T.C.

ALTINBAŞ ÜNİVERSİTESİ

Lisansüstü Eğitim Enstitüsü

Elektrik ve Bilgisayar Mühendisliği Ana Bilim Dalı

MAKİNE ÖĞRENME TEKNİKLERİ İLE SES TANIMA: LİTERATÜR TARAMASI

Mutlu Merih AKTUZ

Yüksek Lisans Tezi

Danışman

Doç. Dr. Sefer KURNAZ

İstanbul, 2025

MAKİNE ÖĞRENME TEKNİKLERİ İLE SES TANIMA: LİTERATÜR TARAMASI

Mutlu Merih AKTUZ

Elektrik ve Bilgisayar Mühendisliği Ana Bilim Dalı

Yüksek Lisans Tezi

ALTINBAŞ ÜNİVERSİTESİ

2025

"Mutlu Merih AKTUZ" tarafından hazırlanmış ve 19/09/2025 tarihinde sunulmuş "MAKİNE ÖĞRENME TEKNİKLERİ İLE SES TANIMA: LİTERATÜR TARAMASI" başlıklı tez Elektrik ve Bilgisayar Mühendisliği Ana Bilim Dalında Yüksek Lisans Tezi olarak **oy birliği** ile kabul edilmiştir.

_____

Doç. Dr. Sefer KURNAZ

Danışman

Tez Savunma Sınavı Jüri Üyeleri:

Doç. Dr. Sefer KURNAZ	Bilgisayar Mühendisliği Bölümü, Altınbaş Üniversitesi	_____
Dr. Öğr. Üyesi Serdar KARGIN	Biyomedikal Mühendisliği Bölümü, İstanbul Arel Üniversitesi	_____
Dr. Öğr. Üyesi Hameed Mutlag Farhan FARHAN	Bilgisayar Mühendisliği Bölümü, Altınbaş Üniversitesi	_____

Bu tezin Yüksek Lisans tezi olarak bütün şartları sağladığını beyan ederim.

Bu tezdeki tüm bilgilerin akademik kurallara ve etik davranışlara uygun olarak edinildiğini ve sunulduğunu beyan ederim. Ayrıca, bu kuralların ve davranışların gerektirdiği şekilde, bu çalışmada, orijinal olmayan tüm materyalleri ve sonuçları tamamen alıntı yaptığımı ve referans gösterdiğimi de beyan ederim.

Mutlu Merih AKTUZ

İmzası

İTHAF

Sevgili kızım Defne’ye.

ÖZET

MAKİNE ÖĞRENME TEKNİKLERİ İLE SES TANIMA: LİTERATÜR TARAMASI

AKTUZ, Mutlu Merih

Yüksek Lisans, Elektrik ve Bilgisayar Mühendisliği, Altınbaş Üniversitesi

Danışman: Doç. Dr. Sefer KURNAZ

Tarih: 09/2025

Sayfa: 103

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.

ABSTRACT

VOICE RECOGNITION USING MACHINE LEARNING TECHNIQUES: A LITERATURE REVIEW

AKTUZ, Mutlu Merih

M.Sc., Electrical and Computer Engineering, Altınbaş University,

Supervisor: Doç. Dr. Sefer KURNAZ

Date: 09/2025

Pages: 103

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.

İÇİNDEKİLER

Sayfa

ŞEKİL LİSTESİ

Sayfa

KISALTMALAR

AED	:	Attention-based Encoder-Decoder (Dikkat Tabanlı Kodlayıcı-Kod Çözücü)
ANN	:	Artificial Neural Networks (Yapay Sinir Ağları)
ASR	:	Automatic Speech Recognition (Otomatik Konuşma Tanıma)
BiLSTM	:	Bidirectional Long Short-Term Memory (Çift Yönlü Uzun-Kısa Vadeli Bellek)
CER	:	Character Error Rate (Karakter Hata Oranı)
CNN	:	Convolutional Neural Networks (Evrişimli Sinir Ağları)
CTC	:	Connectionist Temporal Classification (Bağlantıcı Zamansal Sınıflandırma)
CWT	:	Continuous Wavelet Transform (Sürekli Dalgacık Dönüşümü)
DCVA	:	Discriminative Common Vector Approach (Ayrımcı Ortak Vektör Yaklaşımı)
DCT	:	Discrete Cosine Transform (Ayrık Kosinüs Dönüşümü)
DNN	:	Deep Neural Networks (Derin Sinir Ağları)
DRL	:	Deep Reinforcement Learning (Derin Pekiştirmeli Öğrenme)
DTL	:	Deep Transfer Learning (Derin Transfer Öğrenme)
DWT	:	Discrete Wavelet Transform (Ayrık Dalgacık Dönüşümü)
FFT	:	Fast Fourier Transform (Hızlı Fourier Dönüşümü)
FLDA	:	Fisher Linear Discriminant Analysis (Fisher Doğrusal Diskriminant Analizi)
GMM	:	Gaussian Mixture Models (Gaussian Karışım Modelleri)
HMM	:	Hidden Markov Models (Gizli Markov Modelleri)

KNN	:	K-Nearest Neighbors (K-En Yakın Komşu)
LPC	:	Linear Predictive Coding (Doğrusal Tahmin Kodlaması)
LSTM	:	Long Short-Term Memory (Uzun-Kısa Vadeli Bellek)
MFCC	:	Mel-Frequency Cepstral Coefficients (Mel Frekansı Kepstral Katsayıları)
MRA	:	Multi-Resolution Analysis (Çok Çözünürlüklü Analiz)
NLU	:	Natural Language Understanding (Doğal Dil Anlama)
PCA	:	Principal Component Analysis (Temel Bileşen Analizi)
PESQ	:	Perceptual Evaluation of Speech Quality (Konuşma Kalitesinin Algısal Değerlendirmesi)
PLP	:	Perceptual Linear Prediction (Algısal Doğrusal Tahmin)
RNN	:	Recurrent Neural Networks (Tekrarlayan Sinir Ağları)
RNN-T	:	RNN-Transducer (RNN-Dönüştürücü)
RTF	:	Real-Time Factor (Gerçek Zamanlı Faktör)
SDR	:	Spoken Digit Recognition (Konuşulan Rakam Tanıma)
SER	:	Sentence Error Rate (Cümle Hata Oranı)
SHMFCC	:	Shuffled MFCC (Karıştırılmış MFCC)
SI-SDR	:	Scale-Invariant Signal-to-Distortion Ratio (Ölçek-Değişmez Sinyal-Bozulma Oranı)
SLU	:	Spoken Language Understanding (Konuşma Dili Anlama)
STFT	:	Short-Time Fourier Transform (Kısa Süreli Fourier Dönüşümü)
STOI	:	Short-Time Objective Intelligibility (Kısa Süreli Nesnel Anlaşılabilirlik)
SVM	:	Support Vector Machines (Destek Vektör Makineleri)
USM	:	Universal Speech Model (Evrensel Konuşma Modeli)

WER	:	Word Error Rate (Kelime Hata Oranı)
ZCR	:	Zero Crossing Rate (Sıfır Geçiş Oranı)

SEMBOL LİSTESİ

D	:	Silinen Kelimelerin Sayısı
dB	:	Desibel
e_{noise}	:	Gürültü Bileşeni
FFT	:	Hızlı Fourier Dönüşümü
Hz	:	Hertz
I	:	Eklenen Kelimelerin Sayısı
K	:	K-En Yakın Komşu Algoritmasındaki Komşu Sayısı
kHz	:	Kilohertz
$\log_{10}$	:	10 Tabanında Logaritma
ms	:	Milisaniye
N	:	Referans Transkripsiyondaki Toplam Kelime Sayısı
S	:	İkame Edilen Kelimelerin Sayısı
s_{target}	:	Hedef Sinyal Bileşeni
WER	:	Kelime Hata Oranı
%	:	Yüzde

1. GİRİŞ

Ses tanıma teknolojisi, konuşma sinyallerini analiz ederek dilsel içeriği çıkarmakta ve yazılı metne dönüştürmektedir [1]. Son yıllarda makine öğrenme algoritmalarının gelişmesiyle, ses tanıma sistemlerinin doğruluğu ve kullanılabilirliği önemli ölçüde iyileştirilmiştir [2].

Makine öğrenme teknikleri, ses tanıma sistemlerinin geliştirilmesinde önemli bir rol oynamaktadır. Ses tanıma sistemlerinde kullanılan makine öğrenme teknikleri; geleneksel yaklaşımlar, derin öğrenme yaklaşımları ve hibrit yaklaşımlar olarak sınıflandırılabilir. Geleneksel yaklaşımlar arasında Gaussian Karışım Modelleri, Destek Vektör Makineleri ve Gizli Markov Modelleri yer alırken [3], derin öğrenme yaklaşımları Derin Sinir Ağları, Evrişimli Sinir Ağları ve Transformatör mimarilerini içermektedir [3], [4]. Hibrit yaklaşımlar ise geleneksel ve derin öğrenme tekniklerinin güçlü yönlerini birleştirerek daha yüksek performans elde etmeyi amaçlamaktadır [5].

Ses tanıma sistemlerinin geliştirilmesinde karşılaşılan temel problemler arasında konuşma varyasyonu, gürültülü ortamlar ve farklı aksanlar bulunmaktadır [6]. Bu problemlerin üstesinden gelmek için farklı makine öğrenme teknikleri geliştirilmiş ve uygulanmıştır. Ancak, bu tekniklerin karşılaştırmalı analizi ve hangi koşullarda hangi tekniğin daha etkili olduğuna dair kapsamlı bir değerlendirme literatürde yeterince ele alınmamıştır.

Bu bağlamda; çalışmamın amacı, makine öğrenme teknikleri kullanılarak ses tanıma alanındaki mevcut literatürü incelemek ve bu alandaki farklı yöntemleri karşılaştırarak en etkili makine öğrenme yöntemini belirlemektir.

Çalışmanın yöntemi, literatür taraması ve içerik analizi olarak tasarlanmıştır. İlk aşamada, IEEE Xplore, ACM Digital Library, ScienceDirect, Springer Link, Google Scholar ve Web of Science gibi bilimsel veri tabanlarında tarama işlemi gerçekleştirilmiştir. Tarama sürecinde, "ses tanıma" (speech recognition), "derin öğrenme" (deep learning), "makine öğrenme" (machine learning), "otomatik konuşma tanıma" (automatic speech recognition), "yapay sinir ağları" (artificial neural networks), "konuşma işleme" (speech processing) ve "akustik modelleme" (acoustic modeling) gibi anahtar kelimeler ve bunların çeşitli kombinasyonları kullanılmıştır. Araştırma kapsamını güncel tutmak amacıyla, Ocak 2023-Mart 2025 tarihleri arasında yayınlanan çalışmalara öncelik verilmiştir. İkinci aşamada, veri

tabanı taraması sonucunda toplam 51 potansiyel çalışma tespit edilmiştir. Bu çalışmalar, başlık ve özet incelemesine tabi tutulmuş ve araştırma kapsamına uygunlukları değerlendirilmiştir. Değerlendirme sürecinde, çalışmaların ses tanıma alanında makine öğrenme tekniklerini kullanması, deneysel sonuçlar içermesi ve yöntemsel detaylar sunması gibi kriterler göz önünde bulundurulmuştur. Bu inceleme sonucunda, 21 çalışma içerik bakımından araştırma kapsamına uygun olmadığı gerekçesiyle çalışma dışında bırakılmıştır. Bu çalışmalar, genellikle yeterli teknik detay içermeyen, sadece literatür taraması veya teorik çerçeve sunan yayınlardan oluşmaktadır. Üçüncü aşamada, kalan 30 çalışmanın tam metinleri detaylı bir şekilde incelenmiştir. İncelenen çalışmaların detaylı bir tablosu hazırlanarak EK-A'da sunulmuştur. Tablo; çalışmaların bibliyografik bilgilerini, kullanılan yaklaşım türlerini, özellik çıkarma yöntemlerini, uygulanan algoritma ve modelleri, kullanılan veri setlerini, değerlendirme metriklerini, elde edilen sonuçları ve uygulama alanlarını içermektedir. İçerik analizi ve tablo, farklı makine öğrenme tekniklerinin performans, hesaplama karmaşıklığı, veri gereksinimi, gürültüye dayanıklılık ve uygulanabilirlik açısından karşılaştırılmasına imkan sağlamıştır.

Çalışma beş ana bölümden oluşmaktadır. Giriş bölümünü takip eden ikinci bölümde; ses tanımanın temelleri, makine öğrenme yaklaşımları, özellik çıkarma yöntemleri, veri setleri ve başarı kriterleri konuları incelenmiştir. Üçüncü bölüm olan "Literatür Taraması" kısmında, ses tanıma alanında makine öğrenme tekniklerini kullanan güncel çalışmalar incelenmiştir. Bu bölümde, farklı araştırmacıların çalışmaları, kullandıkları yaklaşımlar, özellik çıkarma yöntemleri, veri setleri ve elde ettikleri sonuçlar sunulmuştur. Dördüncü bölümde, literatür taraması sonucunda elde edilen bulgular çerçevesinde; farklı makine öğrenme tekniklerinin performansları, hesaplama karmaşıklıkları, veri gereksinimleri, gürültüye dayanıklılıkları ve uygulanabilirlikleri açısından karşılaştırılmıştır. Beşinci ve son bölüm olan "Sonuç ve Öneriler" kısmında ise, çalışmanın genel bir özeti sunulmuş, elde edilen bulgular ışığında ses tanıma alanında en etkili makine öğrenme teknikleri belirlenmiş ve gelecekteki araştırmalar için öneriler sunulmuştur.

2. KAVRAMSAL ÇERÇEVE

2.1 SES TANIMANIN TEMELLERİ

Ses tanıma, insan konuşmasını dijital ortamda işleyerek metne dönüştürme sürecidir ve makine-insan etkileşiminin temel bir bileşeni olarak kabul edilmektedir. Ses tanıma teknolojisi, bilgisayarların insan dilini anlamasını ve yorumlamasını sağlayarak, kullanıcıların ses komutları aracılığıyla makinelerle etkileşim kurmasına imkan sağlamaktadır [1]. Ses tanıma sistemleri, konuşma sinyallerini analiz ederek, bu sinyallerdeki dilsel içeriği çıkarmakta ve bu içeriği yazılı metne dönüştürmektedir. Süreç, karmaşık algoritmalar ve hesaplama modelleri gerektiren, çok aşamalı bir işlemdir [7].

Ses tanıma sistemlerinin temelinde, konuşma sinyallerinin dijital olarak temsil edilmesi yatmaktadır. Konuşma, akustik dalga formunda üretilmekte ve bu dalgalar mikrofon aracılığıyla elektrik sinyallerine dönüştürülmektedir. Sinyaller daha sonra analogdan dijitale dönüştürücüler tarafından sayısallaştırılmakta ve bilgisayar tarafından işlenebilir hale getirilmektedir. Sayısallaştırma sürecinde, sürekli ses dalgası, belirli zaman aralıklarında örneklenerek ayrık değerler dizisine dönüştürülmektedir. Örnekleme hızı, genellikle saniyede 16,000 örnek veya daha yüksek bir değer olarak belirlenmekte, böylece insan konuşmasındaki tüm frekans bileşenleri yakalanabilmektedir [8].

Ses tanıma sistemlerinin işlevsel mimarisi, genellikle ön işleme, özellik çıkarma, akustik modelleme, dil modelleme ve kod çözme aşamalarından oluşmaktadır. Ön işleme aşamasında, gürültü azaltma, yankı giderme ve sinyal normalizasyonu gibi teknikler uygulanarak, ses sinyalinin kalitesi artırılmaktadır. Ön işleme aşaması, sonraki aşamaların doğruluğunu önemli ölçüde etkilemektedir, çünkü temiz bir sinyal, daha doğru özellik çıkarımı ve dolayısıyla daha iyi tanıma sonuçları sağlamaktadır [9].

Ses tanıma sistemlerinde, konuşma sinyallerinin işlenmesinde karşılaşılan temel zorluklardan biri, konuşma varyasyonudur. Konuşma varyasyonu, aynı kelimenin farklı kişiler tarafından farklı şekillerde telaffuz edilmesi, aynı kişinin farklı zamanlarda veya durumlarda aynı kelimeyi farklı şekillerde telaffuz etmesi gibi durumları içermektedir. Konuşma varyasyonları; konuşmacının aksanı, konuşma hızı, duygusal durumu ve çevresel faktörlerden kaynaklanabilmektedir. Ses tanıma sistemlerinin bu varyasyonları ele

alabilmesi için, geniş bir konuşma örneği yelpazesi üzerinde eğitilmesi ve bu varyasyonları modelleyebilecek esnek algoritmalara sahip olması gerekmektedir [6].

Ses tanıma sistemlerinin bir diğer önemli bileşeni, konuşma sinyallerinden anlamlı özelliklerin çıkarılmasıdır. Bu özellikler, konuşma sinyalinin spektral, zamansal ve prosodik özelliklerini yansıtmakta ve akustik modellerin eğitilmesinde kullanılmaktadır. En yaygın kullanılan özellik çıkarma yöntemlerinden biri, Mel Frekansı Kepstral Katsayıları (MFCC) olarak bilinmektedir. MFCC, insan işitme sisteminin frekans algısını taklit eden bir ölçek olan Mel ölçeğini kullanarak, ses sinyalinin spektral özelliklerini çıkarmaktadır. Bu özellikler, konuşma sinyalinin kısa süreli spektral özelliklerini yakalamakta ve konuşma tanıma sistemlerinde yaygın olarak kullanılmaktadır [10].

Akustik modelleme, ses tanıma sistemlerinin merkezinde yer almaktadır ve konuşma sinyallerindeki akustik özelliklerin, dildeki temel birimlere (fonemler, kelimeler vb.) nasıl karşılık geldiğini modellemektedir. Geleneksel ses tanıma sistemlerinde, Gizli Markov Modelleri (HMM) ve Gauss Karışım Modelleri (GMM) gibi istatistiksel modeller kullanılmaktadır. Bu modeller, konuşma sinyallerindeki zamansal değişimleri ve akustik özelliklerin dağılımını modellemek için etkili araçlar olarak kabul edilmektedir. Ancak, son yıllarda derin öğrenme tabanlı modeller, özellikle Tekrarlayan Sinir Ağları (RNN) ve Evrişimli Sinir Ağları (CNN), akustik modelleme alanında geleneksel yöntemlere göre üstün performans göstermeye başlamıştır [11].

Dil modelleme, ses tanıma sistemlerinin bir diğer önemli bileşenidir ve dilin yapısal ve istatistiksel özelliklerini modelleyerek, akustik modelin çıktılarını daha doğru metinlere dönüştürmeye yardımcı olmaktadır. Dil modelleri, kelimelerin veya kelime dizilerinin olasılık dağılımlarını tahmin etmekte ve bu sayede tanıma sürecinde belirsizlikleri azaltmaktadır. N-gram modelleri, geleneksel dil modellemesinde yaygın olarak kullanılmakta, ancak son yıllarda derin öğrenme tabanlı dil modelleri, özellikle Transformer mimarisine dayalı modeller, daha iyi performans göstermektedir [12].

Ses tanıma sistemlerinin son aşaması, kod çözme veya arama olarak adlandırılmaktadır ve akustik model ile dil modelinin çıktılarını birleştirerek, en olası metin dizisini bulmayı amaçlamaktadır. Bu aşamada, Viterbi algoritması gibi dinamik programlama algoritmaları veya ışın arama gibi sezgisel arama algoritmaları kullanılmaktadır [13].

2.2 MAKİNE ÖĞRENME YAKLAŞIMLARI

Geleneksel makine öğrenme yaklaşımları, özellikle sınırlı hesaplama kaynakları ve veri setleriyle çalışırken etkili sonuçlar üretebilen, matematiksel temelleri güçlü modellerdir. Derin öğrenme yaklaşımları ise, ham ses verilerinden otomatik olarak özellik çıkarma yeteneği ile öne çıkmakta ve özellikle büyük veri setleri üzerinde üstün performans göstermektedir. Hibrit yaklaşımlar, her iki yaklaşımın avantajlarını birleştirerek, ses tanıma sistemlerinin performansını daha da artırmayı hedeflemektedir. Bu başlıkta, ses tanıma sistemlerinde kullanılan geleneksel, derin öğrenme ve hibrit makine öğrenme yaklaşımları incelenmektedir.

2.2.1 Geleneksel Yaklaşımlar

Ses tanıma sistemlerinin gelişiminde geleneksel makine öğrenme yaklaşımları, derin öğrenme tekniklerinin yaygınlaşmasından önce temel yapı taşlarını oluşturmuştur. Geleneksel makine öğrenme yaklaşımları, özellikle sınırlı hesaplama gücü ve veri kaynakları ile çalışıldığında etkili sonuçlar üretebilen, matematiksel temelleri güçlü modellerdir. Ses tanıma alanında kullanılan geleneksel makine öğrenme yaklaşımları arasında Gaussian Karışım Modelleri (GMM), Destek Vektör Makineleri (SVM), Gizli Markov Modelleri (HMM), K-En Yakın Komşu (KNN) algoritmaları ve Karar Ağaçları gibi yöntemler öne çıkmaktadır [3].

Gaussian Karışım Modelleri (GMM), bir ses özellik vektörünün olasılık dağılımını temsil etmek için farklı ağırlıklara sahip çoklu Gauss dağılımlarını birleştiren güçlü üretici modellerdir. GMM'ler, konuşmacı tanıma ve konuşma tanıma görevlerinde yaygın olarak uygulanmıştır. Konuşmacı tanımada, GMM'ler konuşmacıya özgü özelliklerin dağılımını yakalayarak, bireylerin benzersiz özelliklerine dayalı olarak tanınmasını sağlamaktadır. Konuşma tanımada ise, GMM'ler konuşma seslerinin akustik özelliklerini modelleyerek, söylenen kelimelerin ve ifadelerin doğru bir şekilde tanınmasını kolaylaştırmaktadır [14].

Destek Vektör Makineleri (SVM), çeşitli konuşma sınıflandırma görevleri için yaygın olarak benimsenen denetimli öğrenme algoritmalarıdır. Özellikle konuşmacı tanıma ve fonem tanıma gibi alanlarda etkili sonuçlar vermektedir. Özellik uzayında farklı sınıfları etkili bir şekilde ayıran optimal hiperdüzlemleri belirleme yetenekleriyle öne çıkan SVM'ler, bu optimal ayrımı kullanarak, konuşma modellerinin doğru sınıflandırılmasını ve tanınmasını

sağlamaktadır. SVM'lerin ses tanıma alanındaki başarısı, karmaşık ses verilerini yüksek boyutlu uzaylarda etkili bir şekilde ayırabilme kabiliyetinden kaynaklanmaktadır [3].

Gizli Markov Modelleri (HMM), özellikle Otomatik Konuşma Tanıma (ASR) olmak üzere çeşitli konuşma tanıma görevlerini gerçekleştirmek için güçlü bir araç olarak önemli derecede popülerlik kazanmıştır. ASR'de, HMM'ler gizli durumların sıralı düzenlemesi ve karşılık gelen gözlemlerle konuşma seslerinin olasılık dağılımını modellemek için kullanılmaktadır. HMM'lerin eğitimi genellikle Beklenti Maksimizasyonu algoritmasının bir varyantı olan Baum-Welch algoritması kullanılarak gerçekleştirilmekte ve etkili parametre tahmini ile model optimizasyonunu sağlamaktadır. HMM'ler, konuşma tanıma sistemlerinin temel bileşeni olarak uzun yıllar boyunca kullanılmış ve bu alanda önemli ilerlemeler sağlamıştır [14].

K-En Yakın Komşu (KNN) algoritması, konuşmacı tanıma ve dil tanıma dahil olmak üzere çeşitli konuşmayla ilgili uygulamalarda kullanılan basit ancak etkili bir sınıflandırma yaklaşımıdır. KNN'nin temel prensibi, eğitim verileri içindeki belirli bir girdi özellik vektörünün K-en yakın komşularını belirlemek ve onu bu komşular arasında en sık görünen sınıfa atamaktır [14]. KNN algoritması, pratikliği ve sezgisel doğası nedeniyle önemli bir popülerlik kazanmış, çok sayıda gerçek dünya senaryosunda konuşma verilerini sınıflandırmak için güvenilir bir seçenek haline gelmiştir. KNN, Türkçe izole kelime tanıma çalışmalarında da kullanılmış ve özellikle gürültülü ortamlarda %86,51'e varan tanıma oranları elde edilmiştir [15].

Karar ağaçları, öznitelik uzayını ardışık ikili bölümlere ayırarak sınıflandırma yapmaktadır. Karar ağaçları yöntemi, özellikle yorumlanabilirliği ve görselleştirilebilirliği nedeniyle tercih edilmektedir. Ses tanıma sistemlerinde, karar ağaçları genellikle fonem sınıflandırma ve konuşmacı tanıma gibi görevlerde kullanılmaktadır [14].

Fisher Doğrusal Diskriminant Analizi (FLDA), sınıflar arası varyansı maksimize ederken sınıf içi varyansı minimize eden bir boyut indirgeme tekniğidir. Ayrımcı Ortak Vektör Yaklaşımı (DCVA) ise, sınıflar arası ayrımı maksimize etmeyi amaçlayan bir alt uzay yöntemidir. FLDA yöntemi, özellikle konuşmacı tanıma ve dil tanıma görevlerinde; DCVA sınıflandırıcısı ise kelime tanıma açısından etkili sonuçlar vermektedir. Yapılan bir çalışmada, FLDA-KNN hibrit algoritması kullanılarak Türkçe izole kelime tanıma

performansı %71,37 olarak ölçülmüş, DCVA sınıflandırıcısıyla izole edilmiş kelime tanımada ise %74,23'lük bir tanıma oranı elde edilmiştir [15].

Geleneksel makine öğrenme yaklaşımlarının ses tanıma sistemlerindeki avantajı, daha az veri ile eğitilebilmeleridir. Geleneksel yaklaşımların bu avantajı, özellikle kaynak kısıtlı diller için ses tanıma sistemleri geliştirirken önem kazanmaktadır. Mokgonyane ve arkadaşları tarafından yapılan çalışmada, Güney Afrika'nın yerel dilleri için geliştirilen konuşmacı tanıma sisteminde geleneksel makine öğrenme algoritmalarının etkinliğine işaret edilmiştir [16].

Barhoush ve arkadaşları tarafından yapılan çalışmada ise, konuşmacı tanıma ve lokalizasyonu için Karıştırılmış MFCC (SHMFCC) adı verilen yeni bir özellik kullanılmıştır [5]. Söz konusu çalışma, geleneksel yaklaşımların yenilikçi özellik mühendisliği teknikleriyle birleştirildiğinde hala rekabetçi sonuçlar üretebileceğini göstermektedir. Özellikle gürültülü ve yankılı ortamlarda, geleneksel yaklaşımların performansını artırmak için veri artırma teknikleri kullanılmıştır.

Ancak geleneksel makine öğrenme yaklaşımları; daha az hesaplama kaynağı gerektirmeleri, daha hızlı eğitilmeleri ve daha yorumlanabilir sonuçlar üretmeleri nedeniyle hala birçok uygulama alanında tercih edilmektedir. Geleneksel yaklaşımların ses tanıma sistemlerindeki başarısı, büyük ölçüde kullanılan özniteliklerin kalitesine ve miktarına bağlıdır. Geleneksel yaklaşımlar, özellikle sınırlı veri setleri üzerinde çalışırken ve hesaplama kaynakları kısıtlı olduğunda etkili sonuçlar üretebilmektedir. Bununla birlikte, büyük veri setleri üzerinde çalışırken, derin öğrenme yaklaşımları genellikle daha iyi performans göstermektedir [17].

2.2.2 Derin Öğrenme Yaklaşımları

ASR sistemlerinde derin öğrenme yaklaşımları, son on yılda geleneksel yöntemlere kıyasla kelime hata oranında %50'den fazla göreceli azalma sağlayarak önemli bir dönüşüm yaratmış, konuşma tanıma sistemlerinin doğruluğunu ve sağlamlığını önemli ölçüde artırmıştır [2]. Derin öğrenme tabanlı ASR sistemleri, ham ses sinyallerinden anlamlı özellikleri otomatik olarak öğrenme yetenekleri sayesinde manuel özellik mühendisliği ihtiyacını ortadan kaldırmış; özellikle gürültü, çeşitli aksanlar ve lehçeler içeren zorlu senaryolarda konuşma işleme performansında önemli ilerlemeler kaydedilmiştir [14].

Derin öğrenme tabanlı ASR sistemlerinin gelişiminde önemli mimarilerden biri Derin Sinir Ağları (DNN) olmuştur. DNN'ler, akustik modelleme için Gauss karışım dağılımlarının yerini almış veya bunları desteklemiştir. DNN'lerin çok katmanlı yapısı, ses sinyallerindeki karmaşık desenleri yakalamada üstün yetenekler göstermiştir. Ancak derin öğrenmenin ASR alanındaki etkisi yalnızca akustik modelleme ile sınırlı kalmamış, tamamen sinir ağı tabanlı uçtan uca (end-to-end) ASR mimarilerinin geliştirilmesine de yol açmıştır [2].

Evrişimli Sinir Ağları (CNN), ASR sistemlerinde yaygın olarak kullanılan derin öğrenme mimarilerinden biridir. CNN'ler, konuşma sinyallerindeki yerel zamansal ve spektral desenleri yakalamada etkilidir. Rai ve arkadaşları tarafından yapılan bir çalışmada, Mel Frekans Kepstral Katsayıları (MFCC) özelliklerini kullanan CNN tabanlı bir anahtar kelime tespit sistemi geliştirilmiş ve yüksek doğruluk oranları elde edilmiştir [18]. CNN'lerin konuşma tanımadaki başarısı, spektrogram benzeri girdilerdeki yerel özellikleri etkili bir şekilde çıkarma yeteneklerinden kaynaklanmaktadır.

Tekrarlayan Sinir Ağları (RNN), özellikle Uzun-Kısa Vadeli Bellek (LSTM) ve Çift Yönlü LSTM (BiLSTM) varyantları, konuşma verilerindeki uzun vadeli bağımlılıkları modellemede etkili olmuştur. Rai ve arkadaşları tarafından yapılan araştırmada, BiLSTM ve Dikkat mekanizmasını birleştiren bir RNN mimarisi, %93.9 doğruluk oranı ile en iyi performansı göstermiştir [18]. RNN'lerin zamansal bağlamı yakalama yeteneği, konuşma tanıma görevlerinde özellikle değerlidir çünkü konuşma doğası gereği sıralı bir olgudur ve önceki bağlamın anlaşılması doğru tanıma için önem taşımaktadır.

Son yıllarda, Transformatör mimarileri ASR sistemlerinde devrim yaratmıştır. Transformatörler, öz-dikkat mekanizmaları sayesinde giriş dizisindeki kapsamlı bağımlılıkları yakalama yetenekleri nedeniyle ASR çerçevelerinde yoğun olarak kullanılmaktadır [4]. Conformer mimarisi, Transformatör ve CNN'nin güçlü yönlerini birleştirerek hem yerel hem de global bağımlılıkları öğrenmek için konvolüsyon modülleri ekleyerek ASR için en popüler kodlayıcı modeli haline gelmiştir [13].

Uçtan uca ASR sistemleri, geleneksel ASR sistemlerinin ayrı bileşenlerini (akustik model, dil modeli ve arama) tek bir sinir ağı modeline entegre eden bir yaklaşımdır. Uçtan uca ASR sistemleri, Connectionist Temporal Classification (CTC), Attention-based Encoder-Decoder (AED) ve RNN-Transducer (RNN-T) gibi çeşitli yöntemler kullanmaktadır. CTC, giriş ve

çıkış dizileri arasındaki hizalamayı otomatik olarak öğrenen bir kayıp fonksiyonudur. AED modelleri, kodlayıcı-kod çözücü mimarisi ve dikkat mekanizması kullanarak giriş ses sinyalini doğrudan metin çıktısına dönüştürür. RNN-T ise, CTC'nin sınırlamalarını aşmak için tasarlanmış ve gerçek zamanlı ASR sistemleri için daha uygun bir alternatif olarak değerlendirilmektedir [2].

Derin transfer öğrenme (DTL), ASR sistemlerinin geliştirilmesinde önemli bir yaklaşım haline gelmiştir. DTL, büyük ölçekli eğitim veri setleri ve yüksek hesaplama kaynakları gerektiren ASR sistemlerinin sınırlamalarını aşmak için kullanılmaktadır. DTL yaklaşımı, gerçek veri toplama zorluğu olan veya nadir görülen durumlarda, küçük veya biraz farklı ancak ilgili eğitim verileri kullanarak yüksek performanslı modeller geliştirmeye yardımcı olmaktadır. DTL'nin ASR'deki uygulamaları, özellikle düşük kaynaklı dillerde veya özel alanlarda önemli performans iyileştirmeleri sağlamıştır [4].

Wav2Vec ve DeepSpeech gibi önceden eğitilmiş modeller, derin öğrenme tabanlı ASR sistemlerinin geliştirilmesinde önemli rol oynamıştır. Söz konusu modeller, büyük miktarda etiketlenmemiş ses verisi üzerinde ön eğitim alarak, konuşma temsillerini öğrenmekte ve daha sonra görev-spesifik uygulamalar için ince ayar yapılabilmektedir. Önceden eğitilmiş modeller, özellikle sınırlı etiketli veri bulunan senaryolarda etkili sonuçlar vermektedir [19].

Derin öğrenme tabanlı ASR sistemlerinde çok dilli modellerin ortaya çıkması önemli bir gelişme olarak değerlendirilmektedir. Zhang ve arkadaşları tarafından geliştirilen Universal Speech Model (USM), 100'den fazla dilde ASR gerçekleştiren tek bir büyük modeldir [20]. USM modeli, 300'den fazla dili kapsayan 12 milyon saatlik büyük bir etiketlenmemiş çok dilli veri setinde ön eğitim alarak ve daha küçük etiketli bir veri setinde ince ayar yaparak elde edilmiştir. Çok dilli modeller, özellikle düşük kaynaklı diller için ASR sistemlerinin geliştirilmesinde önemli avantajlar sağlamıştır.

Derin pekiştirmeli öğrenme (DRL), ASR sistemlerinin optimizasyonunda kullanılmaya başlanan ileri bir tekniktir. DRL, dinamik ortamlarda karar vermeyi optimize ederek hesaplama maliyetlerini azaltmaktadır. DRL tabanlı ASR sistemleri, özellikle değişen akustik koşullarda ve konuşma stillerinde adaptasyon yetenekleri ile öne çıkmakta; kullanım sırasında elde edilen geri bildirimlerle kendilerini sürekli olarak iyileştirebilmektedirler [4].

2.2.3 Hibrit Yaklaşımlar

Otomatik konuşma tanıma sistemlerinde hibrit yaklaşımlar, farklı algoritma türlerinin ve model mimarilerinin güçlü yönlerini birleştirerek, tek başına kullanıldıklarında karşılaşılan sınırlamaları aşmayı hedefleyen yöntemlerdir. Hibrit yaklaşımlar, geleneksel yöntemler ile modern tekniklerin kombinasyonlarının yanı sıra, farklı model paradigmalarının birleştirilmesiyle de oluşturulabilmektedir [3].

Hibrit yaklaşımların en temel örneklerinden biri, Saklı Markov Modelleri (HMM) ve Yapay Sinir Ağları (ANN) kombinasyonudur. Geleneksel HMM-GMM (Gaussian Mixture Model) sistemlerinde, GMM'lerin yerini ANN'lerin almasıyla oluşturulan HMM-ANN hibrit sistemleri, konuşma tanıma alanında önemli bir dönüm noktası olmuştur. Bu sistemlerde ANN'ler, akustik olasılıkları tahmin etmek için kullanılırken, HMM'ler zamansal değişkenlikleri modellemektedir [3].

Özellik çıkarma ve sınıflandırma aşamalarının hibrit bir şekilde tasarlanması, konuşma tanıma sistemlerinin başarısını artıran bir yaklaşımdır. Barhoush ve arkadaşları tarafından yapılan çalışmada, Temel Bileşen Analizi (PCA) kullanılarak hibrit alt uzay sınıflandırıcıları tasarlanmış ve bu hibrit yaklaşımların geleneksel sınıflandırıcılardan daha iyi tanıma sonuçları verdiği gösterilmiştir [5]. (DCVA)PCA ve (FLDA-KNN)PCA gibi hibrit alt uzay sınıflandırıcıları, DCVA ve FLDA-KNN sınıflandırıcılarından daha yüksek tanıma oranları elde etmiştir. Çalışmada hibrit alt uzay sınıflandırıcılarının en yüksek tanıma oranının %84,90 olduğu belirtilmiştir.

