştır (Kheddar vd., 2023). Wav2Vec ve DeepSpeech gibi önceden eğitilmiş modeller, büyük miktarda etiketlenmemiş ses verisi üzerinde ön eğitim alarak konuşma temsillerini öğrenmekte ve görev-spesifik uygulamalar için ince ayar yapılabilmektedir (Jia, 2023).

Hibrit yaklaşımlar, farklı algoritmaların ve model mimarilerinin güçlü yönlerini birleştirerek tek başına kullanıldıklarında karşılaşılan sınırlamaları aşmayı hedeflemektedir. HMM-ANN hibrit sistemleri, konuşma tanıma alanında önemli bir dönüm noktası olmuştur. Bu sistemlerde ANN'ler akustik olasılıkları tahmin ederken, HMM'ler zamansal değişkenlikleri modellemektedir (Malik vd., 2021). Özellik çıkarma ve sınıflandırma aşamalarının hibrit tasarımı, konuşma tanıma sistemlerinin başarısını artırmaktadır. (DCVA)PCA ve (FLDA-KNN)PCA gibi hibrit alt uzay sınıflandırıcıları, geleneksel sınıflandırıcılardan daha yüksek tanıma oranları elde etmiştir (Barhoush vd., 2023).

Hibrit akustik ve dil modellerinin birleştirilmesi, özellikle alan-spesifik konuşma tanıma sistemlerinde önemli performans artışları sağlamaktadır. Önceden eğitilmiş Wav2Vec2-Large-LV60 akustik modeli ve harici bir KenLM dil modelinin birleştirilmesiyle oluşturulan hibrit sistem, ticari ASR sistemlerini geçen bir performans göstermiştir (Jia, 2023). Çoklu model füzyonu yaklaşımında, farklı mimarilere sahip modellerin çıktıları birleştirilerek daha doğru ve sağlam tanıma sonuçları elde edilmektedir. Gizlilik korumalı hibrit konuşma tanıma sistemleri, federe öğrenme ve diferansiyel gizlilik tekniklerini birleştirerek, kullanıcı verilerinin gizliliğini korurken yüksek tanıma performansı sağlamaktadır (Kheddar vd., 2023).

2.3. Ses Tanıma Sistemlerinde Özellik Çıkarma

Ses tanıma sistemlerinde özellik çıkarma, ham ses sinyallerinden anlamlı ve ayırt edici bilgilerin elde edilmesi sürecidir. Yüksek boyutlu ve gürültülü ses sinyallerinden özellik çıkarma, boyutsallığı azaltırken önemli bilgileri korur ve gürültüye karşı dayanıklılığı artırır.

Zaman alanı özellikleri, ses sinyallerinin zamansal boyutunda doğrudan analiz edilmesiyle elde edilir. Enerji özelliği, sessiz/sesli bölümlerin ayrımında kullanılırken, sıfır geçiş oranı (ZCR) frekans içeriği hakkında bilgi sağlar (Gourisaria vd., 2024). Zaman alanı özellikleri, sınırlı kaynaklara sahip gömülü sistemlerde tercih edilir ancak gürültüye karşı daha hassastır (Ahmed vd., 2024).

Frekans alanı özellikleri, ses sinyallerinin spektral içeriğini temsil eder ve insan işitme sisteminin algısal özelliklerini modelleyerek sistem performansını artırır (Zaman vd., 2023). Kısa Süreli Fourier Dönüşümü (STFT), sinyalin farklı zaman noktalarındaki frekans bileşenlerini analiz eder. Mel-Frekans Kepstral Katsayıları (MFCC), ses tanıma sistemlerinde en yaygın kullanılan frekans alanı özelliklerindendir ve sinyalin spektral zarfını temsil eder. Frekans alanı özellikleri, gürültülü ortamlarda daha dayanıklıdır çünkü gürültü genellikle belirli frekans bantlarını etkiler (Barhoush vd., 2023).

Dalgacık dönüşümü, sinyalin hem zaman hem de frekans bilgisini eş zamanlı sunarak durağan olmayan ses sinyallerinin analizinde avantaj sağlar (Ganchev, 2021). Sürekli Dalgacık Dönüşümü (CWT) ve Ayrık Dalgacık Dönüşümü (DWT) olmak üzere iki ana kategoriye ayrılır. Çok çözünürlüklü analiz (MRA), sinyalin farklı çözünürlük seviyelerinde analiz edilmesini sağlar. Dalgacık dönüşümü tabanlı özellikler, çok çözünürlüklü yapısı sayesinde gürültülü ortamlarda bile ses sinyalinin önemli özelliklerini koruyabilir (Ahmed vd., 2023).

Doğrusal Tahmin Kodlaması (LPC), konuşma sinyallerinin geçmiş örneklerinin doğrusal kombinasyonu olarak modellendiği bir tekniktir. İnsan ses üretim sisteminin fiziksel özelliklerini matematiksel olarak temsil eder (Mehrish vd., 2023). LPC katsayıları, konuşma sinyalinin spektral zarfını temsil eder ve az sayıda parametre ile sinyalin özelliklerini kodlar, ancak gürültülü ortamlarda performansı düşebilir (Barhoush vd., 2023).

Perceptual Linear Prediction (PLP), insan işitme sisteminin psikoakustik özelliklerini dikkate alarak konuşma sinyallerinden anlamlı özellikler çıkarmayı amaçlar. PLP analizi, konuşma sinyalinin spektral temsilini Bark ölçeğine göre yeniden örnekler ve insan algısına göre düzenler. Özellikle gürültülü ortamlarda ve farklı konuşmacıların seslerini tanımada etkilidir (Keser, 2023).

Derin öğrenme tabanlı özellik çıkarma yöntemleri, manuel tasarlanan özellikler yerine ham ses sinyallerinden otomatik olarak anlamlı özellikler öğrenebilme kapasitesiyle üstün performans gösterir (Mehrish vd., 2023). Evrişimli sinir ağları (CNN), spektrogram temsillerinden yerel desenleri tespit ederek hiyerarşik yapıda temsil eder. Tekrarlayan sinir ağları (RNN), ses sinyallerinin zamansal yapısını modellemede kullanılır. Wav2Vec 2.0 gibi kendini denetimli öğrenme yaklaşımları, etiketlenmemiş büyük ses veri kümeleri üzerinde önceden eğitilerek düşük kaynaklı diller için avantaj sağlar (Mehrish vd., 2023).

2.4. Ses Tanıma Sistemlerinde Veri Setleri

Veri setleri, konuşma tanıma sistemlerinin eğitilmesi, test edilmesi ve değerlendirilmesi için temel kaynaktır. Sistemlerin başarısı büyük ölçüde kullanılan veri setlerinin kalitesine, büyüklüğüne ve çeşitliliğine bağlıdır.

Genel amaçlı konuşma veri setleri, çeşitli konuşmacılardan, farklı akustik ortamlardan ve geniş kelime dağarcığından oluşan konuşma örneklerini içerir. Özellikle derin öğrenme tabanlı yaklaşımların başarısı, eğitim için kullanılan veri setlerinin kalitesine ve miktarına bağlıdır (Prabhavalkar vd., 2023). Bu veri setleri genellikle saatler veya binlerce saat uzunluğunda konuşma kayıtları içerir. Google, Universal Speech Model (USM) için 12 milyon saatlik etiketlenmemiş çok dilli veri seti kullanmaktadır (Zhang vd., 2023). Genel amaçlı konuşma veri setlerinin oluşturulmasında demografik çeşitlilik, gürültü koşulları, duygusal durumlar ve iletişimde kullanılan dil gibi faktörler önemlidir (Dhanjal ve Singh, 2024).

Komut tanıma veya anahtar kelime tespiti (keyword spotting) veri setleri, sınırlı sayıda kelime veya kısa ifade içeren özel amaçlı veri setleridir. Bu veri setleri, modellerin belirli komutları hızlı ve doğru bir şekilde tanımasını sağlamak için tasarlanmıştır (Rai vd., 2023). Google Konuşma Komutları Veri Seti, bu alanda yaygın olarak kullanılmaktadır. Fayda-spesifik konuşma tanıma sistemleri geliştirmek için önceden eğitilmiş DeepSpeech2 ve Wav2Vec2 akustik modelleri kullanılmaktadır (Jia, 2023). Qwen2-Audio adlı büyük ölçekli ses-dil modeli, çeşitli ses sinyali girişlerini kabul edebilmekte ve konuşma talimatlarına ilişkin ses analizi veya doğrudan metinsel yanıtlar üretebilmektedir (Chu vd., 2023).

