



SOMALİ DİLİNDE KONUŞMA TANIMA

Abdullatif ALI TAHLİL

Yüksek Lisans Tezi
Elektrik-Elektronik Mühendisliği Anabilim Dalı
Elektronik Bilim Dalı
Doç. Dr. Bülent ÇAVUŞOĞLU
2020
Her hakkı saklıdır

**ATATÜRK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

YÜKSEK LİSANS TEZİ

SOMALİ DİLİNDE KONUŞMA TANIMA

Abdullatif ALI TAHLİL

**ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI
Elektronik Bilim Dalı**

**ERZURUM
2020**

Her hakkı saklıdır

T.C.
ATATÜRK ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü Müdürlüğü

TEZ KABUL VE ONAY TUTANAĞI

SOMALİ DİLİNDE KONUŞMA TANIMA

Doç. Dr. Bülent ÇAVUŞOĞLU danışmanlığında, Abdullatif ALI TAHLİL tarafından hazırlanan bu çalışma, 03/02/2020 tarihinde aşağıdaki jüri tarafından Elektrik Elektronik Mühendisliği, Anabilim Dalı Elektronik Bilim Dalı'nda Yüksek Lisans tezi olarak **oybirliği / oy çokluğu (.../...) ile kabul edilmiştir.**

Jüri Başkanı: Doç. Dr. İ. Yücel ÖZBEK
Atatürk Üniversitesi

Danışman: Doç. Dr. Bülent ÇAVUŞOĞLU
Atatürk Üniversitesi

Jüri Üyesi: Doç. Dr. Nur Hüseyin KAPLAN
Erzurum Teknik Üniversitesi

[Signature]
[Signature]
[Signature]

Bu tezin Atatürk Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliği'nin ilgili maddelerinde belirtilen şartları yerine getirdiğini onaylarım. **07.02.2020 / 06-64**

[Signature]
Prof. Dr. Mehmet KARAKAN
Enstitü Müdürü

ÖZET

Yüksek Lisans Tezi

SOMALİ DİLİNDE KONUŞMA TANIMA

Abdullatif ALI TAHLİL

Atatürk Üniversitesi
Fen Bilimleri Enstitüsü
Elektrik-Elektronik Mühendisliği Anabilim Dalı
Elektronik Bilim Dalı

Danışman: Doç. Dr. Bülent ÇAVUŞOĞLU

Konuşma tanıma teknolojisinin popüleritesi, temel algoritmalarda kaydedilen ilerlemeler ve bu algoritmaları eğitmek için bol miktarda verinin bulunması nedeniyle son on yılda önemli ölçüde artmıştır. Ancak, bu tüm diller için geçerli değildir. Somalice gibi kaynakların yetersiz olduğu diller, konuşma tanıma araştırmacıları için herkese açık bir veri kümesine sahip değildir. Bu tez çalışması halka açık ilk Somali dili konuşma veri kümesini yaratma girişimidir. Bu çalışmada SUUBAN olarak bilinen Somali Konuşma Komutları veri kümesi oluşturulmuştur. Veri kümesi Sayılar, IoT komutları, Robotik Hareket komutları gruplarından herhangi birine ait altmış hedef kelimedenden oluşur. Veri seti 177 Somalice konuşmacının (106 erkek ve 71 kadın) katkısı ile toplam 17 saatten oluşmaktadır. Sesler farklı ortamlarda (iç ve dış mekan) kaydedildi. Konuşma ifadeleri Audacity kullanılarak etiketlendi ve bir MATLAB kodu kullanılarak çıkarıldı. Veri kümesindeki arka plan gürültüsü ve bilinmeyen sözcükler dahil toplam dosya sayısı 62.816'dır. Ses dosyaları 16kHz'de örneklendi ve WAV formatında saklandı. Veri kümesinin alt kümelerinin her birini test etmek için bir Derin Konvolüsyon sinir ağı oluşturularak konuşma komutları veri kümesinin çeşitli modelleri eğitmek ve değerlendirmek için yararlı olduğu gösterilmiştir. Sonuçlar umut vericiydi ve bu derin CNN'lerin her birinin test doğruluğu %95'ten yüksekti.

2020, 62 sayfa

Anahtar Kelimeler: Konuşma tanıma, Somali dili, anahtar kelime belirleme, Konvolüsyonel Sinir Ağı (CNN), konuşma komutu veri kümesi.

ABSTRACT

MS Thesis

SPEECH RECOGNITION IN SOMALI LANGUAGE

Abdullatif ALI TAHLIL

Atatürk University
Graduate School of Natural and Applied Sciences
Department of Electrical-Electronics
Electronic Science

Supervisor : Assoc. Prof. Bülent Çavuşoğlu

The popularity of speech recognition technology has increased dramatically in the past decade due to the advances made in the underlying algorithms and availability of abundant data to train these algorithms. However, this is not the case for all languages. Under-resourced languages, like Somali, do not have a publicly available dataset for speech recognition researchers. In this thesis work, a Somali Speech Commands dataset, known as SUUBAN, was built. The dataset consists of sixty target words belonging to any of three groups: Numbers, IoT commands, Robotic Movement commands. The dataset consists of a total of 17 hours contributed by 177 Somali speakers (106 male and 71 female). The audios were recorded in different environments (indoor and outdoor). Speech utterances were labeled using Audacity and extracted using a MATLAB script. The total number of files in the dataset, including background noise and unknown words, is 62,816. Audio files were sampled at 16kHz and stored in WAV format. The Speech Commands dataset has shown to be useful for training and evaluating a variety of models. A Deep Convolutional neural network was built to test each of the subsets of the dataset. The results were promising and the test accuracy of each of these deep CNNs was higher than 95%.

2020, 62 pages

Keywords: Speech recognition, Somali language, keyword spotting, Convolutional Neural Network (CNN), speech command dataset.

TEŞEKKÜR

Tez çalışmam sırasında kıymetli bilgi, birikim ve tecrübeleri ile bana yol gösterici ve destek olan değerli danışman hocam sayın Doç. Dr. Bülent ÇAVUŞOĞLU'ya, ilgilerini ve önerilerini göstermekten kaçınmayan sayın hocalarıma Doç.Dr. İbrahim Yücel ÖZBEK'e ve Dr. Öğr. Üyesi Emin Argun ORAL'a sonsuz teşekkür ve saygılarımı sunarım.

Çalışmalarım boyunca yardımını hiç esirgemeyen değerli arkadaşlarım Aslı Nur Ömeroğlu, Merve Polat, Nida Kumbasar, Sultan Ertaş ve Hussein Alameri'ye teşekkürü bir borç bilirim.

Ayrıca, bana Türkiye Cumhuriyeti'nde yüksek lisans eğitimimi almamı sağlayan, muhteşem Türk kültürünü deneyimleme fırsatı sunan ve burs veren Türkiye Cumhuriyeti'nin Yurtdışı Türkler ve Akraba Topluluklar Başkanlığı'na (YTB) en derin şükranlarımı sunuyorum.

Çalışmalarım boyunca maddi manevi destekleriyle beni hiçbir zaman yalnız bırakmayan aileme de sonsuz teşekkürler ederim.

Abdullatif ALI TAHLİL

Şubat, 2020

İÇİNDEKİLER

ÖZET.....	i
ABSTRACT.....	ii
TEŞEKKÜR.....	iii
SİMGELER ve KISALTMALAR DİZİNİ	vi
ŞEKİLLER DİZİNİ.....	vii
ÇİZELGELER DİZİNİ	ix
1. GİRİŞ.....	1
1.1. Somali Dili	1
1.2. Konuşma Tanıma	2
1.2.1. Konuşma tanıma sistemlerinin sınıflandırılması	2
1.2.2. Konuşma tanıma zorlukları	3
1.2.3. Tipik bir konuşma tanıma sistemi	4
1.2.4. Konuşma tanıma uygulamaları.....	7
1.3. Makine Öğrenmesi	9
1.3.1. Denetimli öğrenme	12
1.3.2. Denetimsiz öğrenme.....	13
1.3.3. Pekiştirmeli öğrenme.....	15
1.4. Yapay Sinir Ağları.....	15
1.4.1. Aktivasyon fonksiyonları	17
1.4.2. YSA mimarisi ve türleri	19
1.4.3. Evrişimli sinir ağları.....	21
2. KAYNAK ÖZETLERİ	28
3. MATERYAL ve YÖNTEM.....	35
3.1. Veri Kümesi	35
3.1.1. Veri kümesi oluşturma	36
3.1.2. Veri kümesini etiketleme.....	38
3.2. Konuşma Tanıma Algoritması	42
3.2.1. Veri hazırlama ve ön işleme.....	42
3.2.2. Model eğitimi	46

4. ARAŞTIRMA BULGULARI ve TARTIŞMA.....	51
5. SONUÇ.....	58
KAYNAKLAR	59
ÖZGEÇMİŞ	63



SİMGELER ve KISALTMALAR DİZİNİ

AI	Yapay Zeka (Artificial Intelligence)
ANN	Yapay Sinir Ağları (Artificial Neural Network)
ASR	Otomatik Konuşma Tanıma (Automatic Speech Recognition)
CNN	Evrişimsel Sinir Ağları (Convolutional Neural Network)
CTC	Bağlantıcı Zamansal Sınıflandırma (Connectionist Temporal Classification)
FFT	Hızlı Fourier Dönüşümü (Fast Fourier Transform)
HMM	Gizli Markov Modeli (Hidden Markov Model)
IVR	Interactive Voice Response (Etkileşimli Sesli Yanıt)
KNN	K-En Yakın Komşu (K-Nearest Neighbor)
LVCSR	Büyük Kelime Sürekli Konuşma Tanıma (Large Vocabulary Continuous Speech Recognition)
MFCC	Mel Frekans Kepstral Katsayı (Mel Frequency Cepstral Coefficient)
NN	Sinir Ağı (Neural Network)
ReLU	Rektifiyeli Doğrusal Ünite (Rectified Linear Unit)
SR	Konuşma Tanıma (Speech Recognition)
SVM	Destek Vektör Makine (Support Vector Machine)
WER	Kelime Hata Oranı (Word Error Rate)
YSA	Yapay Sinir Ağı

ŞEKİLLER DİZİNİ

Şekil 1.1. Somalice'nin anadil olarak konuşulduğu bölgeler	1
Şekil 1.2. Konuşma tanıma sistemlerinin sınıflandırılması	4
Şekil 1.3. Temel bir konuşma tanıma sistemi	5
Şekil 1.4. MFCC türetimi	6
Şekil 1.5. Yapay zeka, makine öğrenimi ve derin öğrenme arasındaki ilişki.....	11
Şekil 1.6. Makine öğrenimi sınıflandırması.....	12
Şekil 1.7. Denetimli öğrenme için etiketli eğitim seti (örneğin, konuşma tanıma)	13
Şekil 1.8. Denetimsiz öğrenme için etiketlenmemiş bir eğitim seti	14
Şekil 1.9. Pekiştirmeli öğrenme	15
Şekil 1.10. Yapay sinir ağı.....	16
Şekil 1.11. Algılayıcı	16
Şekil 1.12. Sigmoid fonksiyonu.....	18
Şekil 1.13. Tanh fonksiyonu	18
Şekil 1.14. ReLU fonksiyonu	19
Şekil 1.15. MLP (a), RNN (b), CNN (c).....	20
Şekil 1.16. CNN mimarisine ve eğitim sürecine genel bakış	22
Şekil 1.17. Evrişim işlemine bir örnek	23
Şekil 1.18. Sıfır dolgu ile konvolüsyon çalışması	24
Şekil 1.19. Maksimum ortaklama işlemi örneği	26
Şekil 1.20. SoftMax aktivasyon fonksiyonunu	27
Şekil 3.1. Audacity arayüzü	39
Şekil 3.2. Veri kümesinin frekans dağılımı	41
Şekil 3.3. Ses sinyalinin doğası	42
Şekil 3.4. Ses sinyali çerçeveleme örneği.....	43
Şekil 3.5. Ses sinyalinde Hamming penceresi fonksiyonu çalışması	45
Şekil 3.6. Bir 16 Mel filtre bankası.....	45
Şekil 3.7. Bir balina şarkısının ses dalga formu ve Mel spektrogramı	46
Şekil 3.8. Veri kümesi bölümlendirme.....	47
Şekil 3.9. Derin CNN model tasarım süreci	48

Şekil 3.10. Derin CNN'lerin mimarisi	50
Şekil 4.1. Rakam alt kümesinin veri dağılımı.....	51
Şekil 4.2. Rakam tanıyıcı karışıklık matrisi.....	52
Şekil 4.3. IoT komutları alt küme veri dağılımı.....	53
Şekil 4.4. IoT komutları alt kümesinin karışıklık matrisi	54
Şekil 4.5. Robotik Hareket komutları veri seti dağılımı	56
Şekil 4.6. Robotik Hareket Ağı Karışıklık matrisi.....	57



ÇİZELGELER DİZİNİ

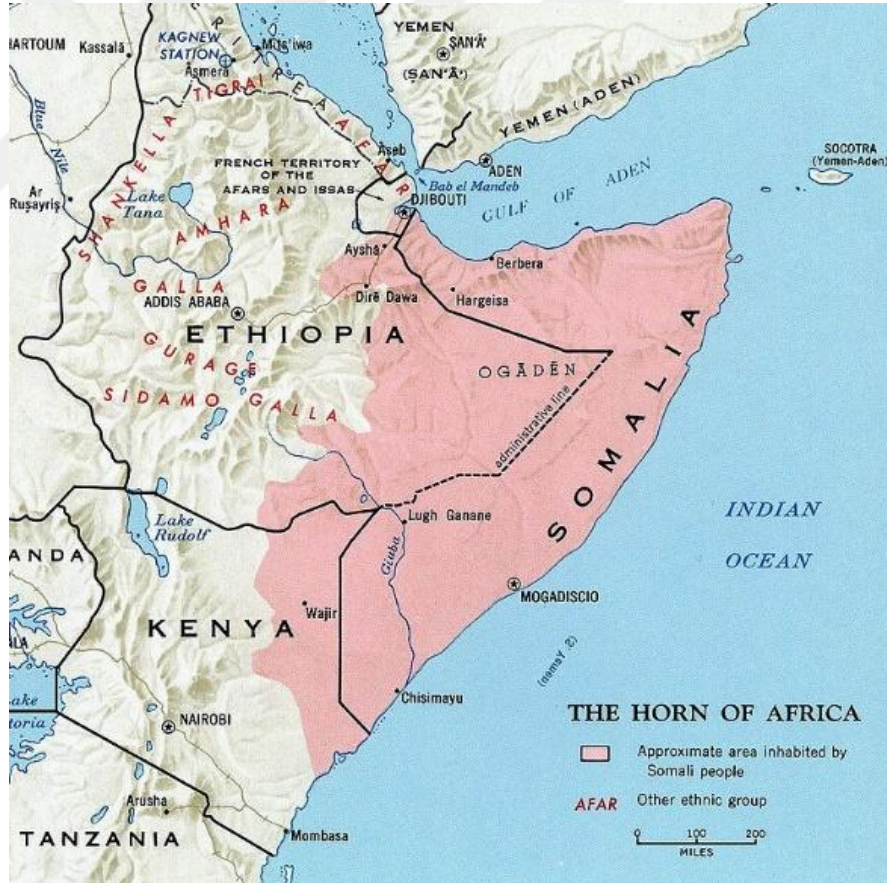
Çizelge 3.1. Bir CNN parametrelerin ve hiperparametrelerin listesi.....	50
Çizelge 4.1. CNN'nin tümünün ağ boyutlarını ve hızlarını	56



1. GİRİŞ

1.1. Somali Dili

Somalice, Afro-asyatik ailesine ait bir dildir. Afrika Boynuzu ve diasporadaki etnik Somalilerin ana dilidir. Somalice, Cibuti'de ulusal Somali'de resmi ve Etiyopya ile Kenya'da çalışma dilidir. Ayrıca birkaç etnik azınlık grubu ve bu alanlardaki bireyler tarafından benimsenen bir dil olarak konuşulmaktadır. Somalice konuşanlarının toplam sayısının 20 milyon civarında olduğu tahmin edilmektedir (Dubnov 2003). Şekil 1.1, Somalice konuşanların Afrika boynuzundaki dağılımını göstermektedir.



Şekil 1.1. Somalice'nin anadil olarak konuşulduğu bölgeler

Somalice, Wadaad'ın yazısı olarak bilinen Arapça tabanlı bir ebced, Latin tabanlı bir alfabe ve Borama, Osmany ve Kaddare alfabeleri dahil olmak üzere birkaç farklı yazı ile yazılmıştır. Somalice'de resmi yazılar için Latin alfabesi kullanılmaktadır. İngilizce latin alfabesinin p, v ve z hariç tüm harflerini kullanır ve 21 ünsüz ve 5 sesli harf içerir. Başlangıçta sözcükten oluşmayan gırtlaksı durağı için kesme işaretinin kullanılması dışında aksan veya başka özel karakterler yoktur. Ek olarak, üç ünsüz digraf vardır ve bunlar DH, KH ve SH'dir (Brown and Ogilvie 2010).

Ton işaretlenmez ve ön ve arka sesli harfler ayırt edilmez. Büyük harfler bir cümle başında ve özel isimler için kullanılır. Somalice'nin hece yapısı (ünsüz)-sesli-(ünsüz) şeklindedir ve kök morfolojileri genellikle tek veya iki heceli bir yapıya sahiptir. Somalice eklemeli bir dildir ve aynı zamanda çekim özellikleri de gösterir. Ekler, görünüş, gerginlik ve durum da dahil olmak üzere birçok gramer anlamını işaret eder (Hyman 1981; Gac 2002).

1.2. Konuşma Tanıma

Konuşma tanıma, bir bilgisayar programının konuşma ifadelerini anlama ve bunları makine tarafından okunabilir bir biçime dönüştürme yeteneğidir. Amacı, doğal dillerin doğal belirsizliğine ve karmaşıklığına rağmen, sözlü bilgileri duyabilen ve anlayabilen akıllı makineler üretmektir. Konuştuğumuz doğal dili kullanarak makinelerle etkileşim kurulmasını sağlar. Temel bir konuşma tanıma yazılımı, birkaç kelime veya kelime öbeğini tanımakla sınırlıdır, ileri düzey olanlar da tıpkı insanlar gibi sürekli doğal konuşmayı tanıyabilir (Gaikwad *et al.* 2010).

1.2.1. Konuşma tanıma sistemlerinin sınıflandırılması

En yaygın konuşma tanıma yaklaşımları “şablon tabanlı yaklaşım” ve “model tabanlı yaklaşım” dır. Doğrusal Zaman Hizalama (LTA) ve Dinamik Zaman Kaydırma (DTW) gibi şablon tabanlı yaklaşımlar en basit tekniklerdir ve düzgün kullanıldıklarında yüksek doğruluk oranına sahiptirler, ancak bazı sınırlamaları vardır. Şablon tabanlı yaklaşımlar

küçük boyutlu kelimeler ve sınırlı miktarda veriler için uygundur. Şu anda en popüler tekniklerden olan Gizli Markov Modeli (HMM) ve Sinir Ağları (NN) gibi model tabanlı yaklaşımlar, konuşma referanslarının güçlü özelliklerini istatistiksel olarak çok miktarda veriden çıkarabilir (Aşlıyan 2008).

Sistemin konuşmayı nasıl ele aldığına bağlı olarak, konuşma tanıma sistemleri ayrık kelime, bağlantılı kelime ve sürekli konuşma tanıma sistemleri olarak sınıflandırılabilir (Tunali 2005). Ayrık kelime tanıma sistemlerinde bir seferde bir kelime söylenip işlenirken, bağlantılı kelime sistemlerinde, dikkatlice ayrılmış kelimeler birkaç kelime veya cümle, sisteme girdi olarak verilir. Bununla birlikte, sürekli konuşma tanıma sistemlerinde, sistemin girdisi, bitişik kelimelerin duraklama olmadan veya kelimeler arasındaki herhangi bir görünür ayrım olmadan birlikte çalıştığı doğal bir konuşmadır.

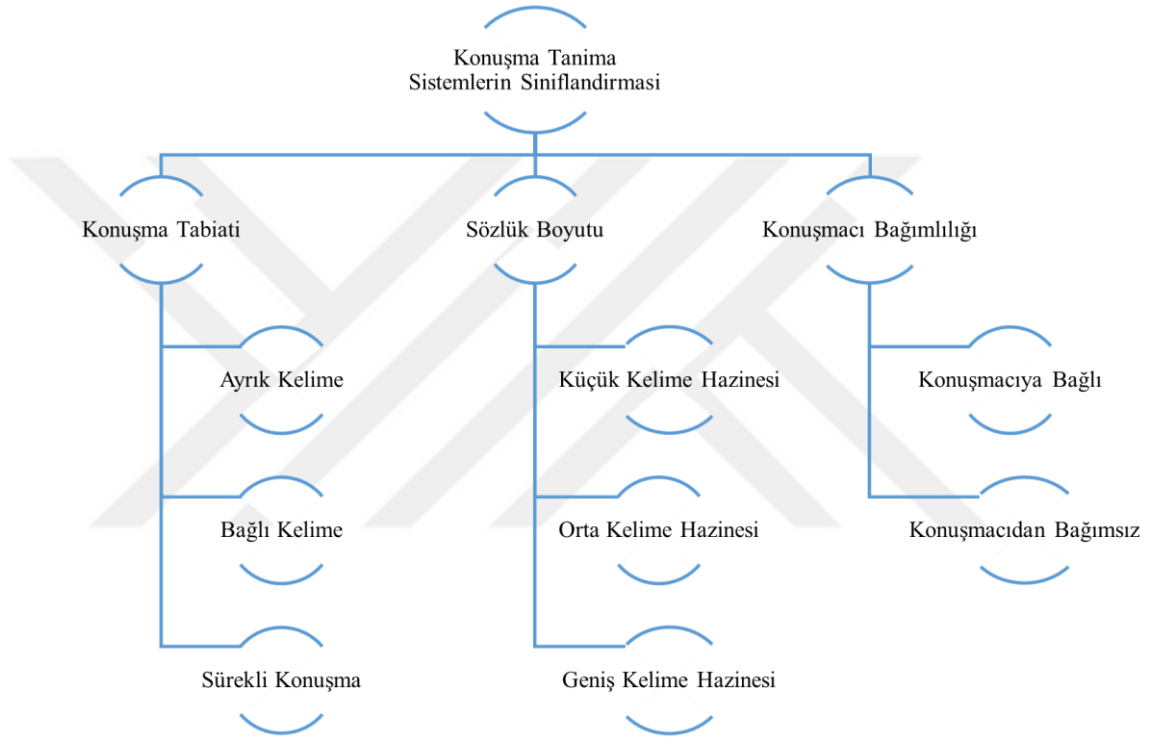
Konuşma tanıma sistemleri, konuşmacıya bağlı ve konuşmacıdan bağımsız olarak da sınıflandırılabilir. Konuşmacıya bağlı sistemlerde, sistem tek bir kullanıcının konuşmasını tanıyacak şekilde tasarlanabilir ve eğitilebilir. Bu tür sistemler mükemmeldir ve %98'lik bir doğruluğa sahip olabilirler. Bununla birlikte, konuşmacıdan bağımsız sistemlerin oluşturulması zor ve daha karmaşıktır. Genellikle konuşmacıdan bağımsız sistemler, konuşmacıya bağlı olanlara kıyasla kötü performanstan muzdariptirler ve yapımı nispeten pahalıdır.

Kelimelerin tanındığı sözlüğün boyutuna bağlı olarak, konuşma tanıma sistemleri küçük, orta ve büyük kelime tanıyıcılar olarak sınıflandırılabilir. Bu sistemler sırasıyla 100'den az kelime, 100-1000 kelime ve 1000'den fazla kelime tanıyabilir (Aşlıyan 2008). Şekil 1.2. konuşma tanıma sistemlerinin sınıflandırılmasını göstermektedir.

