



T.C.
KONYA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

**VİDEOLARDA GEÇEN KONUŞMALARIN
METNE DÖNÜŞTÜRÜLMESİ VE ZAMAN
TABANLI İNDEKSLENMESİ AMACIYLA BİR
UYGULAMA GELİŞTİRİLMESİ**

Oğuzhan Mert KİRAZ

YÜKSEK LİSANS

Bilgisayar Mühendisliği Anabilim Dalı

Nisan-2024
KONYA
Her Hakkı Saklıdır

TEZ KABUL VE ONAYI

Oğuzhan Mert KİRAZ tarafından hazırlanan “Videolardaki sesleri yazıya çevirerek aranılan kelimenin saniyesine götüren program” adlı tez çalışması 29/04/2024 tarihinde aşağıdaki jüri tarafından oy birliği / oy çokluğu ile Konya Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı’nda YÜKSEK LİSANS olarak kabul edilmiştir.

Jüri Üyeleri

İmza

Başkan

Prof.Dr. İsmail BABAOĞLU

.....

Danışman

Prof.Dr. Mustafa Servet KIRAN

.....

Üye

Prof.Dr. İsmail BABAOĞLU

.....

Yukarıdaki sonucu onaylarım.

Prof. Dr. Mevlüt UYAN
Enstitü Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Oğuzhan Mert KİRAZ

Tarih:

ÖZET

YÜKSEK LİSANS TEZİ

VIDEOLARDA GEÇEN KONUŞMALARIN METNE DÖNÜŞTÜRÜLMESİ VE ZAMAN TABANLI İNDEKSLENMESİ AMACIYLA BİR UYGULAMA GELİŞTİRİLMESİ

Oğuzhan Mert KİRAZ

**Konya Teknik Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı**

Danışman: Prof.Dr. Mustafa Servet KIRAN

2024, 74 Sayfa

Jüri

**Prof.Dr.Mustafa Servet KIRAN
Prof.Dr. İsmail BABAOĞLU
Doç.Dr. Hüseyin HAKKI**

Yapay sinir ağlarındaki ilerlemeler, metin tanıma ve metinden sese çevirme gibi birçok veri dönüşümü ve tanıma işlemlerinde önemli gelişmelere yol açmıştır. Özellikle Konuşmadan Metne (Speech to Text) yöntemi, sesli konuşmaları yazıya dönüştürme konusunda popülerlik kazanmıştır. Günümüzde görüntülü ve sesli içeriklerin popülerliği arttıkça insanlar video platformlarında özel içerikler üreterek geniş kitlelere ulaşmakta ve gelir elde etmektedirler. Yapılan çalışma sesin analizinin yanı sıra elde edilen verilerin sorgulanmasını da sağlamaktadır böylece kullanıcılar videoda aradıkları kelimeyle ilgili bölüme erişebilmekte ve bu bölümden itibaren videoları izleyebilmektedir. Bu sayede kullanıcılar aradıkları kelimenin geçtiği saniyelere ulaşmak için bütün bir videoyu izlemek zorunda kalmadan görüntülemek istedikleri kısma ulaşabilmektedir. Geliştirilen yöntemler sayesinde kullanılan STT yöntemine göre videolardaki seslerden elde edilen metinler daha doğru sonuçlar verebilmektedir. Tespit edilen eksiklikler sonucunda geliştirilen Bölümlleme yöntemiyle veri setlerinde konuşmacının durana kadar MFCC algoritması kullanılarak Fourier dönüşümü yöntemi ile frekans değerleri hesaplanmış ve ayrıştırılmıştır. STT yöntemine uygulanan bu işlemlerle YouTube'un kullandığı STT yöntemine göre daha doğru bir süre/metin ilişkisinde veriler kaydedilmiştir. Bu yöntem ile gürültüsüz kaydedilen veri setlerinde saniye sekanslı doğruluk oranı hatalarının önüne geçilmektedir. Tüm bu veri setleri veri tabanına süre/metin eşleştirmesi yapılarak kaydedilmekte böylece kullanıcılar aradıkları kelimelerin geçtiği saniyeleri görebilmekte ve istedikleri zaman o kısmı oynatabilmektedirler. Uygulamanın çalışma süreci iki aşamadan oluşmaktadır: ilk aşama videonun indirilmesi, ikinci aşama ise videodaki seslerin metne çevrimidir. İnternet hızına bağlı olarak videoların indirilme süresi ve kelime çevrim süresi değişiklik göstermektedir. Standart bir internet ağı kullanılarak bir dakikalık bir videonun indirilme süresi yaklaşık 30 saniye, beş dakikalık bir videonun indirilme süresi ise ortalama 75 saniye civarındadır. İkinci aşama