Alan-spesifik veri setleri, belirli bir endüstri, sektör veya uygulama alanına özgü konuşma verilerini içeren özelleştirilmiş koleksiyonlardır. Bu tür veri setleri, genel amaçlı konuşma tanıma sistemlerinin performansını belirli alanlarda artırmak için kullanılır. Alan-spesifik veri setlerinin önemli bir özelliği, hedef alanın terminolojisini ve jargonunu kapsamlı bir şekilde temsil etmeleridir. Tıbbi, hukuk, finans veya mühendislik gibi alanlarda kullanılan ASR sistemleri, ilgili alanın teknik dilini ve terminolojisini doğru bir şekilde tanıyabilmelidir (Dhanjal ve Singh, 2024). Alan-spesifik veri setlerinin oluşturulmasında karşılaşılan zorluklardan biri, düşük kaynaklı diller veya özel alanlar için yeterli miktarda etiketli veri bulmaktır. Bu sorunu aşmak için veri artırma teknikleri, transfer öğrenimi ve yarı-denetimli öğrenme yaklaşımları gibi çözümler önerilmektedir (Prabhavalkar vd., 2023).

Çok dilli veri setleri, birden fazla dilde konuşma örneklerini içerir ve dil-bağımsız veya çok dilli ASR sistemlerinin geliştirilmesinde temel yapı taşları olarak kullanılır. Google'ın Universal Speech Model (USM), 100'den fazla dilde otomatik konuşma tanıma yapabilen bir modeldir ve 12 milyon saatlik etiketlenmemiş çok dilli veri kullanılarak eğitilmiştir (Zhang vd., 2023). Düşük kaynaklı diller için ASR sistemleri geliştirmek daha zordur. Çok dilli öğrenme ve transfer öğrenimi teknikleri, kaynak açısından zengin dillerden düşük kaynaklı dillere bilgi transferini sağlar (Prabhavalkar vd., 2023). Transformatör tabanlı modeller, farklı diller arasındaki ortak özellikleri öğrenebilmekte ve bu bilgiyi düşük kaynaklı dillere transfer edebilmektedir (Mehrisha vd., 2023).

2.5. Ses Tanıma Sistemlerinde Başarı Kriterleri

Otomatik konuşma tanıma sistemlerinin performansını değerlendirmek için çeşitli metrikler kullanılmaktadır. Kelime Hata Oranı (WER), en yaygın kullanılan metriktir ve ikame, ekleme ve silme hatalarını dikkate alarak hesaplanır (Dhanjal ve Singh, 2024). Karakter Hata Oranı (CER) ve Cümle Hata Oranı (SER) da benzer şekilde sistemlerin doğruluğunu ölçmektedir (Kheddar vd., 2023).

Konuşma kalitesi değerlendirmesinde Ölçek-Değişmez Sinyal-Bozulma Oranı (SI-SDR) ve Konuşma Kalitesinin Algısal Değerlendirmesi (PESQ) öne çıkmaktadır. SI-SDR, hedef sinyal ile tahmin edilen sinyal arasındaki ilişkiyi ölçerken, PESQ insan algısını modelleyerek konuşma kalitesini değerlendirir. Kısa Süreli Nesnel Anlaşılabilirlik (STOI), konuşma anlaşılabilirliğini 0-1 aralığında ölçen bir metriktir (Ganchev, 2021).

Anahtar kelime tespiti gibi görevlerde doğruluk ve F1-skoru gibi geleneksel sınıflandırma metrikleri kullanılmaktadır (Rai vd., 2023). Gerçek zamanlı faktör (RTF), sistemin hesaplama verimliliğini ölçer ve özellikle sınırlı kaynaklı cihazlar için önemlidir (Prabhavalkar vd., 2023). Konuşma tanıma sistemleri ayrıca modelin sağlamlığı, genelleştirme yeteneği ve güvenlik değerlendirmeleri gibi kriterlere göre de incelenmektedir.

3. Literatür Taraması

Shukla vd. (2023), otomatik konuşma tanıma (ASR) sistemlerini makine öğrenme teknikleriyle incelemişlerdir. Büyük kelime dağarcıklı sürekli konuşma tanıma (LVCSR) problemlerine odaklanarak HMM-GMM, HMM-DNN hibrit modelleri ve uçtan uca modelleri karşılaştırmışlardır. CTC tabanlı ASR sistemlerinde "soft forgetting" tekniğini önererek aşırı uyumu önlemeyi amaçlamışlardır. MFCC özellik çıkarımı kullanılan çalışmada, 300 saatlik Switchboard veri setiyle yapılan deneylerde, önerilen teknik WER'de %7-9 iyileşme sağlamıştır. Hub5-2000 test setlerinde konuşmacıdan bağımsız modeller için %9.1/%17.4 WER ve adapte edilmiş modeller için %8.7/%16.8 WER elde edilmiştir. Ayrıca, dikkat tabanlı kodlayıcı-kod çözücü modellerinin performansını artırmak için daha küçük modelleme birimleri ve orta düzeyde çerçeve hızı azaltmanın etkili olduğu gösterilmiştir.

Gourisaria vd. (2024), çevresel ses sınıflandırması için makine öğrenme tekniklerinin karşılaştırmalı analizini yapmışlardır. MFCC ve STFT özellik çıkarma yöntemleriyle yedi farklı makine öğrenme modelini (Lojistik Regresyon, KNN, SVM, Naive Bayes, Karar Ağacı, Rastgele Orman ve ANN) karşılaştırmışlardır. UrbanSound8K (10 sınıf, 8732 örnek) ve Sound Event Audio Dataset (8 sınıf, 1288 örnek) veri setleri kullanılmıştır. Her ses örneği için 186 özellik çıkarılmış, 4 saniye uzunluğunda kesilmiş ve gürültü temizleme uygulanmıştır. Sonuçlar, ANN modelinin her iki veri setinde de en iyi performansı gösterdiğini (sırasıyla %91.41 ve %91.27 doğruluk) ortaya koymuştur. Çalışma, basit ANN mimarilerinin dahi düşük hesaplama maliyetiyle yüksek doğruluk sağlayabileceğini göstermiştir.

Gençyılmaz ve Karaoğlan (2024), Türkçe konuşmadan metne dönüşüm (CoST) sistemlerinin optimizasyonu için makine öğrenme yaklaşımlarını analiz etmişlerdir. Türkçe'nin eklemeli yapısı, fonolojik özellikleri ve aksan farklılıkları gibi zorluklara odaklanmışlardır. 26.485 Türkçe ses klibinden oluşan veri setiyle MFCC özellik çıkarımı kullanarak dokuz farklı makine öğrenme modelini (CNN, CRNN, LSTM, RF, SVM, KNN, GNB, DT, GRU)

karşılaştırmışlardır. En yüksek doğruluk CRNN yaklaşımıyla (%96.47), en yüksek F1-skoru (%97.55) ve kesinlik (%97.68) ise CNN modeliyle elde edilmiştir. Çalışma, Türkçe dilinde konuşmadan metne dönüşüm alanında literatürdeki ilk çalışmalardan biridir.

Ayall vd. (2024), Etiyopya'nın resmi dili Amharca için konuşulan rakam tanıma sistemi geliştirmişlerdir. 120 gönüllüden toplanan 12.000 ses kaydı içeren Amharca Konuşulan Rakamlar Veri Seti (AmSDD) oluşturulmuştur. MFCC ve Mel-Spektrogram yöntemleriyle özellik çıkarımı yapılmış, LDA, KNN, SVM ve RF algoritmaları test edilmiştir. Önerilen üç katmanlı CNN mimarisi, MFCC özellikleriyle %99, Mel-Spektrogram özellikleriyle %98 doğruluk elde etmiştir. Geleneksel yöntemler arasında en iyi performansı MFCC ile RF göstermiş (%97). Çalışmanın en önemli katkısı, az kaynaklı bir dil için yüksek doğrulukta konuşma tanıma sistemi geliştirmesidir.

Wojnar vd. (2024), genel amaçlı konuşma tanıma modellerini değerlendirmek için YouTube'u veri kaynağı olarak kullanan "Mi-Go" aracını geliştirmişlerdir. Çalışmada, statik veri setlerinin gerçek dünya senaryolarındaki genelleştirilebilirliği sınırladığı belirtilmiştir. Mi-Go, YouTube'dan veri çıkarma ve değerlendirme sürecini otomatikleştirmektedir. 141 YouTube videosu üzerinde yapılan deneylerde Whisper (tiny.en, base.en, small.en, medium.en, large-v1, large-v3), NVIDIA'nın Conformer-Transducer X-Large ve ESPnet2 modelleri test edilmiştir. Whisper large-v3 %10.58 WER ile en iyi performansı gösterirken, tiny.en %43.64 WER ile en kötü performansı sergilemiştir. NeMo Transducer Xlarge %26.06, ESPnet2 %30.97 WER değerleri elde etmiştir.