1.2.2. Konuşma tanıma zorlukları

Konuşmacıya bağlı konuşma tanıma sistemlerinin doğruluğu neredeyse %98'e ulaşırken, konuşmacıdan bağımsız konuşma tanıma sistemlerinin doğruluğu nispeten zayıftır. Çeşitli değişken kaynaklar konuşma tanımayı zor bir problem haline getirir; bunlardan

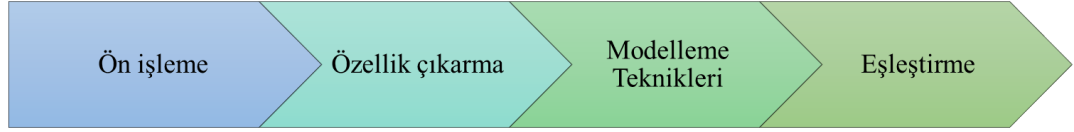
bazıları stil (konuşma/gündelik konuşma veya okuma konuşması, sürekli konuşma veya izole kelimeler vb.), çevre (arka plan gürültüsü, kanal koşulları, oda akustiği, vb.), konuşmacı özellikleri (konuşma hızı, aksan, vb.) ve görev özellikleri (kelime dağarcığındaki kelime sayısı, dil kısıtlamaları vb.) içeren konuşmalar olarak sıralanabilir (Aşlıyan 2008; Petkar 2016).



Şekil 1.2. Konuşma tanıma sistemlerinin sınıflandırılması

1.2.3. Tipik bir konuşma tanıma sistemi

Şekil 1.3'te gösterildiği gibi, temel bir konuşma tanıma sistemi, konuşma sinyalinin ön işlemden geçirilmesi, özellik çıkarma, modelleme teknikleri ve eşleştirme aşamalarından oluşur.



Şekil 1.3. Temel bir konuşma tanıma sistemi

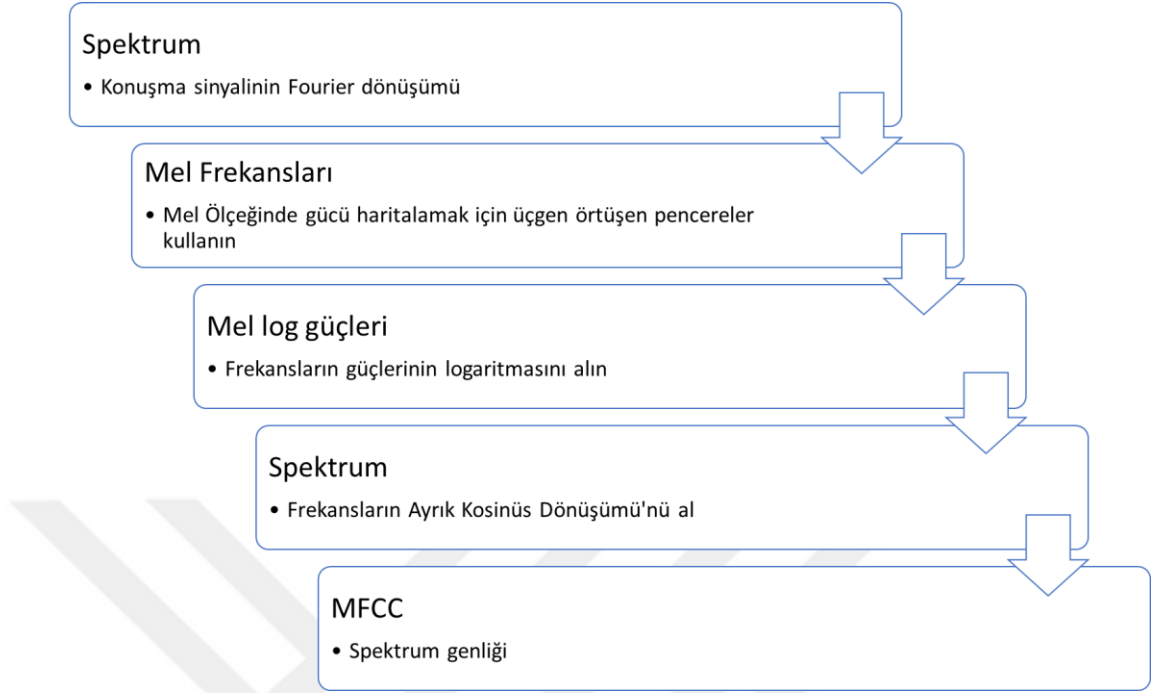
Ön İşleme

Sistem konuşma dalga biçimini girdi olarak alır ve konuşma sinyalini özellik çıkarma analizi için daha kabul edilebilir olacak şekilde ayarlar. Bu aşamada, konuşma sinyaline spektral özellikleri belirlenir ve gürültü azaltma algoritmaları uygulanır. Konuşma sinyalinde yumuşatma, pencereleme ve hece uç nokta algılama burada gerçekleştirilir; konuşma sinyali alt bölümlere (2-5 ms sinyal çerçeve boyutu ve kaydırma boyutu), normal bölümlere (10-30 ms sinyal çerçeve boyutu ve kaydırma boyutu) veya üst bölümlere ayrılabilir (50-200 ms sinyal çerçeve boyutu ve kaydırma boyutu) (İbrahim *et al.* 2017).

Özellik çıkarma

Özellik çıkarma konuşmayı, kelimeyi, foneyi veya konuşmacıyı tanımlamasına yardımcı olabilecek konuşma sinyalinden özellikler çıkarmayı içerdiğinden, sistemin temel parçalarından biridir. Doğrusal Öngörülü Kodlama (LPC), Doğrusal Öngörücü Kepstral Katsayıları (LPCC), Birinci Mertebeden Türev (DELTA), Güç Spektral Analizi (FFT), Mel Ölçeği Kepstral Analizi (MEL), Mel Frekans Kepstral Katsayıları (MFCC'ler), Göreceli Spektral Filtreleme, Olasılıksal Doğrusal Ayırım Analizi (PLDA) vb. gibi, çeşitli özellik çıkarma teknikleri vardır (Alim and Rashid 2018).

MFCC yaygın olarak kullanılan özellik çıkarma tekniğidir. Bunun nedeni, insan kulağının 1000 Hz'nin altındaki frekanslara sahip kritik bant genişliklerinin bilinen varyasyonlarına dayanmasıdır. MFCC'nin temel amacı insan kulağının davranışını taklit etmektir. Şekil 1.4. MFCC türetmenin nasıl yapıldığını gösterir (Tiwari 2010).



Şekil 1.4. MFCC türetimi

Mel frekans ölçeği 1kHz'in altındaki frekanslar için doğrusal ve 1kHz'in üzerindeki frekanslar için logaritmik aralıklıdır (Aşlıyan 2008). İyi bir modelleme tekniğiyle, tanıma doğruluğu yüksek olacaktır. Bu da daha yüksek bir performans oranına yol açacaktır. MFCC'nin diğer özellik çıkarma tekniklerine kıyasla düşük karmaşıklığa sahip olduğunu belirtmek gerekir. MFCC kullanmanın en önemli dezavantajı, spektral forma bağımlı olması nedeniyle gürültüye olan duyarlılığıdır (Muda *et al.* 2010).

Modelleme Teknikleri

Çıkarılan özelliklerin eğitimi ve sınıflandırılması, Gizli Markov Modelleri (HMM), Sinir Ağları (NN) veya hibrit bir HMM-NN kullanılarak farklı biçimlerde uygulanabilir. En başarılı ve popüler konuşma modellerinden biri, Markov zincirine dayanan basitleştirilmiş stokastik bir süreç modeli olan birinci dereceden HMM'dir. NN kullanarak konuşma tanıma (SR) uygulaması ile ilgili literatür kısıtlı olmasına rağmen, çoğu SR sisteminde HMM modeli kullanılmıştır (Riis *et al.* 1999).

Eşleştirme

Bu aşamada telaffuz ve dil modelleri kullanılmaktadır. Telaffuz modeli, alt kelime birimleri (fonemes) ve kelimeler arasında bir bağlantı sağlar. Genellikle basit bir telaffuz sözlüğü tutulur. Sesbirim sekansı ve kelimeler arasındaki bağlantıyı sağlamak için bir konuşma entegrasyonu olması kaçınılmazdır. Telaffuz modeli verilerden öğrenilmez; uzmanlar tarafından türetilmiştir, yani bir dil uzmanı eşleştirmeleri vermektedir (Hain 2002).

Öte yandan, istatistiksel dil modeli, belirli bir dilin morfolojisine göre son kelime bağlamı göz önüne alındığında farklı kelimelerin oluşma olasılığını sağlar. Benzer akustikler arasında ayırım yapmak için de kullanılır. Popüler dil modellemeleri SRILM , KENLM araç setlerini (toolkit), Açık GRM ve NGRAM kütüphanelerini içerir (Alumäe 2008).

Buna ek olarak, bu aşamada bir arama problemi ile ilgilenilir. Her kelime bir Sesbirim sekansı olduğundan ve her Sesbirim bir Gizli Markov Modeli durumuna karşılık geldiğinden arama grafiği kullanılır (Uchat 2006).

Uçtan Uca Nöral ASR Sistemlerinin yeni konuşma tanıma standardı haline geldiğini belirtmek gerekir. Bu yöntemde, sözcükleri onlara karşılık gelen sesbirim ve HMM'lere bölmeden, ancak akustik özelliklerden harflere doğrudan eşleştirerek ve ardından karakter dizisini yeniden vurgulayan karakter dili modelleri kullanılarak ses tanıma görevinin gerçekleştirilmesi amaçlanmaktadır. Burada, telaffuz modeli sesbirimleri almaz; yani dilbilim tarafından yapılan kelime eşleşmesine, foneme gerek yoktur (Chiu *et al.* 2018).

1.2.4. Konuşma tanıma uygulamaları

Donanım, algoritma ve sinyal işlemedeki ilerlemelerden ötürü, konuşma tanıma teknolojisi son 40 yılda büyük gelişim göstermiştir. Konuşma tanıma teknolojileri, mümkün olan hemen hemen her alanda eşi görülmemiş ve yenilikçi yollarla

benimsenmiştir. Konuşma tanımanın en ünlü uygulamaları arasında, telekomünikasyon, imalat, tıbbi, oyun, eğitim alanları ve engellilere yönelik yardımlar bulunmaktadır.

Telekomünikasyon hizmetlerinde konuşma tanıma, çağrı yönlendirme hizmetlerini otomatikleştirerek operatör fonksiyonlarını yerine getirir. Tek tuşla etkileşimli sesli yanıt (IVR) sistemlerinin doğal dil konuşma tanıma sistemleriyle değiştirilmesi, şirketlerin çağrı bekleme sürelerini 30 saniyeye, yanlış yönlendirilmiş çağrıları ise neredeyse sıfıra düşürmelerine yardımcı olmaktadır.

Sesle arama sistemleri, kullanıcıların "Ali'yi ara" gibi anahtar sözcükleri ve adları kullanarak ASR sistemleri sayesinde konuşarak bir düğmeye basmadan arama yapmalarını sağlar. Modern arabalar, sürücülerin araç radyosu, müzik çalar, telefon aygıtı veya araç içi bilgisayarlarla yalnızca kelimeler ve cümleler kullanarak doğal dil konuşma tanıma sistemleri sunar.

Sağlık personeli ve yasal firma çalışanları, tanınan tüm kelimeleri ve cümleleri anında veren konuşma tanıma motorlarına karar verebilirler. Tanınan metin düzenlenir ve oturumu kapatılır ve hedeflenen alıcıya gönderilir. Buna ertelenmiş konuşma tanıma denir ve sağlık sektöründe yaygın olarak kullanılmaktadır.

Konuşma tanıma teknolojisi avcı uçak sistemlerinde yaygın olarak kullanılmaktadır (Chen 2006). Bu tür sistemler, otomatik pilot sistemini kumanda etme, radyo frekanslarını ayarlama, silah bırakma parametrelerini ayarlama, direksiyon noktası koordinatlarını ayarlama ve uçuş ekranını kontrol etme gibi görevleri yerine getirebilirler.

Konuşma tanıma teknolojisi, engelli öğrencilerin (görme bozukluğu olan, üst ekstremitelerde yaralananlar gibi) fiziksel olarak fare ve klavyeyi kullanmak zorunda kalmadan bilgisayar kullanmalarını sağlar. Engellilerin noktalama, yazım denetimi ve diğer yazma mekaniği hakkında endişelenmeden yazmasını mümkün kılar. Dahası, engelli kullanıcıların internette gezinmenin keyfini çıkarmasına, dil öğreniminde

konuşma ve dinleme becerilerini geliştirmesine ve hatta bilgisayarları ve elektronik cihazları kullanmasına yardımcı olur.

Ses tabanlı aramaların en hızlı büyüyen mobil arama türü olduğu ve gelişmiş dünyadaki bir çok teknolojik cihazın %50'sinden fazlasının akıllı ses sistemlerine sahip olacağı ve internet aramalarının yaklaşık %50'sinin 2022'ye kadar sesle aktive edilen kategoride olacağı tahmin edilmektedir.

1.3. Makine Öğrenmesi

Makine öğrenmesi yapay zekanın bir alt kümesidir (Alim and Rashid 2018). Yapay zeka, kısaca, insani ve rasyonel olarak düşünebilen ve hareket edebilen makineler yaratma sanatı ve bilimidir. Bir makinenin insanca düşünebilmesi için öğrenme, problem çözme ve karar verme gibi insan düşüncesiyle ilişkilendirdiğimiz faaliyetleri otomatikleştirmesi gerekir. Rasyonel olarak düşünen makineler algılayabilir, akıl yürütebilir ve harekete geçebilir. İnsanca hareket eden makineler, insanlar tarafından gerçekleştirildiğinde zeka gerektiren fonksiyonları da yerine getirebilir. Rasyonel olarak hareket eden makineler, akıllı davranış sergiler.

İnsanca hareket eden makineler insanlarla başarılı bir şekilde iletişim kurmak (doğal dil işleme); bildiklerini veya duyduklarını depolamak (bilgi temsili); saklanan bilgileri soruları cevaplamak ve yeni sonuçlar çıkarmak için kullanmak (otomatik akıl yürütme); gibi yeteneklere sahiptir. Yeni koşulları öğrenerek bunlara uyum sağlarlar, kalıpları tespit ve tahmin ederler (makine öğrenimi) (Russell and Norvig 2016).

Makine öğrenimi, makinelere açık bir şekilde programlanmadan verilen verilerden otomatik olarak öğrenme yeteneği sağlayan yapay zeka (AI) uygulamasıdır. Makine öğrenimi, tıpkı insanlar ve hayvanlar gibi deneyimlerden öğrenmek için veri analizi tekniklerini kullanır (deneyimlerden öğrenme). Makine öğrenme yöntemleri, önceden belirlenmiş denklemlere dayanmadan, doğrudan verileri kullanarak öğrenmek için

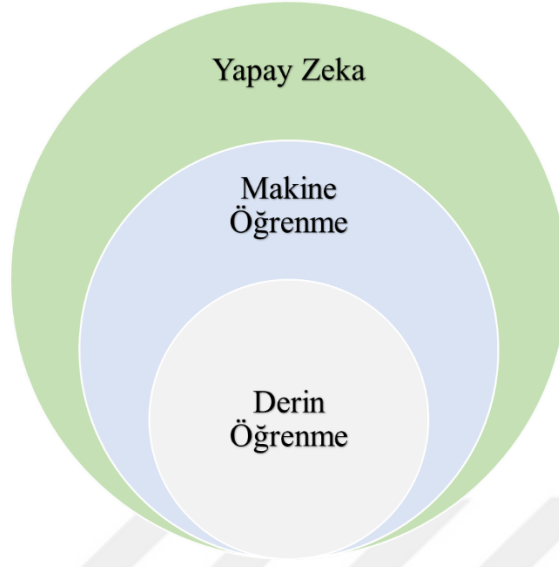
gerekli hesaplama yöntemlerini kullanır. Öğrenme için mevcut veri örneklerinin sayısı arttıkça, algoritmaların performansı uyarlanabilir şekilde artar (Russell and Norvig 2016).

Eğer P tarafından ölçülen T üzerindeki performansı E deneyimiyle iyileşirse, bir bilgisayar programının, bazı görev (T) ve bazı performans ölçüsü (P) ile ilgili olarak E deneyimden (E) ders aldığı söylenir (Tom 1997).

Örneğin, konuşma tanıyıcı, hedef kelimelerin (örn. Konuşmacılar tarafından söylenen) verilen örneklerini tanımayı öğrenebilen bir makine öğrenme programıdır. Sistemin öğrenmek için kullandığı örneklerle eğitim seti denir. Her eğitim örneğine eğitim seti denir. Bu durumda T görevi, tanıyıcı sözlüğünde yeni ses örneklerinin olup olmadığını tanımaktadır. Deneyim E, eğitim verisidir ve performans ölçüsü P doğru tanınan kelimelerin oranıdır. Bu özel performans ölçüsüne doğruluk denir ve genellikle sınıflandırma görevlerinde kullanılır.

Bir makine öğrenme programını eğitmek için iyi etiketlenmiş bir veri kümesi gereklidir. Veri kümesi, sistemin performansında hayati bir rol oynar. Veri kümesi iki veya üç alt kümeye bölünebilir. Bunlar eğitim, validasyon ve test setleridir. Eğitim seti bilinen bir çıktı içerir ve model daha sonra diğer verilerde genelleştirilmek için bu veriler üzerinden öğrenmeyi gerçekleştirir. Test veri seti modelin bu alt kümedeki tahminini test etmek için kullanılır. Validasyon veri seti, model hiperparametrelerini ayarlarken eğitim veri setine uygun bir modelin tarafsız bir değerlendirmesini sağlamak için kullanılan veri örneğidir (Schmidhuber 2015).

Derin öğrenme, makine öğreniminin özel bir şeklidir. Çoğu derin öğrenme yöntemi sinir ağı mimarilerini kullanır, bu yüzden derin öğrenme modelleri genellikle derin sinir ağları olarak adlandırılır. “Derin” terimi genellikle sinir ağındaki gizli katmanların sayısını ifade eder. Geleneksel sinir ağları sadece 2-3 gizli katman içerirken, derin ağlar 150'ye kadar katman içerebilir. Şekil 1.5, Yapay zeka, makine öğrenme ve derin öğrenme arasındaki ilişkiyi gösterir.



Şekil 1.5. Yapay zeka, makine öğrenme ve derin öğrenme arasındaki ilişki

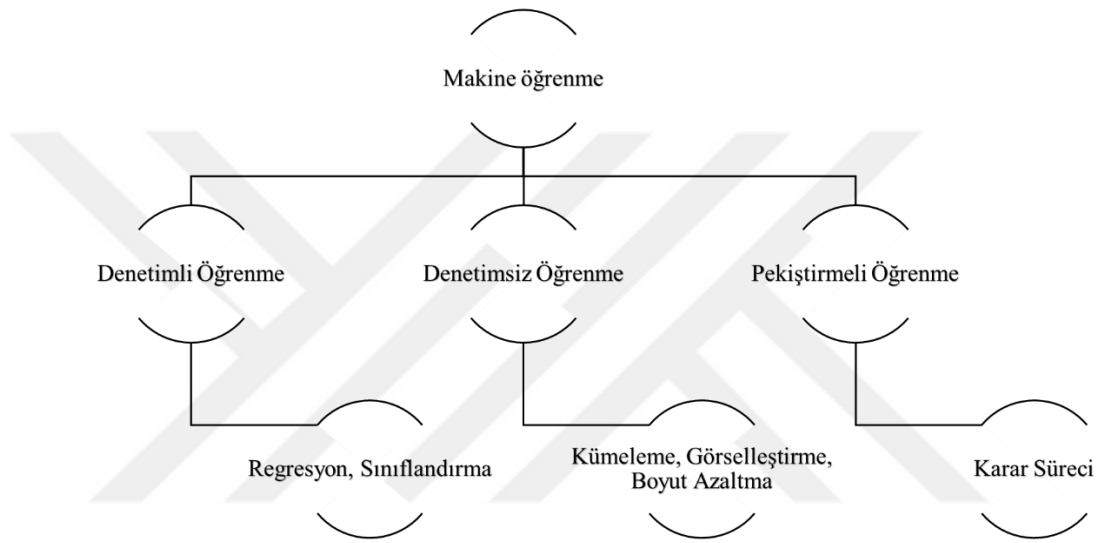
Derin öğrenme modelleri, manuel özelliğe ihtiyaç duymadan özellikleri doğrudan verilerden öğrenen büyük etiketli veri setleri ve sinir ağı mimarileri kullanılarak eğitilir. Derin öğrenme ağlarının temel bir avantajı, eğitim verilerinin boyutu arttıkça genellikle gelişmeye devam etmeleridir (Arel *et al.* 2010).

Makine öğrenimi, aşağıdaki gibi durumlarla başa çıkarken kullanılabilir:

- Verilerin doğası sürekli değişiyor ve sistemin adapte olması gerekiyor; alışveriş eğilimlerini tahmin etme, enerji talebini tahmin etme ve otomatik ticarete olduğu gibi.
- Bir görevin kuralları sürekli değişmektedir; işlem kayıtlarındaki sahtekarlık tespitinde olduğu gibi.
- El yazısı kuralları ve formülleri çok karmaşıktır; yüz tanıma ve konuşma tanıma gibi.

Makine öğrenim sistemleri, sınıflandırmanın dayandığı kriterlere bağlı olarak birçok türe ayrılabilir. İnsan gözetimi ile eğitilmiş olsun ya da olmasın, makine öğrenmesi Denetimli, Denetimsiz, Yarı Denetimli ve Pekiştirmeli öğrenme olarak sınıflandırılabilir. Hızlı bir şekilde adım adım öğrenip öğrenemeyeceğine bağlı olarak, makine öğrenimi çevrimiçi veya toplu öğrenmeye ayrılabilir. Ayrıca, makine öğrenimi yeni veri noktalarını bilinen

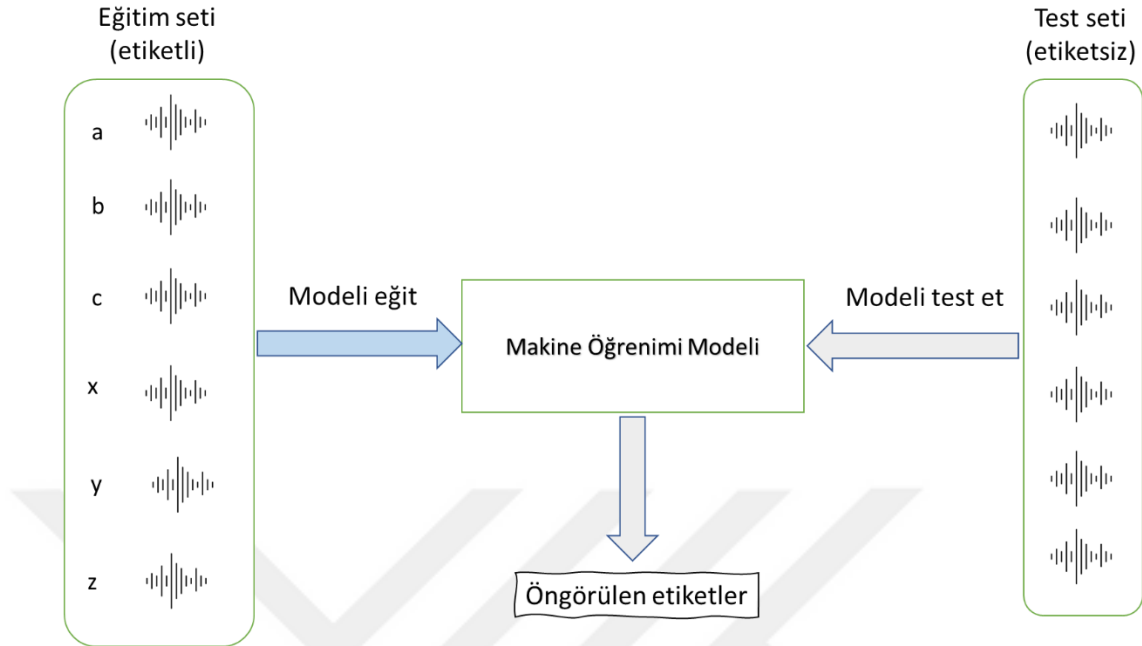
veri noktalarıyla basitçe karşılaştırarak veya bunun yerine eğitim verilerindeki kalıpları tespit edip, bilim adamlarının yaptığı gibi, öngörülü bir model oluşturarak işe yarayıp yaramadığına bakarak, örnek tabanlı ya da model tabanlı öğrenme olarak sınıflandırılabilir. Bu kriterler tek değildir; sistem tasarımcısı tarafından herhangi bir şekilde kombine edilebilirler (Ayodele 2010).