olan videolardaki seslerin metne çevrilme süreci video indirildikten sonra gerçekleştirilir. Bu aşamada bir dakikalık bir videonun kelime çevrimi yaklaşık 30 saniye, 5 dakikalık bir videonun kelime çevrimi ise genellikle ortalama 75 saniye sürmektedir. Geliştirilen program, YouTube otomatik çeviri sistemiyle aynı yöntemi kullanmasına rağmen yapılan deneylerde daha başarılı sonuçlar elde edilmiştir. Yüz videodan oluşan deney sürecinde geliştirilen programın hem oransal olarak daha fazla çevrim gerçekleştirdiği hem de videolardan elde edilen verilere göre rakamsal olarak daha doğru bir kelime çevrimi yaptığı belirlenmiştir. Her iki sistemin aynı yöntemleri kullanmasına rağmen bu farkın geliştirilen programda kullanılan bölümlene yöntemi, MFCC ve Fourier dönüşümü yönteminin uygulanmasından kaynaklanmaktadır. Uygulamanın asıl amacı kullanıcılara zaman tasarrufu sağlamaktır. Fiber internet hızlarının sağlandığı bölgelerde özellikle bir saat üzerindeki videolar için kullanıcıya üç ila beş dakika arasında hızlı bir çözüm sunulabilmektedir. Standart bir Wifi ağında ise tüm aşamalar bir saatten uzun videolar için dört ila yedi dakika arasında tamamlanabilmektedir.

Anahtar Kelimeler: Youtube, Sesten Metne, Ses İşleme, Video işleme, Yapay Sinir Ağları, Fourier Dönüşümü

ABSTRACT

MS THESIS

DEVELOPING AN APPLICATION FOR CONVERTING THE CONVERSATIONS IN VIDEOS INTO TEXT AND TIME-BASED INDEXING

Oğuzhan Mert KİRAZ

**Konya Technical University
Institute of Graduate Studies
Department of Computer Engineering**

Advisor: Prof.Dr. Mustafa Servet KIRAN

2024, 74 Pages

Jury

Prof.Dr. Mustafa Servet KIRAN

Prof. Dr. İsmail BABAOĞLU

Assoc. Prof. Hüseyin HAKKI

Advances in artificial neural networks have led to significant improvements in many data transformation and recognition processes such as text recognition and text-to-speech conversion. In particular, the Speech to Text method has gained popularity for transcribing audio conversations. Nowadays, as the popularity of video and audio content increases, people are reaching large audiences and generating revenue by producing special content on video platforms. In addition to analyzing the audio, the study also enables the querying of the data obtained so that users can access the section of the video related to the word they are looking for and watch the videos from this section. In this way, users can reach the part they want to view without having to watch an entire video to reach the seconds in which the word they are looking for occurs. Thanks to the developed methods, the texts obtained from the audio in the videos can give more accurate results compared to the STT method. With the segmentation method developed as a result of the identified deficiencies, frequency values were calculated and separated by Fourier transform method using the MFCC algorithm until the speaker stops in the data sets. With these processes applied to the STT method, data was recorded in a more accurate duration/text relationship than the STT method used by YouTube. With this method, second-sequence accuracy rate errors are prevented in data sets recorded without noise. All these datasets are saved to the database with duration/text matching so that users can see the seconds in which the words they are looking for occur and can play that part whenever they want. The working process of the application consists of two stages: the first stage is the downloading of the video and the second stage is the transcription of the audio from the video. Depending on the internet speed, the download time of the videos and the transcription time varies. Using a standard internet network, a one-minute video takes about 30 seconds to download and a five-minute video takes about 75 seconds to download. The second stage, the transcription of the audio in the videos, takes place after the video is downloaded. At this stage, it takes about 30 seconds to transcribe a one-minute video and 75 seconds to transcribe a 5-minute video. Although the developed program uses the same method as the YouTube automatic translation system, more successful results were obtained in the experiments. In the experimental process consisting of one hundred videos, it was determined that the developed program performed both proportionally more translations and numerically more accurate word translations according to the data obtained from the videos. Although both systems use the same methods, this difference is due to the partitioning method, MFCC and Fourier transform method used in the developed program. The main purpose of the application is to save time for the users. In regions where fiber internet speeds are provided, a fast solution can be offered to the user between three to five minutes, especially for videos over one hour. On a standard Wifi network, all steps can be completed in four to seven minutes for videos longer than one hour.

Keywords: YouTube, Speech to Text, Sound to Text, Video processing, Voice Procesing, Artificial Neural Networks, Fourier Transform



ÖNSÖZ

Çalışmalarım süresince Yüksek Lisans tez konumun seçilmesinde ve tüm aşamasında büyük ilgi ve desteğini gördüğüm danışman hocam Sayın Prof. Dr. Mustafa Servet KIRAN'a araştırma, veri toplama ve değerlendirme çalışmalarında desteğini gördüğüm Kara Kuvvetleri personeli Mehmet YILDIRIM'a, iş hayatım boyunca yüksek lisans eğitimime destek olan Kara Kuvvetleri Komutanlığı personeli şube müdürüm Sayın Piyade Albay Tolga BAŞKAN'a, lisans öğrenimim süresince de öğrencisi olduğum değerli Konya Teknik Üniversitesi / Mühendislik ve Doğa Bilimleri Fakültesi / Bilgisayar Mühendisliği bölümünün öğretim görevlilerine ve tüm hayatım boyunca her türlü maddi ve manevi desteği benden esirgemeyen ve anlayışlarıyla bana rahat bir çalışma ortamı sağlayan çalışmalarım boyunca bana daima destek olan aileme sonsuz teşekkürlerimi sunarım.