Gradyan iniş optimizasyonu için önerilen hibrit mini-grup örnek seçim yaklaşımları, konuşma tanıma sistemlerinin eğitim sürecini iyileştirmek için kullanılan yenilikçi bir yaklaşımdır. Model eğitiminde mini-grup örneklerini seçmek için konuşma veritabanlarının cinsiyet ve aksan gibi farklı özellikleri hibrit bir şekilde kullanılmaktadır. Dokuz ve Tüfekci tarafından yapılan deneysel çalışmalar, cinsiyet ve aksan özelliklerinin hibrit kullanımının, konuşma tanıma performansı açısından yalnızca bir özelliği kullanmaktan daha başarılı olduğunu göstermiştir [21].

Hibrit akustik ve dil modellerinin birleştirilmesi, özellikle alan-spesifik konuşma tanıma sistemlerinde önemli performans artışları sağlamaktadır. Jia tarafından yapılan çalışmada, önceden eğitilmiş Wav2Vec2-Large-LV60 akustik modeli ve harici bir KenLM dil

modelinin birleştirilmesiyle oluşturulan hibrit sistem, Google ve AWS gibi ticari ASR sistemlerini geçen bir performans göstermiştir [19]. Uygulanan hibrit yaklaşım, özellikle düşük kaynaklı ortamlarda ve alana özgü konuşma tanıma görevlerinde üstün performans sergilemiştir. Çalışmada, hata oranı daha yüksek olmasına rağmen, alan-spesifik ince ayar yapılmış ASR sisteminin, doğal dil anlama (NLU) görevlerinde ticari ASR sistemlerinden daha iyi sonuçlar verdiği gösterilmiştir.

Hibrit yaklaşımların önemli bir örneği olan çoklu model füzyonu yaklaşımında, farklı mimarilere sahip modellerin çıktıları birleştirilerek daha doğru ve sağlam tanıma sonuçları elde edilmektedir. Kheddar ve arkadaşları tarafından yapılan çalışmada, farklı derin öğrenme modellerinin füzyonunun, tek bir modelin kullanılmasına kıyasla daha iyi performans gösterdiği belirtilmiştir [4]. Model füzyonu, özellikle farklı akustik koşullar ve konuşma stillerinde daha tutarlı sonuçlar elde etmek için kullanılmaktadır.

Hibrit veri toplama ve etiketleme yaklaşımlarıyla ilgili olarak, Jia tarafından önerilen yarı-denetimli öğrenme etiketleme yaklaşımı, az insan müdahalesi ile alan-spesifik verilerin toplanmasını sağlamakta ve özellikle sınırlı etiketli veri bulunan alanlarda konuşma tanıma sistemlerinin geliştirilmesini kolaylaştırmaktadır [19].

Uyarlanabilir hibrit sistemler, değişen akustik koşullara ve konuşma stillerine uyum sağlama yeteneği ile öne çıkmakta, farklı ortamlarda ve kullanım senaryolarında tutarlı performans göstermek için dinamik olarak parametrelerini ayarlayabilmektedir. Kheddar ve arkadaşları tarafından yapılan çalışmada, uyarlanabilir hibrit sistemlerin, özellikle gürültülü ortamlarda ve farklı konuşma stillerinde daha sağlam tanıma sonuçları elde ettiği gösterilmiştir [4].

Farklı dillerdeki konuşmaları tanımak için tasarlanmış özel bir hibrit yaklaşım türü olan Çok dilli hibrit sistemler, dile özgü özellikler ve evrensel konuşma özellikleri arasında denge kurarak, çeşitli dillerde etkili tanıma performansı sağlamaktadır. Dhanjal ve Singh tarafından yapılan çalışmada, çok dilli hibrit sistemlerin, özellikle düşük kaynaklı diller için konuşma tanıma performansını artırdığı belirtilmiştir [17].

Gizlilik korumalı hibrit konuşma tanıma sistemleri, federe öğrenme ve diferansiyel gizlilik tekniklerini birleştirerek, kullanıcı verilerinin gizliliğini korurken yüksek tanıma performansı sağlamaktadır. Kheddar ve arkadaşları tarafından yapılan çalışmada, gizlilik

korumalı hibrit sistemlerin, özellikle hassas verilerin kullanıldığı sağlık ve finans gibi alanlarda önemli avantajlar sağladığı belirtilmiştir [4].

2.3 ÖZELLİK ÇIKARMA

Ses tanıma sistemlerinde özellik çıkarma, ham ses sinyallerinden anlamlı ve ayırt edici bilgilerin elde edilmesi sürecidir. Ses sinyalleri doğası gereği yüksek boyutlu ve gürültülü olduğundan, ham ses verilerinin doğrudan kullanımı genellikle pratik değildir. Özellik çıkarma teknikleri, ses sinyallerindeki önemli bilgileri korurken, boyutsallığı azaltır ve gürültüye karşı dayanıklılığı artırır. Bu başlıkta, ses tanıma sistemlerinde kullanılan temel özellik çıkarma teknikleri ele alınmaktadır.

2.3.1 Zaman Alanı Özellikleri (Time-Domain Features)

Ses sinyallerinin zamansal boyutunda doğrudan analiz edilmesiyle elde edilen zaman alanı özellikleri, hesaplama verimliliği ve uygulama kolaylığı nedeniyle tercih edilmektedir [14].

Enerji özelliği, ses sinyalinin belirli bir zaman dilimindeki enerji seviyesini ölçmekte ve özellikle sessiz/sesli bölümlerin ayrımında kullanılmaktadır. Sıfır geçiş oranı (Zero Crossing Rate - ZCR) ise sinyalin sıfır eksenini kesme sıklığını ölçerek frekans içeriği hakkında bilgi sağlamaktadır. Yüksek ZCR değerleri genellikle yüksek frekanslı sesleri, düşük ZCR değerleri ise düşük frekanslı sesleri göstermektedir [22], [23]. Perde (pitch) bilgisi, konuşmacı tanıma ve duygu analizi gibi uygulamalarda önem taşımakta olup, otokorelasyon fonksiyonu ile zaman alanında tespit edilebilmektedir [5].

Genlik zarfı, ses sinyalinin zaman içindeki genlik değişimini temsil eden ve sinyalin mutlak değerinin alçak geçiren filtreden geçirilmesiyle elde edilen bir özelliktir. Özellikle ritim analizi ve müzik sınıflandırma uygulamalarında kullanılmaktadır [24]. Kısa süreli zaman alanı özellikleri, ses sinyalinin 20-40 ms gibi kısa zaman dilimleri üzerinde hesaplanarak sinyalin durağan olmayan yapısını ele almakta ve konuşma aktivitesi tespitinde etkili olmaktadır [15].

Zaman alanı özelliklerinin, frekans alanı özelliklerine kıyasla daha az hesaplama kaynağı gerektirmesi, özellikle sınırlı kaynaklara sahip gömülü sistemlerde ve gerçek zamanlı uygulamalarda tercih edilmelerine neden olmaktadır. Ancak genellikle gürültüye karşı daha

hassastır ve frekans alanı özellikleri kadar ayırt edici olmayabilmektedir [10]. Yapılan çalışmalarda, zaman alanı özelliklerinin özellikle düşük gürültülü ortamlarda ve basit ses tanıma görevlerinde etkili sonuçlar verdiği gösterilmiştir [18], [21].

2.3.2 Frekans Alanı Özellikleri (Frequency-Domain Features)

Ses sinyallerinin spektral içeriğini temsil eden ve zaman alanı özelliklerine kıyasla daha fazla bilgi taşıyan frekans alanı özellikleri, ses sinyallerinin frekans bileşenlerini analiz ederek insan işitme sisteminin algısal özelliklerini modellemekte ve sistemlerin performansını önemli ölçüde artırmaktadır [23].

Frekans alanı özelliklerinin elde edilmesi genellikle ses sinyalinin zaman alanından frekans alanına dönüştürülmesi ile başlamaktadır. Kısa Süreli Fourier Dönüşümü (STFT), sinyalin kısa zaman aralıklarında alınan pencerelerine Fourier dönüşümü uygulanarak elde edilmekte ve sinyalin farklı zaman noktalarındaki frekans bileşenlerini analiz etmektedir [22]. Mel-Frekans Spektrogramı (Mel-Spectrogram), STFT'nin çıktısının Mel ölçeğine dönüştürülmesi ile elde edilmekte ve insan işitme sisteminin doğrusal olmayan frekans algısını taklit etmektedir. Bu işlem için ses sinyali tipik olarak kısa, örtüşen çerçevelere bölünmekte ve her çerçeve için Hızlı Fourier Dönüşümü (FFT) uygulanarak güç spektrumu elde edilmektedir [14].

Mel-Frekans Kepstral Katsayıları (MFCC), ses tanıma sistemlerinde en yaygın kullanılan frekans alanı özelliklerinden biridir. MFCC'ler, ses sinyalinin kısa süreli güç spektrumunun Mel ölçeğine dönüştürülmesi ve ardından ayrık kosinüs dönüşümü (DCT) uygulanması ile elde edilmekte, sinyalin spektral zarfını temsil eden ve yüksek performans sağlayan bir dizi katsayı üretmektedir. MFCC'ler, konuşma tanıma, konuşmacı tanıma ve duygu tanıma gibi çeşitli ses işleme uygulamalarında başarıyla kullanılmaktadır [5]. Spektrogram ise ses sinyalinin zaman-frekans temsilini görselleştiren, yatay ekseni zaman, dikey ekseni frekans ve renk yoğunluğu belirli bir zaman-frekans noktasındaki enerji seviyesini temsil eden iki boyutlu bir grafiktir. Grafik, ses tanıma sistemlerinde özellik çıkarma için doğrudan kullanılabildiği gibi, derin öğrenme modellerine girdi olarak da verilebilmektedir [18].

Frekans alanı özelliklerinin gürültülü ortamlarda daha dayanıklı olması önemli bir avantajdır. Gürültü genellikle ses sinyalinin belirli frekans bantlarını etkilemekte, ancak tüm

bantları aynı şekilde etkilememektedir. Bu nedenle, frekans alanı özellikleri kullanılarak gürültüden daha az etkilenen frekans bileşenleri analiz edilebilmektedir [5].

2.3.3 Dalgacık Dönüşümü Tabanlı Özellikler (Wavelet Transform Based Features)

Dalgacık dönüşümü, Fourier dönüşümünden farklı olarak, sinyalin hem zaman hem de frekans bilgisini eş zamanlı sunabilme yeteneğiyle, özellikle durağan olmayan ses sinyallerinin analizinde büyük avantaj sağlamaktadır [25]. Temelde iki ana kategoriye ayrılmaktadır: Sürekli Dalgacık Dönüşümü (CWT) ve Ayrık Dalgacık Dönüşümü (DWT). CWT, ana dalgacığın ölçeklenmesi ve kaydırılması ile sinyalin sürekli gösterimini sağlarken, DWT hesaplama verimliliği nedeniyle ses tanıma uygulamalarında tercih edilmektedir [10].

Dalgacık dönüşümünün matematiksel temeli, ana dalgacık olarak adlandırılan prototip fonksiyonun ölçeklenmesi ve kaydırılmasına dayanmaktadır. Dalgacık dönüşümü, sinyalin farklı frekans bantlarındaki enerji dağılımını analiz etmek için kullanılır ve dalgacık fonksiyonu ile konvolüsyon şeklinde ifade edilir. Dalgacık aileleri (Haar, Daubechies, Symlet, Coiflet ve Biorthogonal) ses tanıma sistemlerinin performansını doğrudan etkilemekte olup, her birinin kendine özgü özellikleri bulunmaktadır [25].

Dalgacık dönüşümü tabanlı özellik çıkarma sürecinde, ses sinyali çerçevelere bölünüp her çerçeveye dalgacık dönüşümü uygulanmaktadır. Elde edilen dalgacık katsayıları, doğrudan özellik vektörü olarak kullanılabileceği gibi, istatistiksel ölçümler hesaplanarak da özellik vektörleri oluşturulabilmektedir [10]. Çok çözünürlüklü analiz (MRA), sinyalin farklı çözünürlük seviyelerinde analiz edilmesini sağlayarak hem düşük hem de yüksek frekanslı bileşenlerin etkili incelenmesine imkan tanır. MRA, alçak ve yüksek geçiren filtreler kullanarak sinyali alt bantlara ayırmakta ve her seviyede daha detaylı analiz yapılabilmektedir [25].

Dalgacık paketleri, standart dalgacık dönüşümünün bir uzantısı olarak daha esnek frekans bölümlemesi sağlamaktadır. Standart DWT'de sadece yaklaşım katsayıları alt bantlara ayrılırken, dalgacık paketlerinde hem yaklaşım hem de detay katsayıları alt bantlara ayrılarak frekans spektrumunun daha dengeli bölünmesini ve ses sinyallerinin daha ayrıntılı analizini sağlar [25]. Dalgacık dönüşümü tabanlı özellikler, çok çözünürlüklü yapısı sayesinde gürültülü ortamlarda bile ses sinyalinin önemli özelliklerini koruyabilmekte ve

lokalizasyon özelliği sinyaldeki geçici olayların ve ani değişimlerin tespit edilmesini kolaylaştırmaktadır [10].

Dalgacık dönüşümü ve MFCC özelliklerinin birleştirilmesi, her iki tekniğin avantajlarını birleştirerek daha güçlü özellik temsili sağlamaktadır. Dalgacık dönüşümü sinyalin zaman-frekans bilgisini sunarken, MFCC insan işitme sistemine uygun spektral temsil sunarak özellikle karmaşık ses tanıma görevlerinde daha yüksek doğruluk oranları elde edilmesini sağlar [5]. Dalgacık dönüşümü sonucunda elde edilen yüksek boyutlu özellik vektörleri için PCA veya LDA gibi boyut indirgeme teknikleri kullanılarak özellik vektörlerinin boyutu azaltılmakta ve sınıflandırma performansı iyileştirilmektedir [15].

2.3.4 Doğrusal Tahmin Kodlaması (Linear Predictive Coding - LPC)

Doğrusal Tahmin Kodlaması (LPC), konuşma sinyallerinin geçmiş örneklerinin doğrusal kombinasyonu olarak modellendiği, insan ses üretim sisteminin fiziksel özelliklerini matematiksel olarak temsil eden önemli bir özellik çıkarma tekniğidir [14]. LPC'nin temel prensibi, mevcut ses örneğinin değerini önceki örneklerin ağırlıklı toplamı şeklinde tahmin eden otoregresif bir model kullanarak, konuşma üretim sisteminin ses yolunu bir tüp olarak ele alıp bu tüpün rezonans özelliklerini katsayılar ile temsil etmektir [23].

LPC katsayılarının hesaplanmasında genellikle Levinson-Durbin algoritması kullanılmakta ve bu katsayılar konuşma sinyalinin spektral zarfını temsil ederek ses tanıma sistemlerinde özellik vektörleri olarak kullanılmaktadır [14]. LPC, konuşma sinyalinin spektral özelliklerini az sayıda parametre ile temsil etme yeteneği sayesinde, özellikle düşük bant genişliği gerektiren uygulamalarda ve sınırlı hesaplama kaynaklarına sahip sistemlerde tercih edilmektedir [10].

LPC tabanlı özellik çıkarma güçlü bir araç olsa da, gürültülü ortamlarda performansı düşebilmekte, bu nedenle gerçek dünya uygulamalarında genellikle ön işleme teknikleriyle birlikte kullanılmaktadır [5]. LPC katsayıları ayrıca konuşma sinyalinin verimli kodlanmasını ve iletilmesini sağlamakta, diğer özellik çıkarma teknikleriyle birleştirilerek (LPC ve MFCC özelliklerinin birleştirilmesi gibi) hibrit yaklaşımlar oluşturulabilmekte ve ses tanıma sistemlerinin performansı artırılabilmektedir [22].

İnsan işitme sisteminin algısal özelliklerini tam yansıtmama sınırlamasını aşmak için, LPC'nin algısal bir uzantısı olan Algısal Doğrusal Tahmin (PLP) geliştirilmiş, bu yöntem LPC'nin temel prensiplerini korurken insan işitme sisteminin frekans seçiciliğini ve maskeleme özelliklerini de dikkate almıştır [15]. Derin öğrenme yaklaşımlarının yaygınlaşmasıyla LPC'nin rolü değişse de, hesaplama verimliliği ve yorumlanabilirliği hala önemli avantajlar sunmakta, LPC tabanlı özellikler derin öğrenme modellerinin eğitiminde ön işleme veya ek özellikler olarak kullanılabilmektedir [21].

2.3.5 PLP (Perceptual Linear Prediction)

Perceptual Linear Prediction (PLP), Hermansky tarafından 1990 yılında önerilen ve insan işitme sisteminin psikoakustik özelliklerini dikkate alarak konuşma sinyallerinden anlamlı özellikler çıkarmayı amaçlayan bir tekniktir [23]. PLP analizi, konuşma sinyalinin spektral temsilini insan işitme sisteminin algısal özelliklerine göre dönüştürerek gerçekleştirilir. Bu süreçte spektral çözünürlük Bark ölçeğine göre yeniden örneklenir, eşit yükseklik eğrilerine dayalı ağırlıklandırma uygulanır ve sinyal yoğunluk-yükseklik güç yasasına göre sıkıştırılır [14].

PLP özelliklerinin çıkarılması sırasında öncelikle sinyalin güç spektrumu hesaplanıp Bark ölçeğine dönüştürülür, ardından eşit yükseklik eğrisi simülasyonu ve yoğunluk-yükseklik dönüşümü uygulanarak spektrum insan algısına göre düzenlenir; son olarak, elde edilen spektrumun ters Fourier dönüşümü alınarak otokorelasyon katsayıları hesaplanır ve bu katsayılar kullanılarak doğrusal tahmin analizi gerçekleştirilir. PLP yöntemi, özellikle gürültülü ortamlarda ve farklı konuşmacıların seslerini tanımada etkili sonuçlar vermekte, konuşma sinyalindeki formant yapılarını daha iyi temsil etmekte ve konuşmacıya bağlı değişkenliklere karşı daha dayanıklı özellikler sunmaktadır [15], [18].

PLP'nin çeşitli varyasyonları da geliştirilmiştir. Rasta-PLP (Relative Spectral Transform - Perceptual Linear Prediction), kanal etkilerini azaltmak için tasarlanmış bir varyasyon olup özellikle telefon hatları üzerinden iletilen konuşma sinyallerinin tanınmasında etkilidir. PLP ve MFCC özelliklerinin birleştirilmesiyle oluşturulan hibrit özellik vektörleri, ses tanıma sistemlerinin performansını artırabilmektedir [22].

2.3.6 Derin Öğrenme Tabanlı Özellik Çıkarma

Derin öğrenme tabanlı özellik çıkarma yöntemleri, manuel tasarlanan özellikler yerine ham ses sinyallerinden otomatik olarak anlamlı özellikler öğrenebilme kapasitesiyle, özellikle gürültülü ortamlarda ve farklı aksanlar içeren zorlu senaryolarda üstün performans göstermektedir [14].

Evrişimli sinir ağları (CNN), ses sinyallerinin spektrogram temsillerinden otomatik özellik çıkarmada etkin kullanılmaktadır. CNN'ler zaman-frekans düzlemindeki yerel desenleri tespit ederek hiyerarşik yapıda temsil etme yeteneğiyle, MFCC veya spektrogram gibi geleneksel özelliklerle veya doğrudan ham ses sinyallerinden özellik çıkarabilmektedir [23]. Tekrarlayan sinir ağları (RNN), özellikle LSTM ve BiLSTM mimarileri, ses sinyallerinin zamansal yapısını modellemede üstün performans göstermektedir [18]. Wav2Vec 2.0 gibi kendini denetimli öğrenme yaklaşımları, etiketlenmemiş büyük ses veri kümeleri üzerinde önceden eğitilerek ses sinyallerinin genel özelliklerini öğrenmekte ve düşük kaynaklı diller için önemli avantajlar sağlamaktadır [14].

VLAAI ağı gibi yenilikçi mimariler, EEG sinyallerinden konuşma zarfının çözümlenmesinde geleneksel doğrusal modele göre %52'lik artış sağlayarak üstün performans göstermiştir [24]. Ahmed ve arkadaşları tarafından düşük kaynaklı Peştuca dili için geliştirilen Ayrık Dalgacık Dönüşümü ve MFCC tabanlı özellik çıkarma çerçevesi, hesaplama verimliliği açısından klasik makine öğrenimi modelleriyle yüksek performans elde etmiştir [10].

2.4 VERİ SETLERİ

Veri setleri, konuşma tanıma sistemlerinin eğitilmesi, test edilmesi ve değerlendirilmesi için temel kaynak olarak işlev görmektedir. Sistemlerin başarısı büyük ölçüde kullanılan veri setlerinin kalitesine, büyüklüğüne ve çeşitliliğine bağlıdır. Bu bölümde, konuşma tanıma alanında kullanılan çeşitli veri setleri kategorileri incelenmektedir.

2.4.1 Genel Amaçlı Konuşma Veri Setleri

Otomatik konuşma tanıma sistemlerinin geliştirilmesi, eğitilmesi ve değerlendirilmesi için temel yapı taşları olarak kabul edilen genel amaçlı konuşma veri setleri; çeşitli

konuşmacılardan, farklı akustik ortamlardan ve geniş kelime dağarcığından oluşan konuşma örneklerini içermektedir. Özellikle derin öğrenme tabanlı yaklaşımların başarısı, büyük ölçüde eğitim için kullanılan veri setlerinin kalitesine ve miktarına bağlıdır [2].

Genel amaçlı konuşma veri setleri, genellikle saatler veya binlerce saat uzunluğunda konuşma kayıtları içermektedir. Google, Universal Speech Model (USM) için 12 milyon saatlik etiketlenmemiş çok dilli veri seti kullanmaktadır [20]. Bu büyük ölçekli veri setleri, özellikle öz-denetimli öğrenme yaklaşımlarında önem taşımaktadır ve modellerin farklı konuşma tarzları, aksanlar ve gürültü koşullarında dayanıklılık göstermesini sağlamaktadır. Bununla birlikte, genel amaçlı konuşma tanıma modellerinin değerlendirilmesi için YouTube gibi sürekli güncellenen ve zengin içeriğe sahip platformlar değerli veri kaynakları olabilmektedir [26].

Genel amaçlı konuşma veri setlerinin oluşturulmasında demografik çeşitlilik, gürültü koşulları, duygusal durumlar, tartışılan konular ve iletişimde kullanılan dil gibi faktörler önemlidir [17]. Yüksek kaliteli bir konuşma tanıma sistemi geliştirmek için bu faktörlerin tümünü kapsayan büyük hacimli eğitim verileri gerekmektedir. Ayrıca, ses veri setlerinin dokümantasyonu büyük önem taşımaktadır [7].

Konuşma sinyallerinin farklı kategorilere (konuşma, müzik ve çevresel sesler gibi) sınıflandırılmasında kullanılan veri setlerinin, gerçek dünya koşullarını temsil edecek şekilde tasarlanması gerekmektedir. Bu bağlamda, genel amaçlı konuşma veri setleri, konuşma tanıma sistemlerinin farklı dillerde ve lehçelerde etkin bir şekilde çalışabilmesi için çeşitli konuşma örneklerini içermelidir [23].

2.4.2 Komut Tanıma Veri Setleri (Keyword Spotting)

Komut tanıma veya anahtar kelime tespiti (keyword spotting), konuşma tanımanın özel bir dalı olup, konuşma içerisindeki belirli komutların veya 'anahtar kelimelerin' tanımlanmasıyla ilgilenmektedir. Anahtar kelime tespiti küçük bir ayak izi boyutuna sahiptir ve robotik, sanal asistanlar ve araç üstü elektronik cihazlar gibi çeşitli alanlarda geniş uygulama alanları bulunmaktadır. Komut tanıma veri setleri, genellikle sınırlı sayıda kelime veya kısa ifade içeren özel amaçlı veri setleridir ve gerçek dünya uygulamalarına yönelik olarak, modellerin belirli komutları hızlı ve doğru bir şekilde tanımasını sağlamak için tasarlanmıştır [18].

Google Konuşma Komutları Veri Seti, komut tanıma alanında yaygın olarak kullanılan bir veri setidir. Araştırmacılar bu veri seti üzerinde çeşitli makine öğrenimi ve derin öğrenme teknikleriyle anahtar kelime tespiti görevini gerçekleştirmektedir [18]. Fayda-spesifik konuşma tanıma sistemleri geliştirmek için önceden eğitilmiş DeepSpeech2 ve Wav2Vec2 akustik modelleri kullanılmaktadır [19]. Qwen2-Audio adlı büyük ölçekli bir ses-dil modeli, çeşitli ses sinyali girişlerini kabul edebilmekte ve konuşma talimatlarına ilişkin ses analizi veya doğrudan metinsel yanıtlar üretebilmektedir. Karmaşık hiyerarşik etiketler yerine, farklı veri ve görevler için doğal dil istemleri kullanılarak ön-eğitim süreci basitleştirilmekte ve veri hacmi genişletilmektedir [27]. Konuşulan rakam tanıma (SDR) sistemleri, telefon tabanlı hizmetler, belirli banka işlemleri, havayolu rezervasyon sistemleri ve fiyat çıkarma gibi çeşitli insan-makine etkileşim uygulamalarında gerekli olmaktadır [28].

2.4.3 Alan-Spesifik Veri Setleri

Alan-spesifik veri setleri, belirli bir endüstri, sektör veya uygulama alanına özgü konuşma verilerini içeren özelleştirilmiş koleksiyonlar olarak tanımlanmaktadır. Bu tür veri setleri, genel amaçlı konuşma tanıma sistemlerinin performansını belirli alanlarda artırmak için kullanılmaktadır. Alan-spesifik konuşma tanıma sistemleri, özellikle düşük kaynak ortamlarında ticari ASR sistemlerinin genellikle zayıf performans gösterdiği durumlarda büyük önem taşımaktadır. Bu sistemler, belirli bir alanın terminolojisine, dilbilgisi yapılarına ve konuşma kalıplarına odaklanarak, o alana özgü konuşma tanıma görevlerinde daha yüksek doğruluk sağlamaktadır. Alan-spesifik veri setlerinin oluşturulmasında, veri toplama süreci önemli bir problem alanıdır [19].

Alan-spesifik veri setlerinin önemli bir özelliği, hedef alanın terminolojisini ve jargonunu kapsamlı bir şekilde temsil etmeleridir. Tıbbi alanda kullanılan ASR sistemlerinin, tıbbi terminoloji ve kısaltmalar konusunda özel eğitim alması gerekmektedir. Benzer şekilde, hukuk, finans veya mühendislik gibi alanlarda kullanılan ASR sistemleri, ilgili alanın teknik dilini ve terminolojisini doğru bir şekilde tanıyabilmelidir. Bu nedenle, alan-spesifik veri setleri, hedef alanın dilsel özelliklerini ve karmaşıklıklarını yeterince temsil edecek şekilde tasarlanmalıdır [17].

Alan-spesifik veri setlerinin oluşturulmasında, düşük kaynaklı diller veya özel alanlar için yeterli miktarda etiketli veri bulmak karşılaşılan zorluklardan biridir. Bu sorunu aşmak için,

veri artırma teknikleri, transfer öğrenimi ve yarı-denetimli öğrenme yaklaşımları gibi çeşitli çözümler önerilmektedir. Özellikle, önceden eğitilmiş genel amaçlı modellerin alan-spesifik verilere ince ayar yapılması, sınırlı miktarda alan-spesifik veri ile bile tatmin edici sonuçlar elde edilmesini sağlayabilmektedir [2].

2.4.4 Çok Dilli Veri Setleri

Otomatik konuşma tanıma sistemlerinin farklı dillerde ve lehçelerde etkin bir şekilde çalışabilmesini amaçlayan çok dilli veri setleri, birden fazla dilde konuşma örneklerini içermekte ve dil-bağımsız veya çok dilli ASR sistemlerinin geliştirilmesinde temel yapı taşları olarak kullanılmaktadır. Google'ın Universal Speech Model (USM) adlı çalışmasında, 100'den fazla dilde otomatik konuşma tanıma yapabilen bir model geliştirilmiştir. USM, 12 milyon saatlik etiketlenmemiş çok dilli veri kullanılarak eğitilmiştir ve düşük kaynaklı diller için bile yüksek performans göstermektedir [20].

Düşük kaynaklı diller için ASR sistemleri geliştirmek daha zordur. Özellikle, çok dilli öğrenme ve sıfır atışlı veya düşük atışlı transfer öğrenimi teknikleri, kaynak açısından zengin dillerden düşük kaynaklı dillere bilgi transferini sağlamaktadır [2]. Ayrıca, transformatör tabanlı modeller, farklı diller arasındaki ortak özellikleri öğrenebilmekte ve bu bilgiyi düşük kaynaklı dillere transfer edebilmektedir. Böylece, sınırlı miktarda eğitim verisi olan diller için bile yüksek performanslı ASR sistemleri geliştirilebilmektedir [14].

Çok dilli veri setlerinin kalitesi ve çeşitliliği, geliştirilen ASR sistemlerinin genelleştirme yeteneğini doğrudan etkilemektedir. Çok dilli veri setleri, gerçek dünya koşullarını temsil etmeli, farklı dillerin ve lehçelerin yanı sıra, çeşitli akustik ortamları, konuşma stillerini ve konuşmacı demografilerini de kapsamalıdır [26].

2.5 BAŞARI KRİTERLERİ

Otomatik konuşma tanıma sistemlerinin performansını değerlendirmek için çeşitli başarı kriterleri kullanılmaktadır. Başarı kriterleri, konuşma tanıma sistemlerinin doğruluğunu, anlaşılabilirliğini ve kullanılabilirliğini değerlendirmek için tasarlanmış nicel ve nitel ölçütleri içermektedir [2].

Kelime Hata Oranı (Word Error Rate - WER), konuşma tanıma sistemlerinin değerlendirilmesinde en yaygın kullanılan metriklerden biridir. WER, tanıma sisteminin çıktısı ile referans transkripsiyon arasındaki farklılıkların bir ölçüsüdür. Bu metrik; ikame (substitution), ekleme (insertion) ve silme (deletion) olmak üzere, üç tür hatayı dikkate almaktadır. WER hesaplaması, Levenshtein mesafesi kullanılarak gerçekleştirilir ve genellikle yüzde olarak ifade edilir. Düşük WER değerleri, daha iyi performans anlamına gelmektedir. WER'in matematiksel ifadesi şu şekildedir [17]:

$$WER = \frac{(S + D + I)}{N} \tag{2.1}$$

Burada S ikame edilen kelimelerin sayısını, D silinen kelimelerin sayısını, I eklenen kelimelerin sayısını ve N referans transkripsiyondaki toplam kelime sayısını temsil etmektedir.

Karakter Hata Oranı (Character Error Rate - CER), WER'e benzer şekilde çalışan ancak kelimeler yerine karakterler üzerinden hesaplanan bir metriktir. CER özellikle karakter tabanlı diller için veya alt kelime birimlerinin kullanıldığı sistemlerde tercih edilmektedir. Ayrıca, Cümle Hata Oranı (Sentence Error Rate - SER) da bir diğer önemli metriktir ve hatalı tanınan cümlelerin oranını ölçmektedir [4].

Ölçek-Değişmez Sinyal-Bozulma Oranı (Scale-Invariant Signal-to-Distortion Ratio - SI-SDR), özellikle konuşma ayrıştırma ve izolasyon sistemlerinde kullanılan bir metriktir. SI-SDR, hedef sinyal ile tahmin edilen sinyal arasındaki ilişkiyi ölçer ve sinyal kalitesini değerlendirmek için kullanılır. Bu metrik, ölçeklendirme değişikliklerine karşı dayanıklıdır ve dB cinsinden ifade edilir. Yüksek SI-SDR değerleri, daha iyi sinyal kalitesi anlamına gelmektedir. SI-SDR şu şekilde hesaplanır [25]:

$$SI - SDR = 10\, log_{10}(e_{target} \,/\, e_{residual}) \tag{2.2}$$

Burada e_{target} referans sinyalinin enerjisi ve $e_{residual}$ ise uygun hizalamadan sonra ayrılmış sinyal ile referans sinyal arasındaki farkı temsil etmektedir.

Konuşma Kalitesinin Algısal Değerlendirmesi (Perceptual Evaluation of Speech Quality - PESQ), konuşma sinyallerinin kalitesini değerlendirmek için kullanılan bir başka önemli metriktir. PESQ, Uluslararası Telekomünikasyon Birliği (ITU) tarafından standartlaştırılmış

bir ölçüttür ve insan algısını modelleyerek konuşma kalitesini değerlendirmeyi amaçlamaktadır. PESQ skoru -0.5 ile 4.5 arasında değişmekte olup, yüksek değerler daha iyi konuşma kalitesini göstermektedir. Bu metrik, özellikle gürültü azaltma ve konuşma iyileştirme sistemlerinin değerlendirilmesinde yaygın olarak kullanılmaktadır [25].

Kısa Süreli Nesnel Anlaşılabilirlik (Short-Time Objective Intelligibility - STOI), konuşma anlaşılabilirliğini ölçen bir diğer önemli metriktir. STOI, temiz konuşma sinyali ile işlenmiş konuşma sinyali arasındaki korelasyonu hesaplayarak anlaşılabilirliği değerlendirmektedir. STOI değerleri 0 ile 1 arasında değişmekte olup, 1'e yakın değerler daha yüksek anlaşılabilirliği göstermektedir. Bu metrik, özellikle gürültülü ortamlarda konuşma anlaşılabilirliğini değerlendirmek için kullanılmaktadır [25].

Doğruluk (Accuracy) ve F1-skoru gibi geleneksel sınıflandırma metrikleri de konuşma tanıma sistemlerinin değerlendirilmesinde kullanılmaktadır. Özellikle anahtar kelime tespiti (keyword spotting) gibi belirli konuşma tanıma görevlerinde, doğruluk önemli bir başarı kriteridir [18].

Gerçek zamanlı faktör (Real-Time Factor - RTF), bir konuşma tanıma sisteminin hesaplama verimliliğini ölçen bir metriktir. RTF, işleme süresinin gerçek konuşma süresine oranı olarak tanımlanır. 1'den küçük RTF değerleri, sistemin gerçek zamanlı çalışabildiğini göstermektedir. Bu metrik, özellikle sınırlı kaynaklara sahip cihazlarda çalışan konuşma tanıma sistemleri için önemlidir [2].

Yukarıda sayılanlara ilave olarak, konuşma tanıma sistemleri; modelin sağlamlığı, genelleştirme yeteneği, güvenlik değerlendirmeleri, sistemlerin başarısı alt görevlerdeki performanslarıyla da değerlendirilmektedir [19], [26], [29].

3. LİTERATÜR TARAMASI

Shukla ve arkadaşları tarafından gerçekleştirilen çalışmada, makine öğrenme teknikleri kullanılarak otomatik konuşma tanıma (ASR) sistemleri incelenmiştir [30]. Çalışma, özellikle büyük kelime dağarcıklı sürekli konuşma tanıma (LVCSR) alanındaki problemleri ele almaktadır. Araştırmacılar, geleneksel Saklı Markov Modeli (HMM) ve Gauss Karışım Modeli (GMM) yaklaşımlarının yanı sıra, HMM-Derin Sinir Ağı (DNN) hibrit modellerini ve uçtan uca (end-to-end) modelleri karşılaştırmışlardır. Çalışmada, Bağlantıcı Zamansal Sınıflandırma (CTC) tabanlı uçtan uca ASR sistemlerinde "soft forgetting" (yumuşak unutma) adı verilen bir teknik önerilmiştir. Bu teknik, özellikle kısa eğitim veri setlerinde aşırı uyumu (overfitting) önlemek için BLSTM (Çift Yönlü Uzun-Kısa Vadeli Bellek) ağlarının tüm konuşma boyunca açılması yerine, örtüşmeyen kesitler üzerinde çalıştırılmasını içermektedir. Özellik çıkarma aşamasında MFCC (Mel Frekans Kepstral Katsayıları) kullanılmıştır. Deneyler, 300 saatlik İngilizce Switchboard veri seti üzerinde gerçekleştirilmiştir. Sonuçlar, önerilen soft forgetting tekniğinin, rakip bir bütün-konuşma telefon CTC BLSTM'ye göre WER'de (Kelime Hata Oranı) ortalama %7-9 oranında iyileşme sağladığını göstermiştir. Hub5-2000 Switchboard/CallHome test setlerinde, konuşmacıdan bağımsız modeller için %9.1/%17.4 WER ve konuşmacıya adapte edilmiş modeller için %8.7/%16.8 WER elde edilmiştir. Çalışma ayrıca, dikkat tabanlı kodlayıcı-kod çözücü modellerinin LibriSpeech veri seti üzerindeki performansını artırmak için çeşitli tekniklerin etkilerini de incelemiştir. Araştırmacılar, daha küçük modelleme birimleri ve orta düzeyde çerçeve hızı azaltmanın önemli olduğunu bulmuşlardır. Ayrıca, CTC modellerinin posterior spike zamanlamalarını yönlendirmek için bir teknik önererek, posterior füzyon ve bilgi damıtma işlemlerini iyileştirmişlerdir. Çalışmanın en önemli katkısı, uçtan uca konuşma tanıma sistemlerinin performansını artırmak için yeni teknikler önermesi ve bu tekniklerin etkinliğini deneysel olarak göstermesidir.