Şekil 1.6. Makine öğrenimi sınıflandırması

1.3.1. Denetimli öğrenme

Denetimli öğrenmede, algoritmaya verilen eğitim verileri örnek giriş-çıkış çiftlerini içerir. Bu değişkenler setini kullanarak, girişleri istenen çıkışlarla eşleştiren bir fonksiyon üretilir. Eğitim süreci, model eğitim verileri üzerinde istenen bir doğruluk seviyesine ulaşana kadar devam eder. Tipik denetimli öğrenme görevi sınıflandırmadır. Konuşma tanıma sistemi buna iyi bir örnektir. Bir konuşma tanıma sistemi, sınıfları (etiketleri) ile birlikte birçok ses örneği ile eğitilir ve öğrendikten sonra, konuşma ifadelerini karşılık gelen sınıflara sınıflandırabilir. Denetimli Öğrenme algoritmalarına örnek olarak, k-En Yakın Komşular, Doğrusal Regresyon, Lojistik Regresyon, Destek Vektör Makineleri (SVM'ler), Karar Ağaçları ve Rasgele Ormanlar ve Sinir ağları verilebilir (Kotsiantis 2007).



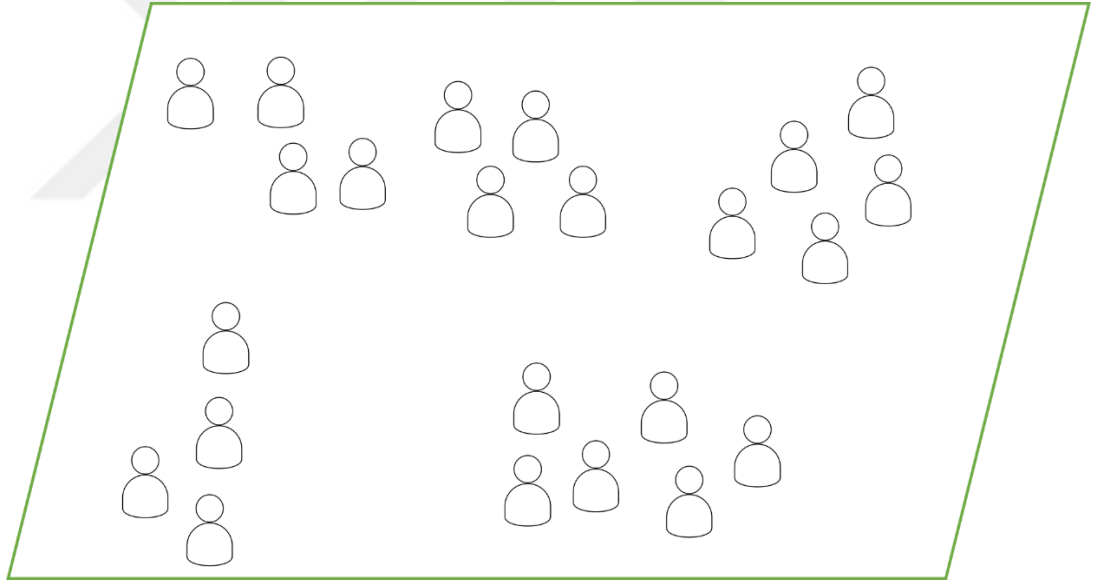
Şekil 1.7. Denetimli öğrenme için etiketli eğitim seti (örneğin, konuşma tanıma)

1.3.2. Denetimsiz öğrenme

Denetimsiz öğrenme, algoritmanın sınıflandırma ve etiket bilgileri üzerinde rehberlik yapmadan hareket etmesine izin verdiği bir makine öğrenme kategorisidir (Earnshaw *et al.* 2019). Etiketlenmemiş veriler üzerinde eğitilen bir algoritmadır. En yaygın kullanılan denetimsiz öğrenme algoritmaları şunlardır: Kümeleme algoritmaları (örn. K-Ortalamlar, Hiyerarşik Küme Analizi (HCA) ve Beklenti Maksimizasyonu (Expectation Maximization)), Görselleştirme ve boyutsallık azaltma (Ana Bileşen Analizi (PCA), Çekirdek PCA, Yerel Doğrusal Gömme (LLE) , t-dağılımlı Stokastik Komşu Gömme (t-SNE)) ve Birliktelik kural öğrenmesi (Ayodele 2010).

Kümeleme, veri noktalarının gruplanmasını içerir. Bir veri noktası seti verildiğinde, her veri noktasını belirli bir grupta sınıflandırmak için kümeleme algoritması kullanılabilir. Teorik olarak, aynı gruptaki veri noktaları genellikle benzer özelliklere sahipken, farklı gruplardaki veri noktaları oldukça farklı özelliklere sahiptir. Kümeleme algoritmaları, noktaların hangi gruplara düştüğünü görerek verilerden bazı değerli bilgiler edinilmesine yardımcı olabilir. Görselleştirme algoritmalarına çok sayıda karmaşık ve etiketsiz veri

verildiğinde, verilerin kolayca çizilebilen 2B veya 3B gösterimini çıkarırlar. Bu algoritmalar, mümkün olduğunca çok yapıyı korumaya çalışırlar (örneğin, giriş alanındaki ayrı kümelerin görselleştirmede çakışmasını engellemeye çalışır), böylece verilerin nasıl düzenlendiğini anlayabilir ve belki de beklenmeyen kalıplar tanımlanabilir. Boyutsal küçültmenin amacı, çok fazla bilgi kaybetmeden verileri basitleştirmektir. Bunu yapmanın bir yolu, birbiriyle ilişkili birkaç özelliği bir araya getirmektir. Başka bir makine öğrenme algoritmasına (denetimli öğrenme algoritması gibi) girdi olarak vermeden önce, bir boyutsallık azaltma algoritması kullanarak eğitim verisinin boyutunun azaltılması genellikle algoritmanın çok daha hızlı çalışmasına, verilerin daha az disk ve bellek alanı kaplamasına ve bazı durumlarda daha iyi performans göstermesine yol açtığı için tavsiye edilmektedir (Sathya *et al.* 2013).

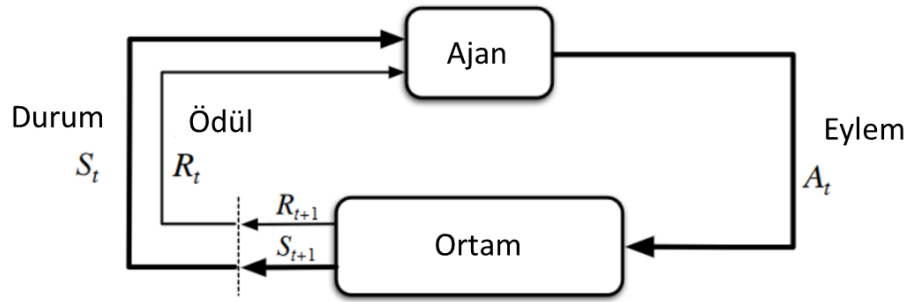


Şekil 1.8. Denetimsiz öğrenme için etiketlenmemiş bir eğitim seti

Denetimsiz öğrenmede, kategoriler olmasa bile, sıralanmamış bilgiler benzerliklere ve farklılıklara göre gruplandırılabilir. Denetimsiz öğrenme algoritmaları, denetimli algoritmalarından daha karmaşık görevleri gerçekleştirebilirler. Denetimli öğrenmenin aksine, denetimsiz bir öğrenme algoritmasının ne kadar iyi performans gösterdiğine ilişkin metrikler bulmak her zaman kolay değildir. “Performans” genellikle öznel ve alana özgüdür.

1.3.3. Pekiştirmeli öğrenme

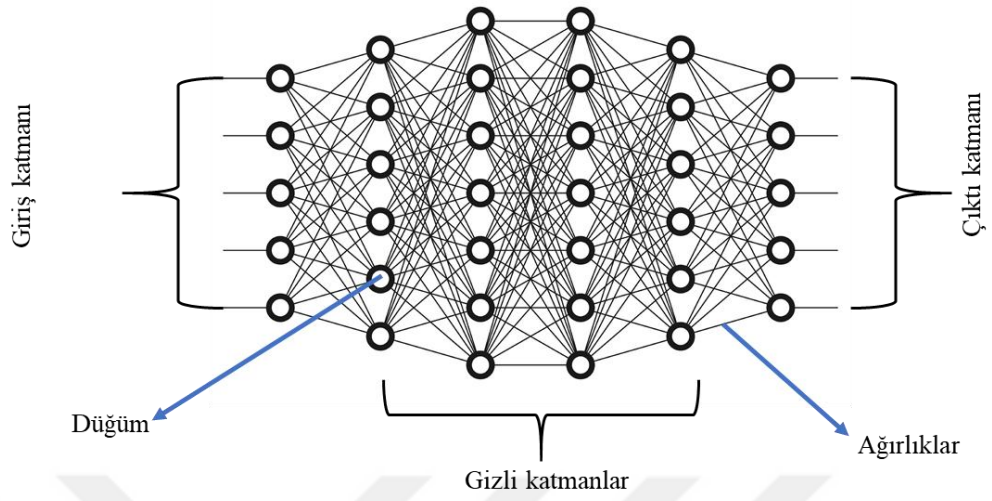
Pekiştirmeli Öğrenme, bir ajanın kendi eylem ve deneyimlerinden gelen geribildirimleri kullanarak deneme yanılma yoluyla etkileşimli bir ortamda öğrenmesini sağlayan bir tür makine öğrenme tekniğidir. Belirsiz ve potansiyel olarak karmaşık bir ortamda, ajan bir hedefe ulaşmayı öğrenir. Bu bağlamda temsilci olarak adlandırılan öğrenme sistemi, çevreyi gözlemleyebilir, eylemleri seçebilir, gerçekleştirebilir ve karşılığında ödül veya olumsuz ödüller biçiminde ceza alabilir. Daha sonra zaman içinde en fazla ödülü almak için politika olarak adlandırılan en iyi stratejinin ne olduğunu kendi başına öğrenmelidir. Politika, belirli bir durumda ajanın hangi eylemi seçmesi gerektiğini tanımlar (Sutton and Barto 1998).



Şekil 1.9. Pekiştirmeli öğrenme

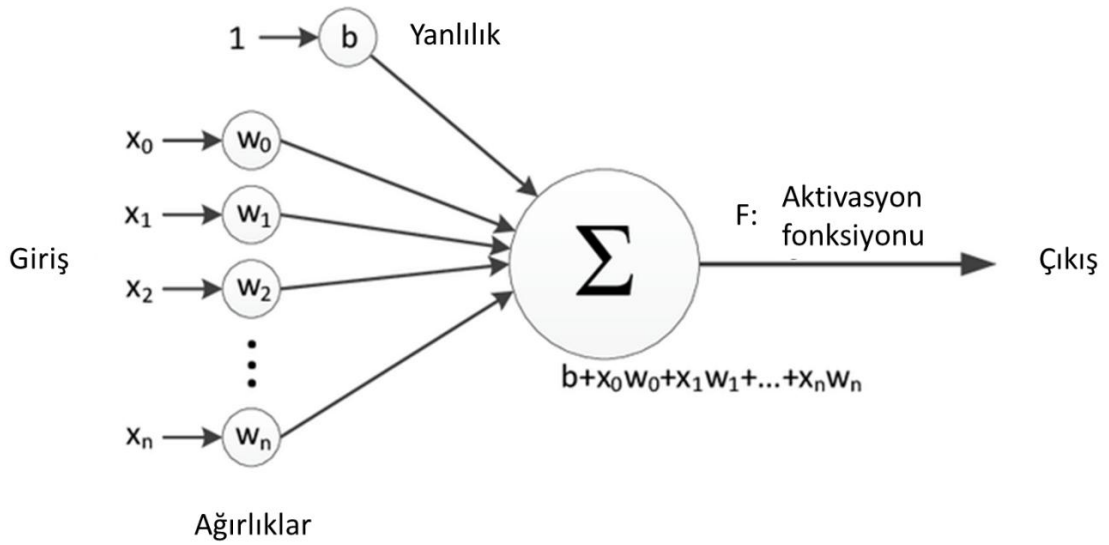
1.4. Yapay Sinir Ağları

Yapay Sinir Ağları veya YSA, beyin süreç bilgisi gibi biyolojik sinir sisteminin yolundan ilham alan bir bilgi işleme paradigmasıdır. Belirli bir sorunu çözmek için birlikte çalışan çok sayıda birbiriyle bağlantılı işleme elemanlarından (nöronlar) oluşur (Russell and Norvig 2016).



Şekil 1.10. Yapay sinir ağı

Yapay nöron ağlarında nöron adı verilen hesaplama birimleri vardır. Bu yapay nöronlar, ağırlıklı değerler olan sinapslarla bağlanır. Bu, bir değer verildiğinde bir nöronun bir çeşit hesaplama (örn. Sigmoid fonksiyonu) gerçekleştireceği ve daha sonra bu hesaplamanın sonucu “seyahat ederken” bir ağırlıkla çarpılacağı anlamına gelir. Ağırlıklı sonuç bazen sinir ağı veya katmanlar halinde yapılandırılmış daha fazla nöron olabilir. Tek katmanlı sinir ağına Algılayıcı (Perceptron) denir ve tek bir çıktı verir. Şekil 1.11 bir algılayıcının yapısını göstermektedir.



Şekil 1.11. Algılayıcı yapısı

Yukarıdaki şekilde, tek bir gözlem için, $x_1, x_2, x_3, \dots, x_n$ ağırlıklı girişleri (bağımsız değişkenler) temsil eder. Bu girişlerin her biri bir bağlantı ağırlığı veya sinaps ile çarpılır. Ağırlıklar $w_1, w_2, w_3, \dots, w_n$ olarak temsil edilir. Ağırlık, belirli bir düğümün gücünü gösterir. b bir yanlılık değeridir. Yanlılık değeri, aktivasyon fonksiyonunu yukarı veya aşağı kaydırmanıza olanak tanır. En basit durumda, bu bağlantılar toplanır, bir sonuç üretmek için bir transfer fonksiyonuna (aktivasyon fonksiyonu) beslenir ve bu sonuç çıktı olarak gönderilir.

Matematiksel olarak, $x_0 \cdot w_0 + x_1 \cdot w_1 + x_2 \cdot w_2 + x_3 \cdot w_3 + \dots + x_n \cdot w_n + b = \sum (x_i w_i) + b$ şeklinde gösterilir.

1.4.1. Aktivasyon fonksiyonları

Bir aktivasyon fonksiyonu, bir nöronun aktif olup olmadığına karar veren, girdilerinin ağırlıklı toplamını hesaplayıp ona bir yanlılık ekleyerek karar verir. Aktivasyon fonksiyonları yapay sinir ağlarında öğrenmenin temel parçasıdır. Yapay bir sinir ağının temel amacı, bir düğümün giriş sinyalini bir çıkış sinyaline dönüştürmektir. Bu çıkış sinyali daha sonra nöral ağ katmanları yığınındaki bir sonraki katmana girdi olarak kullanılır. Amaç, bir nöronun çıkışını doğrusal olmayan hale getirmektir. Bir aktivasyon fonksiyonu, giriş değerleri değiştikçe yumuşak bir geçiş sağlar, yani girişteki küçük bir değişiklik çıktıda küçük bir değişiklik üretir.

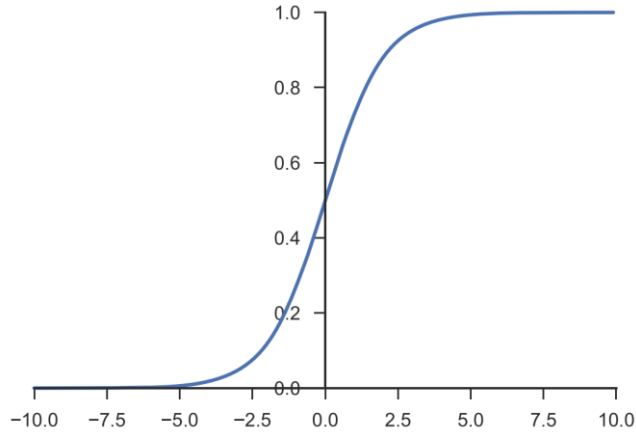
$$y = \sum (ağırlık * giriş) + yanlılık \quad (1.1)$$

En yaygın kullanılan aktivasyon fonksiyonları: Sigmoid Aktivasyon Fonksiyonu, Hiperbolik Tanjant Fonksiyonu ve Rektifiye Doğrusal Birimler (ReLU) (Yamashita *et al.* 2018).

Sigmoid fonksiyonu, bir tür aktivasyon fonksiyonudur ve daha özel olarak bir ezme(squashing) fonksiyonu olarak tanımlanır. Sigmoid fonksiyonları, çıktıyı 0 ile 1

arasında bir sınırla sınırlar ve bu çıktıyı olasılıksal olarak gösterebilmek için kullanışlı hale getirir. Sigmoid aktivasyon fonksiyonunun matematiksel ifadesi (1.2) denklemi ile verilmiştir. Şekil 1.12’de gösterildiği gibi S şeklinde bir eğridir.

$$f(x) = \frac{1}{1 + e^{-x}} \quad (1.2)$$

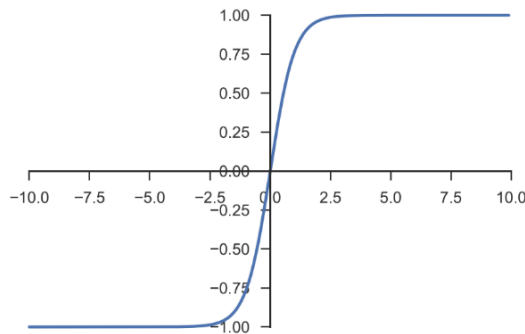


Şekil 1.12. Sigmoid fonksiyonu

Son zamanlarda Hiperbolik tanjant (tanh) fonksiyonunun yakınsamada 6 kat iyileşme sağladığı kanıtlanmıştır. Tanh fonksiyonu, aşağıdaki matematiksel formülle tanımlanır.

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (1.3)$$

Tanh fonksiyonunun aralığı -1 ile +1 arasındadır ve bu nedenle çıkışı sıfır merkezlidir ve grafiği Şekil 1.13’te gösterilmiştir.

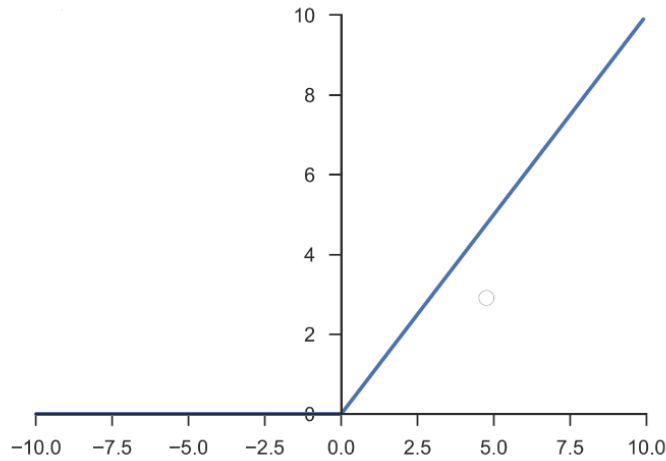


Şekil 1.13. Tanh fonksiyonu

ReLU son birkaç yılda çok popüler olmuştur. ReLU $[0, \infty)$ aralığındadır ve doğrusal değildir. ReLU matematiksel olarak aşağıdaki gibi tanımlanır.

$$f(x) = \max(0, x) \quad (1.4)$$

x pozitifse x , aksi takdirde 0 çıktısı verir. Bu fonksiyonun matematiksel biçiminden çok basit ve etkili olduğu çıkarılabilir. Makine öğrenimi ve bilgisayar bilimlerinde çoğu zaman, en basit ve tutarlı teknik, yöntemlerin sadece tercih edildiğini ve en iyisi olduğunu görüyoruz. Bu nedenle, neredeyse tüm derin öğrenme modellerinde bugünlerde ReLU kullanılmaktadır. ReLU fonksiyonunun grafiği Şekil 1.14'te gösterilmiştir.



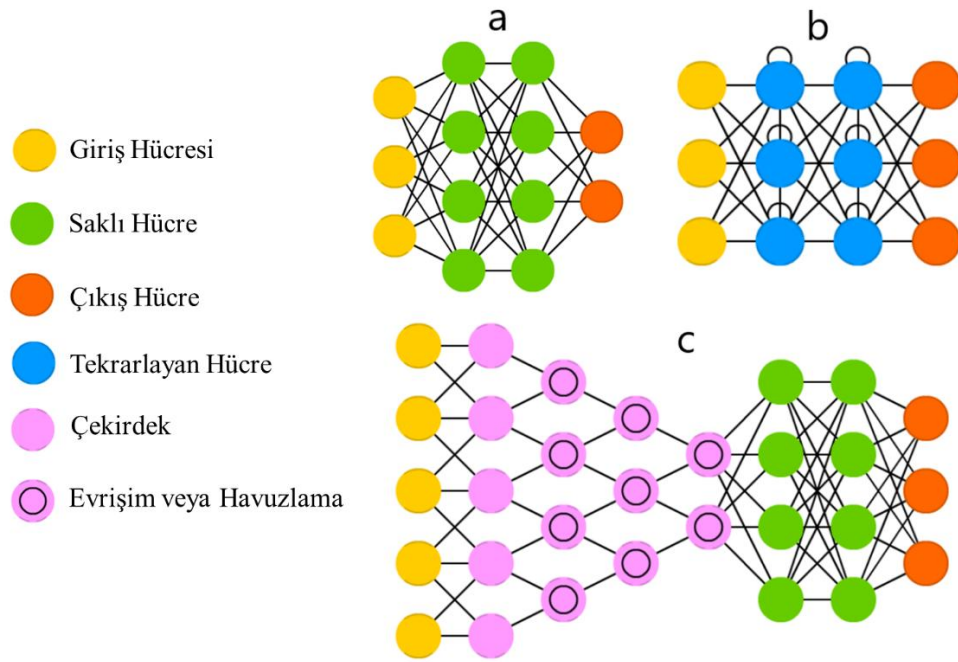
Şekil 1.14. ReLU fonksiyonu

1.4.2. YSA mimarisi ve türleri

Bir yapay sinir ağı yapay nöronlardan oluşur ve birbirine bağlı üç katman halinde düzenlenir: giriş katmanı, birden fazla katman içerebilen gizli katman ve çıkış katmanı. Giriş katmanı, gizli katmana bilgi gönderen giriş nöronlarını içerir. Gizli katman, verileri çıkış katmanına gönderir. Her nöron ağırlıklı girişlere, bir aktivasyon fonksiyonuna (bir giriş verilen çıkışı tanımlar) ve bir çıkışa sahiptir. Ağırlıklar, bir sinir ağını parametrelili bir sisteme dönüştüren ayarlanabilir parametrelerdir. Girişlerin ağırlıklı toplamı, nöronun bir çıkış elde etmek için aktivasyon fonksiyonuna geçirilen aktivasyon sinyalini

üretir. Eğitim, tahmin hatasının en aza indirildiği ve ağı belirli bir doğruluk seviyesine ulaştığı ağırlık optimizasyonu sürecidir. Her bir nöronun hata katkısını belirlemek için, kayıp fonksiyonunun gradyanını hesaplayan en çok kullanılan yöntem geri yayılım algoritmasıdır (backpropagation) (Leondes 2018).