Alzaabi vd. (2024), engelli bireylerin ev aletlerini kontrol edebilecekleri İngilizce ve Arapça (Emirlik lehçesi) komutları tanıyan düşük maliyetli gömülü bir konuşma tanıma sistemi geliştirmişlerdir. 40 katılımcıdan toplanan 527 İngilizce ve 680 Arapça ses örneğinden oluşan veri seti, çeşitli arka plan gürültüleriyle zenginleştirilmiştir. Ses sinyalleri spektrogramlara dönüştürülerek CNN modelleri ile işlenmiştir. Test edilen farklı CNN mimarileri arasında, 30 epoch boyunca eğitilen 3 konvolüsyonel katmanlı model en iyi performansı göstermiştir (%99.84 doğruluk, %99.87 kesinlik, %99.82 duyarlılık). Model ESP32 mikrodenetleyicisine yerleştirilerek gerçek dünya senaryolarında test edilmiştir. Sonuçlar, spektrogram tabanlı görüntü tanıma tekniklerinin düşük kaynaklı donanımlarda bile yüksek doğrulukta çalışan çok dilli konuşma tanıma sistemleri geliştirmede etkili olduğunu göstermiştir.

Barhoush vd. (2023), konuşmacı tanıma ve lokalizasyonu için basit bir tam bağlantılı derin sinir ağı (FC-DNN) ve sadece iki mikrofon kullanan yeni bir model önermişlerdir. Geliştirdikleri Karıştırılmış MFCC (SHMFCC) ve Fark Karıştırılmış MFCC (DSHMFCC) özellikleriyle, veri artırma için ön işleme ve test aşamasında son işleme adımları içeren bir yaklaşım sunmuşlardır. LibriSpeech veri setiyle yapılan testlerde, tek konuşmacı senaryosunda %98.5 tanıma ve %99.6 lokalizasyon doğruluğu; çoklu konuşmacı senaryosunda ise %95.2 tanıma ve %98.7 lokalizasyon doğruluğu elde edilmiştir. Sistem, düşük gürültü seviyelerinde (0 dB SNR) bile %90'ın üzerinde doğruluk sağlamış ve geleneksel yöntemlerden daha iyi performans göstermiştir.

Chen (2023), uç nokta algılama algoritmasına dayalı bir konuşma tanıma sistemini İngilizce kelime öğrenimine entegre etmiştir. HTK konuşma tanıma fonksiyonu ve Markov modeli kullanılarak geliştirilen sistemde, enerji, sıfır geçiş oranı, MFCC ve LPC tabanlı özellikler kullanılmıştır. Konuşma sinyallerinin işlenmesinde çerçeveleme, pencereleme ve uç nokta algılama teknikleri uygulanmıştır. Sistem özellikle fonem bazında hata tespiti ve eksik

okuma tespitine odaklanmış, sesli ünsüzlerin telaffuzundaki hataları tespit etmiş ve farklı fonemlere yönelik hata tolerans mekanizmaları önermiştir.

Xie (2023), makine öğrenme tabanlı konuşma tanıma teknolojisini ağ tabanlı İngilizce konuşma öğretimine uygulamıştır. Öğrencilerin öğrenme stillerini ve davranışlarını tahmin eden sistem, kısa süreli enerji, ZCR, MFCC ve LPC tabanlı özellikler kullanmıştır. Konuşma sinyalleri çerçeveleme, pencereleme ve uç nokta algılama teknikleriyle işlenmiş, sınıflandırma için Naive Bayes algoritması kullanılmıştır. Sekiz öğrenci üzerinde yapılan değerlendirmede, öğrenme tercihlerine göre içerik sunumunda %59.50 ile %92.63 arasında doğruluk oranları elde edilmiştir.

Keser (2023), gürültülü ortamlarda farklı makine öğrenme ve hibrit altuzay sınıflandırıcılarının yalıtık kelime tanıma performanslarını karşılaştırmıştır. Çalışmada KNN, FLDA-KNN, DCVA, SVM, CNN ve RNN-LSTM sınıflandırıcıları kullanılmış, özellik çıkarımı için MFCC ve PLP katsayıları tercih edilmiştir. Literatürde ilk kez DCVA sınıflandırıcısı yalıtık kelime tanıma için derinlemesine test edilmiştir. Temel Bileşen Analizi (PCA) kullanılarak tasarlanan hibrit sınıflandırıcılar ((DCVA)PCA ve (FLDA-KNN)PCA), orijinal sınıflandırıcılardan daha iyi performans göstermiştir. En yüksek tanıma oranı RNN-LSTM ile %93.22 olarak elde edilmiştir.

Guo (2023), üniversite İngilizce eğitiminde gürültülü ortamlarda konuşma tanıma sorununu çözmek için çift sensörlü bir sistem önermiştir. Sistem, çene derisi titreşim basıncı ve hava yoluyla iletilen konuşma sinyallerini toplayarak derin makine öğrenme algoritması ile konuşma tanıma gerçekleştirmektedir. Özellik çıkarımında doğrusal tahmin katsayıları kullanılmış, eğitim için LSTM mimarisi tercih edilmiştir. Önerilen algoritma, geleneksel ve DNN-HMM gibi hibrit yaklaşımlara kıyasla daha yüksek doğruluk ve daha hızlı yakınsama sağlamıştır.

Jia (2023), düşük kaynak koşullarında alan-spesifik konuşma tanıma için derin öğrenme tabanlı bir sistem geliştirmiştir. Çalışan hakları ve sağlık sigortası alanına özgü ASR sistemi için DeepSpeech2 ve Wav2Vec2 akustik modelleri kullanılmıştır. Alan-spesifik veriler, minimal insan müdahalesi gerektiren yarı denetimli öğrenme etiketleme yöntemi ile toplanmıştır. Harici bir KenLM dil modeli ile ince ayar yapılmış Wav2Vec2-Large-LV60 akustik modeli, çalışan hakları alanına özgü konuşmalarda Google ve AWS ASR sistemlerinden daha iyi performans göstermiştir. Çalışma, alan-spesifik verilere ince ayar yapılan önceden eğitilmiş akustik modellerin ticari ASR sistemlerinden daha iyi sonuç verebileceğini kanıtlamıştır.

Rahate vd. (2023), felçli hastalar için EEG sinyalleri kullanarak sessiz konuşma tanıma sistemi geliştirmiştir. 10 denekten 6 farklı kelime için toplanan EEG verileri, Butterworth filtre ve çentik filtre ile önişlemeye tabi tutulmuş, Ampirik Mod Ayrıştırma ile özellikler çıkarılmıştır. ANOVA testi ile seçilen özellikler üzerinde yedi farklı algoritma test edilmiş; KNN %61.45 doğruluk ve %60.93 duyarlılık, LDA ise %79.47 kesinlik ve %65.26 F1-skoru ile öne çıkmıştır.

Accou vd. (2023), EEG'den konuşma zarfını çözümlemek için VLAAI adlı derin sinir ağını geliştirmiştir. Çoklu CNN blokları ve özel bağlam katmanlarından oluşan mimari, 80 katılımcıdan toplanan 141 saatlik veri üzerinde eğitilmiştir. VLAAI, mevcut kişiden bağımsız doğrusal modellere göre %52 performans artışı sağlayarak 0.19

Pearson korelasyon değerine ulaşmış, kişiye özel ince ayarla bu değer 0.25'e yükselmiştir. Doğrusal olmayan bileşenler ve çıktı bağlam modülü performansa en büyük katkıyı (%10) sağlamıştır.

Wang (2023), İngilizce konuşma tanıma için Tekrarlayan Sinir Ağı (RNN) ve Bağlantıcı Zamansal Sınıflandırma (CTC) algoritmasını birleştiren bir yaklaşım önermiştir. MFCC özellik çıkarma yöntemi kullanılan sistem, 630 katılımcılı TIMIT veri seti üzerinde test edilmiştir. RNN-CTC modeli, 2500 test örneği için %10.4 kelime hata oranı ile CNN-CTC (%15.8) ve GMM-HMM (%21.1) modellerini geride bırakmıştır. Eğitim süreleri açısından da RNN-CTC (351.4 saniye), CNN-CTC (410.7 saniye) ve GMM-HMM (498.7 saniye) modellerine göre daha verimli bulunmuştur. Çalışmada benimsenen yaklaşım, konuşma-metin hizalama sorununu çözerek hem doğruluk hem de hesaplama verimliliği açısından üstünlük sağlamaktadır.

Hamian vd. (2023), otomatik rakam ses tanıma için çok amaçlı optimizasyon algoritmalarıyla derin darbeli sinir ağları (SNN) yaklaşımı önermiştir. TIDIGITS veri setinde MFCC ve ZCR özelliklerinden seçilen 7 optimal özellik kullanılmıştır. SNN-WHO modeli %98.2 doğruluk ve 0.0182 RMS hata ile en iyi performansı göstermiş, bunu SNN-GBO (%96.4), ANN (%93.6) ve ANFIS (%90.9) izlemiştir. Benzer üstünlük IRIS ve Trip veri setlerinde de gözlenmiştir. Çalışma, optimize edilmiş SNN'lerin geleneksel yapay sinir ağlarına göre daha verimli çalıştığını kanıtlamıştır.