Oğuzhan Mert KİRAZ
KONYA-2024



İÇİNDEKİLER

ÖZET	iv
ABSTRACT	vi
ÖNSÖZ	viii
İÇİNDEKİLER	ix
SİMGELER VE KISALTMALAR.....	xi
1. GİRİŞ.....	1
1.1. Tezin Amacı	3
1.2. Tezin Önemi	4
2. KAYNAK ARAŞTIRMASI	6
2.1. Yapay Sinir Ağları ile Konuşma Tanıma alanında yapılan çalışmalar	6
2.2. Konuşma Tanıma Süreci.....	7
2.3. Konuşmadan Metne Yöntemi ile ilgili yapılan çalışmalar.....	9
3. MATERYAL VE YÖNTEM.....	16
3.1. Yapay Sinir Ağları.....	16
3.2. Konuşmadan Metne Yöntemi	18
3.3. Konuşma Tanıma Yönteminin Çalışma Mantığı	18
3.4. Derin Öğrenme Mimarileri	22
3.4.1. Tekrarlayan Sinir Ağları (RNN).....	22
3.4.2. Uzun Kısa Vadeli Hafıza Ağları (LSTM Ağları)	23
3.4.3. Dikkat Mekanizması (Self Attention Mechanism)	25
3.5. Fourier Dönüşümü.....	26
3.6. Mel Frekans Kepstral Sayıları Algoritması	27
3.7. Google Speech Recognition yöntemi nasıl çalışır?	30
3.8. Bölümleme Yöntemi.....	30
4. ARAŞTIRMA SONUÇLARI VE TARTIŞMA	32
4.1. STT Yöntemi ve Geliştirilmesi	32
4.2. Uygulamanın Karşılaştırılması.....	32
4.3. Uygulama Arayüzü.....	50
5. SONUÇLAR VE ÖNERİLER	55
5.1 Sonuçlar	55
5.2 Öneriler	57

KAYNAKLAR	59
------------------------	-----------



SİMGELER VE KISALTMALAR

MLP: Çok Katmanlı Algılayıcı (Multi Layer Perceptron)
RNN: Rekurrent Sinir Ağı (Recurrent Neural Network)
CNN: Evrişimli Sinir Ağı (Convolutional Neural Network)
SSD: Katı Hal Sürücüsü (Solid State Drive)
MFCC : Mel Frekans Cepstral Katsayıları (Mel Frequency Cepstral Coefficients)
NLP: Doğal Dil İşleme (Natural Language Processing)
LTSM: Uzun Kısa Vadeli Hafıza (Long Short-Term Memory)
AM: Akustik Modelleme (Acoustic Modeling)
ASR: Otomatik Konuşma Tanıma (Automatic Speech Recognition)
CPU: Merkezi İşlem Birimi (Central Processing Unit)
DNN: Derin Sinir Ağı (Deep Neural Network)
IT: Bilgi Teknolojileri (Information Technology)
RBM: Sınırlı Boltzmann Makineleri (Restricted Boltzmann Machines)
SRN: Basit Tekrarlayan Ağ (Simple Recurrent Network)
AI: Yapay Zekâ (Artificial Intelligence)
MB: Megabayt
ASR: Otomatik Konuşma Tanıma (Automatic Speech Recognition)
STT: Speech to Text (Konuşmadan Metne)
SR: Speech Recognition (Konuşma Tanıma)
XML: Extensible Markup Language (Genişletilebilir İşaretleme Dili)
ÇKA: Çok Katmanlı Algılayıcı
MFC: Mel Frekans Cepstral
FFT: Fast Fourier Transform
DCT: Ayrık Kosinüs Dönüşümü

1. GİRİŞ

Video, ses ve görüntüden meydana gelmektedir. Video, herhangi bir görüntünün iletmek veya saklanmak için elektromanyetik enerjiye çevrilmiş halidir. Ses, insan kulağını etkileyerek işitme duyusu oluşturan hava molekülleri ya da bunların neden olduğu ufak basınç değişimleri ya da bu fiziksel olayın neden olduğu işitsel izlenimdir. Gelişen teknolojiyle birlikte video ve ses analizi alanları arasındaki sınırları aşarak multimodal veri anlayışını artırmaya yönelik büyük bir ilerleme kaydedilmektedir. Video ve ses verilerinin birleştirilerek incelenmesi olayların daha derinlemesine ve bütünsel bir şekilde anlaşılmasına olanak tanımaktadır. Görsel ve işitsel bilgilerin bir araya getirilmesi bilgisayar sistemlerinin çevresel bağlamı daha etkili bir şekilde kavramasına imkân sağlamakta bu da yapay zekâ uygulamalarında daha sofistike ve duyarlı sonuçlara yol açmaktadır. Video ile ses arasındaki bu bağlantı hem eğlence hem de profesyonel alanlarda kullanılan uygulamalarda daha zengin ve anlamlı deneyimler sunmaktadır. Bu bağlamda video-ses tanımları, bilgisayar görüşü ve ses işleme alanlarını birleştirerek daha kapsamlı bir veri analizi yaklaşımı imkânı oluşmaktadır. Bu proje video ve ses verileri arasındaki bu önemli bağlantıyı anlamaya odaklanarak daha derinlemesine ve bütünsel bir anlayışın nasıl elde edilebileceği konusunda kilit bir katkı sağlamayı amaçlamaktadır. Günümüzde derin öğrenme ve onun bir alt dalı olan yapay sinir ağlarının kullanımıyla birlikte gündelik işlerimiz daha da kolaylaşmıştır. Bu gelişmelere örnek verecek olursak özellikle evlerimizde kullandığımız temizleme robotu gibi sürekli öğrenerek kendini yenileyen ve öğrendiği bilgileri işleyerek gelişen sistemler hayatımızı kolaylaştırmaktadır. Video paylaşım platformları da kullanıcıların bilgi paylaşımı, bilgiye ulaşması ve değerlendirmesi açısından birçok insan için fırsat konumundadır. Bu platformlarda eğitimden sanata, sanattan gündeme kadar birçok video bulunmaktadır. Özellikle eğitim amaçlı videolar birçok öğrenci, eğitici ve katkı sağlayıcılar için önemlidir. Video paylaşım platformlarının bazılarında (YouTube gibi) altyazı açıklamaları da bulunmaktadır. YouTube platformunda bulunan altyazılar videodaki seslere göre çevrim sağlamasına rağmen kelime arama imkânı sunmamaktadır. Bu platformlarda gerçekleştirilen kelime çevirilerinde videonun orijinal sesine bağlı olarak farklı çevirimler yapıldığı ve mevcut yöntemlerin frekans değerlerine dayalı olarak doğru çeviri imkânının düşük ihtimalde olduğu göz önünde bulundurularak, yapay sinir ağları alanındaki teknolojik ilerlemelerle dil karmaşıklıklarını anlama ve işleme yeteneklerini artırmayı amaçlayan bir program geliştirilmesi hedeflenmiştir. Sesten metne çevrim imkânı sağlayan ses tanıma