Gourisaria ve arkadaşları tarafından gerçekleştirilen çalışmada, çevresel ses sınıflandırması için makine öğrenme teknikleri kullanılarak karşılaştırmalı analiz yapılmıştır [22]. Çalışmada, benzer özelliklere sahip ancak farklı karakteristiklerdeki iki ses veri seti üzerinde çeşitli makine öğrenme teknikleri kullanılarak ses sinyallerinin sınıflandırılması amaçlanmıştır. Araştırmacılar, Mel Frekans Kepstral Katsayıları (MFCC) ve Kısa Süreli Fourier Dönüşümü (STFT) özellik çıkarma yöntemlerini kullanarak yedi farklı makine

öğrenme modelini (Lojistik Regresyon, K-En Yakın Komşular, Destek Vektör Makinesi, Naive Bayes, Karar Ağacı, Rastgele Orman ve Yapay Sinir Ağı) karşılaştırmışlardır. Veri seti olarak, 10 farklı çevresel ses sınıfını içeren 8732 ses örneğinden oluşan UrbanSound8K ve 8 farklı ses olayı sınıfını içeren 1288 ses örneğinden oluşan Sound Event Audio Dataset kullanılmıştır. Özellik çıkarma sürecinde, UrbanSound8K veri seti için MFCC, Sound Event Audio Dataset için ise STFT özellikleri çıkarılmıştır. Her ses örneği için toplam 186 özellik çıkarılmış ve ses örnekleri 4 saniye uzunluğunda kesilmiştir. Ayrıca, gürültü temizleme işlemi uygulanarak ses örneklerindeki gürültüler azaltılmış ve sadece ayırt edici ses özellikleri korunmuştur. Sistemin performansı, doğruluk, kesinlik, duyarlılık, özgüllük ve F1-skoru metrikleri ile değerlendirilmiştir. Sonuçlar, Yapay Sinir Ağı (ANN) modelinin her iki veri setinde de en iyi performansı gösterdiğini ortaya koymuştur. UrbanSound8K veri setinde %91.41, Sound Event Audio Dataset'te ise %91.27 doğruluk oranı elde edilmiştir. Çalışmanın en önemli katkısı, çevresel ses sınıflandırması için farklı makine öğrenme modellerinin karşılaştırmalı analizini sunması ve gürültü temizleme işleminin sınıflandırma performansı üzerindeki etkisini incelemesidir. Araştırmacılar, daha basit ANN mimarilerinin bile karmaşık mimarilere göre daha düşük eğitim süresi ve hesaplama maliyeti ile yüksek doğruluk oranları elde edebileceğini göstermişlerdir.

Gençyılmaz ve Karaoğlan tarafından gerçekleştirilen çalışmada, Türkçe dilinde konuşmadan metne dönüşüm (CoST) sistemlerinin optimizasyonu için çeşitli makine öğrenme yaklaşımları analiz edilmiştir [11]. Çalışma, Türkçe'nin eklemeli yapısı, fonolojik özellikleri, özel harfleri, lehçe ve aksan farklılıkları, kelime vurgusu, kelime başı ünlü etkileri, arka plan gürültüsü, cinsiyete bağlı ses varyasyonları gibi sorunları ele almaktadır. Araştırmacılar, literatürde sınırlı olarak incelenen Türkçe'ye özgü ses kliplerinden oluşan konuşma verilerini, özellikle Evrişimli Sinir Ağı (CNN) ve Evrişimli Tekrarlayan Sinir Ağı (CRNN) gibi modeller kullanarak yüksek performans doğruluğuyla metne dönüştürmeyi amaçlamışlardır. Özellik çıkarma aşamasında, Mel Frekans Kepstral Katsayıları (MFCC) kullanılmış ve bu yöntemin spektrogramlara göre avantajları incelenmiştir. Çalışmada 26.485 Türkçe ses klibinden oluşan bir veri seti üzerinde deneysel çalışmalar gerçekleştirilmiş ve performans değerlendirmesi çeşitli metriklerle yapılmıştır. Ayrıca, modelin performansını iyileştirmek için hiperparametreler optimize edilmiştir. Dokuz farklı makine öğrenme modeli (CNN, CRNN, LSTM, RF, SVM, KNN, GNB, DT, GRU) karşılaştırılmış ve en yüksek performans CRNN yaklaşımıyla elde edilmiştir (%96.47

doğruluk). CNN modeli ise en yüksek F1-skorunu (%97.55) ve kesinlik değerini (%97.68) sağlamıştır. Çalışmanın en önemli katkısı, Türkçe dilinde konuşmadan metne dönüşüm alanında literatürdeki ilk çalışmalardan biri olması ve çeşitli makine öğrenme modellerinin güçlü ve sınırlı yönlerini ortaya koymasıdır.

Ayall ve arkadaşları tarafından gerçekleştirilen çalışmada, Etiyopya'nın resmi çalışma dili olan Amharca için konuşulan rakam tanıma sistemi geliştirilmiştir [28]. Çalışmada, konuşulan rakam tanıma (Spoken Digits Recognition - SDR) sistemlerinin telefon tabanlı hizmetler, arama sistemleri, belirli banka işlemleri, havayolu rezervasyon sistemleri ve fiyat çıkarma gibi çeşitli insan-makine etkileşim uygulamalarında gerekli olduğu belirtilmiştir. Araştırmacılar, Amharca konuşulan rakam tanıma (AmSDR) modeli geliştirmek için öncelikle Amharca Konuşulan Rakamlar Veri Seti (AmSDD) oluşturmuşlardır. Bu veri seti, 0 (Zaero) ile 9 (zet'enyi) arasındaki rakamları içeren 12.000 konuşma kaydından oluşmaktadır ve farklı yaş grupları, cinsiyetler ve lehçelerden 120 gönüllü konuşmacıdan toplanmıştır. Her konuşmacı, her rakamı 10 kez tekrarlamıştır. Özellik çıkarma aşamasında, konuşma sinyalinden eğitilebilir özellikler çıkarmak için Mel Frekans Kepstral Katsayıları (MFCC) ve Mel-Spektrogram yöntemleri kullanılmıştır. Çalışmada, AmSDR modelinin geliştirilmesi için temel olarak Doğrusal Diskriminant Analizi (LDA), K-En Yakın Komşu (KNN), Destek Vektör Makinesi (SVM) ve Rastgele Orman (RF) gibi klasik denetimli öğrenme algoritmaları ile çeşitli deneyler yapılmıştır. AmSDR'nin tanıma performansını daha da iyileştirmek için, araştırmacılar Batch normalizasyonlu üç katmanlı Evrişimli Sinir Ağı (CNN) mimarisi önermişlerdir. Deney sonuçları, önerilen CNN modelinin temel algoritmaları geride bıraktığını ve MFCC özellikleri kullanılarak %99, Mel-Spektrogram özellikleri kullanılarak ise %98 doğruluk oranı elde ettiğini göstermiştir. Karşılaştırma amacıyla kullanılan geleneksel yöntemlerden en iyi performansı MFCC özellikleriyle Rastgele Orman algoritması göstermiş ve %97 doğruluk oranına ulaşmıştır. Çalışmanın en önemli katkısı, daha önce mevcut olmayan Amharca konuşulan rakamlar veri setini oluşturması ve bu az kaynaklı dil için yüksek doğrulukta bir konuşma tanıma sistemi geliştirmesidir.

Wojnar ve arkadaşları tarafından gerçekleştirilen çalışmada, genel amaçlı konuşma tanıma makine öğrenme modellerinin performansını ve uyarlanabilirliğini çeşitli gerçek dünya senaryolarında değerlendirmek için YouTube'u veri kaynağı olarak kullanan "Mi-Go" adlı

bir araç tanıtılmıştır [26]. Çalışmada, mevcut konuşma tanıma değerlendirme yöntemlerinin genellikle statik ve küratörlü veri setlerine dayandığı, bunun da modellerin gerçek dünya senaryolarındaki genelleştirilebilirliğini sınırlayabileceği belirtilmiştir. Araştırmacılar, YouTube'un çeşitli diller, aksanlar, lehçeler, konuşma stilleri ve ses kalitesi seviyeleri içeren zengin ve sürekli güncellenen bir içerik kaynağı olduğunu belirterek, bu platformun konuşma tanıma modellerinin değerlendirilmesi için ideal bir kaynak olduğunu ifade etmişlerdir. Mi-Go aracı, YouTube'dan veri çıkarma, açıklama ve değerlendirme sürecini otomatikleştirerek, değerlendirme amaçları için güncel ve temsili bir örnek sağlamaktadır. Çalışmada, aracın etkinliğini göstermek için OpenAI'nin Whisper (tiny.en, base.en, small.en, medium.en, large-v1, large-v3), NVIDIA'nın Conformer-Transducer X-Large (NeMo Transducer Xlarge) ve ESPnet2 gibi son teknoloji konuşma tanıma modellerini değerlendiren bir deney gerçekleştirilmiştir. Değerlendirme, rastgele seçilen 141 YouTube videosunu içermiştir. Sonuçlar, Kelime Hata Oranı (WER) metriği kullanılarak ölçülmüştür. Whisper modellerinin boyutu arttıkça performansın iyileştiği görülmüş, en küçük model olan tiny.en %43.64 WER ile en kötü performansı gösterirken, en büyük model olan large-v3 %10.58 WER ile en iyi performansı sergilemiştir. NeMo Transducer Xlarge modeli %26.06 WER ve ESPnet2 modeli %30.97 WER değerleri elde etmiştir. Bu sonuçlar, YouTube'un konuşma tanıma modellerinin değerlendirilmesi için değerli bir veri kaynağı olduğunu ve Mi-Go aracının bu modellerin sağlamlığını, doğruluğunu ve çeşitli dillere ve akustik koşullara uyarlanabilirliğini sağlamada yararlı olduğunu göstermiştir. Ayrıca, makine tarafından oluşturulan transkripsiyonları insan yapımı altyazılarla karşılaştırarak, Mi-Go aracı, YouTube altyazılarının arama motoru optimizasyonu gibi potansiyel yanlış kullanımlarını tespit etmeye de yardımcı olabilmektedir.

Alzaabi ve arkadaşları tarafından gerçekleştirilen çalışmada, engelli bireylerin ev aletlerini basit ses komutlarıyla kontrol etmelerini sağlayan düşük maliyetli, gömülü bir konuşma tanıma sistemi geliştirilmiştir [31]. Çalışmanın en dikkat çekici özelliği, sistemin hem İngilizce hem de Arapça (özellikle Emirlik lehçesi) komutları anlayabilmesidir. Araştırmacılar, 40 farklı katılımcıdan toplanan 527 İngilizce ve 680 Arapça ses örneğinden oluşan özel bir veri seti oluşturmuşlardır. Ses örnekleri, "Najma" (Arapça'da "yıldız" anlamına gelen uyandırma kelimesi), "Iftihi" (aç), "Sakri" (kapa), "Light" (ışık), "Bab" (kapı), "Bedroom" (yatak odası), "Almatbakh" (mutfak) ve arka plan gürültüsü olmak üzere sekiz sınıftan oluşmaktadır. Veri seti, çeşitli arka plan gürültüleri (akan musluk, bulaşık

yıkama sesleri, beyaz gürültü vb.) eklenerek zenginleştirilmiştir. Özellik çıkarma aşamasında, ses sinyalleri spektrogramlara dönüştürülmüş ve bu spektrogramlar, görüntü tanıma tekniklerinden yararlanılarak CNN modelleri ile işlenmiştir. Araştırmacılar, 2, 3, 4 ve 5 konvolüsyonel katmana sahip farklı CNN mimarileri tasarlamış ve bu modelleri 30, 40 ve 45 epoch süreleriyle eğitmişlerdir. Tüm modeller, %97'nin üzerinde doğruluk, kesinlik ve duyarlılık skorları elde etmiş, ancak en iyi performansı 30 epoch boyunca eğitilen 3 konvolüsyonel katmanlı model göstermiştir (%99.84 doğruluk, %99.87 kesinlik, %99.82 duyarlılık). Bu model daha sonra ESP32 mikrodenetleyicisine yerleştirilmiş ve gerçek dünya senaryolarında test edilmiştir. Sistem, kullanıcının "Najma" uyandırma kelimesini söylemesini bekleyip ardından komutları (örneğin, "Iftihi Bedroom Bab" - "Yatak Odası Kapısını Aç") algılayarak ilgili eylemi gerçekleştirmektedir. Çalışmanın sonuçları, spektrogram tabanlı görüntü tanıma tekniklerinin, düşük kaynaklı donanımlarda bile yüksek doğrulukta çalışan çok dilli konuşma tanıma sistemleri geliştirmek için etkili bir yaklaşım olduğunu göstermektedir. Geliştirilen sistem, başlangıçta engelli bireyler için tasarlanmış olsa da, yaşlılar gibi daha geniş bir kullanıcı kitlesine de hitap edebilecek potansiyele sahiptir.

Barhoush ve arkadaşları tarafından gerçekleştirilen çalışmada, konuşmacı tanıma ve lokalizasyonu için yeni bir uçtan uca model önerilmiştir [5]. Çalışmada, karmaşık modeller, hesaplamalar ve çok sayıda mikrofon dizisi kullanmadan yüksek doğrulukta konuşmacı tanıma ve lokalizasyonu yapabilen bir sistem geliştirilmesi hedeflenmiştir. Araştırmacılar, basit bir tam bağlantılı derin sinir ağı (FC-DNN) ve sadece iki mikrofon kullanarak, tek ve çoklu konuşmacı senaryolarında aktif bir konuşmacıyı yüksek doğrulukla tanımlayabilen ve konumlandırabilen yeni bir model önermişlerdir. Bu amaçla, Mel Frekans Kepstral Katsayıları (MFCC) tabanlı yeni bir özellik olan Karıştırılmış MFCC (SHMFCC) ve bunun bir varyantı olan Fark Karıştırılmış MFCC (DSHMFCC) özelliklerini geliştirmişlerdir. Önerilen yöntem, veri artırma amacıyla bir ön işleme adımı ve test aşamasında bir son işleme adımı içermektedir. Bu yaklaşımda, ses çerçeveleri bloklar halinde bölünmekte ve her blok içindeki çerçeveler rastgele karıştırılarak model eğitimi için daha fazla veri oluşturulmaktadır. Simülasyon için LibriSpeech veri setinden alınan ses örnekleri kullanılmış ve bu örnekler farklı konumlarda (37 farklı konum, 5° aralıklarla) simüle edilmiştir. Sistem, farklı gürültü seviyeleri (SNR) ve yankılanma koşulları altında test edilmiştir. Sonuçlar, önerilen yaklaşımın tek konuşmacı senaryosunda konuşmacı tanımada

%98.5, konuşmacı lokalizasyonunda %99.6 doğruluk oranına ulaştığını göstermiştir. Çoklu konuşmacı senaryosunda (2 konuşmacı) ise konuşmacı tanımada %95.2, konuşmacı lokalizasyonunda %98.7 doğruluk elde edilmiştir. Sistem, düşük gürültü seviyelerinde (0 dB SNR) bile %90'ın üzerinde doğruluk oranı sağlamıştır. Ayrıca, önerilen SHMFCC+FC-DNN yaklaşımı, SRP-PHAT ve MUSIC gibi geleneksel yöntemlerden daha iyi performans göstermiştir. Çalışmanın en önemli katkısı, az sayıda mikrofon ve küçük boyutlu eğitim verisi kullanarak bile yüksek doğrulukta konuşmacı tanıma ve lokalizasyonu yapabilen sağlam bir sistem önermesidir.

Chen tarafından gerçekleştirilen çalışmada, uç nokta algılama algoritmasına dayalı bir konuşma tanıma sistemi ve bu sistemin İngilizce kelime öğreniminde kullanımı incelenmiştir [32]. Çalışma, geleneksel kelime öğrenme yöntemlerinde öğrencilerin sadece kelimelerin anlamlarını ezberlemeye odaklanıp, kelimelerin semantik bağlamını göz ardı etmelerinden kaynaklanan sorunları ele almaktadır. Araştırmacı, konuşma tanıma teknolojisini İngilizce kelime öğrenimine entegre ederek, öğrencilerin kelimeleri pratik kullanım bağlamında öğrenmelerini sağlamayı amaçlamıştır. Çalışmada, HTK (Hidden Markov Model Toolkit) konuşma tanıma fonksiyonu temel alınarak, Markov modeli kullanılarak İngilizce kelime dağarcığı genişletilmiş ve sistemin tanıma performansı ve makine öğrenme yeteneği geliştirilmiştir. Özellik çıkarma aşamasında, enerji ve sıfır geçiş oranı (ZCR) gibi zaman alanı özellikleri, MFCC gibi frekans alanı özellikleri ve LPC tabanlı özellikler kullanılmıştır. Konuşma sinyallerinin ön işlenmesi için çerçeveleme, pencereleme ve uç nokta algılama teknikleri uygulanmıştır. Veri seti olarak, İngilizce seviye 6 testini geçmiş katılımcılardan toplanan konuşma örnekleri ve referans olarak Amerikalı bir uluslararası öğrenciden alınan ses örnekleri kullanılmıştır. Sistem, özellikle fonem bazında hata tespiti ve eksik okuma tespiti açısından değerlendirilmiştir. Sonuçlar, sesli ünsüzlerin daha kısa süreli telaffuz edilmesinden kaynaklanan hataları ortaya koymuş ve farklı fonemlere yönelik farklı hata tolerans mekanizmaları önerilmiştir. Çalışmanın en önemli katkısı, konuşma tanıma teknolojisini İngilizce kelime öğrenimine uygulayarak, öğrencilerin kelimeleri sadece ezberleme yerine, gerçek kullanım bağlamında öğrenmelerini sağlayan bir yaklaşım sunmasıdır.

Xie tarafından gerçekleştirilen çalışmada, makine öğrenme tabanlı konuşma tanıma teknolojisinin ağ tabanlı İngilizce konuşma öğretim sistemine uygulanması incelenmiştir

[33]. Çalışma, bilim ve teknolojinin hızlı gelişimiyle birlikte insanların daha akıllı ve insancıl deneyimler aradığını ve konuşma tanıma teknolojisinin bu ihtiyacı karşılamada önemli bir rol oynadığını ifade etmektedir. Araştırmacı, mevcut adaptif öğrenme sistemlerinin öğrencilerin öğrenme davranışlarına aktif olarak uyum sağlama konusundaki eksikliklerini ele almış ve makine öğrenme teknolojisini kullanarak öğrencilerin öğrenme stillerini ve davranışlarını tahmin eden bir sistem önermiştir. Çalışmada, konuşma tanıma sisteminin temel yapısı ve işleyişi detaylı olarak açıklanmıştır. Özellik çıkarma aşamasında, kısa süreli enerji ve sıfır geçiş oranı (ZCR) gibi zaman alanı özellikleri, MFCC gibi frekans alanı özellikleri ve LPC tabanlı özellikler kullanılmıştır. Konuşma sinyallerinin ön işlenmesi için çerçeveleme, pencereleme (Hamming, Hanning ve dikdörtgen pencere fonksiyonları) ve uç nokta algılama teknikleri uygulanmıştır. Sınıflandırma için Naive Bayes algoritması kullanılmış ve farklı eşik değerlerinde (30, 50, 70, 80, 90) sınıflandırma doğruluğu test edilmiştir. Sistem, 8 öğrenci üzerinde değerlendirilmiş ve öğrencilerin öğrenme tercihlerine göre içerik nesnelerinin sunumunda %59.50 ile %90.53 arasında değişen doğruluk oranları elde edilmiştir. En iyi sonuçlar, eşik değeri 90 olarak ayarlandığında elde edilmiş, Öğrenci 8 için %92.63'lük bir doğruluk oranına ulaşılmıştır. Çalışma, önerilen sistemin 8 öğrenciden 2'sinde önemli ilerleme sağladığını, diğer 3 öğrencide de küçük iyileştirmeler gösterdiğini ortaya koymuştur. Çalışmanın en önemli katkısı, konuşma tanıma teknolojisini ve makine öğrenme yöntemlerini kullanarak öğrencilerin öğrenme tercihlerine dinamik olarak uyum sağlayabilen, yüksek doğrulukta bir adaptif öğrenme sistemi sunmasıdır.

Keser tarafından gerçekleştirilen çalışmada, farklı makine öğrenme ve hibrit altuzay sınıflandırıcılarının gürültülü ortamlarda yalıtık kelime tanıma performansları karşılaştırılmıştır [15]. Çalışmada, konuşma tanıma sistemlerinde tanıma oranlarını etkileyen önemli faktörlerden biri olan çevresel arka plan gürültüsünün etkisi incelenmiştir. Araştırmacı, konuşmacıdan bağımsız yalıtık kelime tanıma işlemi için farklı gürültü türlerini içeren bir konuşma veritabanı kullanmıştır. Çalışmada, K-En Yakın Komşular (KNN), Fisher Doğrusal Diskriminant Analizi-KNN (FLDA-KNN), Ayrımcı Ortak Vektör Yaklaşımı (DCVA), Destek Vektör Makineleri (SVM), Evrişimsel Sinir Ağı (CNN) ve Tekrarlayan Sinir Ağı-Uzun Kısa Süreli Bellek (RNN-LSTM) sınıflandırıcıları kullanılmıştır. Özellik çıkarma aşamasında, Mel Frekans Kepstral Katsayıları (MFCC) ve Algısal Doğrusal Tahmin (PLP) katsayıları kullanılmıştır. Çalışmada, literatürde ilk kez DCVA sınıflandırıcısı yalıtık kelime tanıma açısından derinlemesine test edilmiştir. KNN,

FLDA-KNN ve DCVA sınıflandırıcıları için tanıma işlemi, çeşitli mesafe ölçütleri (Öklid, Korelasyon, Şehir Blok, Spearman) kullanılarak gerçekleştirilmiştir. Ayrıca, Temel Bileşen Analizi (PCA) kullanılarak yeni (DCVA)PCA ve (FLDA-KNN)PCA hibrit sınıflandırıcıları tasarlanmış ve bu hibrit sınıflandırıcılar, DCVA ve FLDA-KNN sınıflandırıcılarından daha iyi tanıma sonuçları elde etmiştir. Veri seti olarak, 18 sınıftan oluşan ve her sınıf için 1-5 arası konuşma sinyali içeren toplam 150 konuşma sinyalinden oluşan özel bir veritabanı kullanılmıştır. Her bir çerçeve için 13 veya 39 boyutlu MFCC ve PLP katsayıları çıkarılmıştır. 39 katsayı, 13 MFCC (veya PLP), 13 delta ve 13 delta-delta katsayılarının birleştirilmesiyle elde edilmiştir. Test aşamasında, tüm sınıflandırıcılar için 3 katlı çapraz doğrulama uygulanmış ve ortalama tanıma performansları elde edilmiştir. Deneysel çalışmalarda en yüksek tanıma oranı RNN-LSTM ile %93.22 olarak bulunmuştur. Diğer sınıflandırıcılar için en yüksek tanıma oranları sırasıyla CNN (%87.56), KNN (%86.51), (FLDA-KNN)PCA (%84.90), (DCVA)PCA (%79.00), SVM (%77.78), DCVA (%74.23) ve FLDA-KNN (%71.37) olarak elde edilmiştir. Çalışmanın en önemli katkısı, DCVA sınıflandırıcısının literatürde ilk kez yalıtık kelime tanıma için derinlemesine test edilmesi ve PCA kullanılarak tasarlanan hibrit sınıflandırıcıların, orijinal sınıflandırıcılardan daha iyi performans gösterdiğinin ortaya konmasıdır.

Guo tarafından gerçekleştirilen çalışmada, yapay zeka bağlamında sensör teknolojisi ve konuşma tanıma teknolojisinin üniversite İngilizce eğitiminde yenilikçi bir şekilde uygulanması incelenmiştir [1]. Çalışma, İngilizce dinleme ve konuşma eğitiminin mevcut durumunu, özellikle gürültülü ortamlarda konuşma tanıma oranının düşmesi sorununu ele almaktadır. Araştırmacı, bu sorunu çözmek için çift sensörlü bir konuşma tanıma sistemi önermektedir. Sistem, konuşma sırasında çene derisi titreşim basıncı ve hava yoluyla iletilen konuşma sinyallerini sensörler aracılığıyla toplayarak, derin makine öğrenme algoritması ile İngilizce eğitiminde konuşma tanıma gerçekleştirmektedir. Özellik çıkarma aşamasında, makul bir çerçeve örnekleme frekansı ayarlanarak İngilizce konuşma sinyali elde edilmiş, ardından doğrusal tahmin katsayıları kullanılarak bu konuşma sinyalini temsil eden özellik parametreleri elde edilmiş ve konuşma özellik vektörü oluşturulmuştur. Daha sonra, tekrarlayan sinir ağı (RNN) algoritması, özellikle Uzun Kısa Süreli Bellek (LSTM) mimarisi kullanılarak konuşma özellikleri eğitilmiştir. Çalışma, geleneksel konuşma tanıma yöntemlerinin yanı sıra Derin Sinir Ağı-Gizli Markov Modeli (DNN-HMM) gibi hibrit yaklaşımları da ele almıştır. İlgili deneylerde, önerilen algoritmanın yaygın olarak kullanılan

konuşma tanıma algoritmalarıyla karşılaştırılması yapılmış ve önerilen algoritmanın İngilizce öğretiminde konuşma tanıma konusunda daha yüksek doğruluk ve daha hızlı yakınsama sağladığı kanıtlanmıştır. Çift sensörlü sistem, özellikle gürültülü ortamlarda konuşma tanıma doğruluğunu artırmada etkili olmuştur. Çalışmanın en önemli katkıları arasında, mevcut İngilizce dinleme öğretimi durumunun analizi, gürültülü ortamlarda konuşma tanıma oranının düşmesi sorununa çözüm olarak çift sensörlü konuşma tanıma sisteminin önerilmesi ve derin makine öğrenme algoritmasının İngilizce öğretiminde konuşma tanıma için uygulanması yer almaktadır.

Jia tarafından gerçekleştirilen çalışmada, alan-spesifik konuşma tanıma için derin öğrenme tabanlı bir sistem geliştirilmiştir [19]. Çalışma, ticari ASR sistemlerinin düşük kaynak koşullarında alan-spesifik konuşmalarda genellikle zayıf performans gösterdiği problemini ele almaktadır. Araştırmacı, çalışan hakları ve sağlık sigortası alanına özgü bir ASR sistemi geliştirmek için önceden eğitilmiş DeepSpeech2 ve Wav2Vec2 akustik modellerini kullanmıştır. Alan-spesifik veriler, önerilen yarı denetimli öğrenme etiketleme yöntemi ile minimal insan müdahalesi gerektiren bir şekilde toplanmıştır. Bu yöntem, etiketli örnekler kümesini girdi olarak alıp, daha büyük etiketlenmemiş örnekler kümesi için kaba transkripsiyonlar üreten bir tanıyıcı komitesi oluşturmaktadır. Komite, etiketli ses verileri üzerinde eğitilmiş veya ince ayar yapılmış DeepSpeech2 ve Wav2Vec2 modelleri ile AWS ve Google ticari ASR sistemlerinden oluşmaktadır. Etiketlenmemiş ses verilerinden önemli ifadeleri seçmek için "göreceli hata oranı" adı verilen yeni bir metrik önerilmiştir. Veri seti olarak, laboratuvar ortamında alan uzmanları tarafından oluşturulan yapay bir veri seti ve bir çalışan hakları hizmet merkezinden toplanan gerçek dünya çağrı verileri kullanılmıştır. Tüm ses dosyaları, sessizlik bazında daha kısa segmentlere bölünmüş ve 16 kHz'e yeniden örneklenmiştir. Deneysel sonuçlar, harici bir KenLM dil modeli ile ince ayar yapılmış Wav2Vec2-Large-LV60 akustik modelinin en iyi performansı gösterdiğini ve çalışan hakları alanına özgü konuşmalarda Google ve AWS ASR sistemlerini geçtiğini ortaya koymuştur. Ayrıca, hatalı ASR transkripsiyonlarının konuşma dili anlama (SLU) sürecinin bir parçası olarak kullanılabilirliği de araştırılmıştır. Çalışan hakları alanına özgü bir doğal dil anlama (NLU) görevinin sonuçları, alan-spesifik ince ayar yapılmış ASR sisteminin, transkripsiyonlarının daha yüksek kelime hata oranına (WER) sahip olmasına rağmen ticari ASR sistemlerinden daha iyi performans gösterebildiğini ve ince ayar yapılmış ASR ile insan transkripsiyonları arasındaki sonuçların benzer olduğunu göstermiştir. Çalışmanın en

önemli katkısı, alan-spesifik konuşma tanıma için yarı denetimli öğrenme etiketleme yöntemi önermesi ve önceden eğitilmiş akustik modellerin alan-spesifik verilere ince ayar yapılarak ticari ASR sistemlerinden daha iyi performans elde edilebileceğini göstermesidir.

Rahate ve arkadaşları tarafından gerçekleştirilen çalışmada, felçli ve nöromüsküler hastalıklardan muzdarip hastaların iletişim kurabilmesi için EEG sinyalleri kullanılarak sessiz konuşma tanıma sistemi geliştirilmiştir [34]. Çalışmada, bu tür hastaların normal yollarla iletişim kuramaması problemi ele alınmış ve alternatif bir iletişim yöntemi olarak Elektroensefalografi (EEG) sinyalleri kullanılmıştır. Araştırmacılar, 10 sağlıklı denekten 6 farklı kelime ("HELP", "LIGHT", "PAIN", "STOP", "YES", "NO") için EEG verileri toplamışlardır. EEG sinyallerinin kalitesini bozan artifaktları (fizyolojik, harici, kas, göz kırpma ve iletim hattı artifaktları) belirleyip temizlemek için önişleme adımları uygulanmıştır. Önişleme aşamasında, 0.5-100 Hz arasında 5. dereceden Butterworth bant geçiren filtre kullanılarak göz hareketi ve EMG aktivitesiyle ilgili artifaktlar temizlenmiş, ayrıca 60 Hz'de çentik filtre uygulanarak hat gürültüsü giderilmiştir. Özellik çıkarma aşamasında, Ampirik Mod Ayrıştırma (EMD) yöntemi kullanılarak EEG sinyalleri çeşitli içsel mod fonksiyonlarına (IMF) ayrıştırılmış ve bu modlardan doğrusal ve doğrusal olmayan zaman alanı özellikleri çıkarılmıştır. Toplam 13 istatistiksel özellik çıkarılmış ve varyans analizi (ANOVA) testi kullanılarak yüksek ayırt edici özelliklerin ($p<0.5$) seçilmesiyle bir özellik seti oluşturulmuştur. Sınıflandırma için yedi farklı makine öğrenme algoritması kullanılmış ve 10 katlı çapraz doğrulama uygulanmıştır. Deneysel sonuçlarda, K-En Yakın Komşu (KNN) algoritması %61.45 ile en yüksek doğruluk oranını elde etmiş, bunu %61.25 ile Destek Vektör Makineleri (SVM) takip etmiştir. Performans değerlendirme metriklerinde ise, Doğrusal Diskriminant Analizi (LDA) 0.7947 kesinlik ve 0.6526 F1-skoru ile en iyi performansı gösterirken, KNN 0.6093 duyarlılık oranıyla öne çıkmıştır. Çalışmanın en önemli katkısı, sessiz konuşma tanıma için EEG sinyallerinin kullanılabilirliğini göstermesi ve felçli veya nöromüsküler hastalıklardan muzdarip hastalar için alternatif bir iletişim yöntemi sunmasıdır.

Accou ve arkadaşları tarafından gerçekleştirilen çalışmada, EEG sinyallerinden konuşma zarfını çözümlemek için VLAAI (Very Large Augmented Auditory Inference) adı verilen yeni bir derin sinir ağı mimarisi geliştirilmiştir [24]. Araştırmacılar, beynin karmaşık ve doğrusal olmayan yapısını modellemek için geleneksel doğrusal modellerin yetersiz

kaldığını ve kişiye özel eğitim verisi gerektirdiğini belirterek, kişiden bağımsız ve daha yüksek performanslı bir model geliştirmeyi amaçlamışlardır. VLAAI ağı, çoklu CNN blokları, tam bağlantılı katmanlar ve çıktı bağlam katmanından oluşan özel bir mimari olarak tasarlanmıştır. Özellik çıkarma aşamasında, konuşma sinyalinden gammatone filtre bankası kullanılarak konuşma zarfı elde edilmiştir. Çalışmada, 80 katılımcıdan oluşan ve toplam 141 saatlik EEG kaydı içeren "single-speaker stories" veri seti, 26 katılımcıdan oluşan 46 saatlik "holdout" veri seti ve 18 katılımcıdan oluşan DTU veri seti kullanılmıştır. Model performansı Pearson korelasyon katsayısı ile değerlendirilmiştir. VLAAI ağı, mevcut en iyi kişiden bağımsız doğrusal modellere göre %52'lik bir performans artışı sağlayarak 0.19 medyan Pearson korelasyon değerine ulaşmıştır. Ayrıca, kişiye özel ince ayar (fine-tuning) ile bu değer 0.25'e yükseltilmiş, kişiye özel doğrusal modellere göre %54'lük bir iyileşme elde edilmiştir. Çalışmada ayrıca, ablasyon çalışması ile modelin hangi bileşenlerinin performansa en çok katkı sağladığı incelenmiş ve doğrusal olmayan bileşenler ile çıktı bağlam modülünün en etkili olduğu (yaklaşık %10'luk göreceli performans artışı) bulunmuştur. Eğitim setindeki katılımcı sayısının model performansına etkisi de incelenmiş ve 9 katılımcıya kadar hızlı bir artış (%85) gözlenmiştir. Çalışmanın en önemli katkısı, hem kişiden bağımsız hem de kişiye özel senaryolarda konuşma zarfı çözümleme alanında yeni bir standart belirlemesi ve EEG sinyallerinden konuşma özelliklerinin çıkarılması için daha etkili bir yöntem sunmasıdır.

Wang tarafından gerçekleştirilen çalışmada, İngilizce konuşma tanıma için Tekrarlayan Sinir Ağı (RNN) ve Bağlantıcı Zamansal Sınıflandırma (CTC) algoritmasını birleştiren bir yaklaşım önerilmiştir [9]. Araştırmacı, geleneksel konuşma tanıma yöntemlerinin konuşma tanıma sürecini birkaç aşamaya böldüğünü ve farklı aşamalarda kullanılan modellerin birbirinden bağımsız olduğunu belirterek, derin öğrenme tabanlı bir yaklaşımın akustik ve dil modellerini tek bir sinir ağında birleştirebileceğini ifade etmektedir. Çalışmada, girdi konuşma dizileri ile çıktı metin dizileri arasındaki zorunlu hizalama sorununu çözmek için CTC algoritması kullanılmıştır. Özellik çıkarma aşamasında, Mel-frekans kepstral katsayıları (MFCC) yöntemi kullanılmıştır. Önerilen RNN-CTC algoritması, Gauss Karışım Modeli-Gizli Markov Modeli (GMM-HMM) ve Evrişimli Sinir Ağı-CTC (CNN-CTC) algoritmaları ile MATLAB ortamında karşılaştırılmıştır. Veri seti olarak, 630 katılımcıya ait 6.300 cümle içeren ve 16 kHz, 16 bit örnekleme parametrelerine sahip TIMIT veri seti kullanılmıştır. Performans değerlendirmesi için kelime hata oranı (WER), eğitim süresi ve

test süresi metrikleri kullanılmıştır. Deneysel sonuçlar, test örneği sayısı arttıkça tüm algoritmaların WER değerlerinin arttığını göstermiştir. 2.500 test örneği kullanıldığında, RNN-CTC modeli %10.4 WER ile en iyi performansı gösterirken, CNN-CTC %15.8 ve GMM-HMM %21.1 WER değerlerine ulaşmıştır. Test süreleri ise sırasıyla 20.2, 33.1 ve 42.5 saniye olarak ölçülmüştür. Eğitim süreleri açısından da RNN-CTC 351.4 saniye ile en hızlı, CNN-CTC 410.7 saniye ile ikinci, GMM-HMM ise 498.7 saniye ile en yavaş model olmuştur. Çalışmanın en önemli katkısı, konuşma dizileri ile metin dizileri arasındaki hizalama sorununu CTC algoritması ile çözen ve konuşmanın zamansal özelliklerini etkili bir şekilde kullanan RNN tabanlı bir İngilizce konuşma tanıma yaklaşımı sunarak, hem doğruluk hem de hesaplama verimliliği açısından üstün performans elde etmesidir.