Ameen ve Kadhim (2023), elektrolainks kullanan kanser hastaları için Arapça fonem tanıma sisteminde çeşitli makine öğrenme modellerini karşılaştırmıştır. 8 kişiden kaydedilen 27 Arapça fonem sınıfı için 40 MFCC özelliği çıkarılmıştır. ANN modeli %75 doğruluk, %77 kesinlik ve %21.85 PER ile en iyi performansı göstermiş, bunu CNN-LSTM (%72 doğruluk) ve CNN-XGB (%71 doğruluk) izlemiştir. Çalışma, gürültülü elektrolainks konuşması için ANN'nin en uygun model olduğunu ortaya koymuştur.

Shisode vd. (2023), konuşma engelli kişiler için yüzey elektromiyografi (SEMG) yaklaşımı kullanan bir konuşma tanıma sistemi geliştirmiştir. Yüz kaslarından alınan EMG sinyalleri ve dudak hareketleri videoları kullanılarak GMM ve CNN modelleri uygulanmıştır. EMG sistemi "Emergency" kelimesini %80, "Heart" kelimesini %75, "Water" kelimesini %60 ve "Better" kelimesini %65 doğrulukla tanımıştır. Dudak okuma sistemi eğitim verisi üzerinde %90, test verisi üzerinde %53 doğruluk sağlamıştır. Sistem özellikle larenjektomi geçiren hastalarda kullanılabilecek bir çözüm sunmaktadır.

Ahmed vd. (2024), az kaynaklı Peştuca dili için verimli bir konuşma tanıma sistemi geliştirmiştir. Çalışmada MFCC ve DWT özellik çıkarma teknikleri ile SVM ve k-NN sınıflandırıcıları karşılaştırılmıştır. 30 konuşmacıdan toplanan 161 izole kelimelik veri setinde (312 eğitim, 130 test örneği) yapılan deneylerde, MFCC+SVM kombinasyonu %96.15 doğrulukla en iyi performansı göstermiştir. Bunu MFCC+KNN (%92.31), DWT+SVM (%88.46) ve DWT+KNN (%84.62) izlemiştir. Sonuçlar, MFCC özellik çıkarma ve SVM sınıflandırma kombinasyonunun Peştuca izole kelime tanıma için en etkili yaklaşım olduğunu ortaya koymuştur.

Froiz-Míguez vd. (2023), az kaynaklı Galiçya dili için wav2vec 2.0 tabanlı bir ASR sistemi geliştirmiştir. Sistem, OpenSLR ve Common Voice'dan alınan yaklaşık 20 saatlik etiketli ses verisiyle ince ayar yapılmıştır. Galiçya Parlamentosu'ndan alınan spontane konuşma verileriyle yapılan değerlendirmede, sadece wav2vec 2.0 modeli

%28.67 WER sağlarken, ışın dekoder eklenmesi WER'i %27.42'ye düşürmüş, 51.1 milyon kelimelik dil modelinin eklenmesiyle WER %18.61'e kadar iyileşmiştir. Çalışma, diller arası transfer öğreniminin az kaynaklı dillerde etkin bir yaklaşım olduğunu göstermiştir.

Kwon (2024), ses tanıma sistemlerini optimize edilmiş karşıt örneklere karşı korumak için "AudioGuard" adlı bir savunma yöntemi önermiştir. Yöntem, ayrı bir modül gerektirmeden, gürültü vektörü kullanarak modelin normal örnekler üzerindeki doğruluğunu korurken karşıt örnekleri tespit etmektedir. Mozilla Common Voice veri seti ve DeepSpeech modeliyle yapılan deneylerde, önerilen yöntem karşıt örnekleri %84.2 doğrulukla tespit ederken, normal örnekler üzerindeki model doğruluğunu %94.3 olarak korumuştur.

Rai vd. (2023) konuşma içindeki komutları tespit eden derin öğrenme tabanlı anahtar kelime tanıma sistemleri geliştirmiştir. Çalışmada MFCC özellik mühendisliği ve CNN için 8000 Hz ile yeniden örneklenmiş ham ses dalgaları kullanılmıştır. Google Speech Commands Dataset v2 üzerinde yapılan deneylerde RNN-BiLSTM-Attention modeli %93.9 doğrulukla en iyi performansı göstermiş, bunu RNN-BiLSTM (%92.5), RNN-LSTM (%91.2), CNN (%89.7) ve HMM-GMM (%65.2) izlemiştir. Sonuçlar, dikkat mekanizmalı çift yönlü LSTM modellerinin anahtar kelime tanımada üstün performans sergilediğini ve derin sinir ağlarının geleneksel modellerden önemli ölçüde daha iyi çalıştığını göstermiştir.

Kumar vd. (2023) gürültülü ortamlarda veya ses sinyalinin olmadığı durumlarda kullanılmak üzere dikkat mekanizması tabanlı kodlayıcı-kod çözücü mimarisi kullanan bir görsel konuşma tanıma (VSR) modeli geliştirmiştir. Sistemde 68-şekil kestiricisi yöntemiyle dudak bölgesi tespiti yapılmış ve BERT tabanlı dil modeli entegre edilmiştir. GRID Corpus üzerinde %2.8 WER ve LRS2 Corpus üzerinde %40.1 WER değerleri elde edilmiştir. Çalışma, dikkat mekanizması tabanlı mimarinin dudak okuma görevinde etkili olduğunu göstermiştir.

Weng vd. (2023) konuşma tanıma ve sentezi için derin öğrenme tabanlı bir semantik iletişim sistemi (DeepSC-ST) geliştirmiştir. Sistem, konuşma sinyallerinden metin ile ilgili semantik özellikleri çıkarmakta ve bu özellikleri iletişim kanalları üzerinden aktarmaktadır. Düşük SNR koşullarında (0 dB) DeepSC-ST yaklaşık %10 CER ve %22 WER değerleri elde ederken, geleneksel sistem %85 CER ve %95 WER değerlerine ulaşmıştır. 8 dB SNR değerinde DeepSC-ST'nin performansı %3 CER ve %7 WER değerlerine kadar iyileşmiştir. Konuşma sentezi kalitesi açısından, sistem Tacotron 2 modeline yakın sonuçlar vermiştir (FDSD: ~0.3, KDSD: ~0.05).

Chen vd. (2023) konuşma tanıma ve çevirisi için bağlam içi öğrenme yetenekli bir Konuşma Destekli Dil Modeli (SALM) geliştirmiştir. Model, dondurulmuş metin tabanlı LLM, ses kodlayıcı, modalite adaptörü ve LoRA katmanlarından oluşmaktadır. 110M parametreli Fast Conformer ses kodlayıcısı ve 2B parametreli Megatron LLM kullanılmıştır. SALM, ASR görevinde test-clean için %2.3, test-other için %4.8 WER değerlerine; AST görevinde İngilizce-Almanca için 29.6, İngilizce-Japonca için 16.5 BLEU puanlarına ulaşmıştır. Çalışmanın en önemli katkısı, konuşmadan metne modelleri ilk kez sıfır atımlı bağlam içi öğrenme yeteneği ile donatmasıdır.

Gong vd. (2023) hem konuşma hem de konuşma dışı sesleri eş zamanlı tanıyabilen LTU-AS modelini geliştirmiştir. Model, Whisper'ı algılama modülü ve LLaMA'yı akıl yürütme modülü olarak entegre etmektedir. Mimari olarak Whisper-large ses kodlayıcı-kod çözücü, AudioSet üzerinde önceden eğitilmiş TLTR ve LLaMA-7B kullanılmıştır.

9.6 milyon örnekli Open-ASQA veri setiyle eğitilen model, toplam 8.5 milyar parametreye sahip olmasına rağmen sadece 49 milyon parametre eğitilebilir durumdadır. LTU-AS, ses ve konuşma hakkında serbest biçimli soruları %95'in üzerinde talimat takip oranıyla yanıtlayabilmektedir. Çalışma, hem konuşma hem de konuşma dışı sesleri tanıyıp aralarındaki ilişkileri anlayabilen ilk modeli sunmuştur.

Wang vd. (2023) konuşma tanıma, sentezi ve çevirisi görevlerini tek bir model altında birleştiren VIOLA'yı geliştirmiştir. Çalışmada, tüm konuşma ifadeleri çevrimdışı bir sinir codec kodlayıcısı (EnCodec) kullanılarak ayrık belirteçlere dönüştürülmüştür. Farklı dilleri ve görevleri ele alabilmek için modele görev kimlikleri ve dil kimlikleri entegre edilmiştir. VIOLA, İngilizce ASR'de %9.4, Çince ASR'de %10.2 WER elde etmiştir. Makine çevirisinde İngilizce-Çince için 28.7, Çince-İngilizce için 24.3 BLEU puanı sağlamıştır. TTS görevlerinde 0.85 konuşmacı benzerliği ve 4.1 MOS değeri elde edilmiştir. Çalışma, codec tabanlı otomatik regresif Transformer decoder modellerini çeşitli konuşma işleme görevleri için nicel olarak araştıran ilk çalışmadır.