yöntemleri yapay sinir ağlarına dayanmaktadır. Ses tanıma yöntemleri videolardaki konuşmaları anlık olarak metin formatına dönüştürme yeteneği sağlar. Daha sonra metin analizi yöntemleriyle kullanıcıların belirli bir kelimeyi içeren bölümleri hızlı ve etkili bir şekilde tespit etmesine olanak tanır. Ses tanıma yöntemlerini kullanarak videolardaki konuşmaları metne dönüştüren bu proje ardından metin analizi yöntemleriyle videoda aranmak istenilen kelimeye, o kelimenin video geçtiği zamana erişebilme ve bunları belirleyebilmeyi hedeflemektedir. YouTube altyazı sistemlerinin kullanımında tespit edilen eksikliklerin tamamlanması ile geliştirilen bu proje aynı zamanda video platformlarından faydalanmak isteyen kullanıcıların zaman tasarrufu yapmalarına yönelik ses ve metin teknolojilerini bir araya getirmektedir. Bu sayede kullanıcılar video içeriğini daha etkili bir şekilde keşfetme, araştırma ve anlama konusunda daha fazla destek alabilmektedir. Ses tanıma yöntemlerinde kullanılan bu teknolojik ilerlemeler yaşanan dil karmaşıklıklarını anlama ve işleme yeteneklerimizi artırarak, iletişim ve veri analizleri alanlarda önemli gelişmelere yol açmıştır. (Jurafsky ve ark., 2009) tarafından, bu alanlardaki çalışmaları incelenmiştir. İlgili çalışma doğal dil işleme, dil modelleri, konuşma tanıma yöntemleri ve sesleri yazıya çevirme gibi konuları kapsayarak bu konularda araştırmalar yapmak için kapsamlı bir kaynak niteliği taşımaktadır. Konuşma tanıma teknolojisi sesi otomatik olarak metne dönüştürerek konuşma ve yazılı metinler arasındaki dönüşümü sağlamayı hedeflemektedir. Yapılan çalışmalar ses dalgalarının matematiksel olarak analizi, dil modelleri, makine öğrenimi ve dil öğrenimi üzerinedir. (Chollet ve ark., 2019) tarafından, derin öğrenme modellerinin konuşma tanıma ve sesleri yazıya çevirme alanında büyük bir etki yarattığını vurgulanmaktadır. Yapay sinir ağları ve derin öğrenme algoritmaları büyük veri setleri üzerinde eğitilerek daha doğru ve kesin sonuçlar elde etmeye yönelik çözümler sunmaktadır. Bu çözümler sayesinde otomatik sesli yanıt sistemleri, dikte edilen metinlerin transkriptleri ve toplantıların otomatik olarak yazılı metinlere dökülmesi gibi birçok uygulama mümkün hale gelmiştir.

(Mikolov ve ark., 2013), kelime gömme (word embedding) tekniklerini tanıtarak, dilin semantik yapısını temsil etmede önemli bir adım atılmış olduğunu göstermişlerdir. Kelime gömme kelimelerin vektörlerle temsil edilmesini sağlayarak dil işleme algoritmalarının daha etkili ve anlamlı sonuçlar üretmesini sağlamaktadır. Bu teknikler metin analizi, duygu analizi, metin sınıflandırma, öneri sistemleri ve otomatik çeviri gibi birçok uygulama alanında kullanılmaktadır. Bu sayede metin verilerinin daha etkili bir şekilde anlaşılması ve kullanılması mümkün hale gelmektedir. Bunun yanı sıra