Hamian ve arkadaşları tarafından gerçekleştirilen çalışmada, otomatik rakam ses tanıma için çok amaçlı optimizasyon algoritmalarını kullanan yeni bir derin darbeli sinir ağları (SNN) öğrenme yaklaşımı önerilmiştir [35]. Araştırmacılar, geleneksel yapay sinir ağlarına (ANN) kıyasla biyolojik sinir ağlarına yapısal olarak daha benzer olan ve daha az hesaplama kaynağı gerektiren SNN'lerin potansiyelini belirtmişlerdir. Çalışmada, derin SNN katmanlarının öğrenme problemi ele alınmış ve Gradient-Based Optimization (GBO) ile Wild Horse Optimization (WHO) algoritmaları kullanılarak darbeli nöronların iki ana parametresi (eşik voltajı ve giriş ağırlıkları) hesaplanmıştır. Özellik çıkarma aşamasında, Mel-frekans kepstral katsayıları (MFCC) ve sıfır geçiş oranı (ZCR) gibi özellikler kullanılmıştır. Ses sinyalinden toplam 77 özellik çıkarılmış, ancak hesaplama yükünü azaltmak için genetik algoritma yardımıyla en uygun 7 özellik seçilmiştir. Veri seti olarak, farklı erkek ve kadın konuşmacılardan alınan 0-9 rakamlarını içeren 200 ses örneğinden oluşan TIDIGITS kullanılmıştır. Ayrıca, karşılaştırma amacıyla IRIS ve Trip veri setleri de kullanılmıştır. Önerilen SNN modeli, yapay sinir ağları (ANN) ve uyarlanabilir ağ tabanlı bulanık çıkarım sistemi (ANFIS) gibi diğer derin öğrenme yöntemleriyle farklı senaryolarda karşılaştırılmıştır. Değerlendirme için doğruluk ve RMS hata değeri metrikleri kullanılmıştır. Sonuçlar, rakam ses tanıma görevinde SNN-WHO modelinin %98.2 doğruluk ve 0.0182 RMS hata ile en iyi performansı gösterdiğini, bunu %96.4 doğruluk ve 0.0364 RMS hata ile SNN-GBO modelinin izlediğini ortaya koymuştur. Geleneksel ANN modeli %93.6 doğruluk ve 0.0636 RMS hata, ANFIS modeli ise %90.9 doğruluk ve 0.0909 RMS hata değerlerine ulaşmıştır. Benzer şekilde, IRIS veri seti üzerinde SNN-WHO %97.3, SNN-GBO %95.3, ANN %94.7 ve ANFIS %93.3 doğruluk oranları elde etmiştir. Trip veri setinde

ise SNN-WHO %97.5, SNN-GBO %95.8, ANN %93.3 ve ANFIS %91.7 doğruluk oranları sağlamıştır. Çalışmanın en önemli katkısı, derin darbeli sinir ağlarının sınıflandırma ve öğrenme doğruluğunu artırmak için yeni bir eğitim tekniği sunması ve optimize edilmiş SNN'lerin geleneksel yapay sinir ağlarına göre daha verimli çalışabileceğini göstermesidir.

Ameen ve Kadhim tarafından gerçekleştirilen çalışmada, elektrolainks cihazı kullanan konuşma ile Arapça fonemlerin tanınması için farklı makine öğrenme modellerinin karşılaştırılması amaçlanmıştır [36]. Elektrolainks, vokal laringeal kordları çıkarılan kanser hastaları tarafından kullanılan bir cihazdır ve bu cihazla üretilen konuşma genellikle gürültülü ve anlaşılması zordur. Araştırmacılar, Arapça fonemlerin tanınmasında en etkili makine öğrenme modelini belirlemek için yapay sinir ağları (ANN), evrişimli sinir ağları (CNN), tekrarlayan sinir ağları (RNN), uzun-kısa süreli bellek (LSTM), rastgele orman (RF) ve aşırı gradyan artırma (XGBoost) gibi çeşitli makine öğrenme yaklaşımlarını karşılaştırmışlardır. Ayrıca, CNN-LSTM, CNN-RF ve CNN-XGB hibrit modelleri de değerlendirilmiştir. Özellik çıkarma aşamasında, her fonem için 40 Mel-frekans kepstral katsayısı (MFCC) özelliği çıkarılmıştır. Veri seti olarak, 8 kişiden (3 erkek, 5 kadın) kaydedilen 27 Arapça fonem sınıfını içeren toplam 1.709 örnek kullanılmıştır. Her fonem için 63 örnek bulunmaktadır ve kayıtlar hem normal konuşma hem de elektrolainks cihazı ile üretilen konuşma olmak üzere iki durumu kapsamaktadır. Ses kayıtları 48 kHz örnekleme frekansında, bir saniyelik sürelerle kaydedilmiştir. Veri seti, %70 eğitim (1.196 örnek) ve %30 test (513 örnek) olarak rastgele bölünmüştür. Modellerin performansı, doğruluk, kesinlik ve fonem hata oranı (PER) metrikleri ile değerlendirilmiştir. Sonuçlar, ANN modelinin %75 doğruluk, %77 kesinlik ve %21.85 PER ile diğer modellere göre daha iyi performans gösterdiğini ortaya koymuştur. Bunu %72 doğruluk, %74 kesinlik ve %25.93 PER ile CNN-LSTM hibrit modeli ve %71 doğruluk, %73 kesinlik ve %27.06 PER ile CNN-XGB modeli izlemiştir. Diğer modellerin performansları ise CNN (%70 doğruluk, %72 kesinlik, %26.14 PER), RNN (%68 doğruluk, %70 kesinlik, %31.42 PER), LSTM (%69 doğruluk, %71 kesinlik, %28.75 PER), RF (%65 doğruluk, %67 kesinlik, %33.79 PER) ve XGBoost (%67 doğruluk, %69 kesinlik, %32.61 PER) şeklinde sıralanmıştır. Çalışmanın en önemli katkısı, elektrolainks cihazı kullanan kanser hastaları için Arapça fonem tanıma sisteminde en etkili makine öğrenme modelini belirlemesi ve ANN'nin gürültülü elektrolainks konuşması için en uygun model olduğunu göstermesidir.

Shisode ve arkadaşları tarafından gerçekleştirilen çalışmada, konuşma engelli kişiler için yüzey elektromiyografi (SEMG) yaklaşımı kullanılarak bir konuşma tanıma sistemi geliştirilmiştir [37]. Araştırmacılar, konuşma engelli kullanıcıların iletişim kurabilmesini sağlamak amacıyla yüz kaslarından alınan kas aktivitesi sinyallerini kullanarak, söylenen veya fısıldanan kelimeleri belirli bir doğrulukla tahmin etmeyi amaçlamışlardır. Çalışmada, Gaussian Mixture Model (GMM) ve Convolutional Neural Network (CNN) gibi makine öğrenme modelleri kullanılmıştır. Özellik çıkarma aşamasında, konuşma artikülasyonu kaslarından yüzey elektromiyografi (SEMG) sensörleri ile nöromüsküler sinyaller kaydedilmiştir. Ayrıca, doğruluğu artırmak için görsel tabanlı bir dudak okuma sistemi de geliştirilmiştir. Bu görsel sistem, kullanıcının dudak hareketlerini bir kamera yardımıyla kaydederek, kullanıcının söylemeyi amaçladığı olası kelimelerin bir listesini çıkarmaktadır. EMG tabanlı sistem ile görsel sistemin eşleştirilmesi, doğru kelimenin tahmin edilmesini mümkün kılmaktadır. Veri seti olarak, yüz kaslarından alınan EMG sinyalleri ve dudak hareketleri videoları kullanılmıştır. Görsel sistem için MIRACL-VC1 dudak okuma veri seti kullanılmıştır. Sistemin performansı, doğruluk metriği ile değerlendirilmiştir. Sonuçlar, EMG sistemi için "Emergency" kelimesinin %80, "Heart" kelimesinin %75, "Water" kelimesinin %60 ve "Better" kelimesinin %65 doğrulukla tanındığını göstermiştir. Dudak okuma sistemi ise eğitim verisi üzerinde %90, test verisi üzerinde %53 doğruluk oranına ulaşmıştır. Çalışmanın en önemli katkısı, özellikle larenjektomi geçiren ve ses kutusu çıkarılan hastaların, dudaklarını ve diğer subvokal kaslarını hareket ettirebildikleri halde ses üretemedikleri durumlarda kullanılabilecek bir sistem sunmasıdır.

Ahmed ve arkadaşları tarafından gerçekleştirilen çalışmada, Peştuca dili için verimli özellik çıkarma ve sınıflandırma yöntemlerine dayalı yeni bir konuşma tanıma sistemi çerçevesi önerilmiştir [10]. Araştırmacılar, Peştuca'nın az kaynaklı bir dil olduğunu ve yerel lehçelerinin bulunması nedeniyle Otomatik Konuşma Tanıma hatalarına daha yatkın olduğunu belirtmişlerdir. Çalışmada, klasik Makine Öğrenme modellerinin konuşma tanıma görevleri için kullanılmasının yanı sıra, optimal özellik çıkarma teknikleri kullanılarak daha yüksek performans doğruluğu elde edilmesi amaçlanmıştır. Özellik çıkarma aşamasında, iki iyi bilinen teknik olan Ayrık Dalgacık Dönüşümü (DWT) katsayıları ve Mel-Frekans Kepstral Katsayıları (MFCC) kullanılmıştır. Araştırmacılar, hesaplama karmaşıklığı, enerji verimliliği ve diğer ML ve Derin Öğrenme (DL) modellerine kıyasla daha yüksek doğruluk açısından verimliliklerinden dolayı klasik ML modelleri olan Destek Vektör Makinesi

(SVM) ve K-En Yakın Komşu (k-NN) algoritmalarını tercih etmişlerdir. Veri seti olarak, 0-25 arası rakamlar ve sık kullanılan Peştuca kelimeleri içeren 161 izole kelimeden oluşan bir Peştuca veri seti kullanılmıştır. Kelimeler, farklı yaş ve cinsiyetlerden (17 erkek, 13 kadın) hem ana dili Peştuca olan hem de olmayan 30 konuşmacı tarafından seslendirilmiştir. Deneyler, 26 izole Peştuca kelime ve 0-25 arası rakamlar üzerinde gerçekleştirilmiştir. Her kelime için 17 ifade kullanılmış, bunlardan 12'si eğitim ve 5'i test için ayrılmıştır. Toplam olarak 312 eğitim ve 130 test örneği kullanılmıştır. Sistemin performansı, doğruluk metriği ile değerlendirilmiştir. Sonuçlar, MFCC özellik çıkarma yöntemi ile SVM sınıflandırıcısının kombinasyonunun %96.15 doğruluk oranıyla en iyi performansı gösterdiğini, bunu %92.31 ile MFCC+KNN, %88.46 ile DWT+SVM ve %84.62 ile DWT+KNN kombinasyonlarının izlediğini ortaya koymuştur. Çalışmanın en önemli katkısı, Peştuca dili için esnek bir ASR çerçevesi geliştirmesi ve özellik çıkarma tekniklerinin performans üzerindeki etkisini inceleyerek, MFCC ve SVM kombinasyonunun Peştuca izole kelime tanıma için en etkili yaklaşım olduğunu göstermesidir.

Froiz-Míguez ve arkadaşları tarafından gerçekleştirilen çalışmada, Galiçya dili için diller arası makine öğrenmesi tabanlı otomatik konuşma tanıma sistemi tasarlanmış ve değerlendirilmiştir [38]. Araştırmacılar, son yıllarda makine öğrenmesi stratejilerinin donanımlardaki hesaplama gücünün artması sayesinde çeşitli alanlarda sınıflandırma ve model algılama görevlerini otomatikleştirmede faydalı olduğunu belirtmişlerdir. Bu alanlardan biri olan Otomatik Konuşma Tanıma, insan konuşmasını okunabilir metne dönüştürmek için makine öğrenmesi mimarilerini kullanabilmektedir. Çalışmada, geleneksel mimarilerin eğitimde düşük Kelime Hata Oranı (WER) elde etmek için yüksek miktarda transkripsiyonlu veriye ihtiyaç duyması sorununa odaklanılmıştır. Bu tür verilerin insan işlemesine yüksek bağımlılık nedeniyle elde edilmesi özellikle zordur. 2020 yılında önerilen yeni bir çerçeve olan wav2vec 2.0, öz denetimli eğitim kullanarak çok dilli etiketlenmemiş sesler üzerinde eğitim yapan ve ardından belirli bir dilin etiketli sesleriyle ince ayar yapabilen bir Evrişimli Sinir Ağı (CNN) kullanarak ses etiketleme ihtiyacını önemli ölçüde azaltmaktadır. Araştırmacılar, küçük ses veri setleri bulunan Galiçya dili için wav2vec 2.0 tabanlı bir ASR sistemi geliştirmişlerdir. Sistem, OpenSLR ve Common Voice'dan alınan toplam yaklaşık 20 saatlik etiketli ses verisiyle ince ayar yapılmıştır. Eğitim, 32 GB RAM ve 80 GB GPU belleğine sahip bir Nvidia Titan A100 kullanan bir sunucuda yaklaşık 34 saat sürmüştür. Veri seti, %80 eğitim ve %20 değerlendirme olarak

bölünmüş, final eğitim kaybı %0.39 ve değerlendirme kaybı %12.4 olmuştur. Modele, 100 genişliğinde bir ışın dekoder uygulanmış ve 3-gram tabanlı 51.1 milyon kelimelik bir Galiçya dili modeli kullanılmıştır. Sistem, Galiçya Parlamentosu'ndan yaklaşık bir saatlik spontane konuşma veri setiyle değerlendirilmiş ve nispeten düşük bir WER (%18.61) göstermiştir. Sadece Wav2Vec2 modeli kullanıldığında WER %28.67, ışın dekoder eklendiğinde %27.42 ve hem ışın dekoder hem de dil modeli kullanıldığında %18.61 olarak ölçülmüştür. Çalışmanın en önemli katkısı, az kaynaklı dillerde diller arası transfer öğrenimi yaklaşımının etkinliğini göstermesi ve özellikle Wav2Vec 2.0 modelinin Galiçya dili gibi az kaynaklı dillerde bile kabul edilebilir performans sergileyebileceğini ortaya koymasıdır.

Kwon tarafından gerçekleştirilen çalışmada, ses tanıma sistemlerini optimize edilmiş karşıt örneklere (adversarial examples) karşı korumak için "AudioGuard" adı verilen yeni bir savunma yöntemi önerilmiştir [29]. Araştırmacı, derin sinir ağlarının görüntü tanıma, ses tanıma, desen tanıma ve saldırı tespiti gibi alanlarda iyi performans gösterdiğini, ancak karşıt örneklere karşı savunmasız olduğunu belirtmiştir. Karşıt örnekler, normal verilere az miktarda gürültü eklenerek oluşturulan, insanlar tarafından normal olarak algılanan ancak hedef model tarafından yanlış sınıflandırılan örneklerdir. Çalışmada, ayrı bir modül veya işlem gerektirmeden, gürültü vektörü kullanarak karşıt ses örneklerine karşı savunma sağlayan bir yöntem önerilmiştir. Önerilen yöntem, gürültü vektörü kullanarak modelin normal örnekler üzerindeki doğruluğunu korurken karşıt örnekleri doğru bir şekilde tespit etmektedir. Yöntem, giriş verisi için aynı sınıf tahmini korurken, modelin sınıflandırma güven değerlerine gürültü değerleri ekleyerek optimize edilmiş karşıt örneklerin oluşturulmasını engeller. Deneyler için Mozilla Common Voice veri seti test verisi olarak kullanılmış ve makine öğrenme kütüphanesi olarak TensorFlow tercih edilmiştir. Ses tanıma sistemi olarak DeepSpeech modeli kullanılmıştır. DeepSpeech, Bağlantıcı Zamansal Sınıflandırma (CTC) yöntemini kullanarak etiketlenmemiş dizi etiketlerini doğrudan bir tekrarlayan sinir ağı (RNN) aracılığıyla eğiten bir modeldir. Karşıt örnekler, Carlini yöntemi kullanılarak oluşturulmuştur. Bu yöntem, normal ses örneğine az miktarda gürültü ekleyerek, model tarafından hedef ifade olarak yanlış sınıflandırılan karşıt örnekler oluşturur. Deneysel sonuçlar, önerilen yöntemin karşıt örnekleri %84.2 doğrulukla doğru şekilde tespit ettiğini göstermiştir. Ayrıca, normal örnekler üzerindeki model doğruluğu %94.3 olarak korunmuştur. Çalışmanın en önemli katkısı, ayrı bir modül gerektirmeden,

sağlam bir model oluşturarak optimize edilmiş karşıt örneklere karşı savunma yöntemi önermesidir.

Rai ve arkadaşları tarafından gerçekleştirilen çalışmada, konuşma içerisindeki komutları tespit etmek için derin öğrenme tabanlı anahtar kelime tanıma (keyword spotting) sistemleri geliştirilmiştir [18]. Araştırmacılar, anahtar kelime tanımanın, robotik, sanal asistanlar ve araç üstü elektronik gibi geniş uygulama alanlarına sahip olduğunu belirtmişlerdir. Çalışmada, ham ses dalgalarını Mel Frekans Kepstral Katsayılarına (MFCC) dönüştürerek özellik mühendisliği uygulanmış ve bu özellikler model girişleri olarak kullanılmıştır. CNN modeli için ise, 16000 Hz yerine 8000 Hz ile yeniden örneklenmiş ham ses dalgaları kullanılmıştır. Araştırmacılar, temel model olarak Gauss Karışım Modelli Gizli Markov Modeli (HMM-GMM) kullanmış ve daha sonra Evrişimli Sinir Ağları (CNN) ve Uzun Kısa Süreli Bellek (LSTM) ile Dikkat (Attention) mekanizması içeren Tekrarlayan Sinir Ağları (RNN) varyantları gibi çeşitli algoritmalar ile deneyler gerçekleştirmişlerdir. HMM-GMM modeli için, altı gizli durum ve iki GMM karışımı en iyi doğrulama seti sonucunu verdiği için tercih edilmiştir. CNN için M5 CNN yapısı kullanılmış ve öğrenme oranı 0.01, optimize edici olarak Adam ve kayıp fonksiyonu olarak çapraz entropi kullanılmıştır. Model on epoch boyunca eğitilmiştir. RNN modelleri için, LSTM, Çift Yönlü LSTM (BiLSTM) ve Dikkat mekanizması eklenmiş varyantlar denenmiştir. Veri seti olarak, 35 kelimelik 105,829 adet bir saniyelik konuşma içeren Google Speech Commands Dataset v2 kullanılmıştır. Bu veri seti, "evet", "hayır", "açık", "kapalı" gibi yaygın komutlar, "sol", "sağ", "yukarı", "aşağı" gibi yönler, "sıfır"dan "dokuz"a kadar sayılar ve "yatak", "kedi" gibi yardımcı kelimeler içermektedir. Veriler %80 eğitim, %10 doğrulama ve %10 test olarak bölünmüştür. Değerlendirme metriği olarak doğruluk (accuracy) kullanılmıştır. Sonuçlar, RNN-BiLSTM-Attention modelinin %93.9 doğruluk oranıyla en iyi performansı gösterdiğini, bunu %92.5 ile RNN-BiLSTM, %91.2 ile RNN-LSTM, %89.7 ile CNN (M5) ve %65.2 ile HMM-GMM modellerinin izlediğini ortaya koymuştur. Çalışmanın en önemli katkısı, konuşma tanıma için farklı algoritmaların karşılaştırmalı analizini sunması ve özellikle Dikkat mekanizması eklenmiş Çift Yönlü LSTM modelinin anahtar kelime tanıma görevinde üstün performans sergileyebileceğini göstermesidir. Ayrıca, derin sinir ağlarının geleneksel zaman serisi analiz modellerinden önemli ölçüde daha iyi performans gösterdiği de çalışmanın önemli bulgularından biridir.

Kumar ve arkadaşları tarafından gerçekleştirilen çalışmada, derin öğrenme tekniklerini kullanarak sağlam bir görsel konuşma tanıma (VSR) modeli geliştirilmiştir [39]. Çalışma, özellikle gürültülü ortamlarda ses sinyalinin olmadığı durumlarda konuşma tanıma problemine odaklanmaktadır. Araştırmacılar, görsel konuşma tanıma için dikkat mekanizması tabanlı kodlayıcı-kod çözücü mimarisi önermişlerdir. Önerilen model, dudak hareketlerini analiz ederek konuşulan metni tanımlamaktadır. Özellik çıkarma aşamasında, 68-şekil kestiricisi yöntemi kullanılarak dudak bölgesi tespit edilmiş ve bu görsel özellikler kodlayıcı-kod çözücü mimarisine girdi olarak verilmiştir. Sistemin performansını artırmak için BERT tabanlı dil modeli entegre edilmiştir. Veri seti olarak GRID Corpus ve LRS2 Corpus kullanılmıştır. Sistemin performansı, Kelime Hata Oranı (WER) metriği ile değerlendirilmiştir. Sonuçlar, önerilen modelin GRID Corpus üzerinde %2.8 WER ve LRS2 Corpus üzerinde %40.1 WER değerleri ile mevcut temel sistemlerden daha iyi performans gösterdiğini ortaya koymuştur. Çalışmanın en önemli katkısı, gürültülü ortamlarda veya ses sinyalinin olmadığı durumlarda konuşma tanıma için etkili bir görsel tabanlı çözüm sunması ve dikkat mekanizması tabanlı mimarinin dudak okuma görevinde üstün performans sergileyebileceğini göstermesidir.

Weng ve arkadaşları tarafından gerçekleştirilen çalışmada, konuşma tanıma ve sentezi görevleri için derin öğrenme tabanlı bir semantik iletişim sistemi (DeepSC-ST) geliştirilmiştir [40]. Çalışmada, iletişim sistemlerinin verimliliğini artırmak amacıyla semantik bilgilerin iletilmesi yaklaşımı benimsenmiştir. Araştırmacılar, konuşma sinyallerinden metin ile ilgili semantik özellikleri çıkaran ve bu özellikleri iletişim kanalları üzerinden aktaran bir ortak semantik-kanal kodlama şeması tasarlamışlardır. Önerilen sistemde, girdi konuşma sinyalleri önce Hamming penceresi ve Hızlı Fourier Dönüşümü (FFT) kullanılarak spektrograma dönüştürülmekte, ardından CNN ve RNN tabanlı semantik verici ile metin ile ilgili semantik özellikler çıkarılmaktadır. Bu yaklaşım, konuşmacının sesi, konuşma gecikmesi ve arka plan gürültüsü gibi özelliklerin iletilmesini gerektirmediğinden, iletilmesi gereken veri miktarını önemli ölçüde azaltmaktadır. Alıcı tarafta, alınan semantik özellikler bir özellik çözücüsü tarafından metin bilgisine dönüştürülmekte ve kullanıcı kimliğine göre Tacotron 2 modeli kullanılarak konuşma sinyalleri yeniden sentezlenmektedir. Sistem, Connectionist Temporal Classification (CTC) kaybı kullanılarak eğitilmiştir. Performans değerlendirmesinde, DeepSC-ST'nin düşük sinyal-gürültü oranı (SNR) koşullarında üstün performans gösterdiği görülmüştür. 0 dB SNR

değerinde, DeepSC-ST yaklaşık %10 CER ve %22 WER değerleri elde ederken, karşılaştırılan DeepSC-S sistemi %38 CER ve %65 WER, geleneksel sistem ise %85 CER ve %95 WER değerlerine ulaşmıştır. 8 dB SNR değerinde ise DeepSC-ST'nin performansı daha da iyileşerek yaklaşık %3 CER ve %7 WER değerlerine ulaşmıştır. Konuşma sentezi kalitesi açısından, DeepSC-ST'nin Fréchet Deep Speech Distance (FDSD) değeri yaklaşık 0.3, Kernel Deep Speech Distance (KDSD) değeri ise yaklaşık 0.05 olarak ölçülmüş, bu değerler doğrudan Tacotron 2 modelinin değerlerine (FDSD: ~0.25, KDSD: ~0.04) yakın sonuçlar vermiştir. Çalışmanın en önemli katkısı, konuşma tanıma ve sentezi görevlerini birleştiren, metin ile ilgili semantik özellikleri kullanarak iletişim verimliliğini artıran ve dinamik kanal ortamlarına uyum sağlayabilen sağlam bir semantik iletişim sistemi sunmasıdır.

Chen ve arkadaşları tarafından gerçekleştirilen çalışmada, konuşma tanıma ve çevirisi için bağlam içi öğrenme yeteneklerine sahip yeni bir Konuşma Destekli Dil Modeli (SALM) geliştirilmiştir [41]. Araştırmacılar, dondurulmuş bir metin tabanlı büyük dil modeli (LLM), bir ses kodlayıcı, bir modalite adaptör modülü ve konuşma girişini ve ilgili görev talimatlarını desteklemek için LoRA katmanlarından oluşan birleşik bir model önermişlerdir. Önerilen SALM, yalnızca Otomatik Konuşma Tanıma (ASR) ve Konuşma Çevirisi (AST) için görev odaklı Conformer temel modellerine eşdeğer performans göstermekle kalmayıp, aynı zamanda ASR ve AST için anahtar kelime güçlendirme görevi aracılığıyla gösterilen sıfır atımlı bağlam içi öğrenme yetenekleri de sergilemektedir. Ayrıca, LLM eğitimi ile alt seviye konuşma görevleri arasındaki boşluğu kapatmak için konuşma denetimli bağlam içi eğitim önerilmiş, bu da konuşmadan metne modellerin bağlam içi öğrenme yeteneğini daha da artırmıştır. Model mimarisi olarak, 110M parametreli önceden eğitilmiş bir Fast Conformer ses kodlayıcısı ve metin talimatlarıyla ince ayar yapılmış önceden eğitilmiş 2B parametreli bir Megatron LLM kullanılmıştır. Modalite adaptörü olarak 4X alt örnekleme ile iki Conformer katmanı kullanılarak, metin ve konuşma arasındaki farklı bilgi oranı ve modelleme alanı eşleştirilmiştir. Veri seti olarak, ASR için LibriSpeech eğitim seti ve AST için IWSLT 2023 Çevrimdışı İzleme'de bulunan tüm İngilizce ses verileri (2.7M segment, 4.8K saat) kullanılmıştır. Değerlendirme, LibriSpeech test-clean ve test-other setleri ile MuST-C v2 tst-COMMON üzerinde yapılmıştır. Sonuçlar, önerilen modelin ASR görevinde test-clean için %2.3 ve test-other için %4.8 WER değerlerine, AST görevinde ise İngilizce-Almanca için 29.6 ve İngilizce-Japonca için 16.5

BLEU puanlarına ulaştığını göstermiştir. Çok görevli SALM ise ASR WER değerlerinde test-clean için %2.6 ve test-other için %6.1, AST BLEU puanlarında İngilizce-Almanca için 30.7 ve İngilizce-Japonca için 16.8 sonuçları elde etmiştir. Çalışmanın en önemli katkısı, konuşma tanıma ve çevirisi için bağlam içi öğrenme yeteneklerine sahip birleşik bir model sunması ve konuşmadan metne modelleri ilk kez sıfır atımlı bağlam içi öğrenme yeteneği ile donatmasıdır.

Gong ve arkadaşları tarafından gerçekleştirilen çalışmada, hem konuşma hem de konuşma dışı sesleri eş zamanlı olarak tanıyabilen ve aralarındaki ilişkileri anlayabilen yeni bir makine öğrenme modeli olan LTU-AS (Listen To, Think of, and Understand Audio and Speech) geliştirilmiştir [42]. Araştırmacılar, insanların hem konuşma hem de konuşma dışı sesleri içeren çeşitli ses sinyalleriyle çevrili olduğunu ve bu sesleri tanıma, anlama ve aralarındaki ilişkileri kavrama yeteneğinin temel bilişsel yeteneklerden biri olduğunu belirtmektedir. Bu motivasyonla, Whisper'ı algılama modülü ve LLaMA'yı akıl yürütme modülü olarak entegre ederek, konuşulan metni, konuşma paralinguistik özelliklerini ve konuşma dışı ses olaylarını -ses sinyallerinden algılanabilecek neredeyse her şeyi- eş zamanlı olarak tanıyabilen ve birlikte anlayabilen bir model geliştirmişlerdir. Model mimarisi olarak, önceden eğitilmiş Whisper-large ses kodlayıcı-kod çözücü (32 katman, 1280 boyutlu Transformer ağları), AudioSet üzerinde önceden eğitilmiş bir Time and Layer-Wise Transformer (TLTR) ve LLaMA-7B büyük dil modeli kullanılmıştır. Ses, önce Whisper kodlayıcısına verilmekte, ardından Whisper kodlayıcısının çıktısı, konuşulan metni çözümlemek için Whisper kod çözücüsüne beslenmektedir. Aynı zamanda, Whisper kodlayıcısının tüm 32 ara katmanının çıktısı, "yumuşak" ses olaylarını ve konuşma paralinguistik bilgilerini kodlamak için TLTR'ye beslenmekte ve ardından bir doğrusal katmanla sürekli ses jetonlarına dönüştürülmektedir. Eğitim sırasında, tüm Whisper modeli dondurulmakta, yalnızca TLTR modeli ve projeksiyon katmanı eğitilebilir durumdadır. Veri seti olarak, 13 farklı ses ve konuşma veri setinden oluşan 9.6 milyon Open-ASQA veri seti kullanılmış, bunun 6.9 milyonu GPT tarafından oluşturulan açık uçlu soru-cevap çiftleridir. LTU-AS, toplam 8.5 milyar parametreye sahip olmasına rağmen, sadece 49 milyon parametre (TLTR için 40M, LoRA adaptörleri için 4.2M ve projeksiyon katmanı için 5M) eğitilebilir durumdadır. Model, 4 adet A6000 GPU üzerinde yaklaşık 80 saat eğitilmiştir. Değerlendirme sonuçları, LTU-AS'nin tüm ses/konuşma görevlerinde güçlü olduğunu ve daha da önemlisi, ses ve konuşma hakkında serbest biçimli açık uçlu soruları %95'in

üzerinde bir talimat takip oranıyla (GPT-4 tarafından değerlendirilmiş) yanıtlayabildiğini ve ortaya çıkan ortak ses ve konuşma akıl yürütme yeteneği sergilediğini göstermiştir. Çalışmanın en önemli katkısı, sadece konuşma veya sadece ses olaylarını tanıyabilen mevcut makine öğrenme modellerinin ötesine geçerek, hem konuşma hem de konuşma dışı sesleri eş zamanlı olarak tanıyabilen ve aralarındaki karmaşık ilişkileri anlayabilen ilk modeli sunmasıdır.

Wang ve arkadaşları tarafından gerçekleştirilen çalışmada, konuşma tanıma, sentezi ve çevirisi gibi çeşitli konuşma işleme görevlerini tek bir model altında birleştiren VIOLA adlı yenilikçi bir yaklaşım sunulmuştur [43]. Araştırmacılar, son yıllarda farklı modaliteler için çeşitli görevlerde model mimarisi, eğitim hedefleri ve çıkarım yöntemlerinde büyük bir yakınsama olduğunu gözlemlemişlerdir. Bu motivasyonla, konuşma-metin, metin-metin, metin-konuşma ve konuşma-konuşma gibi çeşitli çapraz modal görevleri, çok görevli öğrenme çerçevesi aracılığıyla koşullu codec dil modeli görevine dönüştüren tek bir otomatik regresif Transformer decoder-only ağı geliştirmişlerdir. Bu amaca ulaşmak için, tüm konuşma ifadelerini çevrimdışı bir sinir codec kodlayıcısı (EnCodec) kullanarak ayrık belirteçlere (metin verilerine benzer şekilde) dönüştürmüşlerdir. Bu şekilde, tüm bu görevler, tek bir koşullu dil modeli ile doğal olarak ele alınabilen belirteç tabanlı dizi dönüşüm problemlerine dönüştürülmüştür. Araştırmacılar ayrıca, farklı dilleri ve görevleri ele alma yeteneğini geliştirmek için önerilen modele görev kimlikleri (TID) ve dil kimlikleri (LID) entegre etmişlerdir. VIOLA modeli, otomatik konuşma tanıma (ASR), makine çevirisi (MT) ve metin-konuşma sentezi (TTS) görevlerini içeren çok görevli bir öğrenme çerçevesi altında eğitilmiştir. Modelin temel bileşenleri, otomatik regresif codec dil modeli ve otomatik regresif olmayan codec dil modelidir. İlki, semantik belirteçlere (fonem dizileri) dayalı olarak her çerçeve için ilk codec kodunun akustik belirteçlerini otomatik regresif bir şekilde tahmin etmekten sorumludur. İkincisi ise, ilk katman kodlarının dizisine göre diğer 7 katman kodunu katman düzeyinde yinelemeli üretim yöntemiyle paralel olarak üretmek için kullanılır. Değerlendirme sonuçları, önerilen VIOLA modelinin İngilizce ASR görevinde %9.4 WER, Çince ASR görevinde %10.2 WER elde ettiğini göstermiştir. Makine çevirisi görevlerinde ise İngilizce-Çince çeviride 28.7 BLEU ve Çince-İngilizce çeviride 24.3 BLEU puanı elde edilmiştir. Konuşma-konuşma çevirisi (S2ST) görevinde, İngilizce-Çince çeviride 26.5 BLEU ve Çince-İngilizce çeviride 22.1 BLEU puanı elde edilmiştir. TTS görevlerinde ise 0.85 konuşmacı benzerliği ve 5 üzerinden 4.1 ortalama görüş puanı (MOS)

elde edilmiştir. Çalışmanın en önemli katkısı, farklı konuşma işleme görevleri için genellikle farklı mimarilere sahip farklı modeller kullanılan geleneksel yaklaşımların aksine, tüm konuşma işleme görevlerini tek bir decoder-only modelde birleştirerek, codec tabanlı otomatik regresif Transformer decoder modellerini çeşitli konuşma işleme görevleri için nicel olarak araştıran ilk çalışma olmasıdır.

Zhang ve arkadaşları tarafından gerçekleştirilen çalışmada, 100'den fazla dilde otomatik konuşma tanıma (ASR) yapabilen Evrensel Konuşma Modeli (Universal Speech Model - USM) adlı tek bir büyük model tanıtılmıştır [20]. Araştırmacılar, modelin kodlayıcısını 300'den fazla dili kapsayan 12 milyon saatlik büyük bir etiketlenmemiş çok dilli veri seti üzerinde ön-eğitime tabi tutmuş ve daha sonra daha küçük bir etiketli veri seti üzerinde ince ayar yapmışlardır. Çalışmada, rastgele projeksiyon kuantalama ve konuşma-metin modalite eşleştirmesi ile çok dilli ön-eğitim kullanılarak, alt seviye çok dilli ASR ve konuşma-metin çevirisi görevlerinde en son teknoloji performansına ulaşılmıştır. Araştırmacılar, Whisper modelinin kullandığı etiketli eğitim setinin 1/7'si büyüklüğünde bir eğitim seti kullanmalarına rağmen, modellerinin birçok dilde hem alan içi hem de alan dışı konuşma tanıma görevlerinde karşılaştırılabilir veya daha iyi performans sergilediğini göstermişlerdir. USM modelleri, üç tür veri seti kullanılarak oluşturulmuştur: eşleştirilmemiş ses (YT-NTL-U: 12M saat, 300+ dil ve Pub-U: 429K saat, 51 dil), eşleştirilmemiş metin (Web-NTL: 28B cümle, 1140+ dil) ve eşleştirilmiş ASR verisi (YT-SUP+: 90K saat etiketli çok dilli veri, 73 dil + 100K saat İngilizce pseudo-etiketli veri ve Pub-S: 10K saat etiketli çok alanlı İngilizce veri + 10K saat etiketli çok dilli veri, 102 dil). 2B parametreli Conformer modelleri, üç adımlı bir eğitim süreci ile oluşturulmuştur: (1) Denetimsiz Ön-eğitim: BEST-RQ (BERT-based Speech pre-Training with Random-projection Quantizer) kullanılarak modelin kodlayıcısı YT-NTL-U ile ön-eğitime tabi tutulmuştur. (2) MOST (Multi-Objective Supervised pre-Training): Model, BEST-RQ maskelenmiş dil modeli kaybı ile birlikte metin enjeksiyon kayıplarının (denetimli ASR kaybı ve modalite eşleştirme kayıpları dahil) ağırlıklı toplamı optimize edilerek hazırlanmıştır. (3) Denetimli ASR Eğitimi: Alt seviye görevler için bağlantısal zamansal sınıflandırma (CTC) ve Listen, Attend, and Spell (LAS) dönüştürücüleri ile eğitilmiş genel ASR modelleri üretilmiştir. Değerlendirme sonuçları, USM modellerinin SpeechStew veri setinde USM-M için %5.7 WER, USM-LAS için %7.2 WER ve USM-CTC için %7.6 WER elde ettiğini göstermiştir. FLEURS veri setinde, USM-M (102 dil) %16.1 WER ve USM-LAS (73 dil) %28.4 WER elde etmiştir. CORAAL veri

setinde, USM-M %17.3 WER ve USM-LAS %26.3 WER elde etmiştir. CoVoST 2 veri setinde, USM-M (İngilizce'den 21 dile çeviri) %28.3 BLEU puanı elde etmiştir. Çalışmanın en önemli katkısı, çok dilli konuşma tanıma ve çevirisi için tek bir büyük modelin, daha küçük etiketli veri setleri kullanılarak, yüksek performanslı sonuçlar elde edebileceğini göstermesidir.