Zhang vd. (2023) 100'den fazla dilde ASR yapabilen Evrensel Konuşma Modeli'ni (USM) geliştirmiştir. Model, 300+ dili kapsayan 12 milyon saatlik etiketlenmemiş veri ile ön-eğitime tabi tutulmuş ve etiketli veri ile ince ayar yapılmıştır. Rastgele projeksiyon kuantalama ve konuşma-metin modalite eşleştirmesi kullanılarak, Whisper'ın etiketli veri setinin 1/7'si büyüklüğünde bir eğitim seti ile birçok dilde daha iyi performans elde edilmiştir. 2B parametreli Conformer mimarisi kullanan USM, SpeechStew'da %5.7-7.6 WER, FLEURS'de %16.1-28.4 WER ve CoVoST 2'de %28.3 BLEU puanı sağlamıştır. Çalışma, çok dilli konuşma tanıma ve çevirisinde tek bir büyük modelin küçük etiketli veri setleriyle yüksek performans elde edebileceğini göstermiştir.

Yao vd. (2024) otomatik konuşma tanıma için daha hızlı, bellek açısından verimli ve yüksek performanslı Zipformer modelini sunmuştur. Model, (1) Orta katmanların düşük çerçeve hızlarında çalıştığı U-Net benzeri kodlayıcı yapısı, (2) Dikkat ağırlıklarını yeniden kullanan verimli blok yapısı, (3) Uzunluk bilgilerini koruyan BiasNorm, (4) SwooshR ve SwooshL aktivasyon fonksiyonları olmak üzere dört temel yenilik içermektedir. Ayrıca, parametre ölçeğini öğrenen ScaledAdam optimize edici önerilmiştir. Altı kademeli yapıya sahip Zipformer, LibriSpeech'te test-clean için %1.9 WER, test-other için %3.9 WER elde etmiştir. Model, eğitimde daha hızlı yakınsamakta ve çıkarımda %50'den fazla hızlanma sağlarken daha az GPU belleği gerektirmektedir.

Chu vd. (2024) ses sinyallerini işleyip konuşma talimatlarına göre analiz yapabilen veya metinsel yanıtlar verebilen Qwen2-Audio modelini geliştirmiştir. Karmaşık hiyerarşik etiketler yerine doğal dil komutları kullanılarak ön-eğitim süreci basitleştirilmiştir. Model, sesli sohbet ve ses analizi olmak üzere iki etkileşim modu sunmaktadır. Whisper-large-v3 tabanlı ses kodlayıcı ve Qwen-7B dil modelinden oluşan mimari, 16kHz'e örneklenen ham dalga formunu 128 kanallı mel-spektrogramına dönüştürmekte ve iki adımlı havuzlama ile ses temsil uzunluğunu azaltmaktadır. Herhangi bir göreve özgü ince ayar olmadan LibriSpeech'te %92.62 (1-WER%), CoVoST2'de %30.0 (ortalama BLEU) ve AIR-Bench sohbet kıyaslamasında 6.43-6.88/10 puan elde etmiştir. Çalışmanın özgün katkısı, sistem komutu gerektirmeden sesli sohbet ve ses analizi modları arasında akıllı geçiş yapabilen bir model geliştirmesidir.

4. Karşılaştırma

4.1. Başarım Karşılaştırması

Literatürdeki ses tanıma çalışmaları incelendiğinde, derin öğrenme tabanlı yaklaşımların geleneksel yöntemlere kıyasla üstün performans sergilediği görülmektedir. Gourisaria vd. (2024) çevresel ses sınıflandırmasında ANN modeliyle %91.41 doğruluk oranı elde ederken, Gençyılmaz ve Karaoğlan (2024) Türkçe konuşmadan metne dönüşümde CRNN modeliyle %96.47 doğruluk oranına ulaşmıştır. Keser (2023) gürültülü ortamlarda RNN-LSTM modeliyle %93.22 doğruluk oranı elde etmiştir.

Derin öğrenme mimarileri arasında, RNN tabanlı modeller ve dikkat mekanizması içeren yapılar öne çıkmaktadır. Rai vd. (2023) RNN-BiLSTM-Attention modeliyle %93.9 doğruluk oranı elde ederken, Wang (2023) RNN-CTC modeliyle %10.4 WER değerine ulaşmıştır. Transformatör mimarisine dayalı büyük modeller, özellikle Wojnar vd. (2024) tarafından test edilen OpenAI'nin Whisper large-v3 modeli %10.58 WER ile yüksek başarım göstermiştir.

Hibrit yaklaşımlar, ses tanıma performansını daha da artırmaktadır. Hamian vd. (2023) Wild Horse Optimization algoritmasıyla optimize edilen SNN-WHO modeliyle %98.2 doğruluk oranı elde etmiş, Ahmed vd. (2024) Peştuca dili için MFCC ve SVM kombinasyonuyla %96.15 doğruluk oranına ulaşmıştır.

Özel görevlere yönelik modeller, sınırlı kelime dağarcığı olan uygulamalarda etkileyici sonuçlar vermektedir. Ayall vd. (2024) Amharca rakam tanımada CNN modeliyle %99 doğruluk oranı elde ederken, Alzaabi vd. (2024) ev aletleri kontrolü için geliştirdikleri çok dilli sistemde %99.84 doğruluk oranına ulaşmıştır. Barhoush vd. (2023) konuşmacı tanıma ve lokalizasyonunda karıştırılmış MFCC özellikleriyle %98.5 ve %99.6 doğruluk oranları elde etmiştir.

Çok dilli ve transfer öğrenimi yaklaşımları, düşük kaynaklı diller için önemli avantajlar sunmaktadır. Froiz-Míguez vd. (2023) Galiçya dili için Wav2Vec 2.0 modeliyle %18.61 WER değerine ulaşırken, Zhang vd. (2023) geliştirdikleri Evrensel Konuşma Modeli ile 100'den fazla dilde %5.7 WER değeri elde etmiştir. Chen vd. (2023) Konuşma Destekli Dil Modeli ile LibriSpeech veri setinde test-clean için %2.3 ve test-other için %4.8 WER değerlerine ulaşmıştır.

Sonuç olarak, ses tanıma alanında en etkili yöntemler görev ve veri seti özelliklerine göre değişmektedir. Sınırlı kelime dağarcığına sahip görevlerde CNN tabanlı modeller yüksek başarım gösterirken, karmaşık ve geniş kelime dağarcıklı görevlerde RNN, LSTM ve Transformer tabanlı modeller daha etkili olmaktadır.

4.2. Karmaşıklık Karşılaştırması

Ses tanıma sistemlerinde kullanılan makine öğrenme tekniklerinin karmaşıklığı hesaplama gereksinimleri, model boyutu, eğitim süresi ve çıkarım hızı açısından değerlendirilmektedir. Gourisaria vd. (2024) tarafından yapılan çalışmada, Yapay Sinir Ağı (ANN) modeli en yüksek başarımı gösterirken, hesaplama karmaşıklığı da en yüksek değere sahiptir. Karar Ağacı ve Naive Bayes gibi geleneksel modeller daha düşük hesaplama karmaşıklığına sahip olmasına rağmen, başarım açısından daha düşük sonuçlar vermektedir.

Wojnar vd. (2024) çalışmasında, Whisper modellerinin boyutu arttıkça hesaplama karmaşıklığının da arttığı, ancak başarımın da buna paralel olarak iyileştiği gösterilmiştir. Whisper large-v3 modeli 1.55 milyar parametreye sahipken, tiny.en modeli sadece 39 milyon parametreye sahiptir. Alzaabi vd. (2024) çalışmasında, 3 konvolüsyonel

katmanlı CNN modelinin, hesaplama karmaşıklığı ve başarım arasında optimum denge sağladığı ve ESP32 gibi sınırlı kaynaklı mikrodenetleyicilerde çalıştırılabilecek kadar verimli olduğu gösterilmiştir.

Zhang vd. (2023) tarafından geliştirilen Evrensel Konuşma Modeli (USM), 600 milyon parametreye sahip olup yüksek hafıza gereksinimleri göstermektedir. Buna karşın, Hamian vd. (2023) tarafından geliştirilen derin darbeli sinir ağları (SNN) modeli, geleneksel yapay sinir ağlarına göre daha düşük hafıza gereksinimleri sunmaktadır. Ayall vd. (2024) çalışmasında, 3 katmanlı CNN mimarisinin 16-bit kayan noktalı sayı formatında yaklaşık 2.3 MB boyutunda olduğu ve düşük kaynaklı cihazlarda çalıştırılabilecek kadar kompakt olduğu belirtilmiştir.

Eğitim süresi açısından, Gençyılmaz ve Karaoğlan (2024) çalışmasında, CRNN modelinin 100 epoch boyunca eğitilmesi yaklaşık 3 saat sürerken, Rastgele Orman modeli sadece birkaç dakika içinde eğitilebilmiştir. Rai vd. (2023) çalışmasında, dikkat mekanizması eklenmesinin eğitim süresini artırdığı, ancak modelin daha hızlı yakınsadığı gösterilmiştir. Barhoush vd. (2023) çalışmasında, Karıştırılmış MFCC özelliklerinin kullanılmasının modelin daha hızlı yakınsamasını sağladığı belirtilmiştir.