konuşma tanıma, doğal dil işleme ve bilgisayar destekli dil öğrenimi teknolojileri birçok sektörde büyük etkiler yaratmaktadır. Örneğin; çağrı merkezlerinde müşteri hizmetleri için otomatik sesli yanıt sistemleri, kullanıcı arayüzlerinde sesli komut sistemleri ve sanal asistanlar gibi uygulamalar insan makine etkileşimini daha kullanıcı dostu ve verimli hale getirmektedir. Ayrıca veri madenciliği ve büyük veri analizleri gibi alanlarda da doğal dil işleme teknikleri kullanılarak, büyük miktarda metin verisinden anlamlı bilgiler çıkarılmakta ve karar verme süreçlerine katkı sağlamaktadır. Bu alanlardaki ilerlemeler; iletişim, arama motorları, metin madenciliği, sesli komut sistemleri, transkriptler gibi birçok alanda insanların günlük hayatını kolaylaştırmaktadır. Ancak hala bazı zorluklar ve geliştirilmesi gereken alanlar bulunmaktadır. Örneğin; aksanlı konuşmalar, hızlı konuşmalar, arka plan gürültüsü gibi faktörler konuşma tanıma ve sesleri yazıya çevirme süreçlerinde doğruluk oranını etkileyebilmektedir. Bu nedenle daha hassas ve adaptif algoritmaların geliştirilmesi ve daha fazla veri toplanması önemli araştırma konularıdır.

Genel olarak bu teknolojiler, bilgisayar destekli dil öğrenimi ve konuşma tanıma gibi alanlardaki teknolojik gelişmelerle insanların konuşma ve yazılı metin arasındaki dönüşümü daha etkili ve verimli bir şekilde gerçekleştirmesini sağlamaktadır.

1.1.Tezin Amacı

Bu tez çalışmasının temel amacı video içeriklerinde yer alan ses bilgilerini anlamlı metinlere dönüştürerek, kullanıcıların istediği kelimeleri arama süreçlerini optimize eden bir program geliştirmektir. Bu hedef doğrultusunda videolardaki ses verilerini analiz etmek, tanımak ve transkribe etmek amacıyla kullanılan mevcut teknolojilerin sınırlarını aşmak için kapsamlı bir yaklaşım benimsenmiştir. Öncelikle videolardan elde edilen ses bilgilerinin doğru ve güvenilir bir şekilde metin formatına çevrilmesi için ses tanıma yöntemlerinin geliştirilmesi üzerine odaklanılmıştır. Bu yöntemlerin farklı aksanlara, gürültü seviyelerine ve konuşma hızlarına adapte edilebilirliği sağlanmasıyla çeşitli video kaynaklarından gelen ses verilerinin etkili bir şekilde işlenebilmesi mümkün olmuştur. Bu yönteminin kullanımı sonrası elde edilen metin verileri üzerinde dil işleme teknikleri kullanılarak anlamlı kelimelerin belirlenmesi ve bu kelimelerin içeriğe göre endekslenmesi sağlanmıştır. Bu aşama kullanıcıların belirli bir kelimeyi videolarda hızlı ve doğru bir şekilde aramasına olanak tanıyan temel bir bileşeni oluşturacaktır.

Günümüzde video paylaşım sitelerini kullanıcılar video izlemek, yüklemek ve paylaşmak için tercih etmektedir. Bu da insanların bilgiye ulaşmasını kolaylaştırmaktadır. Bazı video sitelerinde video yükleyen kullanıcılar altyazı da kendileri ekleyerek kendi konuştuğu dilin altyazılı halini de kullanıcılara sunmaktadır. YouTube video paylaşım sitesinde video üreterek dünya çapında bilinmiş farklı ülkelerde izleyicileri olan kullanıcılar bulunmaktadır. Video üreticileri birçok farklı ülkelere erişmek için seslendirme ve altyazıları o ülkenin diline göre de çevirmek için birçok personele ihtiyaç duymaktadır. Video paylaşım sitelerinden YouTube üzerinde bulunan videolar altyazı eklenmediği durumda kullanıcının isteği üzerine otomatik olarak çevirme eklentisi de bulunmaktadır. Çevirme butonuna basıldıktan sonra videodaki seslere göre farklı çevrimler ortaya çıktığı ayrıca kullanıcıların herhangi bir kelime arayamadığı da bilinmektedir. Tüm bu araştırmalar sonucu STT yönteminin tespit edilen ihtiyaçlara göre geliştirilmesi ile kullanıcılara kolay ve anlaşılır bir arayüzle kullanıcının, istediği videoyu indirebilmesi, istediği kelimeyi araması, aradığı kelimenin saniyesini gösterilmesi ve aradığı kelimeye tıkladığında kullanıcının kendi medya oynatıcısından o saniyeden itibaren çalışmaya başlayacak program için araştırma ve geliştirme faaliyetleri yapılmıştır. Bu proje ile ses tanıma alanında araştırma ve geliştirme çalışmalarına katkıda bulunmak ve kullanıcılara zaman kaybı yaşatmaksızın kullanışlı bir program sunulması hedeflenmiştir.