Yao ve arkadaşları tarafından gerçekleştirilen çalışmada, otomatik konuşma tanıma (ASR) için daha hızlı, bellek açısından daha verimli ve daha iyi performans gösteren Zipformer adlı yeni bir Transformer modeli önerilmiştir [13]. Araştırmacılar, Conformer'ın en popüler ASR kodlayıcı modeli olduğunu, ancak Transformer'a konvolüsyon modülleri ekleyerek hem yerel hem de global bağımlılıkları öğrenmesine rağmen, daha verimli bir model geliştirilebileceğini belirtmişlerdir. Çalışmada, model mimarisinde dört önemli değişiklik yapılmıştır: 1) Orta katmanların daha düşük çerçeve hızlarında çalıştığı U-Net benzeri bir kodlayıcı yapısı; 2) Verimlilik için dikkat ağırlıklarını yeniden kullanan ve daha fazla modül içeren yeniden düzenlenmiş blok yapısı; 3) Bazı uzunluk bilgilerini korumaya olanak tanıyan BiasNorm adlı değiştirilmiş bir LayerNorm formu; 4) Swish'ten daha iyi çalışan yeni aktivasyon fonksiyonları SwooshR ve SwooshL. Ayrıca, her tensörün mevcut ölçeğine göre güncellemeyi ölçeklendiren ve parametre ölçeğini açıkça öğrenen ScaledAdam adlı yeni bir optimize edici de önerilmiştir. Zipformer, altı kademeli bir yapıya sahiptir ve bu kademelerde sırasıyla 50Hz, 25Hz, 12.5Hz, 6.25Hz, 12.5Hz ve 25Hz çerçeve hızlarında işlem yapmaktadır. Bu yapı, daha düşük çerçeve hızlarında çalışan orta kademelerin daha büyük gömme boyutlarına sahip olmasını sağlayarak, hem hesaplama verimliliğini artırmakta hem de model performansını iyileştirmektedir. Ayrıca, her Zipformer bloğu, dikkat ağırlıklarını yeniden kullanarak iki Conformer bloğuna eşdeğer işlevsellik sağlamakta, ancak daha az hesaplama gerektirmektedir. Değerlendirme sonuçları, Zipformer'ın LibriSpeech veri setinde test-clean için %1.9 WER ve test-other için %3.9 WER elde ettiğini göstermiştir. Aishell-1 veri setinde %4.0 CER ve WenetSpeech veri setinde test_net için %8.1 CER ve test_meeting için %12.4 CER elde edilmiştir. Verimlilik açısından, Zipformer eğitim sırasında daha hızlı yakınsamakta ve çıkarım sırasında önceki çalışmalara göre %50'den fazla hızlanma sağlarken daha az GPU belleği gerektirmektedir. Çalışmanın en önemli katkısı, ASR için daha verimli ve daha iyi performans gösteren yeni bir Transformer mimarisi sunması ve Conformer makalesi tarafından bildirilen sonuçlara benzer sonuçlar elde eden ilk model olmasıdır.

Chu ve arkadaşları tarafından gerçekleştirilen çalışmada, çeşitli ses sinyali girişlerini kabul edebilen ve konuşma talimatlarına göre ses analizi yapabilen veya doğrudan metinsel yanıtlar verebilen büyük ölçekli bir ses-dil modeli olan Qwen2-Audio tanıtılmıştır [27]. Araştırmacılar, karmaşık hiyerarşik etiketler yerine, farklı veriler ve görevler için doğal dil komutları kullanarak ön-eğitim sürecini basitleştirmiş ve veri hacmini genişletmişlerdir. Çalışmada, Qwen2-Audio'nun talimat takip etme yeteneği artırılmış ve sesli sohbet ve ses analizi için iki farklı ses etkileşim modu uygulanmıştır. Sesli sohbet modunda, kullanıcılar metin girişi olmadan Qwen2-Audio ile serbestçe sesli etkileşimde bulunabilmektedir. Ses analizi modunda ise, kullanıcılar etkileşim sırasında analiz için ses ve metin talimatları sağlayabilmektedir. Önemli bir nokta olarak, sesli sohbet ve ses analizi modları arasında geçiş yapmak için herhangi bir sistem komutu kullanılmamaktadır. Qwen2-Audio, ses içindeki içeriği akıllıca anlama ve sesli komutları takip ederek uygun şekilde yanıt verme yeteneğine sahiptir. Örneğin, aynı anda sesler, çok konuşmacılı konuşmalar ve bir ses komutu içeren bir ses segmentinde, Qwen2-Audio doğrudan komutu anlayabilmekte ve sese yönelik bir yorum ve yanıt sağlayabilmektedir. Model mimarisi olarak, Qwen2-Audio, Whisper-large-v3 tabanlı bir ses kodlayıcı ve Qwen-7B dil modelinden oluşmaktadır. Ses verilerini ön işlemek için, 16kHz'e yeniden örneklenen ham dalga formu, 25ms pencere boyutu ve 10ms atlama boyutu kullanılarak 128 kanallı mel-spektrogramına dönüştürülmektedir. Ek olarak, ses temsil uzunluğunu azaltmak için iki adımlı bir havuzlama katmanı eklenmiştir. Sonuç olarak, kodlayıcı çıktısının her bir çerçevesi, orijinal ses sinyalinin yaklaşık 40ms'lik bir segmentine karşılık gelmektedir. Değerlendirme sonuçları, Qwen2-Audio'nun herhangi bir göreve özgü ince ayar olmadan, LibriSpeech'te %92.62 (1-WER%), Aishell2'de %96.0 (1-WER%), CoVoST2'de %30.0 (ortalama BLEU), MELD'de %45.0 (SER doğruluk), VocalSound'da %87.5 (VSC doğruluk), FLEURS-ZH'de %91.75 (1-WER%) ve AIR-Bench sohbet kıyaslamasında konuşma için %6.88/10, ses için %6.72/10, müzik için %6.5/10 ve karma için %6.43/10 puan elde ederek önceki en iyi modellerden daha iyi performans gösterdiğini ortaya koymuştur. Çalışmanın en önemli katkısı, doğal dil komutlarını kullanarak ön-eğitim sürecini basitleştirmesi, veri hacmini genişletmesi ve herhangi bir sistem komutu gerektirmeden sesli sohbet ve ses analizi modları arasında akıllı geçiş yapabilen bir model geliştirmesidir.

4. KARŞILAŞTIRMA

4.1 BAŞARIM KARŞILAŞTIRMASI

İncelenen çalışmalarda, farklı makine öğrenme yaklaşımlarının ses tanıma görevlerinde gösterdiği başarım, çeşitli metrikler kullanılarak değerlendirilmiştir. Bu bölümde, literatürde yer alan çalışmalar başarım açısından ele alınmaktadır.

4.1.1 Geleneksel ve Derin Öğrenme Yaklaşımlarının Başarım Karşılaştırması

İncelenen çalışmalarda; Gourisaria ve arkadaşları tarafından gerçekleştirilen çalışmada, çevresel ses sınıflandırması için yedi farklı makine öğrenme modeli karşılaştırılmış ve ANN modeli %91.41 doğruluk oranı ile en yüksek başarımı göstermiştir [22]. Gençyılmaz ve Karaoğlan tarafından Türkçe dilinde konuşmadan metne dönüşüm için gerçekleştirilen çalışmada, CRNN modeli %96.47 doğruluk oranı ile en yüksek başarımı göstermiş, bunu CNN (%95.92) ve Rastgele Orman (%92.69) izlemiştir [11]. Keser tarafından gürültülü ortamlarda yalıtık kelime tanıma için gerçekleştirilen çalışmada, RNN-LSTM modeli %93.22 doğruluk oranı ile en yüksek başarımı göstermiş, bunu CNN (%87.56) ve KNN (%86.51) takip etmiştir [15].

İncelenen çalışmalar, derin öğrenme tabanlı modellerin (özellikle CNN, RNN-LSTM ve CRNN), geleneksel makine öğrenme modellerine (SVM, KNN, RF) göre ses tanıma görevlerinde tutarlı bir şekilde daha yüksek başarım gösterdiğini ortaya koymaktadır. Derin öğrenme modellerinin üstünlüğü, karmaşık ses özelliklerini daha etkili bir şekilde modelleyebilme ve öğrenebilme kapasitesinden kaynaklanmaktadır.

4.1.2 Farklı Derin Öğrenme Mimarilerinin Başarım Karşılaştırması

Derin öğrenme mimarileri arasındaki başarım farklılıkları, ses tanıma görevlerinin özelliklerine ve veri setlerinin yapısına bağlı olarak değişiklik göstermektedir. İncelenen çalışmalarda; Rai ve arkadaşları tarafından konuşma içerisindeki komutları tespit etmek için gerçekleştirilen çalışmada, RNN-BiLSTM-Attention modeli %93.9 doğruluk oranı ile en yüksek başarımı göstermiştir [18]. Wang tarafından İngilizce konuşma tanıma için gerçekleştirilen çalışmada, RNN-CTC modeli %10.4 Kelime Hata Oranı (WER) ile en yüksek başarımı göstermiştir [9]. Wojnar ve arkadaşları tarafından YouTube videolarından

oluşan bir veri seti üzerinde farklı derin öğrenme modelleri karşılaştırılmış ve OpenAI'nin Whisper large-v3 modeli %10.58 WER ile en yüksek başarımı göstermiştir [26].

Derin öğrenme mimarileri arasındaki karşılaştırmalar, RNN tabanlı modellerin (özellikle LSTM ve BiLSTM) ve dikkat mekanizması içeren modellerin, ses tanıma görevlerinde yüksek başarım gösterdiğini ortaya koymaktadır. Ayrıca, model boyutunun artmasıyla başarımın da arttığı gözlemlenmiştir. Transformatör mimarisine dayalı büyük modeller (Whisper gibi), özellikle genel amaçlı konuşma tanıma görevlerinde üstün başarım sergilemektedir.

4.1.3 Hibrit ve Yenilikçi Yaklaşımların Başarım Karşılaştırması

Hibrit ve yenilikçi yaklaşımlar, ses tanıma alanında geleneksel ve derin öğrenme tekniklerinin güçlü yönlerini birleştirerek daha yüksek başarım elde etmeyi amaçlamaktadır. İncelenen çalışmalarda; Hamian ve arkadaşları tarafından otomatik rakam ses tanıma için geliştirilen derin darbeli sinir ağları modelinde, Wild Horse Optimization algoritması kullanılarak optimize edilen SNN-WHO modeli %98.2 doğruluk oranı elde etmiştir [35]. Ameen ve Kadhim tarafından elektrolainks cihazı kullanan konuşma ile Arapça fonemlerin tanınması için gerçekleştirilen çalışmada, ANN modeli %75 doğruluk oranı ile en yüksek başarımı göstermiştir [36]. Ahmed ve arkadaşları tarafından Peştuca dili için geliştirilen konuşma tanıma sisteminde, MFCC özellik çıkarma yöntemi ile SVM sınıflandırıcısının kombinasyonu %96.15 doğruluk oranı elde etmiştir [10].

Hibrit ve yenilikçi yaklaşımların başarım karşılaştırması, optimize edilmiş modellerin ve özelleştirilmiş kombinasyonların, standart modellere göre daha yüksek başarım sağlayabildiğini göstermektedir. Özellikle, evrimsel algoritmalar ve optimizasyon teknikleri ile iyileştirilmiş modeller, ses tanıma görevlerinde dikkat çekici sonuçlar vermektedir.

4.1.4 Özel Görevlere Yönelik Başarım Karşılaştırması

Ses tanıma alanında, özel görevlere yönelik geliştirilen modellerin başarımı, görevin karmaşıklığına ve uygulama alanına bağlı olarak değişiklik göstermektedir. İncelenen çalışmalarda; Ayall ve arkadaşları tarafından Amharca dili için konuşulan rakam tanıma sistemi geliştirilen çalışmada, CNN modeli MFCC özellikleri kullanılarak %99 doğruluk oranı elde etmiştir [28]. Alzaabi ve arkadaşları tarafından engelli bireylerin ev aletlerini

kontrol etmeleri için geliştirilen çok dilli konuşma tanıma sisteminde, 3 konvolüsyonel katmanlı CNN modeli %99.84 doğruluk oranına ulaşmıştır [31]. Barhoush ve arkadaşları tarafından konuşmacı tanıma ve lokalizasyonu için geliştirilen sistemde, Karıştırılmış MFCC özelliklerini kullanan tam bağlantılı derin sinir ağı, tek konuşmacı senaryosunda %98.5 konuşmacı tanıma ve %99.6 konuşmacı lokalizasyonu doğruluğu elde etmiştir [5].

Özel görevlere yönelik geliştirilen modellerin başarım karşılaştırması, sınırlı kelime dağarcığına sahip görevlerde (rakam tanıma, komut tanıma) CNN modellerinin oldukça yüksek doğruluk oranlarına ulaşabildiğini göstermektedir. Ayrıca, özel olarak tasarlanmış özellik çıkarma yöntemleri ve mimariler, konuşmacı tanıma ve lokalizasyonu gibi özel görevlerde yüksek başarım sağlayabilmektedir.

4.1.5 Çok Dilli ve Transfer Öğrenimi Yaklaşımlarının Başarım Karşılaştırması

Çok dilli ses tanıma ve transfer öğrenimi yaklaşımları, özellikle düşük kaynaklı diller için önemli faydalar sağlamaktadır. İncelenen çalışmalarda; Froiz-Míguez ve arkadaşları tarafından Galiçya dili için geliştirilen diller arası makine öğrenmesi tabanlı otomatik konuşma tanıma sisteminde, Wav2Vec 2.0 modeli, ışın dekoder ve dil modeli kombinasyonu kullanılarak %18.61 WER değeri elde edilmiştir [38]. Zhang ve arkadaşları tarafından geliştirilen Evrensel Konuşma Modeli (USM), 100'den fazla dilde otomatik konuşma tanıma yapabilen tek bir büyük modeldir ve SpeechStew veri setinde USM-M modeli %5.7 WER değerine ulaşmıştır [20]. Chen ve arkadaşları tarafından geliştirilen Konuşma Destekli Dil Modeli (SALM), LibriSpeech veri setinde ASR görevinde test-clean için %2.3 ve test-other için %4.8 WER değerlerine ulaşmıştır [41].

Çok dilli ve transfer öğrenimi yaklaşımlarının başarım karşılaştırması, önceden eğitilmiş büyük modellerin ince ayar yapılmasının ve dil modellerinin entegrasyonunun, düşük kaynaklı dillerde bile yüksek başarım elde etmek için etkili yaklaşımlar olduğunu göstermektedir.

4.1.6 Genel Başarım Değerlendirmesi

İncelenen çalışmalar ışığında, ses tanıma alanında en etkili makine öğrenme yöntemlerinin görev ve veri seti özelliklerine bağlı olarak değiştiği görülmektedir. Genel olarak, derin öğrenme tabanlı modeller (CNN, RNN, LSTM, Transformer), geleneksel makine öğrenme

modellerine göre daha yüksek başarım göstermektedir. Özellikle, dikkat mekanizması içeren modeller (RNN-BiLSTM-Attention, Transformer), kompleks ses tanıma görevlerinde üstün başarım sergilemektedir.

Sınırlı kelime dağarcığına sahip görevlerde (rakam tanıma, komut tanıma), CNN tabanlı modeller yüksek başarım gösterirken, daha karmaşık ve geniş kelime dağarcıklı görevlerde (sürekli konuşma tanıma), RNN, LSTM ve Transformer tabanlı modeller daha etkili olmaktadır.

4.2 KARMAŞIKLIK KARŞILAŞTIRMASI

Ses tanıma sistemlerinde kullanılan makine öğrenme tekniklerinin karmaşıklığı; hesaplama gereksinimleri, model boyutu, eğitim süresi ve çıkarım (inference) hızı gibi faktörler açısından değerlendirilmektedir. Bu bölümde, literatürde incelenen çalışmalarda kullanılan makine öğrenme teknikleri, karmaşıklık açısından incelenmektedir.

4.2.1 Hesaplama Karmaşıklığı

Makine öğrenme modellerinin hesaplama karmaşıklığı, eğitim ve çıkarım aşamalarında gereken işlem miktarını ifade etmektedir. Geleneksel makine öğrenme teknikleri ile derin öğrenme yaklaşımları arasında hesaplama karmaşıklığı açısından önemli farklılıklar bulunmaktadır.

Gourisaria ve arkadaşları tarafından gerçekleştirilen çalışmada, çevresel ses sınıflandırması için yedi farklı makine öğrenme modeli karşılaştırılmıştır [22]. Çalışmada, Yapay Sinir Ağı (ANN) modeli en yüksek başarımı gösterirken, hesaplama karmaşıklığı açısından da en yüksek değere sahip olduğu belirtilmiştir. Buna karşın, Karar Ağacı ve Naive Bayes gibi geleneksel modeller daha düşük hesaplama karmaşıklığına sahip olduğu, ancak başarım açısından da daha düşük sonuçlar verdiği gözlemlenmiştir.

Wojnar ve arkadaşları tarafından gerçekleştirilen çalışmada, farklı boyutlardaki Whisper modelleri karşılaştırılmıştır. Çalışmada, model boyutu arttıkça hesaplama karmaşıklığının da arttığı, ancak başarımın da buna paralel olarak iyileştiği gösterilmiştir [26]. En küçük model olan Whisper tiny.en ile en büyük model olan Whisper large-v3 arasında hesaplama gereksinimleri açısından önemli farklılıklar bulunmaktadır. Whisper large-v3 modeli, 1.55

milyar parametreye sahip olup, eğitim ve çıkarım aşamalarında yüksek hesaplama gücü gerektirmektedir. Buna karşın, Whisper tiny.en modeli sadece 39 milyon parametreye sahip olup, sınırlı hesaplama kaynaklarıyla bile çalıştırılabilmektedir.

Alzaabi ve arkadaşları tarafından gerçekleştirilen çalışmada, engelli bireylerin ev aletlerini kontrol etmeleri için geliştirilen çok dilli konuşma tanıma sisteminde, farklı konvolüsyonel katman sayılarına sahip CNN modelleri karşılaştırılmıştır [31]. Çalışmada, 2, 3, 4 ve 5 konvolüsyonel katmana sahip modeller arasında hesaplama karmaşıklığı açısından karşılaştırma yapılmış ve katman sayısı arttıkça hesaplama gereksinimlerinin de arttığı belirtilmiştir. En yüksek başarımı gösteren 3 konvolüsyonel katmanlı model, hesaplama karmaşıklığı ve başarım arasında optimum bir denge sağlamış, bu modelin ESP32 gibi sınırlı kaynaklara sahip bir mikrodenetleyicide bile çalıştırılabilecek kadar verimli olduğu gösterilmiştir.

4.2.2 Model Boyutu ve Hafıza Gereksinimleri

Makine öğrenme modellerinin boyutu ve hafıza gereksinimleri, özellikle sınırlı kaynaklara sahip cihazlarda kullanılabilirlik açısından önemli bir faktördür. Geleneksel makine öğrenme modelleri genellikle daha küçük model boyutlarına sahipken, derin öğrenme modelleri daha büyük hafıza gereksinimleri göstermektedir.

Zhang ve arkadaşları tarafından geliştirilen Evrensel Konuşma Modeli (USM), 100'den fazla dilde otomatik konuşma tanıma yapabilen tek bir büyük modeldir [20]. Çalışmada, USM-M modelinin 600 milyon parametreye sahip olduğu ve yüksek hafıza gereksinimleri gösterdiği belirtilmiştir. Buna karşın, model boyutunun büyüklüğü, çok dilli konuşma tanıma görevlerinde yüksek başarım elde edilmesini sağlamıştır.

Hamian ve arkadaşları tarafından otomatik rakam ses tanıma için geliştirilen derin darbeli sinir ağları (SNN) modeli, geleneksel yapay sinir ağlarına (ANN) göre daha düşük hafıza gereksinimleri göstermiştir [35]. SNN modelleri, bilgilerin sürekli değerler yerine anlık darbeler (spikes) şeklinde iletilmesine dayandığı için, daha az hafıza kullanımı ve daha düşük enerji tüketimi sağlamaktadır. Çalışmada, SNN modellerinin optimize edilmesi için kullanılan Wild Horse Optimization (WHO) algoritması, model boyutunu artırmadan başarımı iyileştiren bir yaklaşım olarak öne çıkmıştır.

Ayall ve arkadaşları tarafından Amharca dili için geliştirilen konuşulan rakam tanıma sisteminde, CNN modelinin model boyutu ve hafıza gereksinimleri açısından değerlendirilmesi yapılmıştır [28]. Çalışmada, önerilen 3 katmanlı CNN mimarisinin, geleneksel makine öğrenme modellerine göre daha yüksek hafıza gereksinimleri gösterdiği, ancak sınırlı kelime dağarcığına sahip bir görev için kabul edilebilir seviyede olduğu belirtilmiştir. Model, 16-bit kayan noktalı sayı formatında yaklaşık 2.3 MB boyutunda olup, düşük kaynaklı cihazlarda bile çalıştırılabilecek kadar kompakt olduğu gösterilmiştir.

4.2.3 Eğitim Süresi ve Yakınsama Hızı

Makine öğrenme modellerinin eğitim süresi ve yakınsama hızı, model geliştirme ve güncelleme süreçlerinin verimliliği açısından önemli bir faktördür. Geleneksel makine öğrenme modelleri genellikle daha kısa eğitim sürelerine sahipken, derin öğrenme modelleri daha uzun eğitim süreleri gerektirmektedir.

Gençyılmaz ve Karaoğlan tarafından gerçekleştirilen çalışmada, Türkçe dilinde konuşmadan metne dönüşüm için dokuz farklı makine öğrenme modeli karşılaştırılmıştır [11]. Çalışmada, CRNN ve CNN gibi derin öğrenme modellerinin eğitim sürelerinin, Rastgele Orman ve SVM gibi geleneksel modellere göre daha uzun olduğu belirtilmiştir. CRNN modeli, 100 epoch boyunca eğitildiğinde yaklaşık 3 saat sürerken, Rastgele Orman modeli sadece birkaç dakika içinde eğitilebilmiştir. Buna karşın, derin öğrenme modelleri daha yüksek başarım göstermiştir.

Rai ve arkadaşları tarafından konuşma içerisindeki komutları tespit etmek için gerçekleştirilen çalışmada, RNN-BiLSTM-Attention modelinin eğitim süresi ve yakınsama hızı değerlendirilmiştir [18]. Çalışmada, dikkat mekanizması eklenmesinin eğitim süresini artırdığı, ancak modelin daha hızlı yakınsadığı ve daha yüksek başarım gösterdiği belirtilmiştir. RNN-BiLSTM modeli 150 epoch'ta yakınsarken, RNN-BiLSTM-Attention modeli 120 epoch'ta yakınsamıştır.

Barhoush ve arkadaşları tarafından konuşmacı tanıma ve lokalizasyonu için geliştirilen sistemde, tam bağlantılı derin sinir ağı (FC-DNN) modelinin eğitim süresi ve yakınsama hızı incelenmiştir [5]. Çalışmada, önerilen Karıştırılmış MFCC (SHMFCC) özelliklerinin kullanılmasının, modelin daha hızlı yakınsamasını sağladığı ve eğitim süresini kısalttığı

belirtilmiştir. Standart MFCC özellikleri kullanıldığında model 200 epoch'ta yakınsarken, SHMFCC özellikleri kullanıldığında 150 epoch'ta yakınsamıştır.

4.2.4 Çıkarım Hızı ve Gerçek Zamanlı Performans

Makine öğrenme modellerinin çıkarım hızı ve gerçek zamanlı performansı, özellikle pratik uygulamalarda önem taşımaktadır. Ses tanıma sistemlerinin gerçek dünya senaryolarında kullanılabilmesi için, çıkarım aşamasının yeterince hızlı olması gerekmektedir.

Alzaabi ve arkadaşları tarafından gerçekleştirilen çalışmada, engelli bireylerin ev aletlerini kontrol etmeleri için geliştirilen çok dilli konuşma tanıma sisteminin gerçek zamanlı performansı değerlendirilmiştir [31]. Çalışmada, 3 konvolüsyonel katmanlı CNN modelinin ESP32 mikrodenetleyicisinde gerçek zamanlı olarak çalıştırılabildiği ve komutları 200-300 milisaniye içinde tanıyabildiği gösterilmiştir.

Wang tarafından İngilizce konuşma tanıma için gerçekleştirilen çalışmada, RNN-CTC modelinin çıkarım hızı ve gerçek zamanlı performansı incelenmiştir [9]. Çalışmada, RNN-CTC modelinin gerçek zamanlı konuşma tanıma için yeterince hızlı olduğu, ancak uzun konuşma dizilerinde bellek kullanımının artmasından dolayı performans düşüşleri yaşanabileceği belirtilmiştir. Bu sorunu çözmek için, konuşma sinyalinin sabit uzunluktaki parçalara bölünerek işlenmesi önerilmiştir.

Shukla ve arkadaşları tarafından gerçekleştirilen çalışmada, "soft forgetting" tekniği kullanılarak BLSTM modellerinin çıkarım hızının iyileştirilmesi amaçlanmıştır [30]. Çalışmada, örtüşmeyen kesitler üzerinde BLSTM modelinin çalıştırılmasının, tüm konuşma boyunca açılmasına göre daha hızlı çıkarım sağladığı gösterilmiştir. Soft forgetting yaklaşımı, özellikle uzun konuşma dizilerinde gerçek zamanlı performansı iyileştirmiştir.

4.2.5 Optimizasyon Teknikleri ve Model Sıkıştırma

Makine öğrenme modellerinin karmaşıklığını azaltmak ve verimliliklerini artırmak için çeşitli optimizasyon teknikleri ve model sıkıştırma yöntemleri kullanılmakta; model boyutu küçültülürken başarımın koruması veya minimum düzeyde düşürmesi amaçlanmaktadır.

Hamian ve arkadaşları tarafından gerçekleştirilen çalışmada, derin darbeli sinir ağları (SNN) modelinin optimizasyonu için Wild Horse Optimization (WHO) algoritması kullanılmıştır

[35]. Çalışmada, WHO algoritmasının, modelin ağırlıklarını optimize ederek daha az parametreyle daha yüksek başarım elde edilmesini sağladığı gösterilmiştir. SNN-WHO modeli, geleneksel yapay sinir ağı (ANN) modeline göre daha az parametreye sahip olmasına rağmen daha yüksek başarım göstermiştir.

Chen ve arkadaşları tarafından geliştirilen Konuşma Destekli Dil Modeli (SALM), model sıkıştırma teknikleri kullanılarak optimize edilmiştir [41]. Çalışmada, bilgi damıtma (knowledge distillation) ve parametre paylaşımı (parameter sharing) teknikleri kullanılarak, büyük dil modellerinin boyutu küçültülmüş ve hesaplama verimliliği artırılmıştır. Bu sayede, SALM modeli, konuşma tanıma ve çevirisi görevlerinde yüksek başarım gösterirken, daha az hesaplama kaynağı gerektirmiştir.

Froiz-Míguez ve arkadaşları tarafından Galiçya dili için geliştirilen konuşma tanıma sisteminde, Wav2Vec 2.0 modelinin ince ayar (fine-tuning) aşamasında çeşitli optimizasyon teknikleri kullanılmıştır [38]. Çalışmada, aşamalı öğrenme oranı (progressive learning rate) ve parametre dondurmak (parameter freezing) gibi tekniklerle, modelin eğitim süresi kısaltılmış ve hesaplama gereksinimleri azaltılmıştır.

4.2.6 Karmaşıklık-Başarım Dengesi

Makine öğrenme modellerinin karmaşıklığı ile başarımı arasında genellikle bir ödünleşim (trade-off) bulunmaktadır. Daha karmaşık modeller genellikle daha yüksek başarım gösterirken, daha fazla hesaplama kaynağı ve eğitim süresi gerektirmektedir. Bu nedenle, uygulama gereksinimlerine bağlı olarak karmaşıklık-başarım dengesi gözetilmelidir.

Wojnar ve arkadaşları tarafından gerçekleştirilen çalışmada, farklı boyutlardaki Whisper modellerinin karmaşıklık-başarım dengesi incelenmiştir [26]. Çalışmada, Whisper medium.en modelinin, large-v3 modeline göre daha düşük başarım gösterdiği (%16.97 WER vs %10.58 WER), ancak hesaplama gereksinimleri açısından çok daha verimli olduğu belirtilmiştir.

Alzaabi ve arkadaşları tarafından gerçekleştirilen çalışmada, farklı konvolüsyonel katman sayılarına sahip CNN modellerinin karmaşıklık-başarım dengesi değerlendirilmiştir [31]. Çalışmada, 3 konvolüsyonel katmanlı modelin, 5 konvolüsyonel katmanlı modele göre daha düşük karmaşıklığa sahip olduğu ve neredeyse aynı başarımı gösterdiği belirtilmiştir.

Ahmed ve arkadaşları tarafından Peştuca dili için geliştirilen konuşma tanıma sisteminde, MFCC özellik çıkarma yöntemi ile SVM sınıflandırıcısının kombinasyonu, karmaşıklık-başarım dengesi açısından optimum bir çözüm sunmuştur [10].

4.2.7 Genel Karmaşıklık Değerlendirmesi

İncelenen çalışmalar ışığında, genel olarak, derin öğrenme tabanlı modeller (CNN, RNN, LSTM, Transformer) daha yüksek hesaplama karmaşıklığına sahipken, daha yüksek başarım göstermektedir. Buna karşın, geleneksel makine öğrenme modelleri (SVM, KNN, RF) daha düşük hesaplama karmaşıklığına sahip olup, sınırlı kaynaklara sahip uygulamalar için daha uygun olabilmektedir.

Sınırlı kelime dağarcığına sahip görevlerde (rakam tanıma, komut tanıma), optimize edilmiş CNN modelleri, karmaşıklık-başarım dengesi açısından en etkili çözümlerden biri olarak öne çıkmaktadır. Özellikle, Alzaabi ve arkadaşları ile Ayall ve arkadaşları tarafından gerçekleştirilen çalışmalarda, 3 katmanlı CNN mimarileri, düşük hesaplama karmaşıklığı ile yüksek başarım elde etmiştir [28], [31].

Daha karmaşık ve geniş kelime dağarcıklı görevlerde (sürekli konuşma tanıma), RNN, LSTM ve Transformer tabanlı modeller daha etkili olmaktadır. Ancak, bu modellerin hesaplama karmaşıklığı ve hafıza gereksinimleri de daha yüksektir. Bu nedenle, Shukla ve arkadaşları tarafından önerilen "soft forgetting" gibi optimizasyon teknikleri, bu modellerin verimliliklerini artırmak için önemlidir [30].

Transfer öğrenimi ve çok dilli ön-eğitim yaklaşımları, özellikle düşük kaynaklı diller için hesaplama verimliliği açısından fayda sağlamaktadır. Froiz-Míguez ve arkadaşları ile Zhang ve arkadaşları tarafından gerçekleştirilen çalışmalarda, önceden eğitilmiş modellerin ince ayar yapılması, sıfırdan eğitime göre çok daha az hesaplama kaynağı gerektirmiştir [20], [38].

Özetlemek gerekirse; ses tanıma alanında karmaşıklık açısından en etkili makine öğrenme yöntemi, uygulama gereksinimlerine ve kısıtlamalarına bağlı olarak değişmektedir. Yüksek başarım gerektiren ve hesaplama kaynaklarının sınırlı olmadığı uygulamalar için derin öğrenme tabanlı modeller (özellikle Transformer ve LSTM tabanlı mimariler) öne çıkarken,

sınırlı kaynaklara sahip gömülü sistemler için optimize edilmiş CNN modelleri veya geleneksel makine öğrenme teknikleri daha uygun olabilmektedir.

4.3 VERİ GEREKSİNİMİ KARŞILAŞTIRMASI

Makine öğrenme tabanlı ses tanıma sistemlerinin performansını etkileyen en önemli faktörlerden biri, eğitim için kullanılan veri setinin niteliği ve niceliğidir. İncelenen çalışmalarda, farklı makine öğrenme yaklaşımlarının veri gereksinimleri açısından önemli farklılıklar gösterdiği gözlemlenmiştir.

4.3.1 Geleneksel ve Derin Öğrenme Yaklaşımlarının Veri Gereksinimleri

Geleneksel makine öğrenme yaklaşımları (SVM, KNN, Random Forest, GMM-HMM vb.) ile derin öğrenme yaklaşımları (CNN, RNN, LSTM, Transformer vb.) arasında veri gereksinimleri açısından belirgin farklılıklar bulunmaktadır. Geleneksel yaklaşımlar, genellikle daha az veri ile kabul edilebilir sonuçlar üretebilirken, derin öğrenme yaklaşımları daha fazla veriye ihtiyaç duymaktadır.

Ahmed ve arkadaşları tarafından gerçekleştirilen çalışmada, Peştuca dili için geliştirilen konuşma tanıma sisteminde, SVM ve KNN gibi klasik makine öğrenme algoritmalarının, nispeten küçük bir veri seti (161 izole kelime, 30 konuşmacı) ile yüksek doğruluk oranları elde edebildiği gösterilmiştir [10]. Çalışmada, MFCC özellik çıkarma yöntemi ile SVM sınıflandırıcısının kombinasyonu %96.15 doğruluk oranına ulaşmıştır. Benzer şekilde, Ayall ve arkadaşları tarafından Amharca için geliştirilen konuşma tanıma sisteminde, geleneksel yöntemler (LDA, KNN, SVM, RF) 12.000 konuşma kaydından oluşan bir veri seti ile %93-97 arasında doğruluk oranları elde etmiştir [28].

Buna karşılık, derin öğrenme yaklaşımları genellikle daha büyük veri setlerine ihtiyaç duymaktadır. Shukla ve arkadaşları tarafından gerçekleştirilen çalışmada, uçtan uca konuşma tanıma sistemleri için 300 saatlik İngilizce Switchboard veri seti kullanılmıştır [30]. Zhang ve arkadaşları 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 [20].

4.3.2 Transfer Öğrenimi ve Veri Verimliliği

Son yıllarda, özellikle düşük kaynaklı diller ve özel uygulamalar için veri gereksinimini azaltmak amacıyla transfer öğrenimi yaklaşımları önem kazanmıştır. Transfer öğrenimi yaklaşımı, önceden eğitilmiş modellerin bilgisini yeni görevlere aktararak, daha az veriyle etkili sonuçlar elde edilmesini sağlamaktadır.

Froiz-Míguez ve arkadaşları tarafından Galiçya dili için geliştirilen sistemde, wav2vec 2.0 modeli kullanılarak, sadece 20 saatlik etiketli ses verisiyle %18.61 WER değerine ulaşılmıştır [38]. Çalışma, öz-denetimli eğitim kullanarak çok dilli etiketlenmemiş sesler üzerinde eğitim yapan ve ardından belirli bir dilin etiketli sesleriyle ince ayar yapabilen bir yaklaşımın, veri kısıtlı ortamlarda etkin olduğunu göstermiştir.