YSA'ların yapması gereken görevlere bağlı olarak farklı türleri vardır. Yapay sinir ağlarının çoğu, daha karmaşık biyolojik yapıları ile benzerlik gösterir ve amaçlanan görevlerinde çok etkilidir (Russell and Norvig 2016). En yaygın kullanılan YSA'lar Çok Katmanlı Perceptron (MLP), Konvolüsyon sinir ağları (CNN) ve Tekrarlayan Sinir Ağlarıdır (RNN).



Şekil 1.15. MLP (a), RNN (b), CNN (c)

Çok katmanlı bir algılayıcı (MLP) üç veya daha fazla katmana sahiptir. Doğrusal olarak ayrılamayan verilerin sınıflandırmasını sağlayan doğrusal olmayan bir aktivasyon fonksiyonu kullanır. Katmandaki her düğüm, bir sonraki katmandaki her düğüme bağlanır ve ağı tam olarak bağlar. Bir evrişimli sinir ağı (CNN), bir ya da daha fazla evrişimsel katman içerir. Evrişim katmanları, sonucu bir sonraki katmana aktarmak için girişe bir

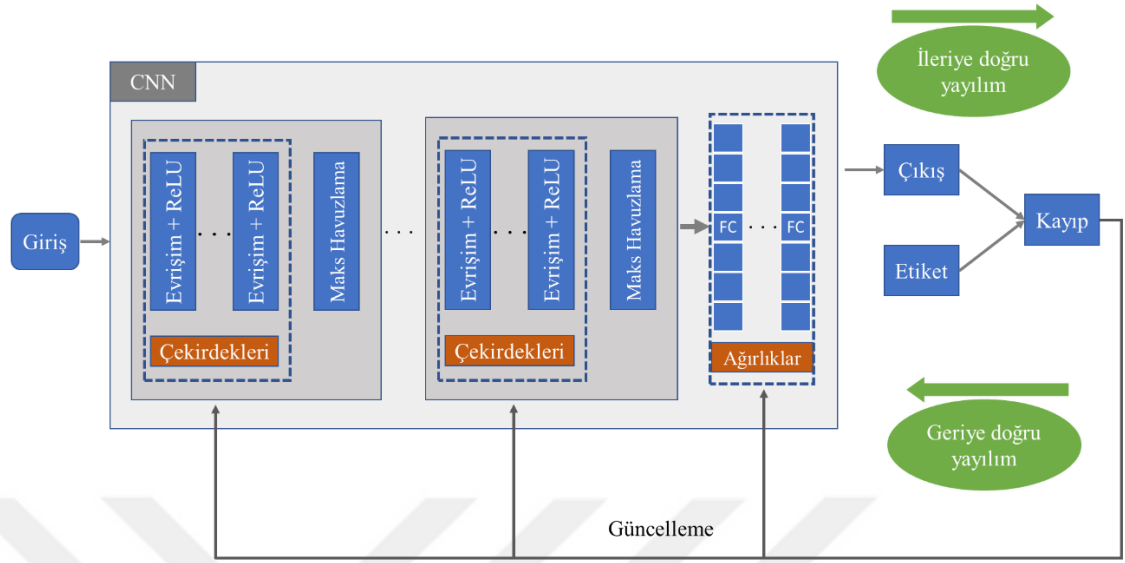
evrişim işlemi uygular. Bu işlem, ağı çok daha az parametre ile daha derin olmasını sağlar. Tekrarlayan sinir ağları, gizli durumlara sahip önceki durumların giriş olarak kullanılmasına izin verir ve ağırlıklar zaman içinde paylaşılarak, hesaplamalarda geçmiş bilgiler dikkate alınmış olur. Ayrıca, model boyutu giriş boyutunun artması ile artmaz ve herhangi bir uzunlukta girişlerin işlenmesi mümkündür. Şekil 1.15 yaygın olarak kullanılan sinir ağlarının mimarisini gösterir.

1.4.3. Evrişimli sinir ağları

Evrişimli sinir ağı, çeşitli bilgisayarlı görme (computer vision) görevlerinde baskın hale gelen ve çeşitli alanlarda ilgi çeken bir derin öğrenme yöntemidir. Nöronların arasındaki bağlantı örüntüsü, görsel korteks sisteminin organizasyonundan esinlenmiştir. Evrişimli sinir ağı, evrişim katmanları, havuzlama katmanları ve tamamen bağlı katmanlar gibi birden fazla yapı taşından oluşur ve bir geri yayılım algoritması aracılığıyla özelliklerin uzamsal hiyerarşilerini otomatik ve uyarlamalı olarak öğrenmek için tasarlanmıştır (Chartrand *et al.* 2017).

CNN mimarisinin yapı taşları

CNN, tipik olarak üç tip katmandan (veya yapı taşlarından) oluşan matematiksel bir yapıdır: evrişim, ortaklama ve tamamen bağlı katmanlar. İlk ikisi, evrişim ve ortaklama katmanları, özellik çıkarma işlemini gerçekleştirirken, üçüncüsü, tamamen bağlı katman, çıkarılan özellikleri sınıflandırmaktadır. Tipik bir mimari, birkaç evrişim katmanından ve bir ortak katmandan oluşan bir yapının ardından bir veya daha fazla tamamen bağlı katmanın tekrarlanmasından oluşur. Giriş verilerinin bu katmanlar üzerinden çıktıya dönüştürüldüğü adıma ileri yayılma denir. Şekil 1.16. evrişimli sinir ağı (CNN) mimarisine ve eğitim sürecine genel bir bakış sunar. Bu bölümde, 2D CNN için açıklanan evrişim ve ortaklama işlemleri, üç boyutlu (3D) -CNN için de benzer işlemler uygulanarak yapılabilir.



Şekil 1.16. CNN mimarisine ve eğitim sürecine genel bakış

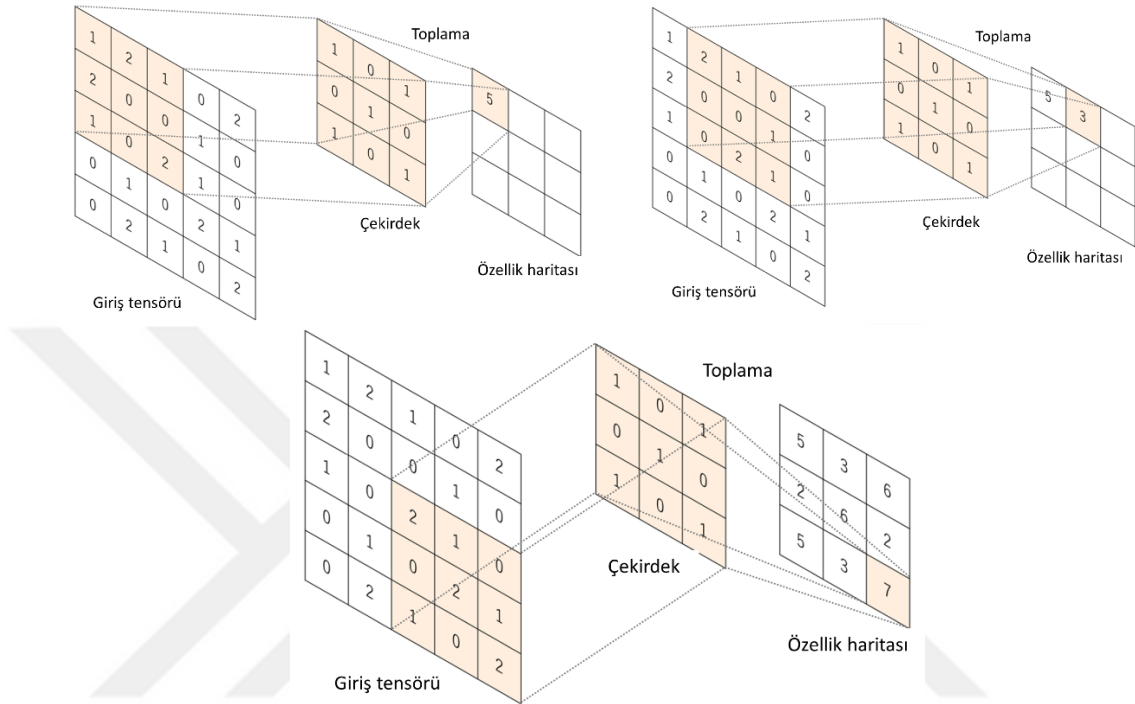
Evrişim katmanı

Bir evrişim katmanı, tipik olarak doğrusal ve doğrusal olmayan işlemlerin, yani evrişim işlemi ve aktivasyon fonksiyonunun bir kombinasyonundan oluşan özellik çıkarmayı gerçekleştiren CNN mimarisinin temel bir bileşenidir.

Konvolüsyon, tensör adı verilen bir sayı dizisi olan girişe, çekirdek adı verilen küçük bir sayı dizisinin uygulandığı özellik çıkarma için kullanılan özel bir doğrusal işlem türüdür. Çekirdeğin her bir elemanı ve giriş tensörü arasındaki bir ürün, tensörün her bir yerinde hesaplanır ve bir özellik haritası adı verilen çıkış tensörünün karşılık gelen pozisyonunda çıkış değerini elde etmek için toplanır (Yamashita *et al.* 2018). Şekil 1.17 çekirdek boyutu 3×3 olan, dolgu olmayan ve adım boyutu 1 olan bir evrişim operasyonu örneğini göstermektedir.

Bu prosedür, giriş tensörlerinin farklı özelliklerini temsil eden gelişigüzel sayıda özellik eşlemesi oluşturmak için çoklu çekirdekler uygulanarak tekrarlanır; farklı çekirdekler, bu nedenle, farklı özellik çıkarıcılar olarak düşünülebilir. Evrişim işlemini tanımlayan iki temel hiperparametre, boyut ve çekirdek sayısıdır. Birincisi tipik olarak 3×3 'tür, ancak

bazen 5×5 veya 7×7 'dir. İkincisi isteğe bağlıdır ve çıktı özellik haritalarının derinliğini belirler.



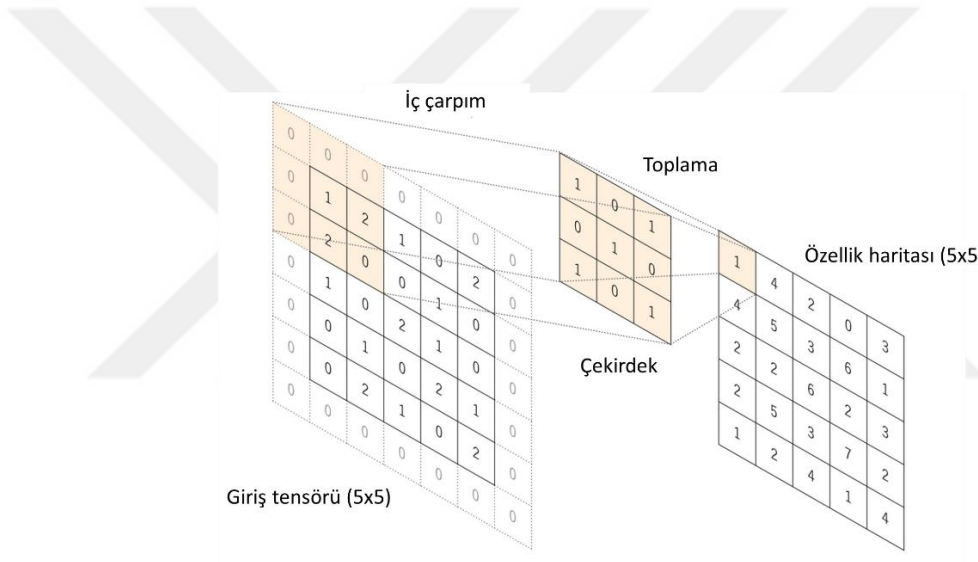
Şekil 1.17. Evrişim işlemine bir örnek

Yukarıda açıklanan evrişim işlemi, her çekirdeğin merkezinin giriş tensörünün en dıştaki elemanı ile örtüşmesine izin vermez ve giriş tensörüne kıyasla çıkış özelliği haritasının yüksekliğini ve genişliğini azaltır. Dolgu, tipik olarak sıfır dolgusu, bu sorunu gidermek için kullanılan bir tekniktir, burada giriş tensörünün her iki tarafına sıfır satırları ve sütunları eklenir, böylece evrişim işlemi bir çekirdeğin merkezini ve en dıştaki öğeye sığdırır ve boyut ile aynı düzlemde tutar.

Şekil 1.18, sırasıyla 5×5 giriş ve çıkış boyutlarına sahip resme çekirdek boyutu 3×3 ve 1 adım sıfırı doldurma ile bir evrişim işlemini göstermektedir.

Modern CNN mimarileri, daha fazla katman uygulamak için düzlem içi boyutları korumak için genellikle sıfır dolgu kullanır. Sıfır dolgusu olmadan, birbirini izleyen her özellik haritası evrişim işleminden sonra küçülür.

Birbirini izleyen iki çekirdek konumu arasındaki mesafeye, evrişim işlemi de tanımlayan adım adı verilir. Bir adımın ortak seçimi 1'dir; bununla birlikte, bazen özellik haritalarının altörneklenmesini sağlamak için 1'den büyük bir adım kullanılır. Alt örnekleme gerçekleştirmek için alternatif bir teknik, aşağıda tarif edildiği gibi bir ortaklama işlemidir.



Şekil 1.18. Sıfır dolgu ile konvolüsyon çalışması

Bir evrişim işleminin temel özelliği ağırlık paylaşımıdır: çekirdekler tüm görüntü konumlarında paylaşılır. Ağırlık paylaşımı, evrişim işlemlerinin aşağıdaki özelliklerini oluşturur:

- Çekirdekler tarafından çıkarılan yerel özellik modellerinin, çekirdekler tüm görüntü konumlarında dolaşırken ve öğrenilen yerel kalıpları tespit ederken değişmez.
- Bir ortaklama işlemiyle birlikte altörnekleme yaparak özellik desenlerinin uzamsal hiyerarşilerini öğrenmek, giderek daha geniş bir görüş alanını yakalamakla sonuçlanır.
- Tamamen bağlı sinir ağlarına kıyasla öğrenilecek parametre sayısını azaltarak model verimliliğini artırmak.

Daha sonra açıklanacağı üzere, bir CNN modelinin evrişim katmanı ile ilgili eğitim süreci, belirli bir eğitim veri kümesine dayanarak belirli bir görev için en iyi çalışan çekirdekleri tanımlamaktır. Çekirdekler, evrişim katmanındaki eğitim sürecinde otomatik olarak öğrenilen tek parametredir; diğer yandan, çekirdeklerin boyutu, çekirdek sayısı, dolgu ve adım, eğitim süreci başlamadan önce ayarlanması gereken hiperparametrelerdir.

Evrişim gibi doğrusal bir işlemin çıktıları daha sonra doğrusal olmayan bir aktivasyon fonksiyonundan geçirilir. Günümüzde kullanılan en yaygın doğrusal olmayan aktivasyon fonksiyonu, önceki bölümlerde açıklanan doğrultulmuş doğrusal ünitedir (ReLU).

Havuzlama katmanı

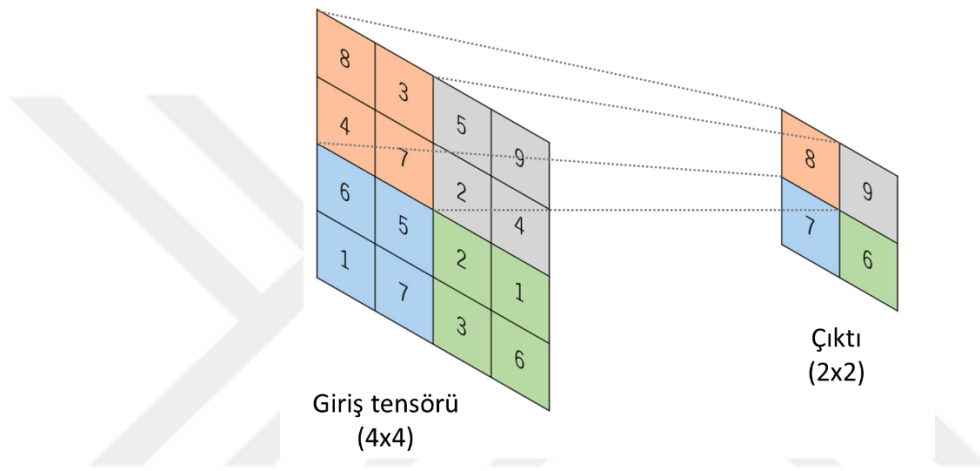
Bir havuzlama katmanı, küçük kaymalara ve bozulmalara bir geçiş değişmezliği kazandırmak ve sonraki öğrenilebilir parametre sayısını azaltmak için özellik haritalarının düzlem içi boyutluluğunu azaltan tipik bir aşağı örnekleme işlemi sağlar. Havuzlama katmanlarının hiçbirinde öğrenilebilir bir parametre bulunmamaktadır. Filtre boyutu, adım ve dolgu, konvolüsyon işlemlerinde olduğu gibi havuzlama işlemlerinde de hiperparametreler kullanılmaktadır.

Havuzlama işleminin en popüler biçimi, giriş özellik haritalarından ilgili bölgeleri alan, her bir bölgede maksimum değeri veren ve diğer tüm değerleri atayan maksimum ortaklamadır. Şekil 1.19, Maksimum havuzlama oluşturma örneğini göstermektedir. Uygulamada yaygın olarak 2 adımlı bir filtreye sahip 2×2 büyüklüğünde bir maksimum havuzlama kullanılır. Bu özellik haritalarının düzlem içi boyutunu yarıya indirir. Yükseklik ve genişliğin aksine özellik haritalarının derinlik boyutu değişmeden kalır.

Tamamen bağlı katman

Havuzlama katmanının çıktı özellik haritaları tipik olarak düzleştirilir, yani bir boyutlu (1D) sayı dizisine (veya vektöre) dönüştürülür ve yoğun katmanlar olarak da bilinen bir veya daha fazla tamamen bağlı katmana bağlanır. Her girdi her bir çıktıya öğrenilebilir

bir ağırlıkla bağlıdır. Evrişim katmanları tarafından çıkarılan ve havuzlama katmanları tarafından alt örneklenen özellikler oluşturulduktan sonra, ağı son çıktılarına tam bağlı katmanlarından her sınıf için elde edilen olasılıklar değerlendirilerek sınıflandırma yapılır. Son tam bağlı katman tipik olarak sınıf sayısı ile aynı sayıda çıkış düğümüne sahiptir. Tamamen bağlı katmanlar, yukarıda tarif edildiği gibi ReLU gibi doğrusal olmayan bir fonksiyon gibi davranır.



Şekil 1.19. Maksimum ortaklama işlemi örneği

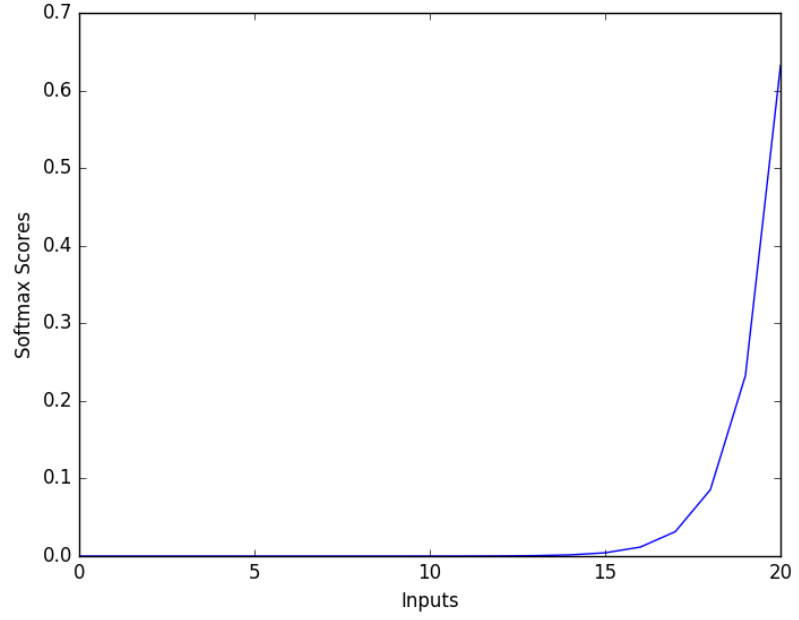
Son katman aktivasyon fonksiyonu

Son tam bağlanan katmana uygulanan aktivasyon fonksiyonu genellikle diğerlerinden farklıdır. Her göreve uygun bir aktivasyon fonksiyonun seçilmesi gerekir. Son tam bağlı katmandan çıktı gerçek değerlerini hedef sınıf olasılıklarına normalleştiren çok sınıflı sınıflandırma görevine uygulanan aktivasyon fonksiyonu, bir SoftMax fonksiyonudur; burada her bir değer 0 ile 1 arasında değişir ve tüm değerlerin toplamı 1'dir. SoftMax fonksiyonu aşağıdaki denklemde matematiksel olarak ifade edilir.

1.5 denklemi, SoftMax aktivasyon fonksiyonunun matematiksel gösterimini gösterir.

$$f(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}} \quad (1.5)$$

SoftMax aktivasyon fonksiyonunun grafiğini Şekil 1.20’de gösterilmiştir.



Şekil 1.20. SoftMax aktivasyon fonksiyonunu

2. KAYNAK ÖZETLERİ

İlk konuşma tanıma çalışmaları ABD ve Avrupa'da 40'lı yılların sonunda ve 50'lerin başında başlamıştır. Konuşmayı fonetik "atomlara" dönüştürebilen bir sistem olan "phonetographe", Dreyfus-Graf (1952) tarafından tasarlanmıştır. Bu sistem ilk konuşmacıya bağımlı, yalıtılmış rakam tanıtıcısıydı ve sıfır geçiş sayımına dayanan sınırlı sayıda akustik parametre kullanmıştır.

Fonetik birimlerde segmentasyona dayalı bir kelime tanıyıcı tasarlamak için fonetik kod çözme yaklaşımı Bell Laboratories'deki bir araştırma grubu tarafından önerilmiştir (Dudley and Balashek 1958). Tek bir konuşmacıyla ayrı ayrı dikte edilen heceleri tanıyabilen ilk “fonetik daktilo”, fonetik bir yaklaşım kullanılarak RCA laboratuvarlarında geliştirilmiştir (Olson and Belar 1956). Dil bilgisini İngilizce iki fonemin izin verilen dizileri hakkında istatistiksel bilgi biçiminde birleştiren ilk sistem, Londra Üniversitesi Koleji'nde geliştirilen temel bir fonem tanıyıcıdır (Denes 1959).

50'li yılların sonlarında ve 60'lı yılların başlarında, MIT Lincoln Laboratuvarı bilgisayar tabanlı konuşmacıdan bağımsız on sesli harf tanıma sistemi üzerinde deneyler yapmıştır (Forgie and Forgie 1959). Aynı dönemde sesli harfler , fonem ve rakam tanıma için özel amaçlı donanım, Japon araştırmacılar tarafından geliştirildi (Nagata *et al.* 1963). Rus araştırmacılar (Slutsker, 1968; Vintsyuk, 1968), dinamik programlama kullanarak dinamik zamanın yanlış yorumlanması ile ilgili temel kavramları önermiştir. Stanford Üniversitesi'nden fonemlerin dinamik takibi ile sürekli konuşmanın tanınması (Reddy 1966) beş kelimelik havuzdan oluşturulmuş 561 kelimeyle cümlelerin konuşmacıya bağlı olarak tanınmasını sağlamıştır (Vicens 1969).