1.2.Tezin Önemi

Günümüzde internet üzerindeki video içerikleri hızla artmakta ve bu içeriklerdeki ses bilgilerine erişim geleneksel metin tabanlı arama yöntemlerine kıyasla daha karmaşık hale gelmektedir. Bu proje videolardaki sesleri metne dönüştürerek kullanıcıların belirli konuları veya kelimeleri içeren video içeriklerini hızlı ve hassas bir şekilde bulabilmesini sağlamak amacıyla bir çözüm sunmayı hedeflemektedir. Videolardaki sesleri metne çeviren bir programın geliştirilmesi eğitimden medyaya, güvenlik uygulamalarından araştırmalara kadar birçok endüstriye uygulanabilir çözümler sunma potansiyeli taşımaktadır. Örneğin; eğitim sektöründe ders içeriklerini daha etkili bir şekilde indeksleme, medya sektöründe içerik arama ve kataloglama ve güvenlik uygulamalarında sesli içerik analizi gibi birçok alanda kullanılabilirlik açısından geniş bir perspektif sağlamaktadır. Videolardaki sesleri metne dönüştürme, büyük veri setlerinin içerisinde gizli kalmış bilgilerin daha etkili bir şekilde analiz edilmesine olanak tanımaktadır. Bu da işletmelerin veya kurumların veri tabanı

sorgulama, veri madenciliđi, veri grselleřtirme ve raporlama gibi iřlevleri gerekleřtirmesi gibi uygulamalarında daha derinlemesine igrler elde etme ve stratejik kararlar alma srelerini desteklemektedir. Bu proje mevcut video paylařım sitelerinde bulunan altyazı sistemlerinde tespit edilen eksiklikler ve bu platformlarda kelime arama zelliđi olmaması nedeniyle bu alanda yapılan arařtırma ve geliřtirmelerle IT (Bilgi Teknolojileri) sektrnn temel amacı olan insanların hayatını yararlı bir řekilde kolaylařtırmaktadır. Video paylařım siteleri kullanıcılara bir bilgi paylařımını sunmaktadır. Videolar bazen bir dakika olabildiđi gibi bazen beř saatten fazla bile olabilmektedir. Programın nemi kullanıcı izlemek istediđi yeri izlemek iin saatlerce aramaya veya video iinde dakika dakika gezerek zaman kaybetmesini engellemek, aradıđı kelimeyi ve kelimenin bulunduđu saniyeyi bulup kendi medya oynatıcısından izlemesini sađlayarak hem zaman kaybını nlemek hem de gndelik hayatımız iin kolaylık sunmak maksadıyla geliřtirilmesi hedeflenmiřtir.

2. KAYNAK ARAŞTIRMASI

2.1. Yapay Sinir Ağları ile Konuşma Tanıma alanında yapılan çalışmalar

Yapay sinir ağları, konuşma tanıma alanında son yıllarda giderek artan bir öneme sahip olmuştur. Bu teknolojiler biyolojik sinir sisteminden esinlenerek tasarlanan ve genellikle derin öğrenme modellerini içeren bilgisayar algoritmalarıdır. İncelenen çalışmalarda da görüldüğü gibi özellikle Rekürsif Sinir Ağları (RNN) ve LSTM ağları, konuşma sinyallerinin karmaşıklığını ve içerdği desenleri anlama konusunda yüksek başarı elde etmiştir.

(Ahad, A. ve ark., 2002) tarafından yapılan çalışmada, yapay sinir ağlarının çok katmanlı algılayıcılarının konuşma tanıma performansı üzerindeki etkisi incelenmiştir. Bu çalışmada 0 ila 9 arasındaki sayılardan oluşan bir veri kümesi kullanılmıştır. Her bir ifade 100 kez tekrarlanarak toplamda 1000 ifade içeren bir veri kümesi elde edilmiştir. Bu veri kümesi üzerinde konuşma tanıma için en uygun 10 yapay sinir ağı modeli araştırılmıştır. Çalışmada iki katmanlı ağ kullanıldığında öğrenme oranının etkisi ve iterasyon sayısının tanıma oranına olan etkisi incelenmiştir. Sistemin en iyi performansı %0,01 hata oranıyla elde edilmiş ve öğrenme oranının farklı değerlerinin sistemin başarısını doğrudan etkilemediği gözlemlenmiştir. Ayrıca sonuca varmak için birkaç yüz adet iterasyon gerektiği belirtilmektedir.

(Uzun, E., 2007) tarafından gerçekleştirilen "Görme Engelliler İçin Basılı Doküman Yorumlama ve Seslendirme Sistemi" çalışmasında sistem metin, grafik ve tablo gibi görüntülerin sesli olarak aktarımını sağlamaktadır. Bu amaçla Yapay Sinir Ağları teknolojisi kullanılmıştır.

(Saijid, A. Marhon ve ark., 2006) tarafından Higher Enstitüsünde gerçekleştirilen yapay sinir ağlarıyla konuşmacı tanıma çalışmasında ortalama %80 tanıma başarısı elde edilmiştir. Bu çalışmada toplam 8 kişiden 5 adet eğitim ve 5 adet test verisi alınarak toplamda 80 veri noktasıyla bir veri tabanı oluşturulmuştur. Konuşma tanıma işlemleri bu veri noktalarıyla gerçekleştirilmiştir.

(Moralı, İ. A. ve ark., 2006) tarafından Dokuz Eylül Üniversitesi Bilgisayar Mühendisliği bölümünde yapılan çalışmada çok katmanlı algılayıcılar kullanılarak 4 kişiden alınan 72 adet ses örneğiyle konuşmacı tanıma çalışması gerçekleştirilmiştir. Bu çalışmada ses örneklerinin MFCC çıkarılarak geri yayılım algoritması uygulandığında %57 oranında başarı sağlanmıştır.