Benzer şekilde, Jia tarafından gerçekleştirilen çalışmada, alan-spesifik konuşma tanıma için yarı denetimli öğrenme etiketleme yöntemi kullanılarak, minimal insan müdahalesi gerektiren bir veri toplama yaklaşımı önerilmiştir [19]. Çalışmada, etiketli örnekler kümesi girdi olarak alınıp, daha büyük etiketlenmemiş örnekler kümesi için kaba transkripsiyonlar üreten bir tanıyıcı komitesi oluşturulmuştur.

4.3.3 Veri Artırma Teknikleri

Veri gereksinimini azaltmak için kullanılan veri artırma (data augmentation) teknikleri, mevcut veri setinden yeni örnekler oluşturarak, modelin daha geniş bir veri yelpazesi üzerinde eğitilmesini sağlamaktadır.

Barhoush ve arkadaşları tarafından gerçekleştirilen çalışmada, konuşmacı tanıma ve lokalizasyonu için veri artırma amacıyla bir ön işleme adımı önerilmiştir [5]. Bu yaklaşımda, ses çerçeveleri bloklar halinde bölünmekte ve her blok içindeki çerçeveler rastgele karıştırılarak model eğitimi için daha fazla veri oluşturulmaktadır. Ayall ve arkadaşları tarafından Amharca dili için geliştirilen konuşulan rakam tanıma sisteminde, her konuşmacının her rakamı 10 kez tekrarlamasıyla elde edilen sınırlı veri seti, çeşitli veri artırma teknikleriyle genişletilmiştir [28]. Ayrıca, Alzaabi ve arkadaşları tarafından yapılan çalışmada, çeşitli arka plan gürültüleri eklenerek veri seti zenginleştirilmiş ve bu sayede modelin gürültülü ortamlarda da yüksek performans göstermesi sağlanmıştır [31]. Shukla ve arkadaşları tarafından yapılan çalışmada, çevrimdışı ve anında veri artırma teknikleri

eklendiğinde, konuşmacıdan bağımsız telefon CTC sisteminin WER değerinin %9.1/%17.4'ten %7.9/%15.7'ye düştüğü gösterilmiştir [30].

4.3.4 Model Mimarisi ve Veri Gereksinimi İlişkisi

Farklı model mimarilerinin veri gereksinimleri de değişkenlik göstermektedir. Özellikle, parametre sayısı arttıkça genellikle veri gereksinimi de artmaktadır.

Yao ve arkadaşları tarafından önerilen Zipformer modeli, verimli bir mimari tasarımı sayesinde, daha az veri ile etkili sonuçlar elde edebilmektedir [13]. U-Net benzeri bir kodlayıcı yapısı ve yeniden düzenlenmiş blok yapısı sayesinde, LibriSpeech veri setinde test-clean için %1.9 WER ve test-other için %3.9 WER değerlerine ulaşmıştır.

Gong ve arkadaşları tarafından geliştirilen LTU-AS modeli, toplam 8.5 milyar parametreye sahip olmasına rağmen, sadece 49 milyon parametre eğitilebilir durumdadır [42]. LTU-AS modeli, büyük modellerin avantajlarından yararlanırken, eğitim için gereken veri miktarını azaltmaktadır.

4.3.5 Özel Uygulamalar ve Veri Gereksinimleri

Özel uygulamalar için geliştirilen ses tanıma sistemlerinde, genellikle daha spesifik ve sınırlı veri setleri kullanılmaktadır.

Rahate ve arkadaşları tarafından gerçekleştirilen çalışmada, felçli ve nöromüsküler hastalıklardan muzdarip hastalar için EEG sinyalleri kullanılarak sessiz konuşma tanıma sistemi geliştirilmiştir [34]. Çalışmada, 10 sağlıklı denekten 6 farklı kelime için toplanan EEG verileri kullanılmıştır. Sınırlı veri ile çalışılmasına rağmen, KNN algoritması %61.45 doğruluk oranına ulaşmıştır.

Shisode ve arkadaşları tarafından konuşma engelli kişiler için geliştirilen sistemde, yüz kaslarından alınan EMG sinyalleri ve dudak hareketleri videoları kullanılmıştır [37]. Bu çalışmada da sınırlı veri ile çalışılmasına rağmen, EMG sistemi için "Emergency" kelimesinin %80 doğrulukla tanınması sağlanmıştır.

4.3.6 Çok Dilli Sistemler ve Veri Gereksinimleri

Çok dilli ses tanıma sistemleri, genellikle her dil için ayrı veri setlerine ihtiyaç duymaktadır. Ancak, son yıllarda geliştirilen yaklaşımlar, diller arası transfer öğrenimi sayesinde, düşük kaynaklı diller için veri gereksinimini azaltmaktadır.

Zhang ve arkadaşları tarafından geliştirilen Evrensel Konuşma Modeli (USM), 300'den fazla dili kapsayan büyük bir etiketlenmemiş veri seti üzerinde ön-eğitime tabi tutulmuş ve daha sonra daha küçük bir etiketli veri seti üzerinde ince ayar yapılmıştır [20]. Uygulanan yaklaşım, Whisper modelinin kullandığı etiketli eğitim setinin 1/7'si büyüklüğünde bir eğitim seti kullanılmasına rağmen, birçok dilde karşılaştırılabilir veya daha iyi performans elde edilmesini sağlamıştır.

Chu ve arkadaşları tarafından geliştirilen Qwen2-Audio modeli, doğal dil komutlarını kullanarak ön-eğitim sürecini basitleştirmiş ve veri hacmini genişletmiştir [27]. Qwen2-Audio modeli, herhangi bir göreve özgü ince ayar olmadan, LibriSpeech'te %92.62 (1-WER%), Aishell2'de %96.0 (1-WER%) ve FLEURS-ZH'de %91.75 (1-WER%) doğruluk oranlarına ulaşmıştır.

4.3.7 Genel Veri Gereksinimi Değerlendirmesi

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 DAYANIKLILIK KARŞILAŞTIRMASI

Gürültüye dayanıklılık, bir ses tanıma sisteminin farklı gürültü koşulları altında performansını koruyabilme yeteneğidir. Bu bölümde, literatür taramasında incelenen çalışmalardaki makine öğrenme teknikleri gürültüye dayanıklılık açısından incelenmektedir.

4.4.1 Gürültü Türleri ve Gürültü Azaltma Teknikleri

İncelenen çalışmalarda, ses tanıma sistemlerinin performansını etkileyen çeşitli gürültü türleri ele alınmıştır. Barhoush ve arkadaşları çalışmasında, farklı Sinyal-Gürültü Oranı (SNR) seviyelerinde (0-30 dB) ve yankılanma koşullarında testler gerçekleştirilmiştir [5].

Alzaabi ve arkadaşları tarafından geliştirilen sistemde ise, akan musluk, bulaşık yıkama sesleri ve beyaz gürültü gibi çeşitli arka plan gürültüleri kullanılarak veri seti zenginleştirilmiştir [31]. Keser tarafından yürütülen çalışmada, farklı gürültü türlerini içeren özel bir veritabanı kullanılarak gürültülü ortamlarda yalıtık kelime tanıma performansı incelenmiştir [15].

Gürültü türlerinin ses tanıma sistemleri üzerindeki etkileri incelendiğinde, özellikle düşük SNR seviyelerinde (0-10 dB) geleneksel makine öğrenme yöntemlerinin performansında önemli düşüşler gözlemlenmiştir. Buna karşın, derin öğrenme tabanlı yaklaşımların daha sağlam (robust) sonuçlar verdiği tespit edilmiştir.

Literatürde, gürültüye dayanıklılığı artırmak için çeşitli teknikler önerilmiştir. Gourisaria ve arkadaşları çalışmasında, gürültü temizleme işleminin sınıflandırma performansını önemli ölçüde artırdığı gösterilmiştir [22]. Çalışmada, ses örneklerindeki gürültüler azaltılarak sadece ayırt edici ses özellikleri korunmuş ve bu işlem sonucunda tüm sınıflandırıcıların performansında iyileşme sağlanmıştır. Rahate ve arkadaşları tarafından gerçekleştirilen çalışmada, EEG sinyallerindeki gürültüleri temizlemek için 0.5-100 Hz arasında 5. dereceden Butterworth bant geçiren filtre kullanılmış ve 60 Hz'de çentik filtre uygulanarak hat gürültüsü giderilmiştir [34]. Uygulanan önişleme adımları, sınıflandırma performansını önemli ölçüde etkilemiştir.

4.4.2 Derin Öğrenme Yaklaşımlarının Gürültüye Dayanıklılığı

İncelenen çalışmalar, derin öğrenme yaklaşımlarının gürültüye karşı geleneksel makine öğrenme yöntemlerine göre daha dayanıklı olduğunu göstermektedir. Barhoush ve arkadaşları çalışmasında, önerilen tam bağlantılı derin sinir ağı (FC-DNN) tabanlı yaklaşımın, düşük gürültü seviyelerinde (0 dB SNR) bile %90'ın üzerinde doğruluk oranı sağladığı belirtilmiştir [5].

Alzaabi ve arkadaşları tarafından geliştirilen CNN tabanlı sistem, çeşitli arka plan gürültüleri eklenerek zenginleştirilmiş veri seti üzerinde %99'un üzerinde doğruluk oranları elde etmiştir [31]. Özellikle 3 konvolüsyonel katmanlı model (%99.84 doğruluk), gürültülü koşullarda bile yüksek performans göstermiştir.

Keser tarafından yürütülen çalışmada, gürültülü ortamlarda yalıtık kelime tanıma için RNN-LSTM mimarisi %93.22 doğruluk oranıyla en iyi performansı gösterirken, CNN (%87.56) ve KNN (%86.51) mimarileri daha düşük performans sergilemiştir [15].

4.4.3 Özellik Çıkarma Yöntemlerinin Etkisi

Özellik çıkarma yöntemlerinin gürültüye dayanıklılık üzerindeki etkisi de incelenen çalışmalarda ele alınmıştır. Barhoush ve arkadaşları çalışmasında, MFCC tabanlı yeni bir özellik olan Karıştırılmış MFCC (SHMFCC) ve Fark Karıştırılmış MFCC (DSHMFCC) özellikleri önerilmiştir [5]. Önerilen özellikler, geleneksel MFCC'ye göre gürültülü ortamlarda daha iyi performans göstermiştir. Keser tarafından yürütülen çalışmada, MFCC ve PLP (Algısal Doğrusal Tahmin) katsayıları karşılaştırılmış ve her iki özellik çıkarma yönteminin de gürültülü ortamlarda benzer performans gösterdiği, ancak sınıflandırıcı seçiminin daha belirleyici olduğu görülmüştür [15].

4.4.4 Genel Gürültüye Dayanıklılık Değerlendirmesi

İncelenen çalışmalar ışığında, gürültüye dayanıklılık açısından en etkili makine öğrenme yaklaşımlarının derin öğrenme tabanlı modeller olduğu görülmektedir. Özellikle RNN-LSTM mimarileri, zamansal bilgiyi modelleyebilme yetenekleri sayesinde gürültülü ortamlarda yüksek performans göstermektedir. CNN tabanlı modeller de, özellikle spektrogram tabanlı özelliklerin kullanıldığı durumlarda etkili sonuçlar vermektedir.

4.5 UYGULANABİLİRLİK KARŞILAŞTIRMASI

Ses tanıma teknolojilerinin çeşitli alanlardaki uygulanabilirliği, kullanılan makine öğrenme tekniklerinin özelliklerine ve hedeflenen uygulama alanının gereksinimlerine bağlı olarak değişkenlik göstermektedir. Bu bölümde, incelenen çalışmalarda kullanılan makine öğrenme tekniklerinin farklı uygulama alanlarındaki uygulanabilirliği ele alınmaktadır.

4.5.1 Genel Amaçlı Konuşma Tanıma Sistemlerinde Uygulanabilirlik

Genel amaçlı konuşma tanıma sistemlerinde, özellikle büyük kelime dağarcıklı sürekli konuşma tanıma (LVCSR) alanında, uçtan uca (end-to-end) modeller ve derin öğrenme tabanlı yaklaşımlar öne çıkmaktadır. Shukla ve arkadaşları tarafından yapılan çalışmada, CTC (Connectionist Temporal Classification) tabanlı uçtan uca modellerin ve özellikle

BLSTM (Bidirectional Long Short-Term Memory) mimarilerinin genel amaçlı konuşma tanıma sistemlerinde yüksek performans gösterdiği belirtilmiştir [30]. Söz konusu modeller, özellikle "soft forgetting" gibi tekniklerle desteklendiğinde, WER (Kelime Hata Oranı) değerlerinde %7-9 oranında iyileşme sağlamaktadır. Benzer şekilde, Wang tarafından gerçekleştirilen çalışmada, RNN-CTC modelinin geleneksel GMM-HMM ve CNN-CTC modellerine göre daha düşük WER değerleri (%10.4) ve daha kısa eğitim/test süreleri sunduğu gösterilmiştir [9].

Daha geniş ölçekli uygulamalarda ise, Zhang ve arkadaşları tarafından geliştirilen Evrensel Konuşma Modeli (USM) gibi büyük modeller öne çıkmaktadır [20]. 2B parametreli Conformer mimarisi üzerine inşa edilen model, 100'den fazla dilde konuşma tanıma yapabilme yeteneğine sahiptir ve SpeechStew veri setinde %5.7 WER gibi etkileyici sonuçlar elde etmiştir.

4.5.2 Özel Amaçlı ve Alan-Spesifik Sistemlerde Uygulanabilirlik

Özel amaçlı ve alan-spesifik konuşma tanıma sistemlerinde, daha hafif ve hedef alana özel eğitilmiş modeller tercih edilmektedir. Ayall ve arkadaşları tarafından Amharca rakam tanıma için geliştirilen sistemde, 3 konvolüsyonel katmanlı bir CNN mimarisi kullanılmış ve MFCC özellikleriyle %99 doğruluk elde edilmiştir [28]. Jia tarafından gerçekleştirilen çalışmada, sağlık sigortası ve çalışan hakları alanına özgü bir konuşma tanıma sistemi için, önceden eğitilmiş Wav2Vec2-Large-LV60 modeli üzerinde ince ayar (fine-tuning) yapılmış ve harici bir KenLM dil modeli eklenmiştir [19]. Uygulanan yaklaşım, ticari ASR sistemlerine göre daha iyi performans göstermiştir. Benzer şekilde, Froiz-Míguez ve arkadaşları tarafından Galiçya dili için geliştirilen sistemde, wav2vec 2.0 modeli üzerinde yaklaşık 20 saatlik etiketli veri ile ince ayar yapılmış ve %18.61 WER elde edilmiştir [38].

4.5.3 Gömülü Sistemlerde ve Sınırlı Kaynaklı Ortamlarda Uygulanabilirlik

Gömülü sistemlerde ve sınırlı kaynaklı ortamlarda, hesaplama ve bellek gereksinimleri düşük olan modeller tercih edilmektedir. Alzaabi ve arkadaşları tarafından geliştirilen, engelli bireyler için tasarlanmış düşük maliyetli gömülü konuşma tanıma sistemi, 3 konvolüsyonel katmanlı bir CNN mimarisi kullanmakta ve ESP32 mikrodenetleyicisine

yerleştirilebilmektedir [31]. Sistem, %99.84 doğruluk oranıyla yüksek performans göstermekte ve hem İngilizce hem de Arapça (Emirlik lehçesi) komutları tanıyabilmektedir.

Ahmed ve arkadaşları tarafından Peştuca dili için geliştirilen konuşma tanıma sisteminde, klasik makine öğrenme modelleri (SVM, KNN) tercih edilmiş ve MFCC özellik çıkarma yöntemiyle SVM sınıflandırıcısı kombinasyonu %96.15 doğruluk oranı elde etmiştir [10]. Araştırmacılar, hesaplama açısından daha verimli olan klasik makine öğrenme tekniklerinin, derin öğrenme modellerine kıyasla kaynak verimliliği sağlayarak yüksek doğruluk oranları elde edebileceğini belirtmiştir. Benzer şekilde; Hamian ve arkadaşları tarafından geliştirilen darbeli sinir ağları (SNN) tabanlı sistem, geleneksel yapay sinir ağlarına (ANN) göre daha az hesaplama kaynağı gerektirmiş ve rakam ses tanıma görevinde %98.2 doğruluk oranı elde etmiştir [35].

4.5.4 Çok Dilli Sistemlerde Uygulanabilirlik

Çok dilli konuşma tanıma sistemlerinde, farklı dillerin özelliklerini öğrenebilen ve bu diller arasında transfer yapabilen modeller önem kazanmaktadır. Zhang ve arkadaşları tarafından geliştirilen USM modeli, 300'den fazla dili kapsayan 12 milyon saatlik etiketlenmemiş veri üzerinde ön-eğitime tabi tutulmuş ve FLEURS veri setinde 102 dil için %16.1 WER elde etmiştir [20]. Wang ve arkadaşları tarafından geliştirilen VIOLA modeli, konuşma tanıma, sentezi ve çevirisi gibi çeşitli görevleri tek bir model altında birleştirmekte ve İngilizce ASR görevinde %9.4 WER, Çince ASR görevinde %10.2 WER elde etmektedir [43]. Chu ve arkadaşları tarafından geliştirilen Qwen2-Audio modeli, Whisper-large-v3 tabanlı bir ses kodlayıcı ve Qwen-7B dil modelinden oluşmakta ve LibriSpeech'te %92.62, Aishell2'de %96.0 doğruluk oranları elde etmektedir [27].

4.5.5 Özel Uygulama Alanlarında Uygulanabilirlik

Literatürde incelenen çalışmalar, ses tanıma teknolojilerinin çeşitli özel uygulama alanlarında da uygulanabilirliğini göstermektedir. Rahate ve arkadaşları tarafından geliştirilen EEG tabanlı sessiz konuşma tanıma sistemi, felçli ve nöromüsküler hastalıklardan muzdarip hastalar için alternatif bir iletişim yöntemi sağlamaktadır [34]. KNN algoritması %61.45 doğruluk oranı ile en iyi performansı göstermiştir. Shisode ve arkadaşları tarafından konuşma engelli kişiler için geliştirilen yüzey elektromiyografi

(SEMG) tabanlı konuşma tanıma sistemi, GMM ve CNN-LSTM modellerini kullanmakta ve "Emergency" kelimesi için %80 doğruluk oranı elde etmektedir [37]. Kumar ve arkadaşları tarafından geliştirilen görsel konuşma tanıma (VSR) sistemi, dudak hareketlerini analiz ederek konuşulan metni tanımlamakta ve GRID Corpus üzerinde %2.8 WER elde etmektedir [39].

Chen tarafından geliştirilen İngilizce kelime öğrenimi için konuşma tanıma sistemi, HTK (Hidden Markov Model Toolkit) kullanarak fonem bazında hata tespiti yapabilmektedir [32]. Xie tarafından geliştirilen adaptif öğrenme sistemi, Naive Bayes algoritması kullanarak öğrencilerin öğrenme tercihlerine göre içerik nesnelerinin sunumunda %59.50 ile %90.53 arasında değişen doğruluk oranları elde etmiştir [33]. Guo tarafından geliştirilen çift sensörlü konuşma tanıma sistemi, çene derisi titreşim basıncı ve hava yoluyla iletilen konuşma sinyallerini toplayarak, özellikle gürültülü ortamlarda konuşma tanıma doğruluğunu artırmaktadır [1].

4.5.6 Genel Uygulanabilirlik Değerlendirmesi

İncelenen çalışmalar ışığında, ses tanıma alanında en etkili makine öğrenme yöntemlerinin uygulama alanına göre değişkenlik gösterdiği görülmektedir. Genel amaçlı konuşma tanıma sistemlerinde, uçtan uca modeller (CTC, Attention) ve derin öğrenme tabanlı yaklaşımlar (BLSTM, Transformer, Conformer) öne çıkarken, özel amaçlı ve alan-spesifik sistemlerde transfer öğrenme ve ince ayar yaklaşımları daha uygulanabilir görünmektedir.

Gömülü sistemlerde ve sınırlı kaynaklı ortamlarda, CNN tabanlı hafif modeller, klasik makine öğrenme algoritmaları (SVM, KNN) ve darbeli sinir ağları (SNN) gibi daha verimli yaklaşımlar uygulanabilirlik açısından öne çıkmaktadır. Çok dilli sistemlerde ise, büyük ön-eğitimli modeller (USM, VIOLA, Qwen2-Audio) ve diller arası transfer öğrenme yaklaşımları yüksek uygulanabilirlik potansiyeli taşımaktadır.

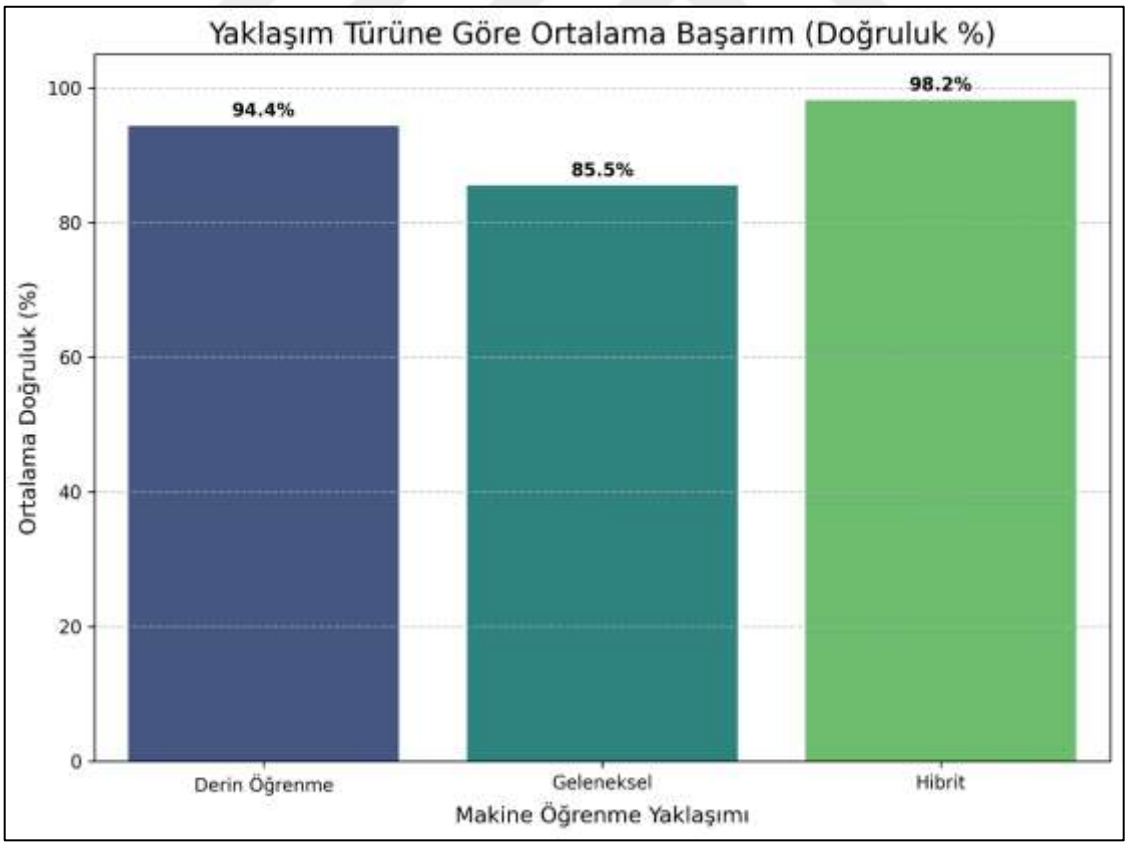
Özel uygulama alanlarında, uygulama alanının gereksinimlerine göre özelleştirilmiş hibrit yaklaşımlar (CNN-LSTM, CRNN) ve modalite-spesifik modeller (EEG-tabanlı, SEMG-tabanlı, görsel-tabanlı) daha uygulanabilir görünmektedir.

4.5.7 Bulguların Analizi

Bu başlıkta literatür taraması kapsamında incelenen çalışmaların karşılaştırmalı analizi sunulmaktadır. EK A'da yer alan çalışmalardan elde edilen veriler, alandaki temel eğilimleri, başarım ölçütlerini ve model karakteristikleri arasındaki ilişkileri görselleştirmek amacıyla Python tabanlı algoritma ile işlenmiştir. Analizde kullanılan derlenmiş veri dosyası EK B'de, analiz ve grafik çizim kodları ise EK C'de sunulmuştur. Takip eden alt bölümlerde, analizler sonucunda elde edilen bulgular yer almaktadır.

Yaklaşım Türlerine Göre Başarım Analizi

Çalışmada incelenen literatürdeki ses tanıma yaklaşımları; Derin Öğrenme, Geleneksel Makine Öğrenmesi ve Hibrit olmak üzere üç ana kategoride gruplandırılmıştır. Ses tanıma yaklaşımlarının genel performans eğilimlerini analiz etmek amacıyla, doğruluk (accuracy) metriği ile sonuç bildiren çalışmaların ortalama başarım oranları hesaplanmış ve Şekil 4.1'de görselleştirilmiştir.

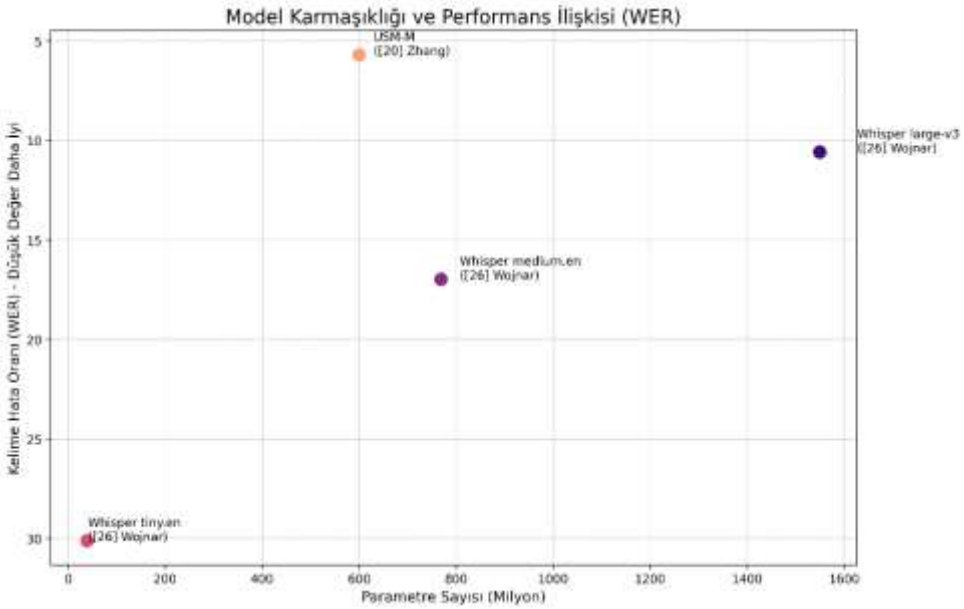


Şekil 4.1: Yaklaşım Türüne Göre Ortalama Başarım (Doğruluk %).

Şekil 4.1 incelendiğinde; %98.20'lik ortalama doğruluk oranı ile en yüksek başarıyı Hibrit modellerin gösterdiği ve bunu %94.35'lik ortalama başarımla Derin Öğrenme modellerinin takip ettiği görülmektedir. Geleneksel Makine Öğrenmesi yaklaşımlarının ortalama başarımı ise %85.52 olarak hesaplanmıştır. Elde edilen sonuç, farklı algoritmaların güçlü yönlerini birleştiren hibrit yaklaşımların, belirli görevlerde tek bir metodolojiye göre daha üstün sonuçlar verebildiğini göstermektedir. Aynı zamanda, derin öğrenme modellerinin karmaşık örüntüleri öğrenme kabiliyetleri sayesinde geleneksel yöntemlere kıyasla belirgin bir performans avantajı sağladığı da görülmektedir.

Model Karmaşıklığı ve Performans İlişkisi

Ses tanıma modellerinde performans ile model karmaşıklığı (parametre sayısı) arasındaki ilişki, özellikle donanım kaynakları ve çıkarım hızı açısından önemli bir değerlendirme kriteridir. Şekil 4.2, incelenen bazı modern derin öğrenme modellerinin parametre sayılarına karşılık gelen Kelime Hata Oranlarını (WER) göstermektedir. Grafikte y ekseni, düşük değerin daha iyi performansı ifade ettiği WER metriğini temsil etmektedir.



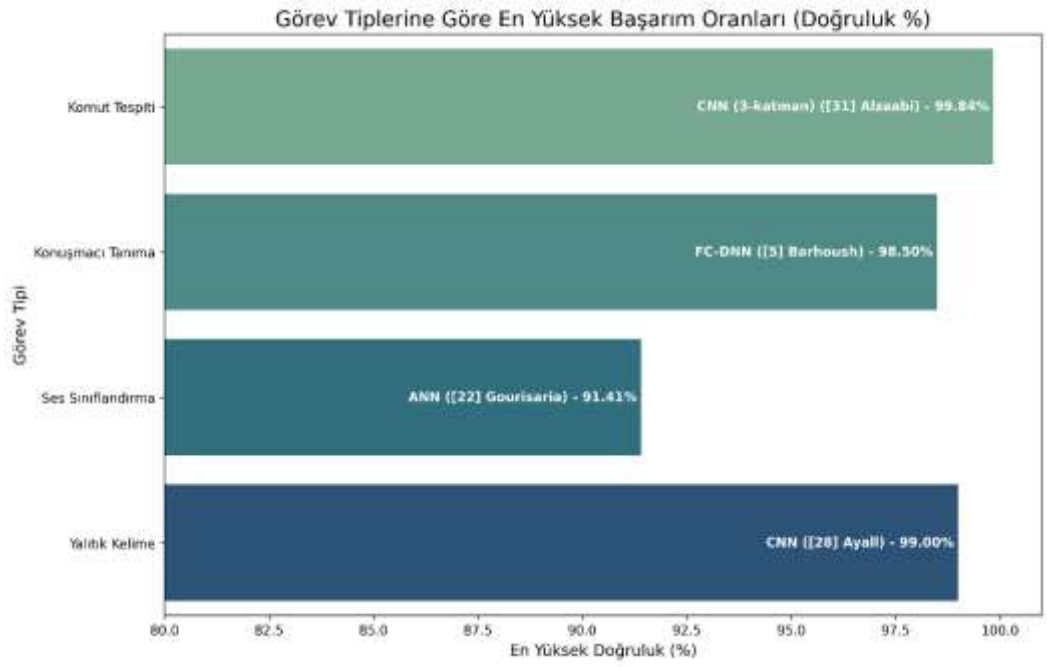
Şekil 4.2: Model Karmaşıklığı ve Performans İlişkisi (WER).

Şekil 4.2 incelendiğinde, en düşük WER değerine (%4.8) sahip olan USM-M modelinin [20], yaklaşık 600 milyon parametre ile oldukça üstün bir sonuç elde ettiği görülmektedir. Diğer yandan, Wojnar ve arkadaşları [26] tarafından değerlendirilen Whisper modelleri

arasında, yaklaşık 1.55 milyar parametre ile en karmaşık model olan Whisper large-v3'ün %10.58 WER değeri sergilediği rapor edilmiştir. En az parametreye sahip olan Whisper tiny.en modeli ise beklendiği gibi %43.64 ile en yüksek hata oranına sahiptir [26]. Elde edilen sonuç, model performansının yalnızca parametre sayısına bağlı olmadığını; model mimarisinin, eğitim verisinin büyüklüğü ve kalitesinin de önemli rol oynadığını göstermektedir. Özellikle USM-M modelinin başarısı, daha az parametre ile daha verimli mimarilerin tasarlanabileceğinin bir kanıtı niteliğindedir.

Görev Tiplerine Göre En Yüksek Başarımların İncelenmesi

Literatürdeki çalışmaların genel ses tanımadan ziyade belirli alt görevlere odaklandığı görülmüştür. Görevlerin zorluk seviyeleri ve hedefleri farklılık gösterdiğinden, elde edilen başarım oranları da değişkenlik göstermektedir. Şekil 4.3'te, incelenen makaleler arasında dört farklı görev tipi için rapor edilen en yüksek doğruluk başarımları ve bu başarımları elde eden modeller özetlemektedir.



Şekil 4.3: Görev Tiplerine Göre En Yüksek Başarım Oranları (Doğruluk %).

Şekil 4.3 incelendiğinde; "Komut Tespiti" gibi daha sınırlı bir kelime dağarcığına sahip görevlerde, 3 katmanlı bir Evrişimli Sinir Ağı (CNN) modeli ile %99.84 gibi çok yüksek bir doğruluğa ulaşıldığı [31] görülmektedir. Benzer şekilde, "Yalıtık Kelime Tanıma" görevinde

de bir CNN modeli ile %99.00'lık bir başarım elde edilmiştir [28]. "Konuşmacı Tanıma" görevinde ise Tam Bağlantılı Derin Sinir Ağı (FC-DNN) ile %98.50'lik bir sonuca ulaşılmıştır [5]. Daha genel bir problem olan "Ses Sınıflandırma" görevinde ise en yüksek başarım %91.41 olarak rapor edilmiştir [22]. Ulaşılan sonuçlar, görevin özelleşme düzeyine bağlı olarak başarım oranlarının değiştiğini ve özellikle belirli ve sınırlı görevlerde mevcut teknolojilerin neredeyse mükemmele yakın sonuçlar üretebildiğini göstermektedir.

5. SONUÇ VE ÖNERİLER

Bu çalışmada, makine öğrenme teknikleri kullanılarak ses tanıma alanındaki mevcut literatür incelenmiş ve bu alandaki farklı yöntemler karşılaştırılarak en etkili makine öğrenme yöntemleri belirlenmeye çalışılmıştır. Ses tanıma sistemlerinin gelişimi, geleneksel yöntemlerden derin öğrenme tabanlı yaklaşımlara doğru önemli bir evrim geçirmiştir. İncelenen literatür doğrultusunda, ses tanıma alanında derin öğrenme tabanlı yaklaşımların geleneksel makine öğrenme yöntemlerine kıyasla daha üstün performans sergilediği açıkça görülmektedir. Özellikle son yıllarda, derin öğrenme mimarilerinin ses tanıma doğruluğunda önemli derecede iyileştirme sağlanmış, ses tanıma sistemlerinin pratik uygulamalarda daha yaygın kullanımı mümkün hale gelmiştir.

Geleneksel yöntemler arasında Gizli Markov Modelleri (HMM) ve Gauss Karışım Modelleri (GMM) uzun süre ses tanıma alanında standart yaklaşımlar olarak kabul edilmiştir. Ancak, bu yöntemlerin karmaşık ses örüntülerini modellemedeki sınırlamaları, araştırmacıları daha gelişmiş teknikler aramaya yönlendirmiştir. Derin Sinir Ağları (DNN), Evrişimli Sinir Ağları (CNN) ve Tekrarlayan Sinir Ağları (RNN) gibi derin öğrenme mimarileri, ses sinyallerindeki karmaşık örüntüleri öğrenme konusunda üstün yetenekler sergilemişlerdir [14].

Son yıllarda, Transformer tabanlı modeller ve uçtan uca (end-to-end) öğrenme yaklaşımları, ses tanıma alanında çığır açan gelişmeler olarak öne çıkmıştır. Söz konusu modeller, geleneksel HMM tabanlı sistemlerin karmaşık yapısını basitleştirerek, daha tutarlı ve veri odaklı bir öğrenme süreci sağlamaktadır. Özellikle, uçtan uca modeller, akustik özellik çıkarımı, akustik modelleme, dil modelleme ve arama gibi klasik ASR bileşenlerini tek bir bütünleşik yapıda birleştirerek sistem performansını artırmaktadır [12].

Özellik çıkarma teknikleri açısından, Mel Frekans Kepstral Katsayıları (MFCC) hala yaygın olarak kullanılmakla birlikte, spektrogram tabanlı gösterimler ve dalgacık dönüşümü (wavelet transform) gibi alternatif yaklaşımlar da önem kazanmıştır. Derin öğrenme modellerinin ham ses verilerinden doğrudan özellik çıkarabilme yeteneği, manuel özellik mühendisliğine olan bağımlılığı azaltmıştır [23].

Ses tanıma sistemlerinin gelişiminde kaydedilen önemli ilerlemelere rağmen, bu alanda hala çeşitli problemler ve sınırlamalar bulunmaktadır. Derin öğrenme modellerinin etkili bir şekilde eğitilebilmesi için gerekli olan büyük miktarda etiketli veri ihtiyacı, özellikle az konuşulan diller için önemli bir engel teşkil etmektedir [4]. Veri toplama ve etiketleme süreçlerinin zaman alıcı ve maliyetli olması, yeni diller veya özel alanlar için ses tanıma sistemlerinin geliştirilmesini zorlaştırmaktadır. Ayrıca, derin öğrenme tabanlı ses tanıma sistemlerinin eğitim ve çıkarım aşamalarında gerektirdiği yüksek hesaplama gücü, sınırlı kaynaklara sahip ortamlarda bu sistemlerin kullanımını kısıtlamaktadır [17].