Geniş sözcük dağarcığının anlaşılması, sürekli konuşma, fonetik kod çözmedeki hataları telafi etmek için sözcüksel ve sözdizimsel gibi üst düzey bilgilerin kullanımına dayanarak çığır açan bir başarı olmuştur. Şablon tabanlı yöntemler üzerine yapılan araştırmalar (Velichko and Zagoruyko 1970) bugün hala kullanılmakta olan bazı temel tekniklerin

geliştirilmesini sağlamıştır. Konuşma kalıplarının dinamik zaman atlama algoritmalarıyla eşleştirilmesi ilk olarak SSCB'de (Slutsker 1968; Vintsyuk 1968) ve Japonya'da (Sakoe and Chiba 1971) geliştirilmiştir. Klasik hızlı Fourier dönüşümü (FFT) veya filtre bankası yöntemleri (Itakura 1975) yerine doğrusal tahmin kodlamasına (LPC) dayanan konuşma analizi de bir dönüm noktası olmuştur (Velichko and Zagoruyko 1970).

İlk düşük maliyetli konuşma tanıyıcıların 1970'lerin sonlarında ticarileştirilmesi, konuşma tanıma sistemlerinin mikroişlemci panolarına uygulanmasıyla mümkün olmuştur. Metodolojide şablon tabanlı yöntemlerden HMM'lere dayanan istatistiksel modele geçiş (Ferguson 1980; Rabiner 1989) ve sinir ağları tekniklerinin (Lippmann 1987) yeniden tanıtılması konuşma tanıma teknolojisinin dönüm noktalarıydı. Algılayıcı olarak ilk sinir ağı modelleri 1950'lerde önerilmiş ve 1980'lerin sonunda yeniden ortaya çıkmıştır.

Fonetikçilerin spektrogram okumadaki uzmanlığına dayanan tasarım fonetik kod çözücüleri (Cole *et al.* 1980) sürekli konuşmanın fonetik kod çözmesini büyük ölçüde geliştirmiştir. Ayrıca, TIMIT (Fisher *et al.* 1986) gibi büyük veritabanlarının kaydedilmesi, konuşma tanımada kaydedilen ilerlemelere doğrudan katkıda bulunmuştur.

TIDIGITS (Leonard and Doddington 1992), 300 farklı konuşmacı tarafından konuşulan ve ücretli katılımcılar tarafından sessiz bir odada kaydedilen 25.000 rakam dizisini içermektedir. Veri kümesi yalnızca Language Data Consortium'dan ticari bir lisans altında edinilebilir ve modern yazılım kullanılarak kodunun çözülmesi zor olan NIST SPHERE dosya biçiminde depolanmıştır.

Konuşma tanıma sistemlerinin 1990 ve 2000'lerde dikte amaçlı geniş kelime dağarcığına (Makhoul and Schwatz 1994) genişletilmesi ve genel telefon şebekelerinde (Wilpon 1994) konuşma tanıma teknolojisinin kullanımı nedeniyle gürültülü veya olumsuz koşulların yanı sıra spontan konuşma tanıma ortaya çıkmıştır.

Tüzün *et al.* (1995) 12 kelimelik bir kelime dağarcığına sahip gizli Markov modellerine (HMM) dayanan, konuşmacıdan bağımsız bir kelime tanıma sistemi sunmuştur. Birkaç LPC temsili özellik vektörleri olarak test edilmiş ve elde edilen sonuçların umut verici olduğu gözlemlenmiştir.

Amrouche *et al.* (2003) Arapça izole edilmiş sözcük tanımasına regresyon sinir ağı (GRNN) uyguladılar. Önerilen mimari iki bölümden oluşuyordu: segmental normalizasyon ve özellik çıkarmadan oluşan bir ön işleme aşaması ve parametrik olmayan yoğunluk tahminine dayalı sinir ağlarını kullanan bir sınıflandırma aşaması. Böyle bir karşılaştırmayı gerçekleştirmek için GRNN ve geleneksel çok katmanlı algılayıcı (MLP) test edilmiştir. Çok sayıda Arapça rakam kullanılarak elde edilen sonuçlar, genel regresyona dayalı sinir ağlarının, ileri besleme geri yayılma hatasına dayanarak tanıma oranını daha fazla geliştirdiğini göstermiştir.

El Choubassi *et al.* (2003) Arapça izole konuşma tanımayı uygulamak için yeni bir yaklaşım tanımlamıştır. Konuşma tanıma (SR) literatürünün çoğu gizli Markov modellerine (HMM) dayanırken, bu çalışmada sistemleri modüler tekrarlayan Elman sinir ağları (MRENN) kullanılmıştır. Bu tasarım sayesinde elde edilen umut verici sonuçlar, bu yeni sinir ağı yaklaşımının geleneksel HMM tabanlı konuşma tanıma yaklaşımlarıyla rekabet edebileceğini göstermiştir.

Tunali (2005), Türkçe için konuşmacıya bağımlı, geniş bir kelime haznesi, yalıtılmış kelime konuşma tanıma sistemi geliştirmiştir. Akustik-fonetik yaklaşım ile konuşma tanımaya stokastik yaklaşımın bir kombinasyonu önerilmiş ve geliştirilmiştir. Akustik-fonetik yaklaşım için foneme dayalı bir modelleme kullanılmıştır. Buna ek olarak Gizli Markov Modelleri kullanılarak stokastik yaklaşım için şablon sınıflandırıcı ve örüntü tanıyıcı oluşturulmuştur. Fonemler tanıma için en küçük birim olarak kullanılır ve gözlem vektörleri olarak 22 boyutlu MFCC özellik matrisleri kullanan 2 durumlu HMM modelleri ile modellenir. Tanıma aşamasında test ifadeleri, en küçük eğitimli fonemden daha büyük olmayan parçalara ayrılmış ve daha sonra bu parçalar eğitimli HMM foneme modellerine karşı araştırılmıştır.

(Seid and Gambäck 2005), HMM/YSA karma yaklaşımına dayanan Amharca konuşmacıdan bağımsız sürekli konuşma tanıyıcı geliştirdi. Model, CSLU Araç Kiti yardımıyla bağlama bağlı bir telefon bölümü alt sözcük düzeyinde oluşturulmuştur. %74,28 kelime ve %39,70 cümle tanıma oranı ile umut verici bir sonuç elde edilmiştir. Bunlar Amharca dili için konuşma tanıma konusunda o güne kadar bildirilen en iyi rakamlar olmuştur.

Abate *et al.* (2005) fonetik olarak zengin ve dengeli bir metin veritabanı kullanarak bir okuma-konuşma Amharca corpus geliştirdi. Metin veritabanını hazırlamak için, Amharca'da yayınlanan tanınmış gazete ve dergilerden seçilen makalelerden oluşan EthioZena web sitesinin arşivi kullanmıştır. Arşiv yarı otomatik olarak temizlenmiştir. Amharca konuşma grubu bir eğitim seti, konuşmacı adaptasyon seti ve test setleri (her biri 5000 ve 20000 kelime büyüklüğüne sahip geliştirme ve değerlendirme test setleri) içerir. Konuşma Etiyopya'da bir ofis ortamında kaydedilmiş ve yarı otomatik olarak bölümlere ayrılmıştır. Kayıtlar, Amharca için hece ve telefon tabanlı LVCSR ile yapılan deneylerde kullanıldı.

Bourouba *et al.* (2006) Arapça izole kelimeleri tanımak için deneyler yaptılar. Tanıma sistemleri, denetimli sınıflandırıcının ve klasik tanıma sistemi tabanlı Markov modellemesinin bir kombinasyonuydu. Hibrit bir GHMM/denetimli sınıflandırıcı kullanılmıştır (SVM veya KNN). GHMM/SVM veya GHMM/KNN tabanlı yeni yaklaşımlar tanıtıldı, değerlendirildi ve izole kelime tanıma sistemi için geleneksel yaklaşım GHMM ile karşılaştırıldı. Bu yaklaşımların her ikisi de özellik çıkarımı ve zaman dizisi modellemesi için aynı ilkeleri uyguladı. Sonuçlar, GHMM-SVM hibritlerinin performanslarının temel sisteme kıyasla daha iyi olduğunu göstermiştir.

Nimaan *et al.* (2006), Cibuti kültürel mirasını indekslemek için Cibuti dillerinin otomatik konuşma tanıma (ASR) için ilk adımı sundu. Çalışmaları, hedeflenen Cibuti ses arşivlerinin yarısını temsil eden Somalice dilini işlemeye adanmıştır. Bir ses (10 saat) ve metinsel (3M kelime) topluluğu toplandı. Bir ASR sistemi geliştirildi ve bu ilk Somalice büyük kelime tanıyıcısının sonuçları umut vericiydi. Somali dilinin belirli özelliklerini

kullanarak (kelimeler bu makalede “kök” adı verilen alt kelimelerin bir araya getirilmesinden oluşur), daha iyi sonuçlar elde edildi.

Raza *et al.* (2009) tarafından Urduca için bağlam temelli fonetik olarak zengin bir konuşma grubu tasarlandı ve geliştirildi. Yapının amacı, Urduca konuşması için akustik modelleri eğitmek ve mümkün olduğunca çok fonetik kapsama sağlamak için temel bir kaynak olarak hareket etmektir. Tri-fonemler temel birim olarak kullanıldı ve geniş ama pratik bir kelime kümesi iki şekilde çıkarıldı: yaklaşık olarak aynı spektral etkiye sahip bağlamı daraltarak ve yüksek frekanslı trifonemlere odaklanarak. Daha sonra, üç fonemleri kapsayacak şekilde korpus türevi kelime listesinden tanıdık kelimelerin minimal bir listesini çıkarmak için açgözlü bir algoritma kullanıldı. Akabinde kelime listesi manuel olarak cümlelere dönüştürüldü ve ihtiyacı karşıladığını doğrulamak için analiz edildi. Bu cümleler anadili tarafından kaydedildi. Bu çalışmada, denge yerine çeşitliliğe odaklanılmıştır.

Kumar and Aggarwal (2011) Gizli Markov Model Araç Seti'ni (HTK) kullanarak Hintçe için bir konuşma tanıma sistemi oluşturdu. Sistem izole kelimeleri akustik bir kelime modeli kullanarak tanıır ve 30 Hintçe kelime için eğitilmiştir. Eğitim verileri sekiz konuşmacıdan toplanmıştır. Deney sonuçları, sunulan sistemin genel doğruluğunun %94,63 olduğunu göstermiştir.

Rousseau *et al.* (2012) TED Konuşmaları'na dayalı Otomatik Konuşma Tanıma yapısındaki TED-LIUM'u geliştirdi. Bu şirket IWSLT 2011 değerlendirme kampanyası sırasında oluşturulmuştur. Bu çalışmada, ses verileri ile altyazılardaki ham metin arasındaki senkronizasyon tekrarlayan yinelemelere dayanarak denetimsiz bir şekilde oluşturulmuştur. Yüzde 18.4'lük bir Global Kelime Hata Oranı için karşılık gelen transkriptlerle toplam 118 saatlik konuşmayı temsil eder. Manuel olarak yazıya dökülmüş bir geliştirme verisi toplam 4 saatlik konuşma için eğitim verisine eşlik eder. Bu veriler kamuya açık olarak piyasaya sürülmüş ve konuşma tanıma için kullanışlı olduğu kanıtlanmıştır.

Herhangi bir konuşma tanıma sisteminin genel performansı büyük ölçüde özellik çıkarma tekniklerine ve kullanılan sınıflandırıcılara bağlıdır (Saksamudre *et al.* 2015). Bu çalışmada, özellik çıkarımı olarak MFCC ve örüntü sınıflandırma tekniği olarak KNN kullanılarak Hintçe Dili için İzole Kelime Tanıma Sistemi üzerinde karşılaştırmalı bir çalışma yapılmıştır. Sistem ilk 10 kelime için eğitildiğinde %89 tanıma oranı ve 100 kelimenin tamamı için eğitildiğinde %62,50 tanıma oranı elde edilmiştir. Bu çalışmada kelime dağarcığı arttıkça performansın düştüğü sonucuna varılmıştır.

LibriSpeech, Creative Commons BY 4.0 lisansı altında yayınlanan ve yaygın olarak desteklenen açık kaynaklı FLAC kodlayıcı kullanılarak saklanan 1.000 saatlik okuma İngilizce konuşmasının bir koleksiyonudur. Etiketler yalnızca cümle düzeyinde hizalanır, böylece kelime düzeyinde hizalama bilgisi yoktur. Bu durum , tam otomatik konuşma tanıma için anahtar kelime tespitinden daha kullanışlıdır (Panayotov *et al* 2015).

Abera (2018), konuşma tanıma amacıyla ilk Tigrinya Dili konuşma grubunu tasarladı ve geliştirdi. Kaynaklarının yetersiz olduğu bir dil olan Tigrinya için bir konuşma tanıma organı oluşturmak için kullanılan prosedürler sağlandı. Eğitim ve test veri setleri için yüz yerli otuz Tigrinya konuşmacısı kaydedildi. Her konuşmacı, hece bakımından zengin ve dengeli cümlelerden oluşan yüz metni okudu. Tigrinya'nın tüm bağlamsal hecelerini ve telefonlarını içeren on bin cümle seti kullanıldı.

Nguyen (2019), Vietnamca konuşma tanıma için Uzun Kısa Süreli Bellek (LSTM) ve Zaman Gecikmesi Derin Sinir Ağı (TDNN) modellerine dayanan Uçtan Uca bir model sunmuştur. Kayıp fonksiyonu olarak Bağlantıcı Zamansal Sınıflandırma (CTC) kullanan iki Vietnam Uçtan Uca mimarisi önerildi. Çalışma ayrıca, CTC modeli uygulanırken herhangi bir transkripsiyon için etiket dizisini üretmek üzere Vietnamca karakterlere veya tonlara dayalı telefon setleri oluşturmak için bir yöntem sundu. Deneysel sonuçlar, CTC tabanlı Uçtan Uca modellerin Vietnamca konuşma tanıma (SR) için WER'in sadece %5'inin daha kötü olduğu geleneksel modellerle rekabet edebildiğini göstermiştir. Bu çalışmanın avantajı ise akustik modelleri eğitmek için zorla hizalama kullanılmasının gerekli olmaması olarak belirtilmiştir. Buna ek olarak, geleneksel modelleri kullanan

alıřmalara benzer, ton bilgisi ve toneme seti Vietnam SR iin performansı optimize etmek iin mlerdir.

Mevcut konuřma tanıma sistemlerinin, konuřma tanıma teknolojisi tarihinde olduka erken ortaya ıkan modellere ve tekniklere dayandıđını belirtmek nemlidir. Ayrıca, dřk hesaplama maliyeti ve insan benzeri performansa sahip ASR sistemlerinin oluřturulması, uzun ve sıkıcı bir ok disiplinli alıřmayı gerektirecektir. Buna ek olarak, sesli giriři bařlatmak iin belirli anahtar kelimeleri srekli olarak dinleyen sistemler geliřtirerek teknoloji kullanıcılarının tamamen eller serbest bir deneyime sahip olmaları kaınılmazdır. Bu sistemler son derece hassas, dřk gecikmeli, ve modern mobil cihazlar gibi hesaplamalı g sınırlı olan ortamlarda alıřmalıdır. Bu sistemleri kk cihazlarda alıřtırmak, sunuculara bađlanıp tanıma esnasındaki gecikmeleri nler.

3. MATERYAL ve YÖNTEM

3.1. Veri Kümesi

Konuşma tanıma araştırması yapmak için araştırmacılar geleneksel olarak üniversiteler veya şirketler gibi büyük organizasyonların kaynaklarına ihtiyaç duyarlar. Bu kuruluşlarda çalışan araştırmacılar genellikle Linguistik Veri Konsorsiyumu gibi gruplarla anlaşarak akademik veriye serbestçe erişebilirler. Konuşma teknolojisi olgunlaştıkça, tanıma modellerini eğitmek ve değerlendirmek isteyenlerin sayısı bu geleneksel grupların ötesine geçmiştir, ancak veri kümelerinin kullanılabilirliği artmamıştır (Warden 2018).

Konuşma veri kümesi, konuşma tanıma ve sentez sistemleri oluşturmanın en temel gereksinimidir. Somali dili söz konusu olduğunda, araştırmacıların konuşma tanıma algoritmalarını eğitmek, test etmek ve karşılaştırmak için herkese açık bir konuşma verisetine (korpus) erişimleri bulunmamaktadır. Bu tez çalışması, basit konuşma tanıma görevleri sınıfı için standart bir eğitim ve değerlendirme veri seti oluşturmayı hedeflemektedir. Çalışmanın birincil amacı, arka plan gürültüsü veya ilgisiz konuşmanın olduğu durumlarda mümkün olduğunca az yanlış pozitifle oranı elde ederek, altmış veya daha az kelime içeren bir veri tabanından tek bir kelime söylendiğinde tespit eden küçük modeller oluşturmak ve test etmek için bir yol sağlamaktır. Bu görev genellikle anahtar kelime belirleme olarak bilinir.

Akıcı konuşmada belirli bir kelime kümesini algılama sorunu, konuşma tanımada en ilgi çekiçi konulardan biridir. Geniş kelime dağarcığına sahip sürekli konuşma tanıma işleme, birçok pratik uygulama için gereksiz ve çok pahalıdır. Örneğin, Örneğin, bir masa lambası için yerinden kıpırdama komutu yerine “Işığı kapat” ve “Işığı aç” komutlarını anlaması oldukça yeterli olacaktır. Anahtar kelime tespit sistemleri, çok basit uygulamalarda (küçük kelime konuşma tanıma) ve "Hey Google!" gibi bir uyandırma komutunu bekleyerek kaynakları kaydetmek istediğimiz ön işleme görevlerinde kullanılır. Uyanık

kelime bulma, ikinci görevdeki daha karmaşık konuşma işleme görevleri için bir tetikleyicidir. İki görev oldukça farklı hedeflere sahip olsa da, temeldeki teknoloji çok benzerlik göstermektedir.

3.1.1. Veri kümesi oluşturma

Kayıt ve veritabanı oluşturma sürecinin kolaylıkla üstesinden gelinebilir olduğundan emin olmak için sınırlı bir kelime dağarcığımız olmasını istedik, ancak bazı uygulamalar için potansiyel olarak yararlı olabilecek veriler konusunda eğitilmiş modeller için yeterli çeşitliliğe de sahip olmak istedik. Ayrıca veri kümesinin, TIDIGITS gibi ortak tescilli koleksiyonlarla karşılaştırılabilir şekilde kullanılabilir olmasını da istedik.

Elektronik cihazlara komut olarak kullanılabilecek altmış ortak kelime seçmemize yol açtı. Veri setimizdeki veri alt kümelerinin Somalice ve Türkçe çiftleri aşağıda bulunmaktadır:

- **Sayılar (rakamlar) Veri alt kümesi:** kow (bir), laba (iki), saddex (üç), afar (dört), shan (beş), lix (altı), toddoba (Yedi), siddeed (sekiz), sagaal (dokuz), toban (on), labaatan (yirmi), soddon (otuz), afartan (kırk), konton (elli), lixdan (altmış), toddobaatan (yetmiş), siddeetan (seksen), sagaashan (doksan), boqol (yüz), kun (bin), milyan (milyon), bilyan (milyar), iyo (ve).
- **IoT komutları Veri alt kümesi:** daar (çal), haki (Duraklat), wadaag (Paylaş), shaandhee (Karıştır), aamusu (sustur), sii wad (devam et), xiga (Sonraki), dhageyso (dinle), fur (aç), xir (kapat), guji (tıkla), raadi (ara), cimilo (hava), wac (ara), qabo (cevapla), aqri (oku), dir (gönder), shid (aç), dami (kapat), baaji (iptal et), xilli (zaman), tirtir (sil).
- **Robotik hareket komutları Veri altkümesi:** hore (ön), dib (geri), kor (yukarı), hoos (aşağı), midig (sağ), bidix (sol), dhaqaaji (hareket et), guri (yerel), jiho (yön), bari (doğu), galbeed (batı), koonfur (güney), waqooyi (kuzey), bilow (başlat), jooji (dur).

Anahtar kelime tanıma ile ilgili en zorlu sorunlardan biri, konuşma tanıma sözlüğünde bulunmayan sözleri yok saymaktır. Bu nedenle algoritmaların veri kümemizde iyi performans gösterme yeteneğini geliştirmek için sözlüğe bir dizi kelime eklememiz gerekiyordu. Bir çok farklı fonemi kapsayan birkaç yüz kelime, bir modelin ayırt edilmesini iyileştirmek için keyfi olarak google anahtar kelime algılama meydan okuma veri kümesinden seçildi.

Somalice konuşma tanıma grubu, gerekli konuşma verilerinin kalitesini ve farklı hoparlör özelliklerine sahip verilerin oransal dağılımını belirten bir dizi yönergeyi karşılamak üzere tasarlanmıştır. Herhangi bir konuşma grubu, her iki cinsiyetin konuşmacılarını kabul edilebilir oranlarda temsil etmeli ve farklı yaşlardaki konuşmacılardan konuşma içermelidir. Bu çalışmada konuşmacılar 18 ila 35 yaşları arasındaydı. Zaman ve sosyal kısıtlamalar nedeniyle çocukların ve yaşlıların konuşmaları veri setinin haricinde bırakıldı.

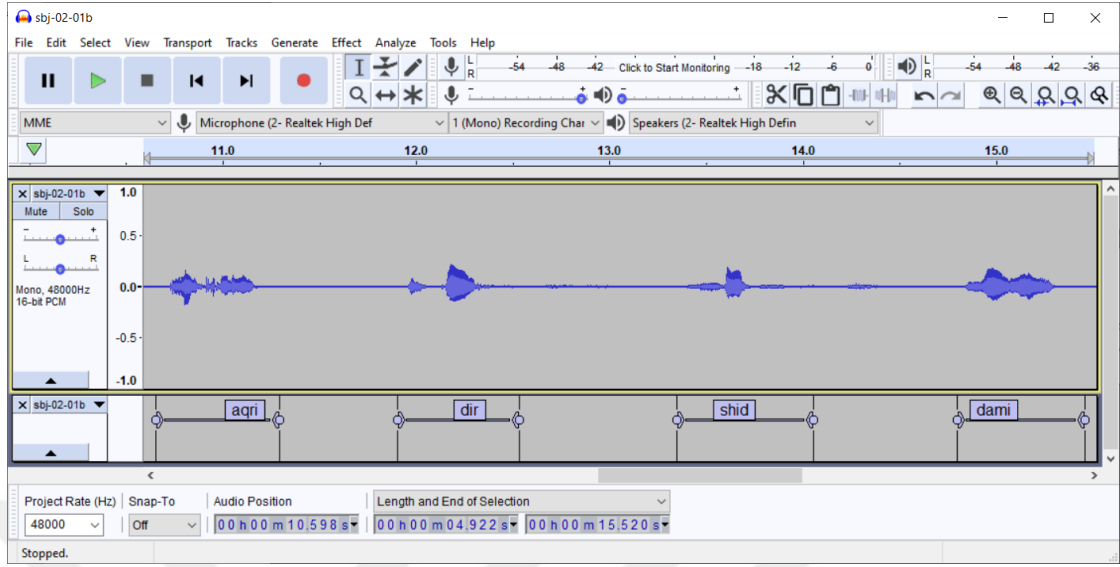
Ayrıca, katkıda bulunanların kişisel olarak tanımlanabilir bilgilerini kişisel verilerin korunması ilkesi kapsamında kaydetmekten kaçındık. Bu, ad, e-posta adresi ve ulusal kimlik numarası gibi hiçbir özellik istemeyeceğimiz anlamına geliyordu. Ayrıca, kullanıcıların katkıda bulunmadan önce bir veri kullanım sözleşmesini kabul etmeleri gerekir. Projenin amacı önce konuşmacıya açıklandı ve tatmin olduktan sonra ne yapılacağı anlatıldı.