(Şentürk, T. ve ark., 2016) tarafından kaleme alınan "Türkçe Metin Seslendirme" isimli makalede iki farklı yöntem incelenmiştir. Çift-ses eklemeli yöntemin

seslendirilen metnin doğallıktan uzak olduğu tespit edilince hece eklemeli YILMAZ ve YAĞCI yöntemi kullanılmıştır. Ayrıca daha önce yapılan çalışmalarla karşılaştırılabilmesi ve yapılarla müdahale edilebilmesi için kullanıcıya XML biçiminde sunulmuştur. Aynı çalışmada görme engelliler için metin düzenleyici program da geliştirilmiştir. Klavyeden girilen her ifade sesli olarak iletilmiştir. Farklı yaş gruplarından dinleyicilerle yapılan deneylerde hece eklemeli yöntemin kullanıldığı sistemin anlaşılabilirlik oranı %96.1 olarak ölçülmüştür.

(Caner, M. ve ark., 2005) tarafından Afyon Kocatepe Üniversitesi ve Celal Bayar Üniversitelerinde gerçekleştirilen çalışmada, 7 farklı kişiden alınan 'a' sesli harfinin öznelikleri belirlenmiş ve bunlar sesli harfin alındığı kişiyi bulacak şekilde eğitilmiş çok katmanlı bir ağ yapısıyla eğitilmiştir. Daha sonra eğitim setinde bulunmayan 11 veri setiyle ağın başarısı test edilmiştir. Test sonucunda 7 kişide ortalama %90 tanıma başarısı elde edilmiştir.

Yapay sinir ağlarının çok katmanlı modellerinin geliştirilmesi ve geri yayılma işleminin sorunları nispeten giderildikten sonra derin öğrenme olarak adlandırılan mevcut olarak sıklıkla kullanılan yöntem 2006'da başlamıştır (Bengio, Y. ve ark, 2007). Bu kapsamda (Morandi, A., 2012) tarafından, Derin İnanç Ağları'nın (DBN) açgözlü yaklaşım ile eğitimini öngören bir strateji kullanarak verimli bir şekilde eğitilebileceğini gösterilmiştir. Stratejinin amacı ağırlıkları rastgele değil daha akıllı bir şekilde başlatmaktır.

2.2. Konuşma Tanıma Süreci

Konuşma teknolojisi günümüzde doğal dil işleme, makine öğrenimi ve ses işleme alanlarının kesişiminde önemli bir araştırma konusu haline gelmiştir. Bu önemli konuları temel alarak geliştirilen proje ses tanıma ve sesleri yazıya çevirme alanında yapılan çalışmalara odaklanarak videolardaki sesleri yazıya çevirerek aranan kelimelerin saniyesine götüren bir program geliştirmeyi amaçlamaktadır. Konuşma tanıma teknolojisinin uygulanabilmesi için veri madenciliği, dil öğrenme sistemleri ve konuşma tabanlı etkileşim yöntemleri önemlidir.

Ses tanıma alanı içerisinde bulunan konuşma tanıma disiplini gelişen teknoloji sürecinde kendine önemli bir yer edinmeye çalışan bir yöntemdir ve insan sesinin bir mikrofon vasıtasıyla bilgisayar tarafından algılanarak tanınması işlemidir (Yalçın, N., 2008). Konuşma tanıma teorisi, Otomatik Konuşma tanıma (Sürekli Ses Tanıma, Sözcük yakalama, Ayırık ses tanıma), konuşmacı tanıma, dil tanıma, ses üretme,

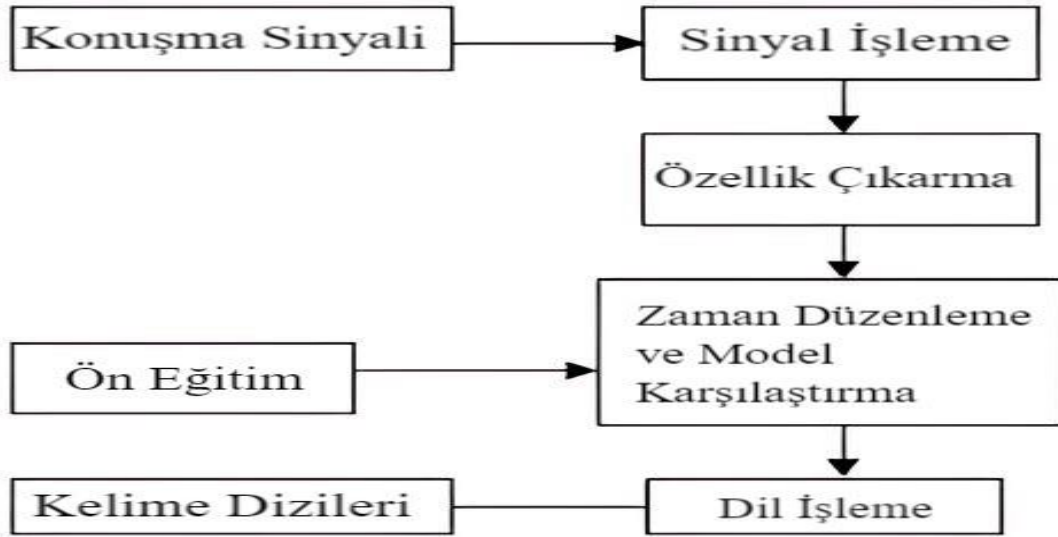
konusmacı ayrımı, ses kodlama uygulamalarını içermektedir. Konuşma tanımda sinyal işleme modülünün amacı örneklenen konuşma sinyalini işlemek ve genlik değişimlerinden, konuşmacı stresinden, iletişim araçlarından veya kanalından gelen gürültüden bağımsız bir gösterim elde etmektir. Sayısala dönüştürme işlemleri sırasıyla örnekleme, nicelendirme ve kodlama aşamalarıdır (Yalçın, N., 2008).