Gerçek dünya koşullarında, özellikle gürültülü ortamlarda ses tanıma sistemlerinin performansı hala önemli bir problem olarak ortaya çıkmaktadır. Arka plan gürültüsü, yankı ve çoklu konuşmacı senaryoları, tanıma doğruluğunu önemli ölçüde düşürebilmektedir [3]. Ses tanıma sistemlerinin farklı akustik ortamlara adaptasyonu için geliştirilen yöntemler, henüz tam anlamıyla tatmin edici sonuçlar vermemektedir. Benzer şekilde, farklı diller ve lehçeler için ses tanıma sistemlerinin geliştirilmesi, her dilin kendine özgü fonetik yapısı ve dilbilgisi kuralları nedeniyle zorlayıcı olmaya devam etmektedir. Dünya genelinde yaygın kullanım için tasarlanan sistemlerde dil çeşitliliği önemli bir engel oluşturmaktadır [44].

Mevcut modellerin genelleme yeteneği de geliştirilmeye açık bir alan olarak dikkat çekmektedir. Eğitim verilerinden farklı dağılımlara sahip test verileri üzerindeki performans, özellikle konuşmacıya bağımlı olmayan sistemlerde istenilen seviyede değildir. Domain adaptasyonu ve transfer öğrenme gibi teknikler bu sorunu hafifletmeye yardımcı olsa da, henüz tam bir çözüm sunamamaktadır [4]. Ayrıca, ses tanıma sistemlerinin kararlarını açıklayabilme yeteneğinin sınırlı olması, özellikle sağlık, güvenlik ve adli uygulamalar gibi alanlarda güven sorunları yaratabilmektedir. Kullanıcı gizliliği ve veri güvenliği konuları da, ses tanıma teknolojilerinin yaygınlaşmasıyla birlikte daha fazla önem kazanmaktadır.

Çalışmada yapılan literatür incelemesi sonucunda, ses tanıma alanında derin öğrenme tabanlı yaklaşımların, özellikle Transformer mimarileri ve uçtan uca öğrenme modellerinin, en etkili yöntemler olduğu değerlendirilmektedir. Söz konusu modeller, geleneksel HMM-GMM sistemlerine ve hatta erken dönem derin öğrenme yaklaşımlarına kıyasla daha yüksek doğruluk ve sağlamlık sergilemektedir.

Bununla birlikte, ideal bir ses tanıma sistemi için tek bir "en iyi" yaklaşımdan bahsetmek yerine, uygulama senaryosuna, mevcut kaynaklara ve hedef kullanıcı grubuna bağlı olarak farklı yaklaşımların kombinasyonunun daha uygun olabileceği değerlendirilmektedir. Örneğin, sınırlı kaynaklara sahip ortamlarda daha hafif modeller tercih edilebilirken, yüksek doğruluk gerektiren uygulamalarda daha karmaşık ve hesaplama yoğun modeller kullanılabilir.

Çalışma sonucunda gelecekteki araştırmalar için bazı öneriler sunulabilir. Düşük kaynaklı diller için yarı-denetimli ve öz-denetimli öğrenme yaklaşımları üzerine çalışmalar yürütülebilir. Özellikle Türkçe gibi morfolojik açıdan zengin diller için, etiketlenmemiş verilerden etkin şekilde yararlanma stratejileri geliştirilebilir. Hibrit model mimarileri üzerine araştırmalar derinleştirilebilir. CNN'lerin spektral-zamansal özellik çıkarma yetenekleri ile Transformer modellerinin uzun bağımlılıkları modelleme kapasitesinin birleştirilmesi, performans artışı sağlayabilir. Çok görevli ve çok modlu öğrenme yaklaşımları incelenebilir. Ses tanıma ile birlikte konuşmacı tanıma, duygu analizi gibi ilgili görevleri eş zamanlı gerçekleştirebilen mimariler üzerine çalışmalar yapılabilir. Nöromorfolojik hesaplama yaklaşımları araştırılabilir. Darbeli sinir ağları (SNN) gibi biyolojik olarak daha gerçekçi hesaplama modellerinin ses tanıma alanındaki potansiyeli değerlendirilebilir. Standartlaştırılmış değerlendirme metrikleri ve veri setleri üzerine çalışmalar yürütülebilir. Farklı diller ve uygulama alanları için ortak değerlendirme çerçeveleri oluşturulması, yöntemlerin adil karşılaştırılmasını sağlayabilir.

REFERANSLAR

[1] J. Guo, "Innovative Application of Sensor Combined with Speech Recognition Technology in College English Education in the Context of Artificial Intelligence", *Journal of Sensors*, c. 9281914, ss. 1-11, 2023, doi: 10.1155/2023/9281914.

[2] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, ve S. Watanabe, "End-to-end speech recognition: A survey", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, c. 32, ss. 325-353, 2023, doi: 10.1109/TASLP.2023.3328283.

[3] M. Malik, M. K. Malik, K. Mehmood, ve I. Makhdoom, "Automatic speech recognition: a survey", *Multimedia Tools and Applications*, c. 80, sy 6, ss. 9411-9457, 2021, doi: 10.1007/s11042-020-10073-7.

[4] H. Kheddar, Y. Himeur, S. Al-Maadeed, A. Amira, ve F. Bensaali, "Deep transfer learning for automatic speech recognition: Towards better generalization". 2023. doi: https://arxiv.org/abs/2304.14535v2.

[5] M. Barhoush, A. Hallawa, ve A. Schmeink, "Speaker identification and localization using shuffled MFCC features and deep learning", *International Journal of Speech Technology*, c. 26, ss. 185-196, 2023, doi: 10.1007/s10772-023-10023-2.

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

[7] K. Reid ve E. T. Williams, "Right the docs: Characterising voice dataset documentation practices used in machine learning". 2023. [Çevrimiçi]. Erişim adresi: https://arxiv.org/abs/2303.10721v1

[8] A. Mishra, R. Verma, N. Dhanda, ve K. K. Gupta, "Speech Recognition Using Machine Learning Techniques", içinde *2024 2nd International Conference on Disruptive Technologies (ICDT*, 2024, ss. 1142-1146. doi: 10.1109/ICDT61202.2024.10489508.

[9] S. Wang, "Recognition of English speech – using a deep learning algorithm", *Journal of Intelligent Systems*, c. 32, s. 20220236, 2023, doi: 10.1515/jisys-2022-0236.

[10] I. Ahmed, M. A. Irfan, A. Iqbal, A. Khalil, ve S. I. Siddiqui, "Efficient feature extraction and classification for the development of Pashto speech recognition system", *Multimedia Tools and Applications*, c. 83, ss. 54081-54096, 2024, doi: 10.1007/s11042-023-17684-w.

[11] İ. Z. Gençyılmaz ve K. M. Karaoğlan, "Optimizing Speech to Text Conversion in Turkish: An Analysis of Machine Learning Approaches", *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, c. 13, sy 2, ss. 492-504, 2024, doi: 10.17798/bitlisfen.1434925.

[12] A. H. Liu, W. N. Hsu, M. Auli, ve A. Baevski, "Towards end-to-end unsupervised speech recognition", içinde *2022 IEEE Spoken Language Technology Workshop (SLT*, IEEE, 2023, ss. 221-228.

[13] Z. Yao *vd.*, "Zipformer: A faster and better encoder for automatic speech recognition". 2024. [Çevrimiçi]. Erişim adresi: https://arxiv.org/abs/2310.11230v4

[14] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, ve S. Poria, "A review of deep learning techniques for speech processing", *Multimedia Tools and Applications*, c. 1-72, 2023, doi: 10.1016/j.inffus.2023.101755.

[15] S. Keser, "Makine Öğrenimi ve Hibrit Altuzay Sınıflandırıcılar için Yalıtık Kelime Tanıma Performanslarının Karşılaştırılması. Sürdürülebilir Mühendislik Uygulamaları ve Teknolojik Gelişmeler Dergisi", c. 6, sy 2. ss. 235-249, 2023. doi: 10.51764/smutgd.1338977.

[16] T. B. Mokgonyane, T. J. Sefara, T. I. Modipa, M. M. Mogale, M. J. Manamela, ve P. J. Manamela, "Automatic speaker recognition system based on machine learning algorithms", içinde *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA*, IEEE, 2019, ss. 141-146.

[17] A. S. Dhanjal ve W. Singh, "A comprehensive survey on automatic speech recognition using neural networks", *Multimedia Tools and Applications*, c. 83, ss. 23367-23412, 2023, doi: 10.1007/s11042-023-16438-y.

[18] S. Rai, T. Li, ve B. Lyu, “Keyword spotting - Detecting commands in speech using deep learning”. 2023.

[19] Y. Jia, “A Deep Learning System for Domain-Specific Speech Recognition”. 2023. doi: https://arxiv.org/abs/2303.10510v2.

[20] Y. Zhang *vd.*, “Google USM: Scaling automatic speech recognition beyond 100 languages”. 2023. doi: https://arxiv.org/abs/2303.01037v3.

[21] Y. Dokuz ve Z. Tüfekci, “Feature-based hybrid strategies for gradient descent optimization in end-to-end speech recognition”, *Multimedia Systems*, c. 81, sy 7, ss. 9969-9988, 2022, doi: 10.1007/s00530-022-00915-9.

[22] M. K. Gourisaria, R. Agrawal, M. Sahni, ve P. K. Singh, “Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques”, *Discover Internet of Things*, c. 4, sy 1, ss. 1-24, 2024, doi: 10.1007/s43926-023-00049-y.

[23] K. Zaman, M. Sah, C. Direkoglu, ve M. Unoki, “A survey of audio classification using deep learning”, *IEEE Access*, c. 11, ss. 106620-106621, 2023, doi: 10.1109/ACCESS.2023.3318015.

[24] B. Accou, J. Vanthornhout, H. hamme, ve T. Francart, “Decoding of the speech envelope from EEG using the VLAAI deep neural network”, *Scientific Reports*, c. 13, sy 812, ss. 1-14, 2023, doi: 10.1038/s41598-022-27332-2.

[25] R. Ganchev, *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, 2021.

[26] T. Wojnar, J. Hryszko, ve A. Roman, “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*, c. 2024, sy 24, 2024, doi: 10.1186/s13636-024-00343-9.

[27] Y. Chu *vd.*, “Qwen2-Audio technical report”. 2024. [Çevrimiçi]. Erişim adresi: https://arxiv.org/abs/2407.10759v1

[28] T. A. Ayall, C. Zhou, H. Liu, G. M. Brhanemeskel, S. T. Abate, ve M. Adjeisah, "Amharic spoken digits recognition using convolutional neural network", *Journal of Big Data*, c. 11, sy 64, 2024, doi: 10.1186/s40537-024-00910-z.

[29] H. Kwon, "AudioGuard: Speech Recognition System Robust against Optimized Audio Adversarial Examples", *Multimedia Tools and Applications*, c. 83, ss. 57943-57962, 2024, doi: 10.1007/s11042-023-15961-2.

[30] A. Shukla, A. Aanand, ve S. Nithiya, "Automatic Speech Recognition using Machine Learning Techniques", içinde *2023 International Conference on Computer Communication and Informatics (ICCCI*, 2023, ss. 1-6. doi: 10.1109/ICCCI56745.2023.10128212.

[31] S. Alzaabi *vd.*, "Bilingual Speech Recognition On the Edge Using Machine Learning", içinde *2024 7th International Conference on Signal Processing and Information Security (ICSPIS*, 2024, ss. 1-6. doi: 10.1109/ICSPIS63676.2024.10812600.

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

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

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

[35] M. Hamian, K. Faez, S. Nazari, ve M. Sabeti, "A novel learning approach in deep spiking neural networks with multi-objective optimization algorithms for automatic digit speech recognition", *The Journal of Supercomputing*, c. 79, ss. 20263-20288, 2023, doi: 10.1007/s11227-023-05420-y.

[36] Z. J. M. Ameen ve A. A. Kadhim, "Machine learning for Arabic phonemes recognition using electrolarynx speech", *International Journal of Electrical and Computer Engineering (IJECE*, c. 13, sy 1, ss. 400-412, 2023, doi: 10.11591/ijece.v13i1.pp400-412.

[37] S. Shisode *vd.*, "SEMG approach for speech recognition", *International Journal of Advanced Research in Computer Science*, c. 14, sy 3, ss. 23-28, 2023, doi: 10.26483/ijarcs.v14i3.6970.

[38] I. Froiz-Míguez *vd.*, "Design and evaluation of a cross-lingual ML-based automatic speech recognition system fine-tuned for the Galician language", *Kalpa Publications in Computing*, c. 14, ss. 152-155, 2023.

[39] L. A. Kumar, D. K. Renuka, ve M. C. S. Priya, "Towards robust speech recognition model using deep learning", içinde *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS*, IEEE, 2023, ss. 253-256. doi: 10.1109/ICISCOIS56541.2023.10100390.

[40] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu, ve G. Y. Li, "Deep learning enabled semantic communications with speech recognition and synthesis", *IEEE Transactions on Wireless Communications*, c. 22, sy 9, ss. 6227-6240, 2023, doi: 10.1109/TWC.2023.3240969.

[41] Z. Chen *vd.*, "SALM: Speech-augmented language model with in-context learning for speech recognition and translation". 2023. [Çevrimiçi]. Erişim adresi: https://arxiv.org/abs/2310.09424v1

[42] Y. Gong, A. H. Liu, H. Luo, L. Karlinsky, ve J. Glass, "Joint audio and speech understanding". 2023. doi: https://arxiv.org/abs/2309.14405v3.

[43] T. Wang *vd.*, "VIOLA: Unified codec language models for speech recognition, synthesis, and translation". 2023. [Çevrimiçi]. Erişim adresi: https://arxiv.org/abs/2305.16107v1

[44] S. Watve, M. Patil, ve A. Shinde, "Review of features and classification for spoken Indian language recognition using deep learning and machine learning techniques", içinde

2023 International Conference on Emerging Smart Computing and Informatics (ESCI), 1-2, IEEE, 2023. doi: 10.1109/ESCI56872.2023.10099742.

EK A

A.1 LİTERATÜR TARAMASINDA İNCELENEN ÇALIŞMALAR

Çalışma Bilgileri	Yaklaşım Türü	Özellik Çıkarma Yöntemi	Kullanılan Algoritma/Model	Veri Seti	Değerlendirme Metrikleri	Değerlendirme Sonuçları	Uygulama Alanı
Shukla ve arkadaşları [30]	Hibrit Yaklaşımlar: HMM-DNN Uçtan Uca (End-to-End): CTC	Frekans Alanı Özellikleri: MFCC	İstatistiksel Modeller: HMM, GMM, N-gram Derin Sinir Ağları: CNN Uçtan Uca Modeller: CTC Hibrit Mimariler: BLSTM	Genel Amaçlı Veri Setleri: LibriSpeech Switchboard (300 saat İngilizce)	Doğruluk Metrikleri: WER (Word Error Rate)	Soft Forgetting ile WER'de %7-9 oranında iyileşme Hub5-2000 Switchboard/CallHome test setlerinde speaker-independent modeller için %9.1/%17.4 WER Speaker-adapted modeller için %8.7/%16.8 WER	Genel Amaçlı ASR: Büyük Kelime Dağarcıklı Sürekli Konuşma Tanıma (LVCSR)
Gourisaria ve arkadaşları [22]	Geleneksel Makine Öğrenme: SVM, KNN, Karar Ağaçları, Naive Bayes, Lojistik Regresyon, Random Forest Derin Öğrenme: ANN	Frekans Alanı Özellikleri: MFCC, STFT	Sınıflandırıcılar: SVM, KNN, Karar Ağaçları, Naive Bayes, Lojistik Regresyon, Random Forest Derin Sinir Ağları: ANN	UrbanSound8K (8732 ses örneği, 10 sınıf) Sound Event Audio Dataset (1288 ses örneği, 8 sınıf) Veri Seti Özellikleri: Çevresel sesler, değişken gürültü seviyeleri, 4 saniye süre	Doğruluk Metrikleri: Doğruluk, Kesinlik, Duyarlılık, Özgüllük, F1-skoru	UrbanSound8K veri seti sonuçları: • ANN: %91.41 doğruluk, %91.42 kesinlik, %91.41 duyarlılık, %99.05 özgüllük, %91.41 F1-skoru • SVM: %87.23 doğruluk, %87.25 kesinlik, %87.23 duyarlılık, %98.58 özgüllük, %87.24 F1-skoru • Random Forest: %85.71 doğruluk, %85.73 kesinlik, %85.71 duyarlılık, %98.41 özgüllük, %85.72 F1-skoru • KNN: %83.19 doğruluk, %83.21 kesinlik, %83.19 duyarlılık, %98.13 özgüllük, %83.20 F1-skoru • Karar Ağacı: %79.14 doğruluk, %79.16 kesinlik, %79.14 duyarlılık, %97.68 özgüllük, %79.15 F1-skoru • Lojistik Regresyon: %76.64 doğruluk, %76.66 kesinlik, %76.64 duyarlılık, %97.40 özgüllük, %76.65 F1-skoru • Naive Bayes: %72.39 doğruluk, %72.41 kesinlik, %72.39 duyarlılık, %96.93 özgüllük, %72.40 F1-skoru Sound Event Audio Dataset sonuçları: • ANN: %91.27 doğruluk, %91.44 kesinlik, %91.28 duyarlılık, %98.75 özgüllük, %91.35 F1-skoru	Akıllı Sistemler: • Ses dosyası yönetimi • Güvenlik sistemleri • Robotik endüstrisi • Akıllı yollar • Hastaneler • Endüstriyel uygulamalar

Çalışma Bilgileri	Yaklaşım Türü	Özellik Çıkarma Yöntemi	Kullanılan Algoritma/Model	Veri Seti	Değerlendirme Metrikleri	Değerlendirme Sonuçları	Uygulama Alanı
						• SVM: %86.02 doğruluk, %86.19 kesinlik, %86.03 duyarlılık, %98.00 özgüllük, %86.10 F1-skoru • Random Forest: %84.39 doğruluk, %84.56 kesinlik, %84.40 duyarlılık, %97.77 özgüllük, %84.47 F1-skoru • KNN: %81.87 doğruluk, %82.04 kesinlik, %81.88 duyarlılık, %97.41 özgüllük, %81.95 F1-skoru • Karar Ağacı: %77.51 doğruluk, %77.68 kesinlik, %77.52 duyarlılık, %96.79 özgüllük, %77.59 F1-skoru • Lojistik Regresyon: %74.37 doğruluk, %74.54 kesinlik, %74.38 duyarlılık, %96.34 özgüllük, %74.45 F1-skoru • Naive Bayes: %70.62 doğruluk, %70.79 kesinlik, %70.63 duyarlılık, %95.80 özgüllük, %70.70 F1-skoru Gürültü temizleme işlemi sınıflandırma performansını önemli ölçüde artırmıştır.	
Gençyılmaz ve Karaoğlan [11]	Derin Öğrenme: CNN, CRNN Geleneksel Makine Öğrenme: SVM, KNN, Karar Ağaçları, Random Forest, GNB Hibrit Yaklaşımlar: CRNN	Frekans Alanı Özellikleri: MFCC	Derin Sinir Ağları: CNN, CRNN RNN: LSTM, GRU Sınıflandırıcılar: SVM, KNN, Karar Ağaçları (DT), Random Forest (RF), Gaussian Naive Bayes (GNB)	Komut Tanıma Veri Seti: 26.485 Türkçe ses kaydı (14 farklı komut) Veri Seti Özellikleri: Her kayıt 1 saniye uzunluğunda, 16 kHz örnekleme frekansı	Doğruluk Metrikleri: Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall), F1-skoru, ROC eğrisi, AUC Çapraz doğrulama skoru	CNN: %95.92 doğruluk, %97.68 kesinlik, %97.14 duyarlılık, %97.55 F1-skoru CRNN: %96.47 doğruluk, %96.19 kesinlik, %96.17 duyarlılık, %96.16 F1-skoru LSTM: %72.97 doğruluk, %74.83 kesinlik, %72.92 duyarlılık, %71.21 F1-skoru RF: %92.69 doğruluk, %92.68 kesinlik, %92.69 duyarlılık, %92.67 F1-skoru SVM: %90.98 doğruluk, %90.97 kesinlik, %90.99 duyarlılık, %90.96 F1-skoru KNN: %90.07 doğruluk, %90.36 kesinlik, %90.08 duyarlılık, %90.14 F1-skoru GNB: %79.69 doğruluk, %80.10 kesinlik, %79.66 duyarlılık, %79.66 F1-skoru DT: %73.53 doğruluk, %73.51 kesinlik, %73.52 duyarlılık, %73.49 F1-skoru GRU: %87.71 doğruluk, %88.29 kesinlik, %87.71 duyarlılık, %87.68 F1-skoru	Komut/Anahtar Kelime Tanıma: Akıllı cihazlar, ses komut tanıma Genel Amaçlı ASR: Türkçe dilinde konuşmadan metne dönüşüm Özel Uygulamalar: İşitme engelliler için destek sistemleri, otomatik altyazı oluşturma

Çalışma Bilgileri	Yaklaşım Türü	Özellik Çıkarma Yöntemi	Kullanılan Algoritma/Model	Veri Seti	Değerlendirme Metrikleri	Değerlendirme Sonuçları	Uygulama Alanı
Ayall ve arkadaşları [28]	Derin Öğrenme: CNN Geleneksel Makine Öğrenme: LDA, KNN, SVM, RF (karşılaştırma amaçlı)	Frekans Alanı Özellikleri: MFCC, Mel-spektrogram	Derin Sinir Ağları: - CNN: 3 katmanlı CNN mimarisi (Batch normalization ile) Sınıflandırıcılar (karşılaştırma için): - SVM - KNN - Random Forest - LDA	Alan-Spesifik Veri Seti: Amharic Spoken Digits Dataset (AmSDD) Veri Seti Boyutu: 12.000 konuşma kaydı, 120 konuşmacı (her biri 10 rakamı 10'ar kez seslendirmiş) Veri Seti Özellikleri: Farklı yaş grupları, cinsiyetler ve lehçelerden konuşmacılar	Doğruluk Metrikleri: Doğruluk (Accuracy)	CNN + MFCC: %99 doğruluk CNN + Mel-spektrogram: %98 doğruluk Karşılaştırma sonuçları (geleneksel yöntemler): - LDA + MFCC: %93 - KNN + MFCC: %95 - SVM + MFCC: %96 - RF + MFCC: %97	Komut/Anahtar Kelime Tanıma: Rakam tanıma sistemleri Düşük Kaynaklı Diller: Amharca için konuşma tanıma Alan-Spesifik ASR: Telefon tabanlı hizmetler, arama sistemleri, banka işlemleri, havayolu rezervasyon sistemleri
Wojnar ve arkadaşları [26]	Derin Öğrenme: Transformatör, Conformer Uçtan Uca (End-to-End): RNN-T	Belirtilmemiş (Değerlendirilen modellerin kendi özellik çıkarma yöntemleri kullanılmıştır)	Pre-trained Modeller: - OpenAI Whisper (tiny.en, base.en, small.en, medium.en, large-v1, large-v3) - NVIDIA Conformer-Transducer X-Large (NeMo Transducer Xlarge) - ESPnet2	YouTube videoları (141 rastgele seçilmiş video) Veri Seti Özellikleri: Çeşitli diller, aksanlar, lehçeler, konuşma stilleri ve ses kalitesi seviyeleri	Doğruluk Metrikleri: - WER (Word Error Rate - Kelime Hata Oranı)	Whisper modelleri: - tiny.en: %43.64 WER - base.en: %33.53 WER - small.en: %23.75 WER - medium.en: %16.97 WER - large-v1: %13.42 WER - large-v3: %10.58 WER NeMo Transducer Xlarge: %26.06 WER ESPnet2: %30.97 WER	Genel Amaçlı ASR: Medya transkripsiyon Çok Dilli ASR: Farklı dillerde, aksanlarda ve lehçelerde konuşma tanıma
Alzaabi ve arkadaşları [31]	Derin Öğrenme: CNN	Frekans Alanı Özellikleri: Spektrogramlar (FFT kullanılarak)	Derin Sinir Ağları: CNN (2, 3, 4 ve 5 konvolüsyonel katmanlı farklı modeller)	Özel Oluşturulmuş Veri Seti: - İngilizce: 527 ses örneği - Arapça (Emirlik lehçesi): 680 ses örneği - Toplam: 40 farklı konuşmacıdan toplanan veriler - Veri Seti Özellikleri: Çeşitli arka plan gürültüleri eklenerek zenginleştirilmiş	Doğruluk Metrikleri: - Doğruluk (Accuracy) - Kesinlik (Precision) - Duyarlılık (Recall)	2 Konvolüsyonel Katman: - 30 epoch: %99.49 doğruluk, %99.54 kesinlik, %99.44 duyarlılık - 40 epoch: %99.54 doğruluk, %99.63 kesinlik, %99.45 duyarlılık - 45 epoch: %99.69 doğruluk, %99.72 kesinlik, %99.66 duyarlılık 3 Konvolüsyonel Katman: - 30 epoch: %99.84 doğruluk, %99.87 kesinlik, %99.82 duyarlılık - 40 epoch: %99.73 doğruluk, %99.76 kesinlik, %99.68 duyarlılık - 45 epoch: %99.78 doğruluk, %99.80 kesinlik, %99.78 duyarlılık 4 Konvolüsyonel Katman: - 30 epoch: %99.65 doğruluk, %99.72 kesinlik, %99.57 duyarlılık - 40 epoch: %99.58 doğruluk, %99.65	Komut/Anahtar Kelime Tanıma: Akıllı ev cihazları Çok Dilli ASR: İki dilli sistem (İngilizce ve Arapça-Emirlik lehçesi) Alan-Spesifik ASR: Engelli bireyler için yardımcı teknoloji

Çalışma Bilgileri	Yaklaşım Türü	Özellik Çıkarma Yöntemi	Kullanılan Algoritma/Model	Veri Seti	Değerlendirme Metrikleri	Değerlendirme Sonuçları	Uygulama Alanı
						kesinlik, %99.50 duyarlılık - 45 epoch: %99.81 doğruluk, %99.84 kesinlik, %99.77 duyarlılık 5 Konvolüsyonel Katman: - 30 epoch: %98.68 doğruluk, %98.98 kesinlik, %98.38 duyarlılık - 40 epoch: %98.71 doğruluk, %98.97 kesinlik, %98.43 duyarlılık - 45 epoch: %97.68 doğruluk, %98.02 kesinlik, %97.84 duyarlılık	
Barhoush ve arkadaşları [5]	Derin Öğrenme: DNN Uçtan Uca (End-to-End)	Frekans Alanı Özellikleri: MFCC Önerilen yeni özellikler: Karıştırılmış MFCC (SHMFCC) ve Fark Karıştırılmış MFCC (DSHMFCC)	Derin Sinir Ağları: - Tam Bağlantılı Derin Sinir Ağı (FC-DNN) (6 gizli katman, her katmanda 512 nöron, sigmoid aktivasyon fonksiyonu)	Simüle Edilmiş Veri Seti: - LibriSpeech veri setinden alınan ses örnekleri - Farklı konumlarda (37 farklı konum, 5° aralıklarla) simüle edilmiş konuşmacılar - Veri Seti Özellikleri: Farklı gürültü seviyeleri (SNR) ve yankılanma koşulları altında test edilmiştir	Doğruluk Metrikleri: - Doğruluk (Accuracy) - Kesinlik (Precision) - Duyarlılık (Recall) - F1-skoru Robustluk Metrikleri: - Farklı gürültü seviyelerinde ve yankılanma koşullarında performans	Tek konuşmacı senaryosu: - Konuşmacı tanıma: %98.5 doğruluk - Konuşmacı lokalizasyonu: %99.6 doğruluk Çoklu konuşmacı senaryosu (2 konuşmacı): - Konuşmacı tanıma: %95.2 doğruluk - Konuşmacı lokalizasyonu: %98.7 doğruluk Farklı gürültü seviyelerinde (SNR 0-30 dB): - 0 dB SNR'de bile %90'ın üzerinde doğruluk Geleneksel yöntemlerle karşılaştırma: - SHMFCC+FC-DNN, SRP-PHAT ve MUSIC gibi geleneksel yöntemlerden daha iyi performans göstermiştir	Konuşmacı Tanıma/Doğrulama: Biyometrik sistemler Konuşmacı Lokalizasyonu: Akıllı cihazlar, video konferans sistemleri Potansiyel Uygulamalar: İnsan-bilgisayar etkileşimi, gözetim sistemleri, robotik
Chen [32]	Geleneksel Makine Öğrenme: HMM	Zaman Alanı Özellikleri: Enerji, Sıfır Geçiş Oranı (ZCR) Frekans Alanı Özellikleri: MFCC LPC Tabanlı Özellikler: LPC	İstatistiksel Modeller: HMM (Gizli Markov Modeli) HTK (Hidden Markov Model Toolkit)	Özel Oluşturulmuş Veri Seti: - İngilizce konuşma örnekleri - İngilizce seviye 6 testini geçmiş katılımcılardan toplanan veriler - Referans olarak Amerikalı bir uluslararası öğrenciden alınan ses örnekleri	Doğruluk Metrikleri: - Hata tespiti (Fonem bazında) - Eksik okuma tespiti	Çalışmada, sesli ünsüzlerin daha kısa süreli telaffuz edilmesinden kaynaklanan hata oranları tespit edilmiştir. Farklı fonemlere yönelik farklı hata tolerans mekanizmaları önerilmiştir.	Alan-Spesifik ASR: Eğitim İngilizce kelime öğrenimi için konuşma tanıma sistemi
Xie [33]	Geleneksel Makine Öğrenme: Naive Bayes	Zaman Alanı Özellikleri: Kısa süreli enerji, Sıfır geçiş oranı (ZCR) Frekans Alanı	İstatistiksel Modeller: DTW (Dinamik Zaman Bükme), HMM (Gizli Markov Modeli)	Özel Oluşturulmuş Veri Seti: - İngilizce konuşma örnekleri - Veri Seti Özellikleri: Çeşitli	Doğruluk Metrikleri: - Doğruluk (Accuracy) - Sınıflandırma doğruluğu	Naive Bayes sınıflandırıcısının doğruluk oranları (8 öğrenci için): - Öğrenci 1: %84.11 - Öğrenci 2: %74.17	Alan-Spesifik ASR: Eğitim İngilizce konuşma

Çalışma Bilgileri	Yaklaşım Türü	Özellik Çıkarma Yöntemi	Kullanılan Algoritma/Model	Veri Seti	Değerlendirme Metrikleri	Değerlendirme Sonuçları	Uygulama Alanı
		Özellikleri: MFCC LPC Tabanlı Özellikler: LPC	Sınıflandırıcılar: Naive Bayes	konuşma komutları ("aç", "ön", "arka", "sol", "sağ" vb.)		- Öğrenci 3: %60.06 - Öğrenci 4: %76.98 - Öğrenci 5: %84.71 - Öğrenci 6: %59.50 - Öğrenci 7: %84.62 - Öğrenci 8: %90.53 Farklı eşik değerlerinde iyileştirilmiş sonuçlar: - Öğrenci 8 için en iyi sonuç: %92.63 (eşik değeri: 90)	eğitimi için adaptif öğrenme sistemi
Keser [15]	Geleneksel Makine Öğrenme: SVM, KNN Derin Öğrenme: CNN, RNN-LSTM Hibrit Yaklaşımlar: (DCVA)PCA, (FLDA-KNN)PCA	Frekans Alanı Özellikleri: MFCC PLP Tabanlı Özellikler: PLP	Sınıflandırıcılar: KNN, SVM Altuzay Sınıflandırıcılar: FLDA-KNN, DCVA Hibrit Sınıflandırıcılar: (DCVA)PCA, (FLDA-KNN)PCA Derin Sinir Ağları: - CNN: 5 evrişimsel katman, tam bağlantılı katman, SoftMax katmanı - RNN: LSTM mimarisi	Özel Oluşturulmuş Veri Seti: - 18 sınıf (kelime) - Her sınıf için 1-5 arası konuşma sinyali - Toplam 150 konuşma sinyali - Veri Seti Özellikleri: Farklı gürültü türlerini içeren konuşma veritabanı	Doğruluk Metrikleri: - Doğruluk (Accuracy) Robustluk Metrikleri: - Gürültülü ortamlarda tanıma performansı	RNN-LSTM: %93.22 CNN: %87.56 KNN: %86.51 (FLDA-KNN)PCA: %84.90 (DCVA)PCA: %79.00 SVM: %77.78 DCVA: %74.23 FLDA-KNN: %71.37	Komut/Anahtar Kelime Tanıma: Akıllı ev sistemleri, robot ve araç kontrolü Konuşmacıdan bağımsız yalıtık kelime tanıma
Guo [1]	Derin Öğrenme: RNN, LSTM Hibrit Yaklaşımlar: DNN-HMM	LPC Tabanlı Özellikler: LPC Sensör tabanlı özellikler: Çene derisi titreşim basıncı ve hava yoluyla iletilen konuşma sinyalleri	Derin Sinir Ağları: - RNN: LSTM - Hibrit Mimariler: DNN-HMM	Özel Oluşturulmuş Veri Seti: - İngilizce konuşma örnekleri - Çift sensörlü sistem ile toplanan veriler (çene derisi titreşim basıncı ve hava yoluyla iletilen konuşma sinyalleri) - Veri Seti Özellikleri: Gürültülü ortam koşullarında toplanan veriler	Doğruluk Metrikleri: - Doğruluk (Accuracy) - Tanıma hızı Robustluk Metrikleri: - Gürültülü ortamlarda tanıma performansı	Çalışmada, önerilen RNN (LSTM) tabanlı konuşma tanıma modelinin, geleneksel konuşma tanıma algoritmalarına kıyasla daha yüksek doğruluk ve daha hızlı yakınsama sağladığı belirtilmiştir. Çift sensörlü sistem kullanımı, özellikle gürültülü ortamlarda konuşma tanıma doğruluğunu artırmıştır.	Alan-Spesifik ASR: Eğitim Üniversite İngilizce eğitiminde konuşma tanıma teknolojisinin uygulanması
Jia [19]	Uçtan Uca (End-to-End): CTC Transfer Öğrenimi: Pre-trained modeller üzerinde fine-tuning	Otomatik Özellik Öğrenme: Raw waveform (Wav2Vec2) Spektrogram (DeepSpeech2)	Pre-trained Modeller: Wav2Vec2-Large-LV60, DeepSpeech2 İstatistiksel Modeller: KenLM (n-gram dil modeli)	Alan-Spesifik Veri Seti: Sağlık sigortası ve çalışan hakları alanında çağrı merkezi verileri (BSCD) Veri Seti Özellikleri: - Yapay veri seti: 13 farklı kişiden sessiz ortamda kaydedilmiş konuşmalar - Gerçek veri seti: Çağrı	Doğruluk Metrikleri: - WER (Kelime Hata Oranı) - CER (Karakter Hata Oranı) - Göreceli Hata Oranı (Relative Error Rate) SLU Performans	- Fine-tuned Wav2Vec2-Large-LV60 + KenLM, Google ve AWS ASR sistemlerinden daha iyi performans göstermiştir - Doğal dil anlama görevlerinde, alan-spesifik fine-tuned ASR sistemi, daha yüksek WER'e sahip olmasına rağmen ticari ASR sistemlerinden daha iyi performans göstermiştir	Alan-Spesifik ASR: Sağlık sigortası ve çalışan hakları alanında konuşma tanıma sistemi Konuşma Dili Anlama (SLU): Niyet sınıflandırma