Metin grubu oluşturduktan sonra, bir okuma konuşma grubu hazırlamanın bir sonraki adımı konuşmayı kaydetmektir. Seçilen ifadelerin kaydedilmesinde, her konuşmacıdan kendisine sunulanları tam olarak okuması istenir. Okunacak metin mobil, tablet veya bilgisayarda sunulur. Kayıt işlemi sırasında katılımcının her ifadesinde hatalar kontrol edilir ve bir hata bulunursa, katılımcıdan metni yeniden okuması istenir. Veri kümesinin kayıtları hem iç hem de dış mekanlarda yapıldı. Metin 177 yerli Somalice konuşmacı tarafından okundu. Okurların 71'i kadın, 106'sı erkekti.

Kayıt için WhatsApp, Facebook, mobil veya dizüstü bilgisayar kayıt uygulamaları kullanılmıştır. Toplanan ses ifadeleri değişken kalitedeydi ve bu nedenle başvuruları kabul etmek veya reddetmek için ölçütlere ihtiyacımız vardı. Kullandığımız resmi kılavuz, eğer bir insan dinleyicisi hangi kelimenin konuşulduğunu söyleyemiyorsa veya yanlış bir kelime gibi geliyorsa, klbin reddedilmesi gerektiği idi. Sesler araştırmacı tarafından dinlendi ve herhangi bir ses kaydında bir sorun varsa, okuyucudan mümkünse söz konusu kötü ifadeyi tekrar okuması istendi. Gerçekleştirmeye çalıştığımız kelime tespit görevine daha uygun olduğundan, bir cümlemin parçası olarak değil, tek başına konuşulan izole kelimeleri kaydetmeye karar verdik. Bu durumda hizalama çok önemli olmadığından, etiketleme işlemi çok daha kolay hale gelmiş oldu. Tüm konuşma verileri 44.1 kHz ve 48 kHz'de örnek başına ayrılmış 16 bit ile stereo ve mono seslerdir. Ses veri kümesi toplam 2 saat 51 dakika 51 saniye içerir. Kaydedilen ses dosyaları mp3, ogg, m4a ve mpeg formatlarındadır.

3.1.2. Veri kümesini etiketleme

Makine öğreniminde, etiketli verilere sahip olmak, verinin hedefi göstermek için işaretlendiği veya açıklandığı anlamına gelir; bu, makine öğrenme modelimizin tahmin etmesini istediğimiz cevaptır. Genel olarak, veri etiketleme, ek açıklama, sınıflandırma, denetleme, transkripsiyon veya işlemeyi içeren görevleri ifade edebilir. Bu tez çalışmasında veri etiketleme görevimiz, ifadeleri ilgili metne yüksek hassasiyetle kopyalamayı içerir. Bu hedefe ulaşmak için Audacity Ekibi tarafından geliştirilen Audacity 2.3.0 uygulamasını kullandık. Audacity, Windows, macOS, Linux ve diğer Unix benzeri işletim sistemleri için kullanılabilen ücretsiz ve açık kaynaklı bir dijital ses düzenleyici ve kayıt uygulama yazılımıdır. Audacity 1999 sonbaharında Carnegie Mellon Üniversitesi'nden Dominic Mazzoni ve Roger Dannenberg tarafından başlatıldı ve 28 Mayıs 2000'de 0.8 sürümü olarak piyasaya sürüldü. Şekil 3.1 Audacity arayüzünü gösterir.



Şekil 3.1. Audacity arayüzü

Etiket Yolu, Audacity projede oluşturulabilecek ek bir parkurdur. Etiket kanalındaki etiketler, projenin ses kanallarındaki noktalara veya bölgelere referans vermek için kullanılabilir, ancak etiket kanalının kendisi ses içermez. Audacity'de etiketleme noktası ve etiketleme bölgesi olmak üzere iki tür etiket vardır. Aralarındaki fark, etiketleme noktasının sesde tek bir an olması, etiketleme bölgesinin ise başlangıç ve bitiş zamanı olan bir açıklık olmasıdır. Etiketler, oynatma veya düzenleme için seçilen noktaları veya ses bölgelerini işaretlemek ve geri yüklemek için kullanılabilir. Ayrıca, etiketler ek açıklama veya transkripsiyon amacıyla metin içerebilir. Etiket metni düzenlenebilir, bölge etiketleri düzenlenebilir ve yeniden boyutlandırılabilir. Ayrıca, nokta etiketleri ve bölge etiketleri de taşınabilir. Bir Audacity projesini kaydederken etiket parçalarının eklendiğini belirtmek gerekir.

Her bir ses dosyasındaki tüm ifadeler dinlenmiş ve dikkatle incelenmiştir. Duyulmayan sözler çıkarılmış ve etiketlenmemiştir. Etiket parçaları da ayrıca kaydedilmiştir. Etiket parçaları düz metin dosyalarına dışa aktarılabilir ve bu dosyalardan içe aktarılabilir. Dosya yapısı, herhangi bir metin düzenleyicisi veya e-tablo uygulaması tarafından açılabilen ve orada düzenlenebilen sekmeyle ayrılmış düz metin biçimidir. Veri setimiz bölge etiketleri kullanılarak etiketlenir. Bölge etiketlerinin standart yapısı basittir. Bir etiket ve zaman çizelgesi konumunu (başlangıç zamanı ve bitiş zamanı) içerir.

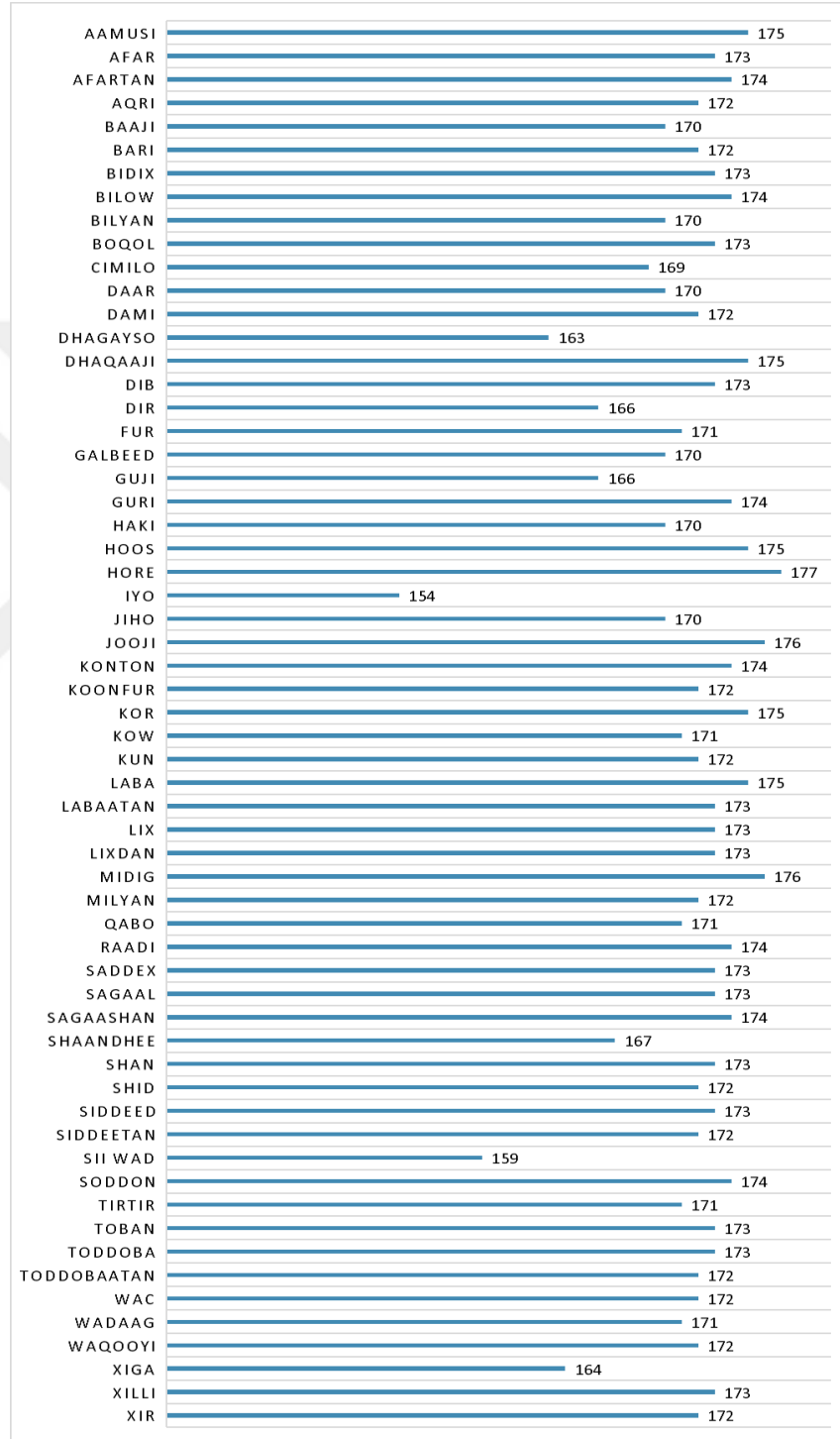
Yaklaşık 190 ses dosyası olan topladığımız tüm veri kümesini etiketledikten sonra, ses dosyalarından ifadeleri çıkarmak için bir MATLAB kodu yazılmıştır. MATLAB, Mathworks tarafından geliştirilen çok paradigmatlı sayısal bir hesaplama ortamı ve özel programlama dilidir. MATLAB, matris manipülasyonlarına, fonksiyonların ve verilerin çizilmesine, algoritmaların uygulanmasına, kullanıcı arayüzlerinin oluşturulmasına ve diğer dillerde yazılmış programlarla arayüz oluşturulmasına izin verir. Bu tez çalışmasında 64 bit akademik lisans sürümü MATLAB 2019a kullanılmıştır. Orijinal ses dosyaları ilk önce mono kanal seslerine dönüştürüldü. Komut dosyası, her ses dosyasını ve karşılık gelen etiket parçası dosyasını okur. Daha sonra her bir etiketi ve karşılık gelen zaman damgasını ses dosyasından çıkarır. Her bir etiket için, karşılık gelen ifade çıkarılır ve uzunluğu bir saniyeden azsa, ifadenin her iki tarafında sıfırlar eşit oranda dolduruldu. Komut dosyası daha sonra konuşma frekanslarını 16kHz'e yeniden örnekledi, çünkü 16kHz'lik bir örnekleme hızı (saniyede 16.000 örnek), insan konuşmasının frekans aralığını karşılamak için yeterlidir. Kayıtlar normalize edildi ve ardından ilgili etiketin klasörü altında WAV ses dosyası formatlarında kaydedildi.

Şekil 3.2 veri kümesinin frekans dağılımını gösterir. Veri kümesindeki ifadelerin ortalaması yaklaşık 172'dir ve standart sapması 3,9'dur.

Komutlar dışındaki tüm kelimelerin dağılımına yaklaşan bir kelime grubu oluşturmak gereklidir. Bunun nedeni, ağın komutları ve diğer tüm kelimeler arasındaki farkı öğrenmek için bu grubu kullanmasıdır. Bu komutlara "bilinmeyen" etiketi verilir. Ayrıca, konuşma komutu tanıma sistemi sadece farklı konuşulan kelimeleri tanımakla kalmamalı, aynı zamanda girişin sessizlik mi yoksa arka plan gürültüsü mü içerdiğini tespit edebilmelidir. Toplam 1.070 sessizlik ve arka plan gürültüsü sesi oluşturularak ve "arka plan" olarak etiketlenmiştir.

Ses veri kümesini zenginleştirmek için her bir ses dosyasına beş farklı arka plan gürültüsü eklediğimizi belirtmek gerekir. Bu arka plan sesleri havaalanı, sokak, hava, ev ve pazar gürültüsü ses dosyalarından alınmıştır. Bu seslerin etkisi kontrol edildi ve her ses

dosyasının yaklaşık %20'sinde tutuldu. Veri kümesindeki toplam ses dosyası sayısı 62.816 oldu..

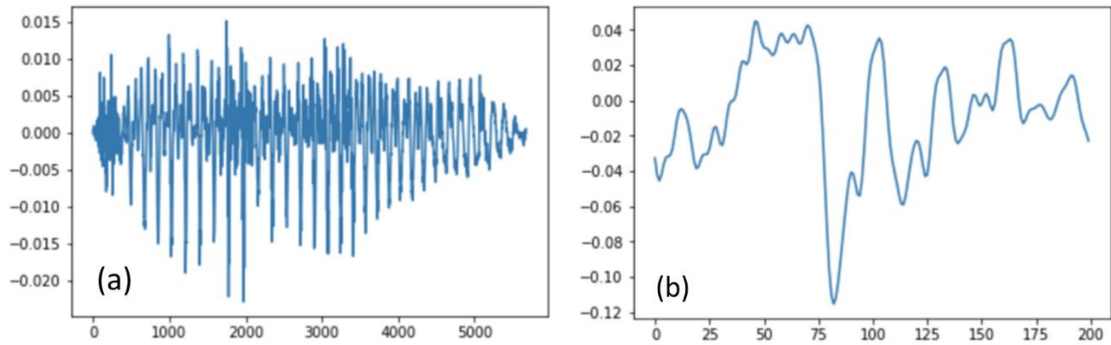


Şekil 3.2. Veri kümesinin frekans dağılımı

3.2. Konuşma Tanıma Algoritması

Veri kümesini toplayıp etiketledikten sonra, yeni oluşturulan veri kümesinin ne kadar iyi olduğunu değerlendirmek için bir konuşma tanıma algoritması oluşturmak gerekir. Bir konuşma tanıma sistemi tasarlamadan ve oluşturmadan önce, bir saniyelik ses kayıtları olan verilerimizin doğasını anlamak gerekir. Sesler doğası gereği analogdur ve dijital biçimde saklanır. Şekil 3.3.a'da gösterildiği gibi, sesler doğada çok karmaşık sinyallerdir ve içlerinde desen bulmak çok zordur. Ses sinyalinin küçük bir kısmını yakınlaştırıp analiz etsek bile, Şekil 3.3.b'de gösterildiği gibi oldukça karmaşıktır.

Pratikte ses dalgaları hiç işlenmeden, yani ham olarak, konuşma tanıma sistemlerini eğitmek için kullanılmaz; bunun yerine, özellikler ham ses dosyalarından çıkarılır ve konuşma tanıma algoritmalarına girdi olarak kullanılır.



Şekil 3.3. (a) ses sinyalinin doğası. (b) yakınlaştırılmış ses sinyali

3.2.1. Veri hazırlama ve ön işleme

Konuşma tanıma sistemlerinde ses dalgalarını temsil etmenin en standart yolu Mel Spectrogram'lardır. Sinyalden spektral katsayıları çıkarmadan önce bazı ön işlemlerin gerekli olduğunu belirtmek gerekir.

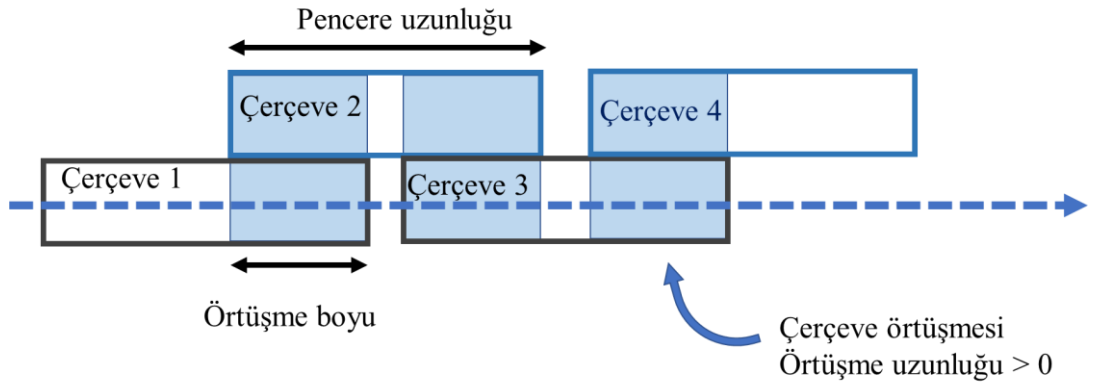
Farklı ses seviyelerinde bir grup ses dosyamız olduğundan, hepsini aynı ses seviyesine olabildiğince yakın hale getirmek iyi bir uygulamadır ve bu normalleştirilerek elde edilir. Normalizasyon, denklem 3.1'de gösterildiği gibi formüle edilir.

$$y = \frac{y}{\max(abs(y))} \quad (3.1)$$

Ön vurgu, yüksek frekanslı bileşenlerin mukavemetini arttırmak için kullanılır. Spektrumu tüm frekans bileşenleri için neredeyse eşit hale getirmek için yapılır. Aşağıdaki denklemle tanımlanır (zaman uzayında):

$$y(n) = x(n) - \alpha x(n - 1) \quad [0.9 \leq \alpha \leq 1.0] \quad (3.2)$$

Konuşma sinyali, kısa bir süre (10 ila 40 ms arasında) incelendiğinde, özelliklerinin kısa süreli durağan olması anlamında zamanla yavaş değişen bir sinyaldir. Pratikte, genellikle 25 ms çerçeve süresi ve %50 çerçeve çakışması alınır. Şekil 3.4 ses sinyali çerçevelemeye bir örneği göstermektedir.



Şekil 3.4. Ses sinyali çerçeveleme örneği

Pencereleme, daha büyük bir sinyalin küçük bir kısmını alma ve bir pencere fonksiyonu ile çarpma işlemidir. Pencere, bir süre sıfır olmayan ve o dönemden önce ve sonra sıfır olan bir fonksiyondur. Başka bir sinyalle çarpıldığında, pencerenin sıfır olmayan kısmı dışında sıfır çıktı üretir. Hamming penceresi her iki uçtaki sinyalin genliğini azaltır.

Pencereli fonksiyonlarda, frekans dönüştürme fonksiyonları çerçevenin orta kısmına daha fazla odaklanacaktır, dolayısıyla sinyalin ortasındaki frekanslar daha fazla tercih edilecektir. En sık kullanılan pencereleme fonksiyonu Hamming penceresidir ve aşağıdaki denklemde gösterildiği gibi matematiksel olarak ifade edilir.

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi}{N-1}\right) \quad (3.3)$$

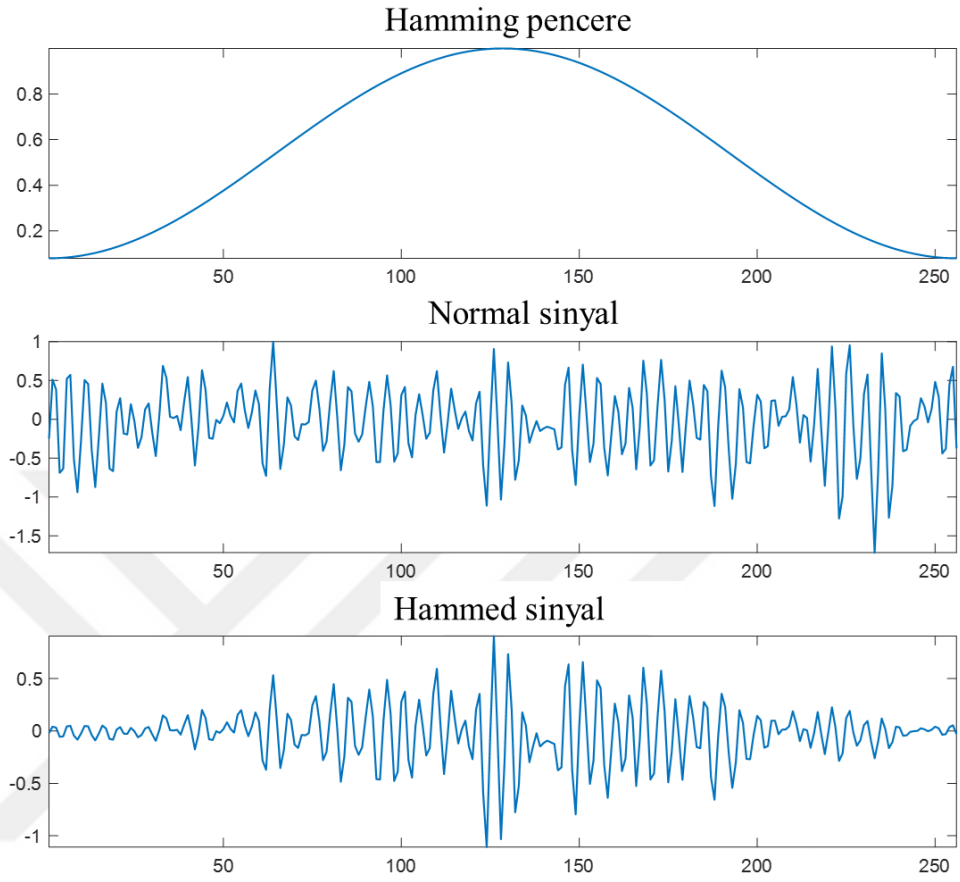
Şekil 3.5. Hamming penceresi fonksiyonun ses sinyalleri üzerinde nasıl çalıştığını gösterir.

Konuşma sinyalini bir dizi akustik özellik vektörüne dönüştürmek için, genellikle Mel spektrogramı kullanılır. Mel spektrogramı insan işitme algısına dayanmaktadır. İnsan kulağı Mel Skalasında ses sinyallerini algılar. Mel skalası 1kHz'e kadar lineerdir ve bunun üzerinde logaritmiktir.

Her bir çerçeve için hesaplanan Mel spektrogramı, Mel frekans bandındaki enerjinin dağılımını verir. Bir ses sinyalinin Mel spektrogramını elde etmek için üç işlem yapılması gerekir. Her ses çerçevesinin Fourier dönüşümü alınır, bunlar Mel filtresinden geçirilir ve logaritmaları alınır. Fourier dönüşümü, her bir frekans bandındaki enerji miktarı hakkında bilgi verir. Hızlı Fourier dönüşümü (FFT), hızlı ve verimli olduğu için çerçevelerin enerji spektrumunu hesaplamak için yaygın olarak kullanılır. Üçgen üst üste binen pencereler kullanılarak, önceki adımda elde edilen güç spektrumu Mel ölçeğine eşlenir. Denklemler 3.4 ve 3.5. Mel skalasına nasıl dönüştürüleceğini göstermektedir.

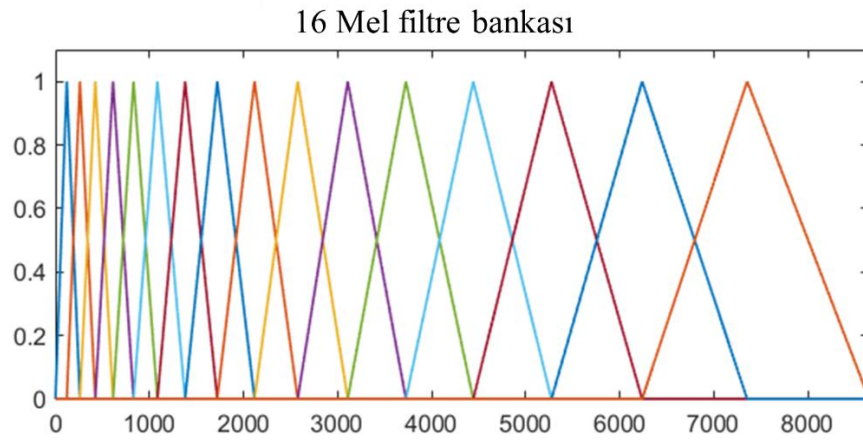
$$F_{\text{mel}} = 2595 * \log_{10}\left(1 + \frac{F_{\text{Hz}}}{700}\right) \quad (3.4)$$

$$F_{\text{Hz}} = 700 \left(10^{\frac{F_{\text{mel}}}{2595}} - 1\right) \quad (3.5)$$



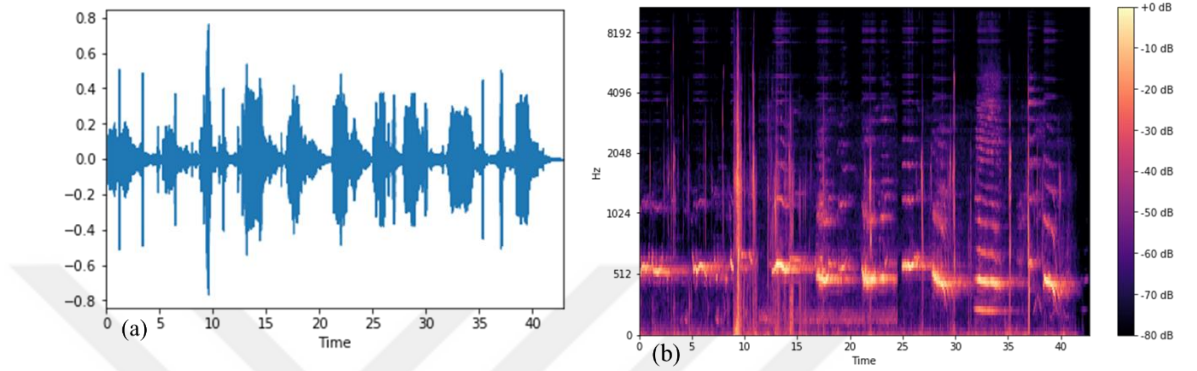
Şekil 3.5. Ses sinyalinde Hamming penceresi fonksiyonu çalışması

Şekil 3.6 16'lı Mel filtre bankasını göstermektedir.