Çıkarım hızı açısından, Alzaabi vd. (2024) çalışmasında, 3 konvolüsyonel katmanlı CNN modelinin komutları 200-300 milisaniye içinde tanıyabildiği gösterilmiştir. Wang (2023) çalışmasında, RNN-CTC modelinin gerçek zamanlı konuşma tanıma için yeterince hızlı olduğu, ancak uzun konuşma dizilerinde bellek kullanımının artmasından dolayı performans düşüşleri yaşanabileceği belirtilmiştir.

Optimizasyon teknikleri açısından, Hamian vd. (2023) çalışmasında, Wild Horse Optimization algoritmasının, modelin ağırlıklarını optimize ederek daha az parametreyle daha yüksek başarım elde edilmesini sağladığı gösterilmiştir. Chen vd. (2023) çalışmasında, bilgi damıtma ve parametre paylaşımı teknikleri kullanılarak büyük dil modellerinin boyutu küçültülmüştür.

En etkili makine öğrenme yöntemi, uygulama gereksinimlerine ve kısıtlamalarına bağlı olarak değişmektedir. Karmaşıklık-başarım dengesi açısından, sınırlı kelime dağarcığına sahip görevlerde optimize edilmiş CNN modelleri etkili çözümler sunarken, daha karmaşık görevlerde RNN, LSTM ve Transformer tabanlı modeller daha etkili olmaktadır. Transfer öğrenimi ve çok dilli ön-eğitim yaklaşımları, özellikle düşük kaynaklı diller için hesaplama verimliliği sağlamaktadır.

4.3. Veri Gereksinimi Karşılaştırması

Makine öğrenme tabanlı ses tanıma sistemlerinin performansını etkileyen en önemli faktörlerden biri eğitim veri setinin niteliği ve niceliğidir. Geleneksel ve derin öğrenme yaklaşımları arasında veri gereksinimleri açısından belirgin farklılıklar bulunmaktadır. Geleneksel yaklaşımlar (SVM, KNN, Random Forest) genellikle daha az veri ile kabul edilebilir sonuçlar üretebilirken, derin öğrenme yaklaşımları daha fazla veriye ihtiyaç duymaktadır.

Ahmed vd. (2024), Peştuca dili için geliştirilen konuşma tanıma sisteminde, SVM ve KNN gibi klasik algoritmaların, küçük bir veri seti (161 izole kelime, 30 konuşmacı) ile %96.15 doğruluk oranına ulaştığını göstermiştir. Buna karşılık, Zhang vd. (2023) tarafından geliştirilen Evrensel Konuşma Modeli (USM), 12 milyon saatlik etiketlenmemiş ses verisi ve 90.000 saatlik etiketli çok dilli veri üzerinde ön-eğitime tabi tutulmuştur.

Transfer öğrenimi yaklaşımları, özellikle düşük kaynaklı diller için veri gereksinimini azaltmaktadır. Froiz-Míguez vd. (2023), Galiçya dili için wav2vec 2.0 modelini kullanarak, sadece 20 saatlik etiketli ses verisiyle %18.61 WER değerine ulaşmıştır. Çalışmada uygulanan yaklaşım, öz-denetimli eğitim kullanarak çok dilli etiketlenmemiş sesler üzerinde eğitim yapan ve ardından belirli bir dilin etiketli sesleriyle ince ayar yapabilen bir yöntem sunmaktadır.

Veri artırma teknikleri, mevcut veri setinden yeni örnekler oluşturarak modelin daha geniş bir veri yelpazesi üzerinde eğitilmesini sağlamaktadır. Barhoush vd. (2023), konuşmacı tanıma için ses çerçevelerini bloklar halinde bölerek ve her blok içindeki çerçeveleri rastgele karıştırarak veri artırma sağlamıştır. Shukla vd. (2023), çevrimdışı ve anında veri artırma teknikleri eklendiğinde, konuşmacıdan bağımsız telefon CTC sisteminin WER değerinin %9.1/%17.4'ten %7.9/%15.7'ye düştüğünü göstermiştir.

Farklı model mimarilerinin veri gereksinimleri de değişkenlik göstermektedir. Yao vd. (2024) tarafından önerilen Zipformer modeli, verimli mimari tasarımı sayesinde daha az veri ile LibriSpeech veri setinde test-clean için %1.9 WER ve test-other için %3.9 WER değerlerine ulaşmıştır. Gong vd. (2023) tarafından geliştirilen LTU-AS modeli, toplam 8.5 milyar parametreye sahip olmasına rağmen, sadece 49 milyon parametre eğitilebilir durumdadır.

Özel uygulamalar için geliştirilen ses tanıma sistemlerinde genellikle daha spesifik ve sınırlı veri setleri kullanılmaktadır. Rahate vd. (2023), felçli hastalar için EEG sinyalleri kullanarak sessiz konuşma tanıma sistemi geliştirmiş ve sınırlı veri ile KNN algoritması %61.45 doğruluk oranına ulaşmıştır.

Çok dilli ses tanıma sistemleri, diller arası transfer öğrenimi sayesinde düşük kaynaklı diller için veri gereksinimini azaltmaktadır. Chu vd. (2024) tarafından geliştirilen Qwen2-Audio modeli, doğal dil komutlarını kullanarak ön-eğitim sürecini basitleştirmiş ve herhangi bir göreve özgü ince ayar olmadan, LibriSpeech'te %92.62, Aishell2'de %96.0 ve FLEURS-ZH'de %91.75 doğruluk oranlarına ulaşmıştır.

İncelenen çalışmalar ışığında; geleneksel makine öğrenme yaklaşımlarının, derin öğrenme yaklaşımlarına göre genellikle daha az veriye ihtiyaç duyduğu görülmektedir. Bu nedenle, veri kısıtlı ortamlarda SVM, KNN gibi geleneksel yaklaşımlar tercih edilebilir.

4.4. Gürültüye Dayanıklılık Karşılaştırması

Ses tanıma sistemlerinin gürültüye dayanıklılığı, farklı ortam koşullarında performans sürdürülebilirliği açısından önem taşımaktadır. İncelenen çalışmalarda, derin öğrenme yaklaşımlarının geleneksel yöntemlere kıyasla gürültüye karşı daha dayanıklı olduğu görülmektedir. Barhoush vd. (2023), tam bağlantılı derin sinir ağı (FC-DNN) modellerinin düşük Sinyal-Gürültü Oranı (SNR) seviyelerinde bile %90'ın üzerinde doğruluk sağladığını göstermiştir. Benzer şekilde, Alzaabi vd. (2024) tarafından geliştirilen CNN tabanlı sistem, akan musluk ve bulaşık yıkama gibi çeşitli arka plan gürültüleriyle zenginleştirilmiş veri setinde %99.84 doğruluk elde etmiştir.

Gürültü azaltma teknikleri, tanıma performansını önemli ölçüde iyileştirmektedir. Gourisaria vd. (2024), gürültü temizleme işleminin sınıflandırma başarısını artırdığını, Rahate vd. (2023) ise EEG sinyallerindeki gürültüleri temizlemek için kullanılan Butterworth bant geçiren filtre ve çentik filtre kombinasyonunun etkili olduğunu

göstermiştir. Keser (2023) çalışmasında, gürültülü ortamlarda RNN-LSTM mimarisi %93.22 doğruluk oranıyla en iyi performansı gösterirken, CNN (%87.56) ve KNN (%86.51) daha düşük performans sergilemiştir.

Özellik çıkarma yöntemleri de gürültüye dayanıklılıkta önemli rol oynamaktadır. Barhoush vd. (2023) tarafından önerilen Karıştırılmış MFCC (SHMFCC) ve Fark Karıştırılmış MFCC (DSHMFCC) özellikleri, geleneksel MFCC'ye göre gürültülü ortamlarda daha iyi sonuçlar vermiştir.

Genel olarak, gürültüye dayanıklılık açısından en etkili makine öğrenme yaklaşımlarının derin öğrenme tabanlı modeller olduğu görülmektedir. RNN-LSTM mimarileri zamansal bilgiyi modelleyebilme yetenekleri sayesinde, CNN tabanlı modeller ise özellikle spektrogram tabanlı özelliklerin kullanıldığı durumlarda gürültüye karşı en dayanıklı yaklaşımlar olarak öne çıkmaktadır.

4.5. Uygulanabilirlik Karşılaştırması

Ses tanıma teknolojilerinin uygulanabilirliği, kullanılan makine öğrenme tekniklerinin özellikleri ve uygulama alanının gereksinimleriyle doğrudan ilişkilidir. Genel amaçlı konuşma tanıma sistemlerinde uçtan uca modeller ve derin öğrenme yaklaşımları öne çıkmaktadır. Shukla vd. (2023) tarafından yapılan çalışmada, CTC tabanlı uçtan uca modeller ve BLSTM mimarilerinin genel amaçlı sistemlerde yüksek performans gösterdiği, "soft forgetting" teknikleriyle WER değerlerinde %7-9 iyileşme sağlandığı belirtilmiştir. Zhang vd. (2023) tarafından geliştirilen Evrensel Konuşma Modeli (USM), 100'den fazla dilde konuşma tanıma yapabilmekte ve SpeechStew veri setinde %5.7 WER elde etmektedir.