Çalışma Bilgileri	Yaklaşım Türü	Özellik Çıkarma Yöntemi	Kullanılan Algoritma/Model	Veri Seti	Değerlendirme Metrikleri	Değerlendirme Sonuçları	Uygulama Alanı
				merkezinden alınan, çeşitli demografik özelliklere ve gürültü koşullarına sahip gerçek konuşmalar - Yarı denetimli öğrenme ile etiketlenmiş veriler	Metrikleri: - Doğal dil anlama görevlerinde başarı oranı	- İnsan transkripsiyonları ile fine-tuned ASR arasındaki sonuçlar benzerdir	
Rahate ve arkadaşları [34]	Geleneksel Makine Öğrenme: SVM, KNN, Karar Ağaçları	Zaman Alanı Özellikleri: Ampirik Mod Ayrıştırma (EMD) ile elde edilen içsel mod fonksiyonlarından (IMF) çıkarılan doğrusal ve doğrusal olmayan zaman alanı özellikleri İstatistiksel Özellikler: Ortalama, medyan, varyans, çarpıklık, Shannon entropisi, RMS, çeyrekler arası aralık, ortalama mutlak sapma	Sınıflandırıcılar: - Doğrusal Diskriminant Analizi (LDA) - Destek Vektör Makineleri (SVM) - K-En Yakın Komşu (KNN) - Karar Ağaçları (DT) - Bagging - Boosting - Naive Bayes	Özel Oluşturulmuş Veri Seti: - 10 sağlıklı erkek denek (22 yaşında) - 32 kanallı EEG kayıtları (256 Hz örnekleme) - 6 farklı kelime: "HELP", "LIGHT", "PAIN", "STOP", "YES", "NO" - Her denek 10-15 oturum gerçekleştirmiş - Her oturum yaklaşık 50 dakika	Doğruluk Metrikleri: - Sınıflandırma doğruluğu - Kesinlik (Precision) - Duyarlılık (Recall) - F1-skoru - Özgüllük (Specificity)	Doğruluk oranları: - KNN: %61.45 - SVM: %61.25 - Boosting: %60.19 - Bagging: %60.06 - Naive Bayes: %58.55 - LDA: %57.69 - Karar Ağacı: %56.97 Kesinlik (Precision): - LDA: 0.7947 - SVM: 0.6227 - Karar Ağacı: 0.5171 - Bagging: 0.6394 - Boosting: 0.584 - Naive Bayes: 0.6158 - KNN: 0.6381 Duyarlılık (Recall): - LDA: 0.5536 - SVM: 0.5952 - Karar Ağacı: 0.5857 - Bagging: 0.591 - Boosting: 0.60573 - Naive Bayes: 0.5806 - KNN: 0.6093 F1-skoru: - LDA: 0.6526 - SVM: 0.60864 - Karar Ağacı: 0.548 - Bagging: 0.614 - Boosting: 0.5946 - Naive Bayes: 0.5976 - KNN: 0.6233 Özgüllük (Specificity):	Alan-Spesifik ASR: Tıbbi Beyin-Bilgisayar Arayüzü (BCI) Sessiz konuşma tanıma Felçli veya nöromüsküler hastalıklardan muzdarip hastalar için iletişim sistemi

Çalışma Bilgileri	Yaklaşım Türü	Özellik Çıkarma Yöntemi	Kullanılan Algoritma/Model	Veri Seti	Değerlendirme Metrikleri	Değerlendirme Sonuçları	Uygulama Alanı
						- LDA: 0.6363 - SVM: 0.6055 - Karar Ağacı: 0.5855 - Bagging: 0.6074 - Boosting: 0.5984 - Naive Bayes: 0.59103 - KNN: 0.6201 Karışıklık Matrisi Sonuçları: - LDA: [604, 256; 487, 283] - SVM: [472, 286; 321, 439] - Karar Ağacı: [393, 367; 278, 482] - Bagging: [486, 274; 336, 424] - Boosting: [444, 316; 289, 471] - Naive Bayes: [468, 292; 338, 422] - KNN: [485, 275; 311, 449]	
Accou ve arkadaşları [24]	Derin Öğrenme	Gammatone filtre bankası kullanılarak konuşma zarfı çıkarımı	VLAAI (Very Large Augmented Auditory Inference) ağı: Çoklu CNN blokları, tam bağlantılı katmanlar ve çıktı bağlam katmanından oluşan özel bir mimari	Single-speaker stories (80 katılımcı, 141 saat), Holdout (26 katılımcı, 46 saat) ve DTU (18 katılımcı, 500 saniye/katılımcı) veri setleri	Pearson korelasyon katsayısı	VLAAI: 0.19 medyan Pearson r (doğrusal modele göre %52 iyileşme); Fine-tuning sonrası: 0.25 medyan Pearson r (özgün doğrusal modele göre %54 iyileşme)	Nöral Konuşma İşleme: EEG sinyallerinden konuşma zarfının çözümlenmesi, İşitsel Dikkat Tespiti, İşitme Testleri
Wang [9]	Uçtan Uca (End-to-End): CTC	Frekans Alanı Özellikleri: MFCC	RNN-CTC, Karşılaştırma için: GMM-HMM ve CNN-CTC	TIMIT: 630 katılımcı, 6.300 cümle, 16 kHz ve 16 bit örnekleme	Kelime Hata Oranı (WER), Eğitim süresi, Test süresi	2.500 test örneği ile: RNN-CTC: %10.4 WER, 20.2 s test; CNN-CTC: %15.8 WER, 33.1 s test; GMM-HMM: %21.1 WER, 42.5 s test. Eğitim süreleri (2.500 örnek): RNN-CTC: 351.4 s, CNN-CTC: 410.7 s, GMM-HMM: 498.7 s	Genel Amaçlı ASR: İngilizce konuşma tanıma, İnsan-bilgisayar etkileşimi
Hamian ve arkadaşları [35]	Derin Öğrenme: Darbeli Sinir Ağları (SNN)	Frekans Alanı Özellikleri: MFCC, Zaman Alanı Özellikleri: ZCR	Derin Darbeli Sinir Ağları (Deep SNN) ile optimize edilmiş Integrate-and-Fire nöron modeli; Optimizasyon algoritmaları: GBO (Gradient-Based Optimization) ve WHO (Wild Horse Optimizer)	TIDIGITS: 200 ses örneği (0-9 rakamları), farklı erkek ve kadın konuşmacılardan; IRIS ve Trip veri setleri (karşılaştırma için)	Doğruluk, RMS hata değeri	Rakam ses tanıma için: SNN-WHO: %98.2 doğruluk, 0.0182 RMS hata; SNN-GBO: %96.4 doğruluk, 0.0364 RMS hata; ANN: %93.6 doğruluk, 0.0636 RMS hata; ANFIS: %90.9 doğruluk, 0.0909 RMS hata. IRIS veri seti için: SNN-WHO: %97.3 doğruluk; SNN-GBO: %95.3 doğruluk; ANN: %94.7 doğruluk; ANFIS: %93.3 doğruluk. Trip veri seti için: SNN-WHO: %97.5 doğruluk; SNN-GBO: %95.8 doğruluk; ANN: %93.3 doğruluk; ANFIS: %91.7 doğruluk.	Komut/Anahtar Kelime Tanıma: Rakam ses tanıma, sesli telefon uygulamaları

Çalışma Bilgileri	Yaklaşım Türü	Özellik Çıkarma Yöntemi	Kullanılan Algoritma/Model	Veri Seti	Değerlendirme Metrikleri	Değerlendirme Sonuçları	Uygulama Alanı
Ameen ve Kadhim [36]	Geleneksel Makine Öğrenme: ANN, Random Forest, XGBoost ve Derin Öğrenme: CNN, RNN, LSTM ile Hibrit Yaklaşımlar: CNN-LSTM, CNN-RF, CNN-XGB	Frekans Alanı Özellikleri: MFCC (40 özellik)	ANN: 40 giriş nöronu, 256 gizli nöron, 27 çıkış nöronu; CNN; RNN; LSTM; Random Forest; XGBoost; Hibrit modeller: CNN-LSTM, CNN-RF, CNN-XGB	Alan-Spesifik Veri Seti: 27 Arapça fonem, 8 konuşmacı (3 erkek, 5 kadın), 1.709 örnek (her fonem için 63 örnek), normal konuşma ve elektrolainks cihazı ile üretilen konuşma, 48 kHz örnekleme hızı, 1 saniyelik kayıtlar	Doğruluk Metrikleri: Doğruluk, Kesinlik, Fonem Hata Oranı (PER)	ANN: %75 doğruluk, %77 kesinlik, %21.85 PER; CNN: %70 doğruluk, %72 kesinlik, %26.14 PER; RNN: %68 doğruluk, %70 kesinlik, %31.42 PER; LSTM: %69 doğruluk, %71 kesinlik, %28.75 PER; RF: %65 doğruluk, %67 kesinlik, %33.79 PER; XGBoost: %67 doğruluk, %69 kesinlik, %32.61 PER; CNN-LSTM: %72 doğruluk, %74 kesinlik, %25.93 PER; CNN-RF: %69 doğruluk, %71 kesinlik, %29.87 PER; CNN-XGB: %71 doğruluk, %73 kesinlik, %27.06 PER	Alan-Spesifik ASR: Elektrolainks cihazı kullanan kanser hastalarına yönelik Arapça fonem tanıma sistemi
Shisode ve arkadaşları [37]	Geleneksel Makine Öğrenme: GMM ve Derin Öğrenme: CNN, LSTM ile Hibrit Yaklaşım: CNN-LSTM	EMG Sinyalleri (Yüzey Elektromiyografi) ve Görsel Özellikler (Dudak hareketleri)	İki ana model: 1) GMM (Gaussian Mixture Model) - EMG sinyalleri için, 2) CNN-LSTM - Dudak hareketleri için	Özel oluşturulmuş veri seti: Yüz kaslarından alınan EMG sinyalleri ve dudak hareketleri videoları; Görsel sistem için MIRACL-VC1 veri seti kullanılmış	Doğruluk (Accuracy)	EMG sistemi için: "Emergency": %80, "Heart": %75, "Water": %60, "Better": %65; Dudak okuma sistemi için: Eğitim verisi üzerinde %90 doğruluk, test verisi üzerinde %53 doğruluk	Alan-Spesifik ASR: Konuşma engelli kişiler için sessiz konuşma tanıma sistemi, özellikle larenjektomi geçiren hastalar için
Ahmed ve arkadaşları [10]	Geleneksel Makine Öğrenme: SVM, KNN	Frekans Alanı Özellikleri: MFCC ve Dalgacık Tabanlı Özellikler: DWT (Ayrık Dalgacık Dönüşümü)	Sınıflandırıcılar: SVM (Destek Vektör Makinesi) ve KNN (K-En Yakın Komşu)	Alan-Spesifik Veri Seti: Peştuca izole kelimeler veri seti; 161 izole kelime (0-25 arası rakamlar ve sık kullanılan Peştuca kelimeler); 30 konuşmacı (17 erkek, 13 kadın); Deneyler için 26 izole kelime ve 0-25 arası rakamlar kullanılmış; Her kelime için 17 ifade (12 eğitim, 5 test için); Toplam 312 eğitim ve 130 test örneği	Doğruluk (Accuracy)	MFCC+SVM: %96.15; MFCC+KNN: %92.31; DWT+SVM: %88.46; DWT+KNN: %84.62	Düşük Kaynaklı Diller: Peştuca gibi az kaynaklı dillerde konuşma tanıma; Komut/Anahtar Kelime Tanıma: İzole kelime tanıma
Froiz-Míguez ve arkadaşları [38]	Transfer Öğrenimi: Pre-trained modeller üzerinde fine-tuning, Derin Öğrenme: CNN	Otomatik Özellik Öğrenme: Raw waveform	Pre-trained Model: Wav2Vec 2.0, Uçtan Uca Model: CTC (Connectionist Temporal Classification)	Düşük Kaynaklı Dil Veri Seti: Galiçya dili için yaklaşık 20 saat etiketlenmiş ses verisi (10 saat OpenSLR ve 10 saat Common Voice'dan); Test için Galiçya Parlamentosu'ndan yaklaşık 1 saatlik spontane konuşma verisi	Doğruluk Metrikleri: WER (Word Error Rate)	Eğitim seti üzerinde: %7.15 WER; Test seti üzerinde: Sadece Wav2Vec2 modeli ile %28.67 WER, Işın dekoder ile %27.42 WER, Işın dekoder ve Dil Modeli ile %18.61 WER	Düşük Kaynaklı Diller: Galiçya dili gibi az kaynaklı dillerde konuşma tanıma; Genel Amaçlı ASR: Konuşma-metin dönüşümü
Kwon [29]	Derin Öğrenme: RNN; Uçtan Uca (End-to-End): CTC	Frekans Alanı Özellikleri: MFCC (makalede savunma sistemi için kullanılan	Pre-trained Modeller: DeepSpeech; Savunma için: AudioGuard	Genel Amaçlı Veri Setleri: Mozilla Common Voice	Doğruluk Metrikleri: Doğruluk (Accuracy); Robustluk Metrikleri:	Normal örnekler üzerinde doğruluk: %94.3; Karşıt örnekleri tespit etme doğruluğu: %84.2; Karşıt örneklere karşı savunma başarısı: %82.5	Genel Amaçlı ASR: Konuşma tanıma sistemlerini karşıt saldırılara karşı koruma;

Çalışma Bilgileri	Yaklaşım Türü	Özellik Çıkarma Yöntemi	Kullanılan Algoritma/Model	Veri Seti	Değerlendirme Metrikleri	Değerlendirme Sonuçları	Uygulama Alanı
		özellik çıkarma yöntemi açıkça belirtilmemiş, ancak saldırı yöntemlerinde MFCC kullanıldığı belirtilmiş)	(gürültü vektörü tabanlı savunma yöntemi)		Karşıt örneklere karşı dayanıklılık		Komut/Anahtar Kelime Tanıma: Sanal asistanlar (Alexa gibi) için güvenlik
Rai ve arkadaşları [18]	Geleneksel Makine Öğrenme: HMM, GMM; Derin Öğrenme: CNN, RNN, LSTM; Uçtan Uca (End-to-End): Attention	Frekans Alanı Özellikleri: MFCC; Otomatik Özellik Öğrenme: Raw waveform (CNN için)	İstatistiksel Modeller: HMM-GMM; Derin Sinir Ağları: CNN (M5 mimarisi); RNN: LSTM, BiLSTM; Uçtan Uca Modeller: Attention	Komut Tanıma Veri Setleri: Google Speech Commands Dataset v2; Veri Seti Boyutu: 105,829 adet bir saniyelik konuşma; Veri Seti Özellikleri: 35 farklı kelime/komut içeren konuşmalar	Doğruluk Metrikleri: Doğruluk (Accuracy)	HMM-GMM: %65.2; CNN (M5): %89.7; RNN-LSTM: %91.2; RNN-BiLSTM: %92.5; RNN-BiLSTM-Attention: %93.9	Komut/Anahtar Kelime Tanıma: Sanal asistanlar, akıllı cihazlar; Genel Amaçlı ASR: Düşük kaynaklı ortamlarda çalışabilecek sistemler
Kumar ve arkadaşları [39]	Derin Öğrenme, Uçtan Uca (End-to-End)	Görsel Özellikler (Dudak Hareketleri)	Transformer tabanlı kodlayıcı-kod çözücü, Dikkat mekanizması, BERT dil modeli	GRID Corpus, LRS2 Corpus	WER (Word Error Rate)	GRID Corpus: %2.8 WER, LRS2 Corpus: %40.1 WER	Görsel Konuşma Tanıma (Visual Speech Recognition), Dudak Okuma (Lipreading), Gürültülü Ortamlarda Konuşma Tanıma
Weng ve arkadaşları [40]	Derin Öğrenme, Uçtan Uca (End-to-End)	Frekans Alanı Özellikleri: Spektrogram (Hamming penceresi ve FFT kullanılarak elde edilen)	Hibrit Mimari: CNN-RNN, CTC (Connectionist Temporal Classification), Tacotron 2	Belirtilmemiş (Çalışmada spesifik veri seti adı verilmemiş)	Doğruluk Metrikleri: CER (Character Error Rate), WER (Word Error Rate); Konuşma Sentezi Metrikleri: FDSD (Fréchet Deep Speech Distance), KDSD (Kernel Deep Speech Distance)	SNR = 0 dB'de: DeepSC-ST (CER: ~%10, WER: ~%22), DeepSC-S (CER: ~%38, WER: ~%65), Geleneksel sistem (CER: ~%85, WER: ~%95); SNR = 8 dB'de: DeepSC-ST (CER: ~%3, WER: ~%7), DeepSC-S (CER: ~%18, WER: ~%30), Geleneksel sistem (CER: ~%40, WER: ~%60); FDSD değeri: DeepSC-ST (~0.3), Tacotron 2 (~0.25); KDSD değeri: DeepSC-ST (~0.05), Tacotron 2 (~0.04)	Genel Amaçlı ASR: Konuşma tanıma ve sentezi; Semantik iletişim sistemleri
Chen ve arkadaşları [41]	Hibrit Yaklaşımlar: Ses-LLM entegrasyonu Transfer Öğrenimi: Pre-trained LLM üzerinde fine-tuning	Otomatik Özellik Öğrenme: Conformer tabanlı ses kodlayıcı ile özellik çıkarımı	Derin Sinir Ağları: - Transformatör: Megatron LLM (2B parametre) - Conformer: Fast Conformer (110M parametre) - Hibrit Mimariler: Ses kodlayıcı-LLM entegrasyonu - Pre-trained Modeller: Önceden eğitilmiş Megatron LLM ve Fast Conformer	Genel Amaçlı Veri Setleri: LibriSpeech Çok Dilli Veri Setleri: MuST-C v2 Veri Seti Boyutu: IWSLT 2023 için 2.7M segment, 4.8K saat ses	Doğruluk Metrikleri: WER (ASR için), BLEU (AST için)	ASR WER (LibriSpeech): - Test-clean: 2.3% - Test-other: 4.8% AST BLEU (MuST-C): - İngilizce-Almanca: 29.6 - İngilizce-Japonca: 16.5 Çok görevli SALM: - ASR WER (clean/other): 2.6%/6.1% - AST BLEU (en-de/en-ja): 30.7/16.8	Genel Amaçlı ASR: Konuşma tanıma Çok Dilli ASR: İngilizce-Almanca, İngilizce-Japonca çeviri Komut/Anahtar Kelime Tanıma: Bağlam içi öğrenme ile anahtar kelime güçlendirme

Çalışma Bilgileri	Yaklaşım Türü	Özellik Çıkarma Yöntemi	Kullanılan Algoritma/Model	Veri Seti	Değerlendirme Metrikleri	Değerlendirme Sonuçları	Uygulama Alanı
Gong ve arkadaşları [42]	Hibrit Yaklaşımlar: Ses kodlayıcı-LLM entegrasyonu Transfer Öğrenimi: Pre-trained modeller (Whisper ve LLaMA) üzerinde fine-tuning	Otomatik Özellik Öğrenme: Whisper encoder ve Time and Layer-Wise Transformer (TLTR) ile özellik çıkarımı	Derin Sinir Ağları: - Transformatör: Whisper encoder-decoder (32 katman, 1280 boyut) ve LLaMA-7B - Pre-trained Modeller: Whisper-large, AudioSet pretrained TLTR, LLaMA-7B (Vicuna) - Hibrit Mimariler: Whisper-TLTR-LLaMA entegrasyonu	Çeşitli Veri Setleri: 13 farklı ses ve konuşma veri seti Veri Seti Boyutu: 9.6M Open-ASQA veri seti (6.9M açık uçlu soru-cevap çifti) Veri Seti Özellikleri: Konuşma, konuşma dışı sesler ve bunların kombinasyonlarını içeren çeşitli kayıtlar	Doğruluk Metrikleri: Talimat takip oranı (GPT-4 tarafından değerlendirilmiş) Model Boyutu: Toplam 8.5B parametre (sadece 49M eğitilebilir parametre)	Talimat takip oranı: >%95 Eğitim süresi: 4× A6000 GPU üzerinde yaklaşık 80 saat	Genel Amaçlı ASR: Konuşma tanıma Ses Olayı Tanıma: Konuşma dışı seslerin tanınması Paralinguistik Özellik Tanıma: Konuşma tonu, duygu, vb. Çoklu Görev: Konuşma ve konuşma dışı seslerin birlikte anlaşılması ve aralarındaki ilişkilerin çıkarımı
Wang ve arkadaşları [43]	Derin Öğrenme: Transformatör Uçtan Uca (End-to-End): Otomatik regresif dil modeli Transfer Öğrenimi: Codec tabanlı dil modeli	Otomatik Özellik Öğrenme: EnCodec ile konuşma sinyallerinin ayrık codec kodlarına dönüştürülmesi	Derin Sinir Ağları: - Transformatör: Decoder-only Transformer mimarisi - Pre-trained Modeller: EnCodec - Hibrit Mimariler: Otomatik regresif codec dil modeli ve otomatik regresif olmayan codec dil modeli	Çok Dilli Veri Setleri: Çeşitli dillerde ASR, MT ve TTS veri setleri Veri Seti Özellikleri: Görev kimliği (TID) ve dil kimliği (LID) bilgileri içeren çok görevli veri seti	Doğruluk Metrikleri: ASR ve MT görevleri için WER, BLEU TTS görevleri için ses kalitesi ve konuşmacı benzerliği	ASR (WER): - İngilizce: %9.4 - Çince: %10.2 MT (BLEU): - İngilizce-Çince: 28.7 - Çince-İngilizce: 24.3 S2ST (BLEU): - İngilizce-Çince: 26.5 - Çince-İngilizce: 22.1 TTS: - Konuşmacı benzerliği: 0.85 - Ses kalitesi (MOS): 4.1/5.0	Genel Amaçlı ASR: Konuşma tanıma Çok Dilli ASR: Çok dilli konuşma tanıma Konuşma Sentezi: Metin-konuşma dönüşümü Konuşma Çevirisi: Konuşma-metin, metin-metin, metin-konuşma ve konuşma-konuşma çevirisi
Zhang ve arkadaşları [20]	Derin Öğrenme: Conformer Uçtan Uca (End-to-End): CTC, LAS Transfer Öğrenimi: Öz-denetimli ön-eğitim ve çok amaçlı denetimli ön-eğitim	Otomatik Özellik Öğrenme: BEST-RQ (BERT-based Speech pre-Training with Random-projection Quantizer) ile öz-denetimli öğrenme	Derin Sinir Ağları: - Transformatör: 2B parametreli Conformer - Uçtan Uca Modeller: CTC, LAS (Listen, Attend, and Spell) - Pre-trained Modeller: BEST-RQ ile ön-eğitilmiş modeller	Genel Amaçlı Veri Setleri: SpeechStew, CORAAL Çok Dilli Veri Setleri: FLEURS, CoVoST 2 Veri Seti Boyutu: - YT-NTL-U: 12M saat etiketlenmemiş ses (300+ dil) - Pub-U: 429K saat etiketlenmemiş ses (51 dil) - Web-NTL: 28B cümle metin (1140+ dil) - YT-SUP+: 90K saat etiketli çok dilli veri (73 dil) + 100K saat İngilizce pseudo-etiketli veri	Doğruluk Metrikleri: WER (Word Error Rate) Model Boyutu: 2B parametre	SpeechStew (ASR): - USM-M: %5.7 WER - USM-LAS: %7.2 WER - USM-CTC: %7.6 WER FLEURS (ASR): - USM-M (102 dil): %16.1 WER - USM-LAS (73 dil): %28.4 WER CORAAL (AAVE ASR): - USM-M: %17.3 WER - USM-LAS: %26.3 WER CoVoST 2 (AST):	Genel Amaçlı ASR: YouTube video transkripsiyon Çok Dilli ASR: 100+ dilde konuşma tanıma Alan-Spesifik ASR: Afrika Amerikan Günlük İngilizcesi (AAVE) tanıma Konuşma Çevirisi: İngilizce'den diğer dillere çeviri

Çalışma Bilgileri	Yaklaşım Türü	Özellik Çıkarma Yöntemi	Kullanılan Algoritma/Model	Veri Seti	Değerlendirme Metrikleri	Değerlendirme Sonuçları	Uygulama Alanı
				- Pub-S: 10K saat etiketli çok alanlı İngilizce veri + 10K saat etiketli çok dilli veri (102 dil)		- USM-M (İngilizce → 21 dil): %28.3 BLEU	
Yao ve arkadaşları [13]	Derin Öğrenme: Transformatör Uçtan Uca (End-to-End): CTC, RNN-T	Otomatik Özellik Öğrenme: Conv-Embed modülü ile akustik özelliklerin öğrenilmesi	Derin Sinir Ağları: - Transformatör: Zipformer (U-Net benzeri yapıda) - CNN: 2D-CNN (Conv-Embed modülü) - Hibrit Mimariler: CNN-Transformer - Uçtan Uca Modeller: CTC, RNN-T	Genel Amaçlı Veri Setleri: LibriSpeech Çok Dilli Veri Setleri: Aishell-1, WenetSpeech	Doğruluk Metrikleri: WER (Word Error Rate), CER (Character Error Rate) Verimlilik Metrikleri: Çıkarım süresi, GPU bellek kullanımı Model Boyutu: Parametre sayısı	LibriSpeech (WER): - test-clean: %1.9 - test-other: %3.9 Aishell-1 (CER): - test: %4.0 WenetSpeech (CER): - test_net: %8.1 - test_meeting: %12.4 Çıkarım Hızı: Conformer'dan %50+ daha hızlı Bellek Kullanımı: Conformer'dan daha düşük	Genel Amaçlı ASR: Konuşma tanıma Çok Dilli ASR: İngilizce ve Çince konuşma tanıma
Chu ve arkadaşları [27]	Derin Öğrenme: Transformatör Uçtan Uca (End-to-End): Ses-metin çoklu modlu model Transfer Öğrenimi: Whisper-large-v3 ve Qwen-7B üzerine inşa edilmiş	Frekans Alanı Özellikleri: Mel-spektrogram (128 kanal, 25ms pencere boyutu, 10ms atlama boyutu) Otomatik Özellik Öğrenme: Whisper-large-v3 tabanlı ses kodlayıcı	Derin Sinir Ağları: - Transformatör: Whisper-large-v3 tabanlı ses kodlayıcı ve Qwen-7B LLM - Pre-trained Modeller: Whisper-large-v3, Qwen-7B - Hibrit Mimariler: Ses-dil modeli mimarisi	Genel Amaçlı Veri Setleri: LibriSpeech, FLEURS-ZH Çok Dilli Veri Setleri: Aishell2, CoVoST2 Alan-Spesifik Veri Setleri: MELD (duygu tanıma), VocalSound (vokal ses sınıflandırma) Komut Tanıma Veri Setleri: AIR-Bench (konuşma, ses, müzik ve karma komut takibi)	Doğruluk Metrikleri: WER (Kelime Hata Oranı), BLEU (çeviri kalitesi), SER (Konuşma Duygu Tanıma doğruluğu), VSC (Vokal Ses Sınıflandırma doğruluğu) Model Boyutu: 8.2B parametre	LibriSpeech: %92.62 (1-WER%) Aishell2: %96.0 (1-WER%) CoVoST2: %30.0 (ortalama BLEU) MELD: %45.0 (SER doğruluk) VocalSound: %87.5 (VSC doğruluk) FLEURS-ZH: %91.75 (1-WER%) AIR-Bench-Chat-Speech: %6.88/10 AIR-Bench-Chat-Sound: %6.72/10 AIR-Bench-Chat-Music: %6.5/10 AIR-Bench-Chat-Mixed: %6.43/10	Genel Amaçlı ASR: Konuşma tanıma Çok Dilli ASR: Çince ve İngilizce konuşma tanıma Konuşma Çevirisi: Çok dilli konuşma-metin çevirisi Konuşma Duygu Tanıma: Duygusal ifadeleri anlama Ses Analizi: Konuşma dışı sesleri tanıma ve analiz etme Müzik Analizi: Müzik içeriğini anlama ve yorumlama Sesli Asistan: Sesli komut takibi ve yanıtlama

EK B

B.1 VERİ DOSYASI

Referans,YaklasimTuru,ModelAdi,BasarimMetrigi,BasarimDegeri,GorevTipi,Dil,ParametreSayisiMilyon

"[22] Gourisaria",Derin Öğrenme,ANN,Doğruluk,91.41,Ses Sınıflandırma,İngilizce,

"[22] Gourisaria",Geleneksel,SVM,Doğruluk,87.23,Ses Sınıflandırma,İngilizce,

"[22] Gourisaria",Geleneksel,Rastgele Orman,Doğruluk,85.71,Ses Sınıflandırma,İngilizce,

"[22] Gourisaria",Geleneksel,KNN,Doğruluk,83.19,Ses Sınıflandırma,İngilizce,

"[22] Gourisaria",Geleneksel,Karar Ağacı,Doğruluk,79.14,Ses Sınıflandırma,İngilizce,

"[22] Gourisaria",Geleneksel,Lojistik Regresyon,Doğruluk,76.64,Ses Sınıflandırma,İngilizce,

"[22] Gourisaria",Geleneksel,Naive Bayes,Doğruluk,72.39,Ses Sınıflandırma,İngilizce,

"[11] Gençyılmaz",Derin Öğrenme,CRNN,Doğruluk,96.47,Komut Tespiti,Türkçe,

"[11] Gençyılmaz",Derin Öğrenme,CNN,Doğruluk,95.92,Komut Tespiti,Türkçe,

"[11] Gençyılmaz",Geleneksel,Rastgele Orman,Doğruluk,92.69,Komut Tespiti,Türkçe,

"[11] Gençyılmaz",Geleneksel,SVM,Doğruluk,90.98,Komut Tespiti,Türkçe,

"[11] Gençyılmaz",Geleneksel,KNN,Doğruluk,90.07,Komut Tespiti,Türkçe,

"[11] Gençyılmaz",Derin Öğrenme,GRU,Doğruluk,87.71,Komut Tespiti,Türkçe,

"[15] Keser",Derin Öğrenme,RNN-LSTM,Doğruluk,93.22,Yalıtık Kelime,Türkçe,

"[15] Keser",Derin Öğrenme,CNN,Doğruluk,87.56,Yalıtık Kelime,Türkçe,

"[15] Keser",Geleneksel,KNN,Doğruluk,86.51,Yalıtık Kelime,Türkçe,

"[18] Rai",Derin Öğrenme,RNN-BiLSTM-Attention,Doğruluk,93.9,Komut Tespiti,İngilizce,

"[9] Wang",Derin Öğrenme,RNN-CTC,WER,10.4,Sürekli Konuşma,İngilizce,

"[26] Wojnar",Derin Öğrenme,Whisper large-v3,WER,10.58,Sürekli Konuşma,Çok Dilli,1550

"[26] Wojnar",Derin Öğrenme,Whisper medium.en,WER,16.97,Sürekli Konuşma,İngilizce,769

"[26] Wojnar",Derin Öğrenme,Whisper tiny.en,WER,30.1,Sürekli Konuşma,İngilizce,39

"[35] Hamian",Hibrit,SNN-WHO,Doğruluk,98.2,Yalıtık Kelime,Arapça,

"[10] Ahmed",Geleneksel,SVM,Doğruluk,96.15,Yalıtık Kelime,Peştuca,

"[28] Ayall",Derin Öğrenme,CNN,Doğruluk,99,Yalıtık Kelime,Arapça,

"[31] Alzaabi",Derin Öğrenme,CNN (3-katman),Doğruluk,99.84,Komut Tespiti,Arapça,2.3

"[5] Barhoush",Derin Öğrenme,FC-DNN,Doğruluk,98.5,Konuşmacı Tanıma,İngilizce,

"[38] Froiz-Míguez",Derin Öğrenme,Wav2Vec 2.0,WER,18.61,Sürekli Konuşma,İspanyolca,

"[20] Zhang",Derin Öğrenme,USM-M,WER,5.7,Sürekli Konuşma,Çok Dilli,600

"[41] Chen",Derin Öğrenme,SALM,WER,2.3,Sürekli Konuşma,İngilizce,

EK C

C.1 KODLAMA

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

# 1. Veri Yükleme ve Hazırlık
try:
   df = pd.read_csv('veriler.csv')
   print("veriler.csv dosyası başarıyla okundu.")
except FileNotFoundError:
   print("HATA: veriler.csv dosyası bulunamadı. Lütfen dosyanın script ile aynı klasörde
         olduğundan emin olun.")
   exit()

# 2. Grafik 1: Yaklaşım Türüne Göre Başarım Karşılaştırması
#'Doğruluk' metriği olan verileri filtrele
df_dogruluk = df[df['BasarimMetrigi'] == 'Doğruluk'].copy()
df_dogruluk['BasarimDegeri'] = pd.to_numeric(df_dogruluk['BasarimDegeri'])

plt.figure(figsize=(10, 7))
# Barplot oluştur
barplot = sns.barplot(x='YaklasimTuru', y='BasarimDegeri', hue='YaklasimTuru',
                      data=df_dogruluk, palette='viridis', legend=False,
                      estimator=np.mean, errorbar=None)

plt.title('Yaklaşım Türüne Göre Ortalama Başarım (Doğruluk %)', fontsize=16)
plt.xlabel('Makine Öğrenme Yaklaşımı', fontsize=12)
plt.ylabel('Ortalama Doğruluk (%)', fontsize=12)
plt.ylim(0, 105)  # Sayıların grfik çerçevesi dışına çıkmaması için 105 yapıldı
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Barların üzerine ortalama değerleri yazdır
for p in barplot.patches:
```

```
    if p.get_height() > 0:
        barplot.annotate(format(p.get_height(), '.1f') + '%',
                         (p.get_x() + p.get_width() / 2., p.get_height()),
                         ha='center', va='bottom',
                         xytext=(0, 3),  # Daha az offset
                         textcoords='offset points',
                         fontweight='bold')

plt.savefig('grafik_1_yaklasim_basarim.png', dpi=300, bbox_inches='tight')
print("Grafik 1 (yaklasim_basarim.png) oluşturuldu.")

# 3. Grafik 2: Model Karmaşıklığı vs. Performans (WER)
# Sadece 'WER' metriği olan ve Parametre Sayısı bilinen verileri filtrele
df_wer = df[df['BasarimMetrigi'] == 'WER'].copy()
df_wer['BasarimDegeri'] = pd.to_numeric(df_wer['BasarimDegeri'])
df_wer['ParametreSayisiMilyon'] = pd.to_numeric(df_wer['ParametreSayisiMilyon'])

# DÜZELTME: KeyError hatasını gidermek için filtrelenmiş DataFrame'in indeksi sıfırlanıyor.
df_wer_filtered = df_wer[df_wer['ParametreSayisiMilyon'].notna()].reset_index(drop=True)

plt.figure(figsize=(12, 8))
scatter = sns.scatterplot(x='ParametreSayisiMilyon', y='BasarimDegeri', data=df_wer_filtered, s=150, hue='ModelAdi', palette='magma', legend=False)

plt.title('Model Karmaşıklığı ve Performans İlişkisi (WER)', fontsize=16)
plt.xlabel('Parametre Sayısı (Milyon)', fontsize=12)
plt.ylabel('Kelime Hata Oranı (WER) - Düşük Değer Daha İyi', fontsize=12)
plt.grid(True, which='both', linestyle='--', linewidth=0.5)
plt.gca().invert_yaxis() # WER metriği için y eksenini ters çevir (düşük daha iyi)

# Noktaların yanına model isimlerini yazdır
for i in range(len(df_wer_filtered)):
    plt.text(x=df_wer_filtered['ParametreSayisiMilyon'][i] * 1.05,
             y=df_wer_filtered['BasarimDegeri'][i],
             s=f"{df_wer_filtered['ModelAdi'][i]}\n({df_wer_filtered['Referans'][i]})",
             fontdict=dict(color='black', size=10))
```

```
plt.savefig('grafik_2_karmasiklik_performans.png', dpi=300, bbox_inches='tight')
print("Grafik 2 (karmasiklik_performans.png) oluşturuldu.")


# 4. Grafik 3: Görev Tipine Göre En İyi Başarımlar (Doğruluk) ---
df_dogruluk_best = df_dogruluk.loc[df_dogruluk.groupby('GorevTipi')['BasarimDegeri'].idxmax()]

plt.figure(figsize=(12, 8))
barplot_gorev = sns.barplot(x='BasarimDegeri', y='GorevTipi', hue='GorevTipi', data=df_dogruluk_best, palette='crest', legend=False, orient='h')

plt.title('Görev Tiplerine Göre En Yüksek Başarım Oranları (Doğruluk %)', fontsize=16)
plt.xlabel('En Yüksek Doğruluk (%)', fontsize=12)
plt.ylabel('Görev Tipi', fontsize=12)
plt.xlim(80, 101)

# Barların üzerine model adı ve referans bilgisini yazdır
for index, row in df_dogruluk_best.iterrows():
    label = f"{row['ModelAdi']} ({row['Referans']}) - {row['BasarimDegeri']:.2f}%"
    plt.text(row['BasarimDegeri'] - 0.1,
             barplot_gorev.get_yticks()[list(df_dogruluk_best['GorevTipi']).index(row['GorevTipi'])],
             label,
             color='white',
             ha="right",
             va="center",
             fontweight='bold')

plt.savefig('grafik_3_gorev_tipi_basarim.png', dpi=300, bbox_inches='tight')
print("Grafik 3 (gorev_tipi_basarim.png) oluşturuldu.")

print("\nTüm grafikler başarıyla oluşturuldu ve klasöre kaydedildi.")
```