Şekil 3.6. Bir 16 Mel filtre bankası

Ses enerjisinin dinamik aralığını azaltmak için filtre bankası enerjilerinin logaritması alınır. Şekil 3.7 bir balina sesinin ses dalga biçimini ve Mel spektrogramını göstermektedir.



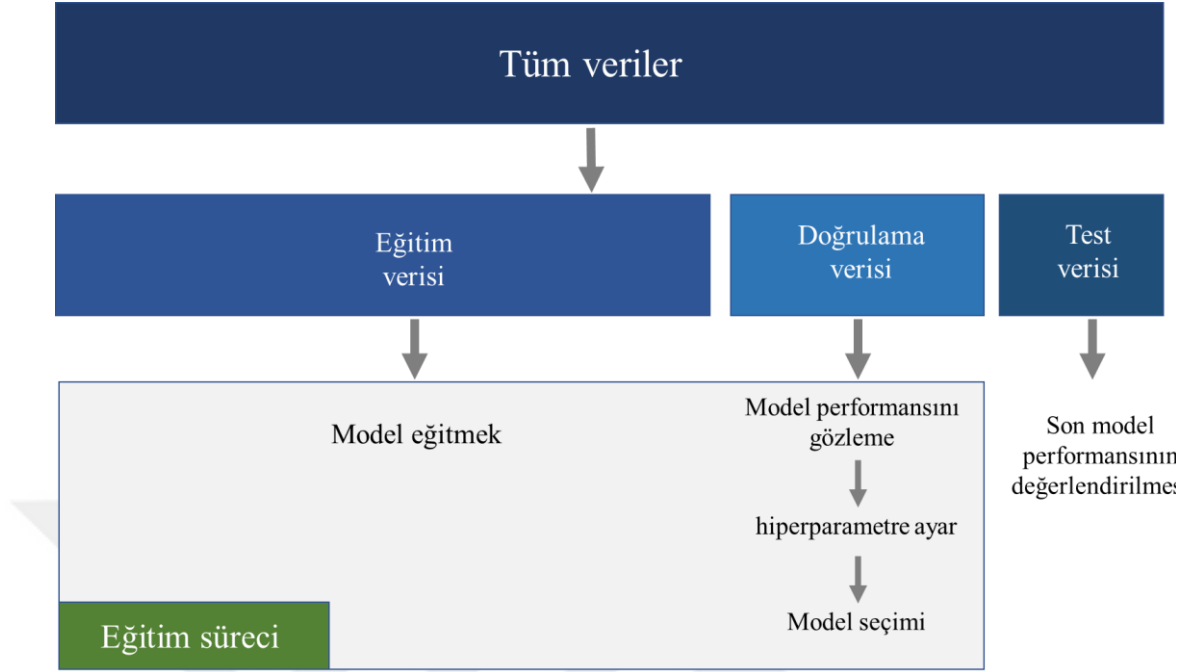
Şekil 3.7. Bir balina sesinin (a) ses dalga formu. (b) Mel spektrogramı (Anonim 2019)

3.2.2. Model eğitimi

Veri ve temel doğruluk etiketleri, derin öğrenme veya diğer makine öğrenme yöntemleri uygulayan araştırmalarda en önemli bileşenlerdir. Başarılı bir derin öğrenme projesi için bir modeli eğitmek ve test etmek için veri ve işaretlenmiş etiketlerinin dikkatli olarak toplanması zorunludur. Ancak yüksek kalitede etiketlenmiş verilerin elde edilmesi maliyetli ve zaman alıcı olabilir.

Şekil 3.8'de gösterildiği gibi, mevcut veri kümesi tipik olarak üç gruba ayrılır: bir eğitim, bir validasyon ve bir test seti.

Kayıp değerlerinin ileri yayılma yoluyla hesaplandığı ve öğrenilebilir parametrelerin geri yayılım yoluyla güncellendiği bir ağı eğitmek için bir eğitim seti kullanılır. Eğitim süreci sırasında modeli değerlendirmek, hiperparametrelerde ince ayar yapmak ve model seçimi yapmak için bir validasyon seti kullanılır. Eğitim sürecinde ince ayarlanmış ve seçilen son modelin performansını eğitim ve validasyon setleri ile değerlendirmek için bir test seti ideal olarak projenin sonunda bir kez kullanılır.



Şekil 3.8. Veri kümesi bölümlenme

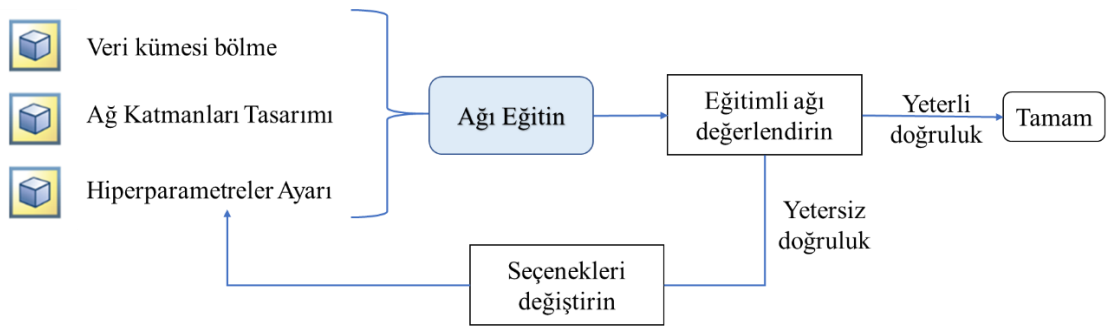
Bir modelin eğitimi her zaman hiperparametrelerinin ince ayarını ve model seçiminin yapılmasını içerdiğinden ayrı validasyon ve test setlerine ihtiyaç vardır. Bu işlem, validasyon setindeki performansa bağlı olarak gerçekleştirildiğinden, bu validasyon seti hakkındaki bazı bilgiler modelin kendisine sızır, yani model öğrenilebilir parametreler için asla doğrudan eğitilmemesine rağmen, validasyon setine aşırı uyum sağlar. Bu nedenle, validasyon setinde ince ayarlı hiperparametreleri olan modelin aynı validasyon setinde iyi performans göstereceği garanti edilir. Bu nedenle, model performansının uygun değerlendirilmesi için tamamen görünmeyen bir veri kümesi, yani ayrı bir test seti gereklidir, çünkü umursadığımız şey daha önce hiç görülmemiş verilerdeki model performansdır, yani genelleştirilebilirlik.

Veri setini değerlendirmek için dört derin evrişimli sinir ağı modeli oluşturulmuştur; birincisi sayıları tanımak için, ikincisi robotik hareket komutlarını tanımak için, üçüncüsü IoT komutlarını tanımak için ve dördüncüsü veri setimizdeki tüm kelimeleri tanımak için tek bir derin sinir ağı.

Rakam (sayı) tanıma CNN 23 sınıftan, robotik hareket komutları tanıma CNN 15 sınıftan ve IoT komutları tanıma CNN 22 sınıftan oluşur. Ayrıca, hepsi bir arada tanıma derin evrimsel sinir ağının 60 sınıfı vardır.

Veri büyütme (augmentation), hem mutlak doğruluk hem de genelleme açısından ağ performansını tutarlı bir şekilde artırdığı kanıtlanmıştır. Ayrıca ağın aşırı uydurma (overfitting) yapmasını önlemeye yardımcı olur. Veri büyütme, Mel spektrogramlarının yeniden boyutlandırılması ve bunların zaman içinde 10 kareye (100 ms) ileri veya geri rastgele çevrilmesi ve spektrogramların zaman eksenini boyunca yüzde 20 yukarı veya aşağı ölçeklendirilmesiyle elde edilir.

Bir evrim sinir ağı tasarımı, insan uzmanlığı ve deneme yanılma içerir. Derin sinir ağı modellerinde model seçimi sadece katman ve gizli birimlerin ve diğer birkaç hiperparametrenin sayısını seçmekle ilgili değildir. Bir modelin mimarisini tasarlamak ayrıca katman türlerini ve bunların düzenlenme ve birbirine bağlanma şeklini seçmeyi de içerir; yani, bir ağı tasarlayanın sonsuz sayıda yolu vardır. Şekil 3.9 veri setimizi değerlendirmek için derin bir CNN'yi modelinin tasarım sürecinin şematik diyagramını gösterir.



Şekil 3.9. Derin CNN model tasarım süreci

Derin bir sinir ağı tasarlarlarken, ilk adım katman sayısını ve katman başına birim sayısını seçmektir. Bu görevi yerine getirmek için iki yaklaşım vardır. Bunlar, tasarıma çok küçük bir modelle başlayarak modeli büyütme veya büyük bir modelle başlayarak modeli küçültme. Büyük bir modelle başlamanın sorunu, ne kadar büyük başlamamız

gerektiğine karar vermenin zor olmasıdır. Öte yandan, tasarıma küçük bir modelle başlarsak, her zaman sıfır noktasına sahibiz. Bu tez çalışmasında bu yaklaşımı takip ettik. Çok küçük bir modelle başladık ve katmanlar ekleyerek ve katman başına birim sayısını ayarlayarak model boyutunu yavaş yavaş artırdık. Ağ modeli tasarımını daha küçük başlatmak ve validasyon hatası düzelene kadar model kapasitesini artırmak genellikle daha fazla fayda sağlar.

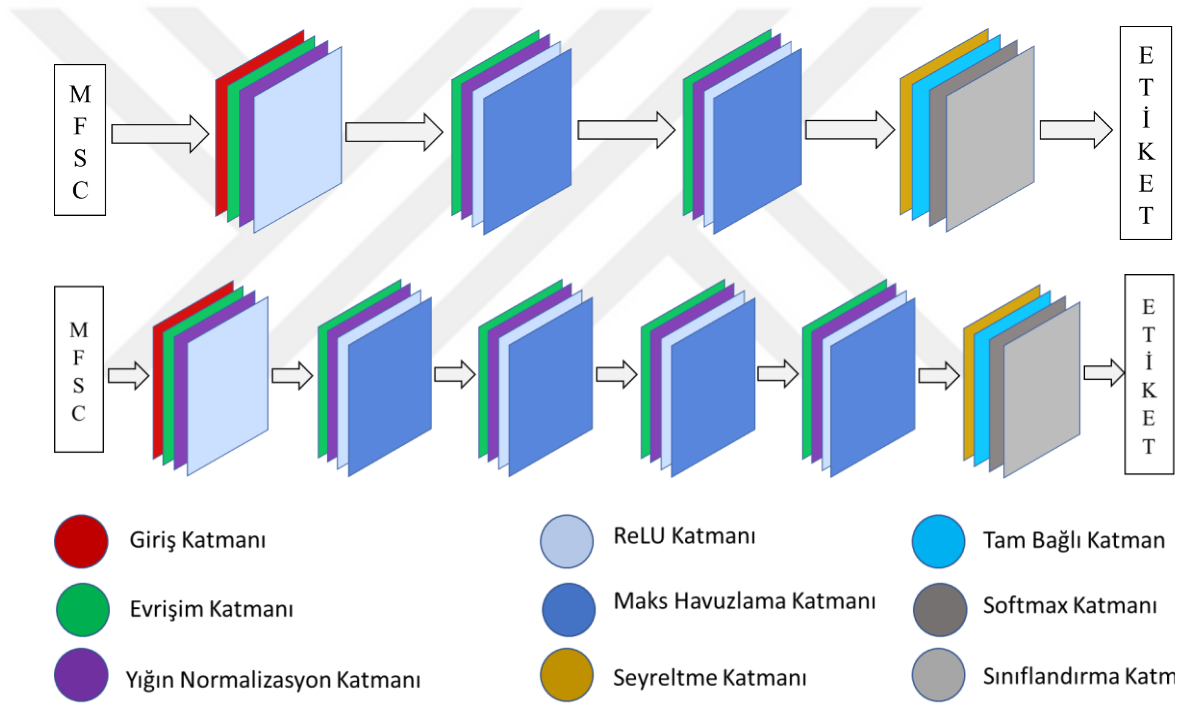
Evrşimsel katmanlardaki çekirdeklerin boyutuna karar vermek de zordur. Küçük gibi görünseler de, 3x3 çekirdekler genellikle en iyi sonucu verir ve 3x3 çekirdekleri birbiri üzerine istifleyerek daha geniş bir alan elde edebiliriz. Adım olarak da bilinen kayan pencere basamak boyutunu seçmek de çok önemlidir. Genellikle, aktivasyonların uzamsal çözünürlüğünü korunması amaçlandığında basamak boyutu 1 seçilir, yada ya da havuzlamanın kullanılmak istenmediği durumlarda 2 seçilir ve alt örnekleme kullanılır.

Buna ek olarak, havuzlama parametrelerini seçerken, aynı dolgulu Maksimum havuzlama ve 2x2 havuzlama boyutu genellikle işe yarar. ReLU aktivasyon fonksiyonu, çıktı katmanı dışında tüm katmanlarda aktivasyon fonksiyonu olarak kullanılır. Çıkış katmanına Softmax aktivasyon fonksiyonu uygulanır. Sinir ağları için en popüler düzenleme tekniği olan bir seyreltme katmanı, aşırı uyma (overfitting) önlemek için Tam Bağlı Katmandan önce kullanılır. Mini yığın büyüklüğü 64 (bütün veri tabanı için) ve 32 (diğerleri için) olan Adam optimizör (yönetimi optimize edici) kullanılarak eğitim seçenekleri belirlendi. Ağlar 12 epok boyunca eğitildi ve öğrenme hızı, 8 epoktan sonra 10 kat azaltıldı.

Tablo 1, evrşimli sinir ağındaki (CNN) parametrelerin ve hiperparametrelerin bir listesini gösterir. Bir parametrenin, eğitim sürecinde otomatik olarak optimize edilen bir değişken olduğunu ve bir hiperparametrenin önceden ayarlanması gereken bir değişken olduğunu belirtmek önemlidir. Şekil 3.10 derin evrşimli sinir ağılarımızın mimarisini gösterir.

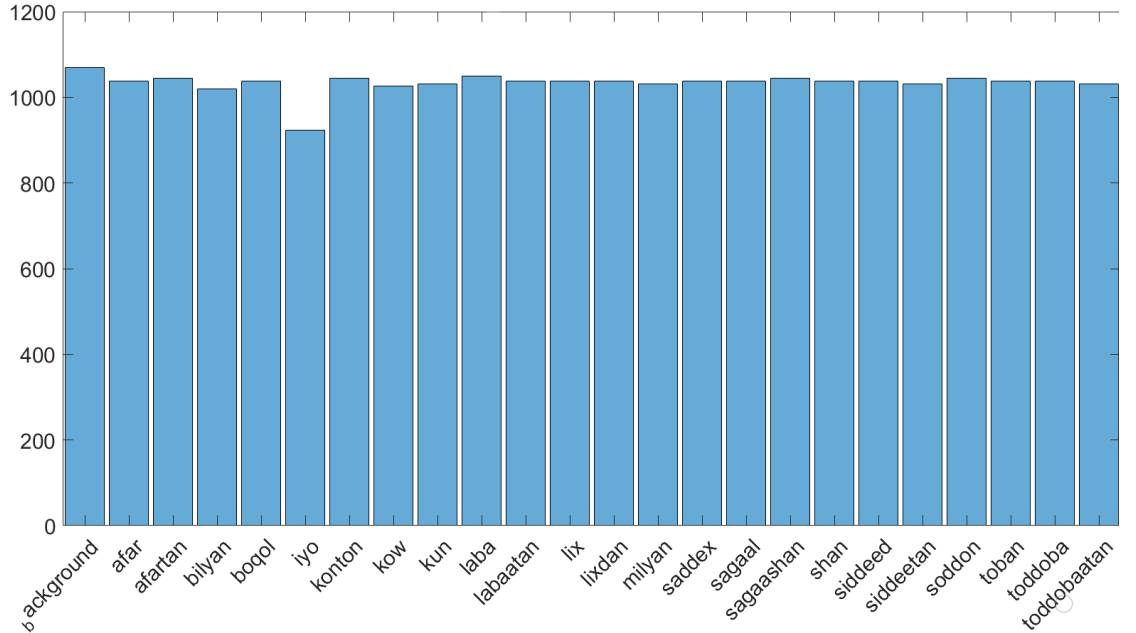
Çizelge 3.1. Bir CNN parametrelerin ve hiperparametrelerin listesi

Katman	Parametre	Hiperparametreler
Evrişim katmanı	Çekirdekleri	Çekirdek boyutu, çekirdek sayısı, adım, dolgu, aktivasyon fonksiyonu
Havuzlama katmanı	---	Havuzlama yöntemi, filtre boyutu, adım, dolgu
Tamamen bağlı katman	Ağırlıklar	Ağırlık sayısı, aktivasyon fonksiyonu
Diğerleri	---	Model mimarisi, iyileştirici, öğrenme oranı, kayıp fonksiyonu, mini yığın boyutu, dönemler, düzenlenme, ağırlık başlatma, veri seti bölme

**Şekil 3.10.** Derin CNN'lerin mimarisi

4. ARAŞTIRMA BULGULARI ve TARTIŞMA

61 sınıf içeren SUUBAN veri kümesine bir kez önışleme yapıldıktan sonra MATLAB matris veri dosyası formatında saklandı. Veri kümesinin performansını değerdendirmek için dört farklı Derin Sinir Ağı tasarlanmıştır. Bu ağlar sırasıyla rakam tanıma, IoT komutları tanıma, robotik hareket komutları tanıma ve tüm veri kümesi için ayrı ayrı uygulanmıştır. Eğitim işlemi sırasında, bu komut kümelerinin her birinin önışleme adımından sonra elde edilen öznitelikler belleğe yüklenmiş ve bu özniteliklerin sayısı artırılmıştır. Bundan sonra, NVIDIA GeForce GTX 1050 Ti kullanılarak ağ eğitimi gerçekleştirilerek eğitimli ağlar MATLAB matris veri dosyası formatında kaydedilmiştir. Şekil 4.1 rakam alt kümesinin veri dağılımını göstermektedir. Arka plan gürültüsü ve biraz daha büyük veri örnekleri olan bilinmeyen sözcükler sınıfı dışında tüm sınıflar benzer bir dağılıma sahipti.



Şekil 4.1. Rakam alt kümesinin veri dağılımı



True Class	_background	afar	afartan	bilyan	boqol	iyo	konton	kow	kun	laba	labaatan	lix	lixdan	milyan	saddex	sagaal	sagaashan	shan	siddeed	siddeetan	soddon	toban	toddoba	toddobaatan
_background	174		1							1										1				
afar		145																						
afartan		3	150																					
bilyan				145				1						13									1	
boqol			1		164											4								
iyo						152																		
konton							153																	
kow								160	1															
kun									156															
laba					1			1		165				1									1	
labaatan											147													5
lix								1				166												
lixdan		1							1				164					1						
milyan				18										127			1							
saddex															152									
sagaal																136								1
sagaashan	1									1							157							
shan																		161						
siddeed															1				148	5				
siddeetan																			3	129				
soddon																					162	2		
toban								1														152	1	
toddoba																						1	145	1
toddobaatan									1	3										3			2	123

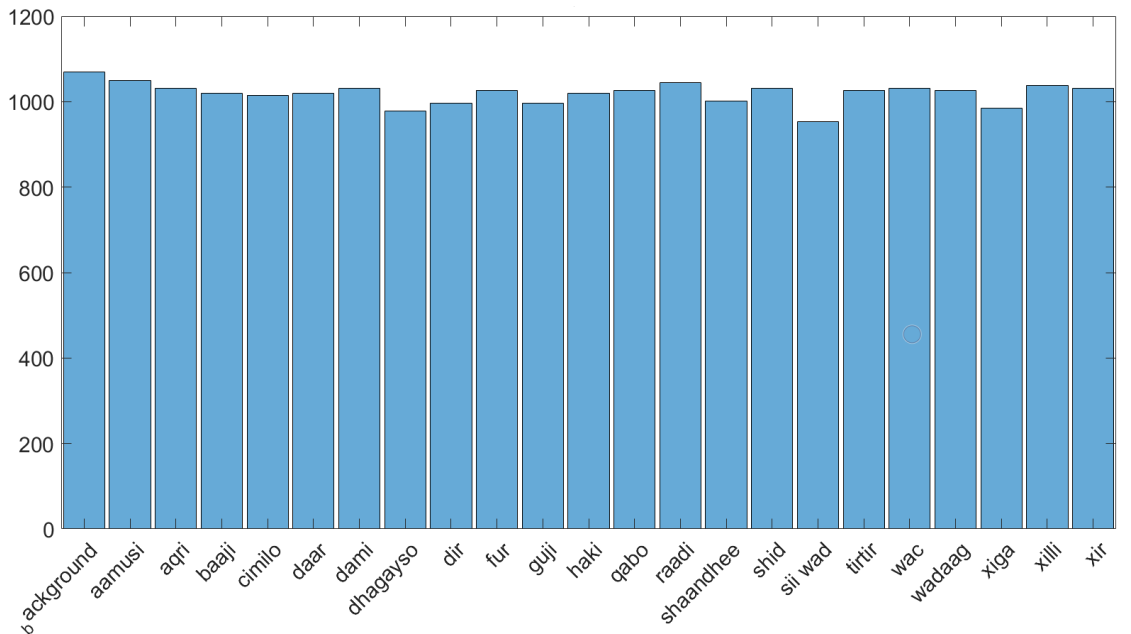
Predicted Class

Şekil 4.2. Rakam tanıyıcı karışıklık matrisi

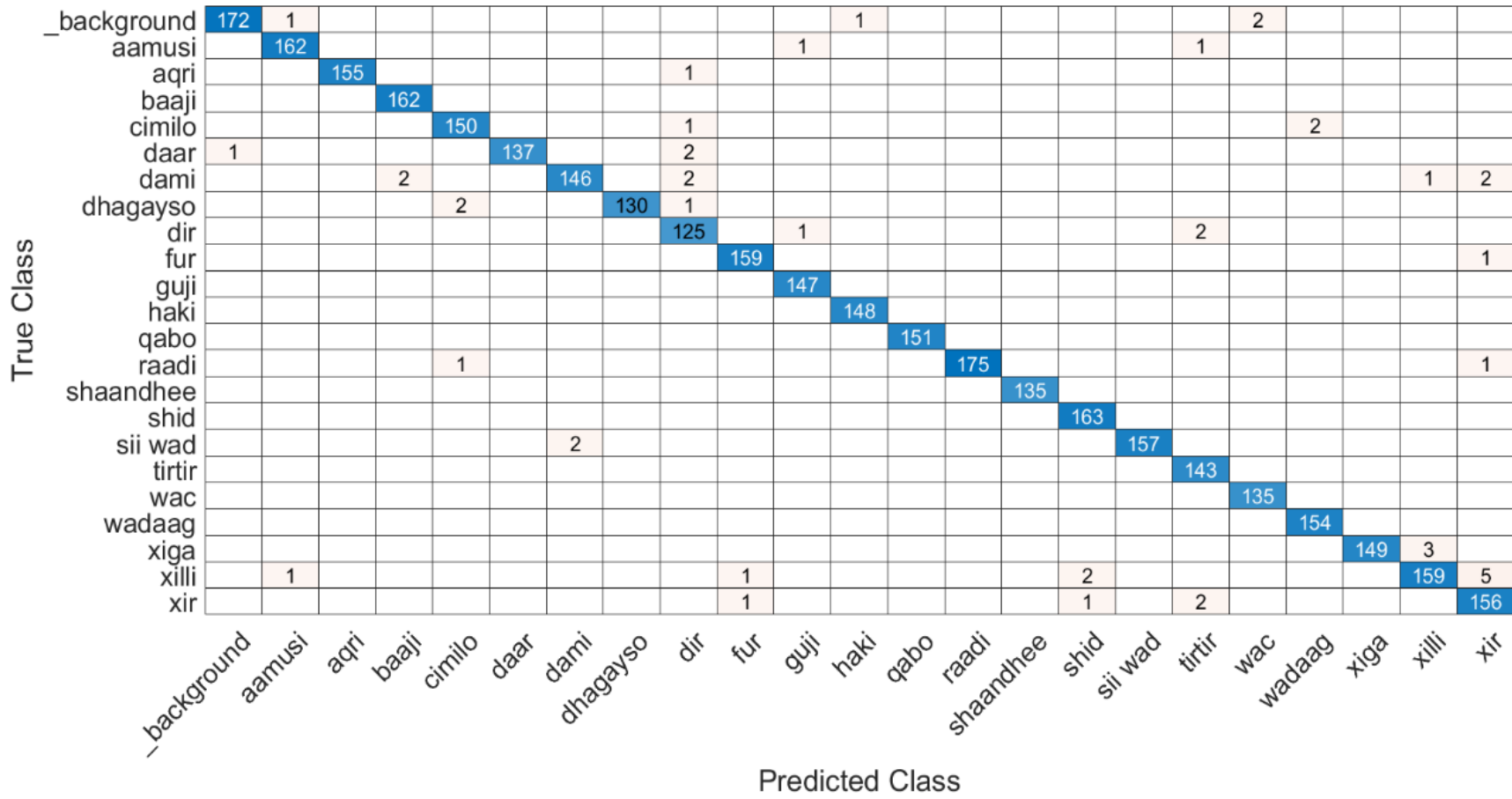
Rakam tanıyıcı ağ 4 katmana sahip ve doğruluğu çok iyi olan bir ağıdır. Eğitim hata oranı %1,763 ve validasyon hatası oranı %1,688 idi. Rakam tanıyıcının test verileri üzerindeki doğruluğu %97,64 iken, tüm verilerdeki (eğitim, validasyon ve test dahil) doğruluğu %98,15'dir. Sınıflandırma doğruluğu ayrıntıları gizlediğinden, sınıflandırma modelimizin performansını daha iyi anlamak için karışıklık matrisini kullanmak gerekmektedir. Rakam tanıyıcının karışıklık matrisi Şekil 4.2'de gösterilmektedir.