Özel amaçlı sistemlerde daha hafif ve alana özgü eğitilmiş modeller tercih edilmektedir. Ayall vd. (2024) tarafından Amharca rakam tanıma için geliştirilen CNN mimarisi, MFCC özellikleriyle %99 doğruluk sağlamıştır. Jia (2023), sağlık sigortası alanına özgü bir sistemde önceden eğitilmiş Wav2Vec2-Large-LV60 modeli üzerinde ince ayar yaparak ticari ASR sistemlerinden daha iyi performans elde etmiştir.

Gömülü sistemlerde hesaplama ve bellek gereksinimleri düşük modeller kullanılmaktadır. Alzaabi vd. (2024) tarafından geliştirilen engelli bireyler için tasarlanmış düşük maliyetli sistem, CNN mimarisi kullanmakta ve ESP32 mikrodenetleyicisine yerleştirilebilmektedir. Ahmed vd. (2024), Peştuca dili için geliştirdikleri sistemde klasik makine öğrenme modellerini (SVM, KNN) tercih etmiş ve MFCC-SVM kombinasyonuyla %96.15 doğruluk elde etmiştir.

Çok dilli sistemlerde, farklı dillerin özelliklerini öğrenebilen modeller önem kazanmaktadır. Wang vd. (2023) tarafından geliştirilen VIOLA modeli, konuşma tanıma, sentezi ve çevirisi görevlerini tek bir model altında birleştirmekte ve İngilizce ASR görevinde %9.4 WER elde etmektedir. Chu vd. (2024) tarafından geliştirilen Qwen2-Audio modeli, LibriSpeech'te %92.62, Aishell2'de %96.0 doğruluk oranlarına ulaşmıştır.

Özel uygulama alanlarında da ses tanıma teknolojileri başarıyla uygulanmaktadır. Rahate vd. (2023) tarafından geliştirilen EEG tabanlı sessiz konuşma tanıma sistemi, felçli hastalar için alternatif bir iletişim yöntemi sunmaktadır. Kumar vd. (2023) tarafından geliştirilen görsel konuşma tanıma sistemi, dudak hareketlerini analiz ederek GRID Corpus üzerinde %2.8 WER elde etmektedir.

Genel olarak değerlendirildiğinde, ses tanıma alanında en etkili makine öğrenme yöntemleri uygulama alanına göre değişmektedir. Genel amaçlı sistemlerde derin öğrenme tabanlı yaklaşımlar, özel amaçlı sistemlerde transfer öğrenme ve ince ayar yaklaşımları, gömülü sistemlerde hafif modeller ve klasik algoritmalar, çok dilli sistemlerde büyük ön-eğitimli modeller, özel uygulama alanlarında ise hibrit yaklaşımlar daha uygulanabilir görünmektedir.

5. Sonuç ve Öneriler

Ses tanıma alanındaki literatür incelendiğinde, derin öğrenme tabanlı yaklaşımların geleneksel yöntemlere kıyasla üstün performans sergilediği görülmektedir. Geleneksel HMM ve GMM yöntemleri uzun süre standart yaklaşımlar olarak kabul edilse de, karmaşık ses örüntülerini modellemedeki sınırlamaları araştırmacıları daha gelişmiş tekniklere yönlendirmiştir. DNN, CNN ve RNN gibi derin öğrenme mimarileri, ses sinyallerindeki karmaşık örüntüleri öğrenmede üstün yetenekler sergilemiştir (Mehrisha vd., 2023).

Son yıllarda, öne çıkan Transformer tabanlı modeller ve uçtan uca öğrenme yaklaşımları, geleneksel HMM tabanlı sistemlerin karmaşık yapısını basitleştirerek daha tutarlı bir öğrenme süreci sağlamaktadır. Uçtan uca modeller, akustik özellik çıkarımı, akustik modelleme, dil modelleme ve arama gibi klasik ASR bileşenlerini tek bir yapıda birleştirerek sistem performansını artırmaktadır (Liu vd., 2023). Özellik çıkarma açısından MFCC hala yaygın kullanılmakla birlikte, spektrogram tabanlı gösterimler ve dalgacık dönüşümü gibi alternatif yaklaşımlar da önem kazanmıştır. Derin öğrenme modellerinin ham ses verilerinden doğrudan özellik çıkarabilme yeteneği, manuel özellik mühendisliğine olan bağımlılığı azaltmıştır (Zaman vd., 2023).

Kaydedilen ilerlemelere rağmen, ses tanıma alanında hala çeşitli problemler bulunmaktadır. Derin öğrenme modellerinin eğitimi için gerekli büyük miktarda etiketli veri ihtiyacı, özellikle az konuşulan diller için engel teşkil etmektedir (Kheddar vd., 2023). Gürültülü ortamlarda ses tanıma sistemlerinin performansı hala önemli bir sorun olarak öne çıkmaktadır. Farklı diller ve lehçeler için sistemlerin geliştirilmesi, her dilin kendine özgü fonetik yapısı nedeniyle zorlayıcı olmaya devam etmektedir (Watve vd., 2023). Mevcut modellerin genelleme yeteneği de geliştirilmeye açık bir alandır. Domain adaptasyonu ve transfer öğrenme gibi teknikler bu sorunu hafifletmeye yardımcı olsa da, henüz tam bir çözüm sunamamaktadır. Ses tanıma sistemlerinin kararlarını açıklayabilme yeteneğinin sınırlı olması, özellikle sağlık ve güvenlik uygulamalarında güven sorunları yaratabilmektedir.

Literatür incelemesi sonucunda, ses tanıma alanında Transformer mimarileri ve uçtan uca öğrenme modellerinin en etkili yöntemler olduğu değerlendirilmektedir. Bununla birlikte, ideal bir ses tanıma sistemi için tek bir "en iyi" yaklaşım yerine, uygulama senaryosuna ve kaynaklara bağlı olarak farklı yaklaşımların kombinasyonu daha uygun olabilir. Gelecekteki araştırmalar için, düşük kaynaklı diller için yarı-denetimli öğrenme yaklaşımları, CNN ve Transformer modellerinin avantajlarını birleştiren hibrit mimariler, çok görevli öğrenme yaklaşımları ve nöromorfolojik hesaplama modelleri üzerine çalışmalar yürütülebilir. Ayrıca, farklı diller ve uygulama alanları için standartlaştırılmış değerlendirme metrikleri geliştirilebilir.

Kaynaklar

Accou, B., Vanthornhout, J., Van hamme, H., & Francart, T. (2023). Decoding of the speech envelope from EEG using the VLAAI deep neural network. *Scientific Reports*, 13(812), 1-14. https://doi.org/10.1038/s41598-022-27332-2

Ahmed, I., Irfan, M. A., Iqbal, A., Khalil, A., & Siddiqui, S. I. (2024). Efficient feature extraction and classification for the development of Pashto speech recognition system. *Multimedia Tools and Applications*, 83, 54081-54096. https://doi.org/10.1007/s11042-023-17684-w

Alzaabi, S., Alzahmi, A., Almheiri, M., Al-Ali, N., Ali, N., Poon, K., & Alteneiji, A. (2024). Bilingual Speech Recognition On the Edge Using Machine Learning. *2024 7th International Conference on Signal Processing and Information Security (ICSPIS)*, 1-6. https://doi.org/10.1109/ICSPIS63676.2024.10812600

Ameen, Z. J. M., & Kadhim, A. A. (2023). Machine learning for Arabic phonemes recognition using electrolarynx speech. *International Journal of Electrical and Computer Engineering (IJECE)*, 13(1), 400-412. https://doi.org/10.11591/ijece.v13i1.pp400-412

Ayall, T. A., Zhou, C., Liu, H., Brhanemeskel, G. M., Abate, S. T., & Adjeisah, M. (2024). Amharic spoken digits recognition using convolutional neural network. *Journal of Big Data, 11*(64). https://doi.org/10.1186/s40537-024-00910-z

Barhoush, M., Hallawa, A., & Schmeink, A. (2023). Speaker identification and localization using shuffled MFCC features and deep learning. *International Journal of Speech Technology*, 26, 185-196. https://doi.org/10.1007/s10772-023-10023-2

Chen, J. (2023). Speech recognition and English corpus vocabulary learning based on endpoint detection algorithm. *International Journal of System Assurance Engineering and Management*. https://doi.org/10.1007/s13198-023-01995-0

Chen, Z., Huang, H., Andrusenko, A., Hrinchuk, O., Puvvada, K. C., Li, J., Ghosh, S., Balam, J., & Ginsburg, B. (2023). SALM: Speech-augmented language model with in-context learning for speech recognition and translation. *arXiv preprint arXiv:2310.09424v1*. https://arxiv.org/abs/2310.09424v1

Chu, H., Jia, W., Liu, Y., Zan, Y., Bai, Y., Xiao, S., Zheng, Y., Xie, Z., Xie, S., Zhang, S., Zhou, M., & Huang, S. (2023). Qwen2-Audio: Advancing universal audio understanding via unified large language models. *arXiv preprint arXiv:2311.07919*.

Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., Zhou, C., Zhou, J., & Qwen Team. (2024). Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759v1*. https://arxiv.org/abs/2407.10759v1

Dhanjal, A. S., & Singh, W. (2023). A comprehensive survey on automatic speech recognition using neural networks. *Multimedia Tools and Applications*, 83, 23367-23412. https://doi.org/10.1007/s11042-023-16438-y

Froiz-Míguez, I., Blanco-Novoa, Ó., Fraga-Lamas, P., Fustes, D., Dafonte, C., Pereira, J., & Fernández-Caramés, T. M. (2023). Design and evaluation of a cross-lingual ML-based automatic speech recognition system fine-tuned for the Galician language. *Kalpa Publications in Computing*, 14, 152-155.

Ganchev, R. (2021). Voice Signal Processing for Machine Learning. The Case of Speaker Isolation: Overview and Evaluation of Decomposition Methods Applied to the Input Signal of Voice Processing ML Models - The Use Case of the Speaker Isolation Problem. Sofia University "St. Kliment Ohridski", Faculty of Mathematics and Informatics.

Gençyılmaz, İ. Z., & Karaoğlan, K. M. (2024). Optimizing Speech to Text Conversion in Turkish: An Analysis of Machine Learning Approaches. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, 13(2), 492-504. https://doi.org/10.17798/bitlisfen.1434925

Gong, Y., Liu, A. H., Luo, H., Karlinsky, L., & Glass, J. (2023). Joint audio and speech understanding. *arXiv preprint arXiv:2309.14405v3*. https://arxiv.org/abs/2309.14405v3

Gourisaria, M. K., Agrawal, R., Sahni, M., & Singh, P. K. (2024). Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. *Discover Internet of Things*, 4(1), 1-24. https://doi.org/10.1007/s43926-023-00049-y

Guo, J. (2023). Innovative Application of Sensor Combined with Speech Recognition Technology in College English Education in the Context of Artificial Intelligence. *Journal of Sensors*, 2023, Article ID 9281914, 1-11. https://doi.org/10.1155/2023/9281914

Hamian, M., Faez, K., Nazari, S., & Sabeti, M. (2023). A novel learning approach in deep spiking neural networks with multi-objective optimization algorithms for automatic digit speech recognition. *The Journal of Supercomputing*, 79, 20263–20288. https://doi.org/10.1007/s11227-023-05420-y

Jia, Y. (2023). A Deep Learning System for Domain-Specific Speech Recognition. *arXiv preprint arXiv:2303.10510v2*. https://arxiv.org/abs/2303.10510v2

Keser, S. (2023). Makine Öğrenimi ve Hibrit Altuzay Sınıflandırıcılar için Yalıtık Kelime Tanıma Performanslarının Karşılaştırılması. *Sürdürülebilir Mühendislik Uygulamaları ve Teknolojik Gelişmeler Dergisi*, 6(2), 235-249. https://doi.org/10.51764/smutgd.1338977

Kheddar, H., Himeur, Y., Al-Maadeed, S., Amira, A., & Bensaali, F. (2023). Deep transfer learning for automatic speech recognition: Towards better generalization. *arXiv preprint arXiv:2304.14535v2*. https://arxiv.org/abs/2304.14535v2

Kumar, L. A., Renuka, D. K., & Priya, M. C. S. (2023). Towards robust speech recognition model using deep learning. In *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)* (pp. 253-256). IEEE. https://doi.org/10.1109/ICISCOIS56541.2023.10100390

Kwon, H. (2024). AudioGuard: Speech Recognition System Robust against Optimized Audio Adversarial Examples. *Multimedia Tools and Applications*, 83, 57943-57962. https://doi.org/10.1007/s11042-023-15961-2

Liu, A. H., Hsu, W. N., Auli, M., & Baevski, A. (2023). Towards end-to-end unsupervised speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)* (pp. 221-228). IEEE.

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6), 9411-9457. https://doi.org/10.1007/s11042-020-10073-7

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). A review of deep learning techniques for speech processing. *Multimedia Tools and Applications*, 1-72. https://doi.org/10.1016/j.inffus.2023.101755

Mishra, A., Verma, R., Dhanda, N., & Gupta, K. K. (2024). Speech Recognition Using Machine Learning Techniques. *2024 2nd International Conference on Disruptive Technologies (ICDT)*, 1142-1146. https://doi.org/10.1109/ICDT61202.2024.10489508

Prabhavalkar, R., Hori, T., Sainath, T. N., Schlüter, R., & Watanabe, S. (2023). End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 325-353. https://doi.org/10.1109/TASLP.2023.3328283

Rahate, J., Tadepalli, S. N. V. R., Saroj, U., Kamble, A., & Ghare, P. (2023). Silent Speech Recognition using EEG Signals. *2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS)*, 1-5. https://doi.org/10.1109/PCEMS58491.2023.10136068

Rai, S., Li, T., & Lyu, B. (2023). Keyword spotting - Detecting commands in speech using deep learning. *arXiv preprint arXiv:2312.05640*.

Rai, S., Li, T., & Lyu, B. (2023). Keyword spotting - Detecting commands in speech using deep learning. *arXiv preprint arXiv:2312.05640*.

Reid, K., & Williams, E. T. (2023). Right the docs: Characterising voice dataset documentation practices used in machine learning. *arXiv preprint arXiv:2303.10721v1*. https://arxiv.org/abs/2303.10721v1

Shisode, S., Mhatre, B., Sikligar, J., Vishwakarma, S., Tupe, S., Jagtap, S., & Nemade, M. (2023). SEMG approach for speech recognition. *International Journal of Advanced Research in Computer Science*, 14(3), 23-28. http://dx.doi.org/10.26483/ijarcs.v14i3.6970

Shukla, A., Aanand, A., & Nithiya, S. (2023). Automatic Speech Recognition using Machine Learning Techniques. *2023 International Conference on Computer Communication and Informatics (ICCCI)*, 1-6. https://doi.org/10.1109/ICCCI56745.2023.10128212

Tandel, N. H., Prajapati, H. B., & Dabhi, V. K. (2020). Voice Recognition and Voice Comparison using Machine Learning Techniques: A Survey. *2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS)*, 459-465. IEEE.

Wang, S. (2023). Recognition of English speech – using a deep learning algorithm. *Journal of Intelligent Systems*, 32, 20220236. https://doi.org/10.1515/jisys-2022-0236

Wang, T., Zhou, L., Zhang, Z., Wu, Y., Liu, S., Gaur, Y., Chen, Z., Li, J., & Wei, F. (2023). VIOLA: Unified codec language models for speech recognition, synthesis, and translation. *arXiv preprint arXiv:2305.16107v1*. https://arxiv.org/abs/2305.16107v1

Weng, Z., Qin, Z., Tao, X., Pan, C., Liu, G., & Li, G. Y. (2023). Deep learning enabled semantic communications with speech recognition and synthesis. *IEEE Transactions on Wireless Communications, 22*(9), 6227-6240. https://doi.org/10.1109/TWC.2023.3240969

Wojnar, T., Hryszko, J., & Roman, A. (2024). Mi-Go: tool which uses YouTube as data source for evaluating general-purpose speech recognition machine learning models. *EURASIP Journal on Audio, Speech, and Music Processing, 2024*(24). https://doi.org/10.1186/s13636-024-00343-9

Xie, Y. (2023). Application of speech recognition technology based on machine learning for network oral English teaching system. *International Journal of System Assurance Engineering and Management*. https://doi.org/10.1007/s13198-023-02143-4

Yao, Z., Guo, L., Yang, X., Kang, W., Kuang, F., Yang, Y., Jin, Z., Lin, L., & Povey, D. (2024). Zipformer: A faster and better encoder for automatic speech recognition. Published as a conference paper at ICLR 2024. https://arxiv.org/abs/2310.11230v4

Zaman, K., Sah, M., Direkoglu, C., & Unoki, M. (2023). A survey of audio classification using deep learning. *IEEE Access*, 11, 106620-106621. https://doi.org/10.1109/ACCESS.2023.3318015

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., Li, B., Axelrod, V., Wang, G., Meng, Z., Hu, K., Rosenberg, A., Prabhavalkar, R., Park, D. S., Haghani, P., Riesa, J., Perng, G., Soltau, H., Strohman, T., Ramabhadran, B., Sainath, T., Moreno, P., Chiu, C.-C., Schalkwyk, J., Beaufays, F., & Wu, Y. (2023). Google USM: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037v3*. https://arxiv.org/abs/2303.01037v3