Karışıklık matrisinden, modelin genel doğruluğunun iyi olduğunu çıkarabiliriz. Sadece “milyan” kelimesi yanlış bir şekilde “bilyan” olarak 18 kez tahmin edilmiştir. Bu kelime için %11,04 hata oranına karşılık gelmektedir. İki kelimenin çok benzer olduğunu belirtmek gerekmektedir.

IoT komutları alt kümesinin veri dağılımı, şekil 4.3'te gösterildiği gibi neredeyse uniformdur. Bu alt kümeyi değerlendirmek için dört katmandan oluşan derin bir CNN tasarlanmıştır. Eğitim ve validasyon hataları sırasıyla %0,667 ve %0,602'dir. Öte yandan, ağıdaki test verilerinin doğruluğu %98,66 ve ağıdaki tüm IoT komutlarının alt kümesinin doğruluğu %99,23'dir.



Şekil 4.3. IoT komutları alt küme veri dağılımı



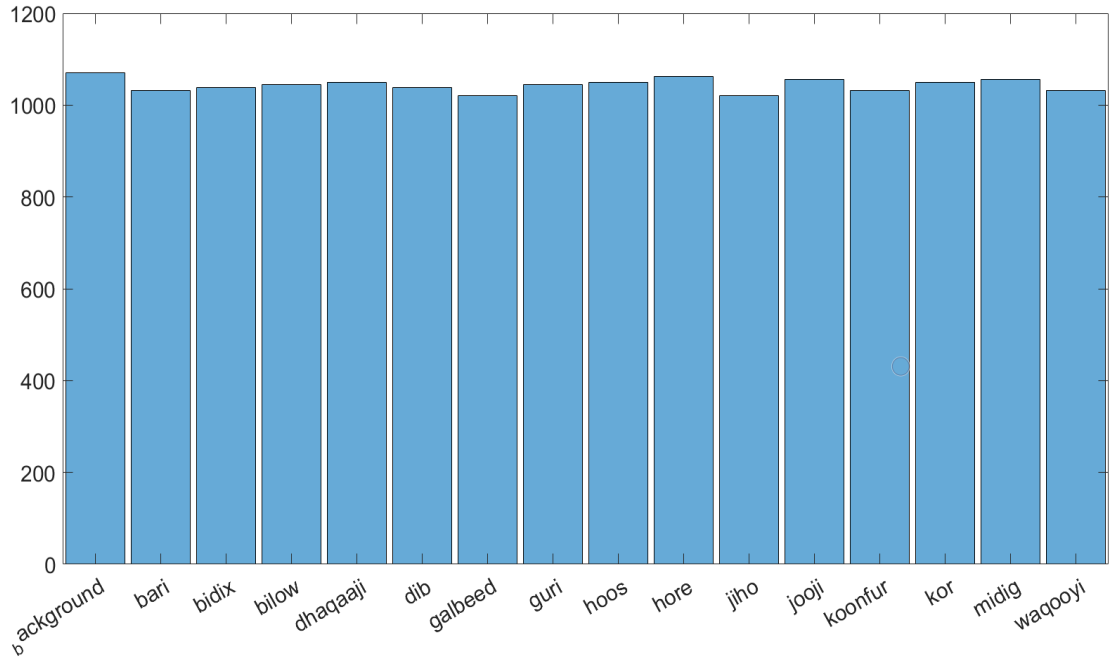
Şekil 4.4. IoT komutları alt kümesinin karışıklık matrisi

Şekil 4.4. IoT komutları alt kümesinin karışıklık matrisini gösterir. Bu veri alt kümesinin de dört katmanlı olan derin bir CNN üzerinde eğitildiğini ve değerlendirildiğini belirtmek gerekir. Ağıın genel doğruluğu iyiydi ve sonuçlar kabul edilebilmektedir.

Benzer şekilde Robotik hareket komutları alt kümesi veri dağılımı şekil 4.5'te gösterilmiştir. Şekilde, veri dağılımının kabaca uniform olduğu açıktır. Diğer veri alt kümelerinde olduğu gibi, dört katmanlı bir CNN, bu Robotik hareket komutları test alt kümesi kullanılarak eğitildi ve test edildi. Test verilerinin doğruluğu %97,40 iken, tüm verilerdeki (test, validasyon ve eğitim dahil) ağıın doğruluğu %98,30'du. Ayrıca, eğitim ve validasyon hataları sırasıyla %1,543 ve %1,325'dir. Şekil 4.6. bu veri alt kümesinde eğitilmiş ve test edilmiş derin CNN'nin karışıklık matrisini gösterir. Karışıklık matris tablosundan, ağıın bu veri alt kümesindeki genel performansının kabul edilebilir olduğunu belirtebiliriz. Bununla birlikte, "guri" kelimesi yaklaşık %3,16 için "midig" olarak sınıflandırılmıştır ve "midig" yaklaşık %7,69 için "dib" olarak yanlış sınıflandırılmıştır. "Dib" kelimesinin "midig" kelimesinin orta kısmına oldukça benzemesi sistemin bu kelimeleri daha sıklıkla karıştırmasına sebep olmuştur.

Ayrıca, tüm SUUBAN veri seti altı katmanlı derin bir Konvolüsyon Sinir Ağı üzerinde eğitilmiş ve test edilmiştir. Eğitim ve validasyon hataları sırasıyla %1,447 ve %1,341 idi. Derin CNN test verileri üzerinde test edildi ve doğruluk %98,15 olarak elde edildi. Ayrıca, tüm SUUBAN veri kümesinde test edildiğinde ağıın doğruluğu %98,49'dir.

Mobil uygulamalar gibi kısıtlı donanım kaynaklarına sahip uygulamalar üzerinde çalışırken, kullanılabilir bellek ve hesaplama kaynaklarındaki sınırlamaları dikkate almak kaçınılmazdır. Ağların toplam büyüklüğü kilobayt cinsinden hesaplandı ve CPU kullanırken tahmin hızları test edildi. Tahmin süresi, tek bir giriş görüntüsünü sınıflandırma süresidir. Çizelge 4.1, dört derin CNN'nin tümünün ağ boyutlarını ve hızlarını özetlemektedir.



Şekil 4.5. Robotik Hareket komutları veri seti dağılımı

Çizelge 4.1. CNN'nin tümünün ağ boyutlarını ve hızlarını

<i>Ağ adı</i>	<i>Rakamlar</i>	<i>IoT komutları</i>	<i>Robotik komutlar</i>	<i>Bütün veriler</i>
<i>Katman sayısı</i>	4	4	4	6
<i>Ağ boyutu</i>	2488 kB	2394 kB	1737 kB	853 kB
<i>Tahmin süresi</i>	7,8261 ms	9,383 ms	8,6976 ms	9,5925 ms
<i>Eğitim hatası</i>	%1,763	%0,667	%1,543	%1,447
<i>Validasyon Hatası</i>	%1,688	%0,602	%1,325	%1,341
<i>Test doğruluğu</i>	%97,64	%98,66	%97,40	%98,15
<i>Tüm veri doğruluğu</i>	%98,15	%99,23	%98,30	%98,49

_background	167				1							1			
bari		165	1				1		2			1			
bidix			151			4			2					1	
bilow				141										1	
dhaqaaji		1			159										1
dib						163								1	
galbeed							136								
guri		2	1			1		137		2				5	
hoos									161			1	2		
hore										152					
jiho	1			2							151				
jooji												161		1	
koonfur													154	1	
kor														162	
midig			5			14		4							149
waqooyi							1			1		3			130

_background

bari

bidix

bilow

dhaqaaji

dib

galbeed

guri

hoos

hore

jiho

jooji

koonfur

kor

midig

waqooyi

Predicted Class

Şekil 4.6. Robotik Hareket Ağı Karışıklık matrisi

5. SONUÇ

Bu tez çalışmasında SUUBAN adlı Somalice Konuşma Komutları Tanıma için yeni bir veri kümesinin genel tasarımını sunduk. Bu veri seti, bu çalışma için kaydedilen her bir ses dosyasını manuel olarak dinlemeye ve etiketlemeye dayalı olarak denetimli bir şekilde oluşturulmuştur. Audacity 2.3.0 ve MATLAB 2019a, ses dosyalarından komutları etiketlemek ve çıkartmak için kullanıldı. Veri setinin kaydedilmesine toplam 177 Somalice konuşmacı katıldı. Veri arttırmadan sonra, veri kümesi toplam 17 saat 09 dakika ve 06 saniye konuşmayı temsil eder. Her ses dosyası adı etiketine karşılık gelen bir klasörün altına yerleştirilir. Veri kümesindeki arka plan gürültüsü ve bilinmeyen sözcükler dahil toplam dosya sayısı 63.224'tür. Somalice, özel bir konuşma tanıma veri kümesine sahip olmayan bir dil olduğundan dolayı, bu veri kümesinin sonraki çalışmalar için yararlı olabileceğine inanıyoruz. Bu yüzden, yakın gelecekte halka açık olarak yayınlamayı planlamaktayız.

Buna ek olarak, Konuşma Komutları veri kümesinin çeşitli modelleri eğitmek ve değerlendirmek için yararlı olduğu gösterilmiştir. Veri setimizi test etmek için dört derin CNN oluşturduk; Rakam tanıma, IoT komutlarını tanıma, Robotik komutları tanıma ve hepsi bir arada. Sonuçlar umut vericiydi ve bu derin CNN'lerin her birinin test doğruluğu sırasıyla %97,18, %98,35, %98,00 ve %97,88 idi. Veri kümesindeki modellerin genel doğruluğunun yanıltıcı olmadığından emin olmak için modellerin karışıklık matrisleri hesaplandı ve yanlış pozitiflerin oranının kabul edilebilir olduğu gözlemlendi. Bu nedenle, oluşturduğumuz veri kümesi dengeli ve komut tanıma sistemlerinin eğitimi için oldukça kullanışlıdır. Veri kümesinin boyutunu ve içerdiği sınıf sayısını artırmak için gelecekte çalışmalar yapılacaktır.

KAYNAKLAR

- Abate, S. T., Menzel, W., and Tafila, B., 2005. An Amharic speech corpus for large vocabulary continuous speech recognition. In Ninth European Conference on Speech Communication and Technology.
- Abera, H., 2018. Design of a Tigrinya Language Speech Corpus for Speech Recognition. In Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing (pp. 78-82).
- Alim, S. A., and Rashid, N. K. A., 2018. Some commonly used speech feature extraction algorithms. From Natural to Artificial Intelligence-Algorithms and Applications.
- Alumäe, T., 2008. Comparison of different modeling units for language model adaptation for inflected languages. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 488-499). Springer, Berlin, Heidelberg.
- Amrouche, A., and Rouvaen, J. M., 2003. Arabic isolated word recognition using general regression neural network. In 2003 46th Midwest Symposium on Circuits and Systems (Vol. 2, pp. 689-692). IEEE.
- Arel, I., Rose, D. C., and Karnowski, T. P., 2010. Deep machine learning-a new frontier in artificial intelligence research. IEEE computational intelligence magazine, 5 (4), 13-18.
- Aşlıyan, R., 2008. Design and implementation of Turkish speech recognition engine (Doctoral dissertation, DEÜ Fen Bilimleri Enstitüsü).
- Ayodele, T. O., 2010. Types of machine learning algorithms. In New advances in machine learning. IntechOpen.
- Bourouba, H., Djemili, R., Bedda, M., and Snani, C., 2006. New hybrid system (supervised classifier/HMM) for isolated Arabic speech recognition. In 2006 2nd International Conference On Information and Communication Technologies (Vol. 1, pp. 1264-1269). IEEE.
- Brown, K., and Ogilvie, S., 2010. Concise encyclopedia of languages of the world. Elsevier. p. 987.
- Chartrand, G., Cheng, P. M., Vorontsov, E., Drozdal, M., Turcotte, S., Pal, C. J. and Tang, A., 2017. Deep learning: a primer for radiologists. Radiographics, 37(7), 2113-2131.
- Chen, F., 2006. Speech Technology in Military Application. Designing Human Interface in Speech Technology, 289-330.
- Chiu, C. C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., and Jaitly, N., 2018. State-of-the-art speech recognition with sequence-to-sequence models. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4774-4778). IEEE.
- Cole, R. A., Rudnicki, A. I., Zue, V. W., and Reddy, D. R., 1980. Speech as patterns on paper. Perception and production of fluent speech, 3-50.
- Denes, P., 1959. The design and operation of the mechanical speech recognizer at University College London. Journal of the British Institution of Radio Engineers, 19(4), 219-229.
- Dreyfus-Graf, J., 1952. Le typosonographe phonétique ou phonétographe. Bulletin Technique Des PTT Suisses. Francia.

- Dubnov, H., 2003. Grammatical Sketch of Somali (Vol. 20). Köppe.
- Dudley, H., and Balashek, S., 1958. Automatic recognition of phonetic patterns in speech. *The Journal of the Acoustical Society of America*, 30(8), 721-732.
- Earnshaw, R., Dill, J., & Kasik, D., 2019. *Data Science and Visual Computing*. Springer International Publishing.
- El Choubassi, M. M., El Khoury, H. E., Alagha, C. J., Skaf, J. A., and Al-Alaoui, M. A., 2003. Arabic speech recognition using recurrent neural networks. In *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology* (IEEE Cat. No. 03EX795) (pp. 543-547). IEEE.
- Ferguson, J. (Ed.), 1980. *Hidden Markov Models for speech*. IDA-CRD, Princeton, New Jersey.
- Fisher William, M., 1986. The DARPA Speech Recognition Research Database: Specifications and Status/William M. Fisher, George R. Doddington, Kathleen M. Goudie-Marshall. In *Proceedings of DARPA Workshop on Speech Recognition* (pp. 93-99).
- Forgie, J. W., and Forgie, C. D., 1959. Results obtained from a vowel recognition computer program. *The Journal of the Acoustical Society of America*, 31(11), 1480-1489.
- Gac, D. L., 2002. Tonal alternations and prosodic structure in Somali. In *Speech Prosody 2002*, International Conference.
- Gaikwad, S. K., Gawali, B. W., and Yannawar, P., 2010. A review on speech recognition technique. *International Journal of Computer Applications*, 10(3), 16-24.
- Hain, T., 2002. *Hidden model sequence models for automatic speech recognition* (Doctoral dissertation, University of Cambridge).
- Hyman, L. M., 1981. Tonal accent in Somali. *Studies in African linguistics*, 12(2), 169.
- Ibrahim, Y. A., Odiketa, J. C., and Ibiyemi, T. S., 2017. Preprocessing technique in automatic speech recognition for human computer interaction: an overview. *Annals. Computer Science Series*, 15(1).
- Itakura, F., 1975. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1), 67-72.
- Kotsiantis, S. B., Zaharakis, I., and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160, 3-24.
- Kumar, K., and Aggarwal, R. K., 2011. Hindi speech recognition system using HTK. *International Journal of Computing and Business Research*, 2(2), 2229-6166.
- Leondes, C. T. (Ed.), 2018. *Intelligent Systems: Technology and Applications*, Six Volume Set. CRC Press.
- Lippmann, R. P., 1987. An introduction to computing with neural nets. *IEEE Assp magazine*, 4(2), 4-22.
- Makhoul, J., and Schwartz, R., 1995. State of the art in continuous speech recognition. *Proceedings of the National Academy of Sciences*, 92(22), 9956-9963.
- Mitchell, T. M., 1997. Does machine learning really work?. *AI magazine*, 18(3), 11-11.
- Muda, L., Begam, M., and Elamvazuthi, I., 2010. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv preprint arXiv:1003.4083*.
- Nagata, K., Kato, Y., and Chiba, S., 1964. Spoken digit recognizer for Japanese language. In *Audio Engineering Society Convention 16*. Audio Engineering Society.

- Nguyen, V. H., 2019. An End-to-End Model for Vietnamese Speech Recognition. In 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF) (pp. 1-6). IEEE.
- Nimaan, A., Nocera, P., and Bonastre, J. F., 2006. Automatic transcription of somali language. In Ninth International Conference on Spoken Language Processing.
- Olson, H. F., and Belar, H., 1956. Phonetic typewriter. *The Journal of the Acoustical Society of America*, 28(6), 1072-1081.
- Petkar, H., 2016. A review of challenges in automatic speech recognition. *International Journal of Computer Applications*, 151(3), 23-26.
- R. G. Leonard and G. R. Doddington., 1992 A speaker-independent connected-digit database. [Online]. Available: <https://catalog.ldc.upenn.edu/docs/LDC93S10/tidigits.readme.html> (22.01.2020)
- Rabiner, L. R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257-286.
- Raj Reddy, D., 1966. An approach to computer speech recognition by direct analysis of the speech wave. Tech. Rept. CS 49. Stanford: Comp. Sci. Dept., Stanford Univ.
- Raza, A. A., Hussain, S., Sarfraz, H., Ullah, I., and Sarfraz, Z., 2009. Design and development of phonetically rich Urdu speech corpus. In 2009 oriental COCOSA international conference on speech database and assessments (pp. 38-43). IEEE.
- Riis, S. K., and Hansen, S. D., 1999. Hidden Markov models and neural networks for speech recognition.
- Rousseau, A., Deléglise, P., and Esteve, Y., 2012. TED-LIUM: an Automatic Speech Recognition dedicated corpus. In LREC (pp. 125-129).
- Russell, S. J., and Norvig, P., 2016. Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited.
- Sakoe, H. and Chiba, S., 1971. A dynamic programming approach to continuous speech recognition. In 7th ICA, 20C-13.
- Sakoe, H., 1971. Dynamic-programming approach to continuous speech recognition. In 1971 Proc. the International Congress of Acoustics, Budapest.
- Saksamudre, S. K., and Deshmukh, R. R., 2015. Comparative study of isolated word recognition system for Hindi language. *International Journal of Engineering Research and Technology*, 4(07).
- Sathya, R., and Abraham, A., 2013. Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(2), 34-38.
- Schmidhuber, J., 2015. Deep learning in neural networks: An overview. *Neural networks*, 61, 85-117.
- Seid, H., and Gambäck, B., 2005. A speaker independent continuous speech recognizer for Amharic. Userware Laboratory Swedish Institute of Computer Science. Kista.
- Slutsker, G. S., 1968. Nelinejnyp method analiza recevych signalov. *Trudy Niir*, 2, 76-82.
- Sutton, R. S., and Barto, A. G., 1998. Introduction to reinforcement learning (Vol. 2, No. 4). Cambridge: MIT press.
- Tiwari, V., 2010. MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies*, 1 (1), 19-22.

- Tunali, V., 2005. A Speaker Dependent Large Vocabulary Isolated Word Speech Recognition System for Turkish. Ms Thesis, Institute For Graduate Studies in Pure and Applied Sciences, İstanbul.
- Tüzün, Ö. B., Erzin, E., Demirekler, M., Memişoğlu, A. T., Uğur, M. S., and Cetin, A. E., 1995. A speaker independent isolated word recognition system for Turkish. In *Speech Recognition and Coding* (pp. 104-107). Springer, Berlin, Heidelberg.
- Uchat, N. S., 2006. Hidden Markov Model and Speech Recognition. In Seminar report, Department of Computer Science and Engineering Indian Institute of Technology, Mumbai.
- V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.
- Velichko, V. M., and Zagoruyko, N. G., 1970. Automatic recognition of 200 words. *International Journal of Man-Machine Studies*, 2(3), 223-234.
- Vicens, P., 1969. *Aspects of Speech Recognition by Computer* (No. CS-127). Stanford Univ Calif Dept OF Computer Science.
- Vintsyuk, T. K., 1968. Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis*, 4(1), 52-57.
- Warden, P., 2018. Speech commands: A dataset for limited-vocabulary speech recognition. arXiv preprint arXiv:1804.03209.
- Warden, P., 2018. Speech commands: A dataset for limited-vocabulary speech recognition. arXiv preprint arXiv:1804.03209.
- Wilpon, J. G., 1994. Applications of voice-processing technology in telecommunications. *Voice communication between humans and machines*, 280.
- Yamashita, R., Nishio, M., Do, R. K. G., and Togashi, K., 2018. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9(4), 611-629.

ÖZGEÇMİŞ

Abdullatif ALI TAHLİL, 01.01.1989'da Mogadişu'da dünyaya geldi. İlk ve orta okulu Albirri'de okudu. Lise öğrenimini 2011 yılında SOS'ta bitirdi. 2011'de Hormuud Üniversitesi, Mühendislik Fakültesi, Telekomünikasyon Bölümü'nü kazandı ve burslu olarak 2015'te tamamladı. 2016 yılında Türkiye Cumhuriyeti'nin düzenlediği burs programıyla (YTB) Atatürk Üniversitesinde Elektrik-Elektronik Mühendisliği Elektronik Ana Bilim Dalı'nda Yüksek Lisans yapmaya hak kazandı.