Konuşma sinyalinin modellenmesi, sesli ifadelerin bilgisayar destekli olarak tanıma sürecine eklenebilmesi için bunların öncelikle bu sürece hazırlanmaları gerekmektedir. Bu amaçla sesli ifadelerin bir mikrofon aracılığıyla örneksel sinyallere dönüştürülmesi, sayısallaştırılması, sayısallaştıran bu sinyallerin gerekirse filtrelenmesi, etiketlenmesi (örneğin sesler, fonemler, sözcükler olarak) ve tanıma işlemlerine taban oluşturacak parametrik yapılar ya da yalın modellerle ifade edilen biçimlere dönüştürülmesi gerekmektedir. Konuşma Teknolojisi, insan sesini matematiksel olarak analiz ederek konuşmayı anlaşılır metinlere dönüştürmeyi amaçlayan bir yöntemdir. Bu teknoloji doğal dil işleme, makine öğrenimi ve derin öğrenme gibi disiplinlerin etkileşimini içermektedir. Konuşma tanıma süreci ses dalgalarının örnekleme ve zamanlama işlemlerinden geçirilerek sayısal bir temsile dönüştürülmesiyle başlamaktadır. Ardından örüntü tanıma algoritmaları ve dil modelleri kullanılarak sesin metne dönüşümü gerçekleştirilir.

Derin sinir ağlarında, RNN ve LSTM gibi yapılar konuşma tanıma yöntemlerinde kullanılan etkili yöntemler olarak bilinmektedir. Bu modeller konuşmacıların farklı aksanları, konuşma hızı, arka plandaki gürültü veya diğer sesler gibi zorlukları aşmak için öğrenme yeteneklerini kullanırlar. Proje, konuşma teknolojisi alanında videolardaki sesleri yazıya çevirerek aranılan kelimenin saniyesine götüren bir programın geliştirilmesini amaçlamaktadır. Bu program derin öğrenme modellerinin kullanımıyla konuşmanın anlaşılır metinlerle dönüştürülmesini sağlayarak kullanıcıların videoda arama yapmasını sağlamaktadır.

Bir konuşma tanıma modeli Şekil 2.1.'de yer almaktadır. Şekil 2.1. adım adım işlenecek olursa aşağıdaki gibidir;

- Konuşma tanıma sinyallerinin gösterimini elde etmek için sinyal işleme,
- Gösterilen anahtar elemanın belirlemek ve fazla bilgiyi çıkarmak amacıyla özellik çıkarma,
- Kelime tespiti çıkarmak için zaman düzenleme ve model karşılaştırma algoritmaları,
- Bir final kelime dizisi seçmek için dil modeli (Yalçın, N., 2008)



Şekil 2.1. Konuşma Tanıma Modeli (Yalçın, N., 2008)

2.3. Konuşmadan Metne Yöntemi ile ilgili yapılan çalışmalar

Konuşmayı metne dönüştürme hesaplamalı dil bilim yoluyla konuşulan dilin tanınmasını ve metne çevrilmesini sağlayan bir konuşma tanıma yazılımıdır. Konuşmayı metne dönüştürme yöntemi herhangi bir cihazda sesi dinleyerek sesin düzenlenebilir bir biçimde kelimesi kelimesine bir metin dökümünü sunar. Yazılım bu işlemi ses tanıma özelliği yoluyla gerçekleştirir. STT yöntemi, dil bilimsel algoritmalarından yararlanarak konuşulan sözcüklere ait ses sinyallerini ayırt eder ve bu sinyalleri Unicode karakterler kullanarak metne aktarır. Konuşmanın metne dönüştürülmesi birkaç adım içeren karmaşık bir makine öğrenimi modeli yoluyla gerçekleştirilir. Adım adım incelenecek olursa:

Konuşma sırasında sözcükler ağzımızdan çıkarken bir dizi titreşim meydana gelir. Konuşmayı metne dönüştürme teknolojisi bu titreşimleri algılayıp bir analog-dijital dönüştürücü yoluyla dijital dile çevirir. Analog-dijital dönüştürücü bir ses dosyasındaki sesleri alır, dalgaları ayrıntılı olarak ölçüp filtreleyerek ilgili sesleri ayırt eder. Sesler daha sonra saniyenin yüzde veya binde biri küçüklüğünde segmentlere ayrılır ve ardından fonemlerle eşleştirilir. Fonem herhangi bir dilde bir sözcüğü başka bir sözcükten ayırt eden bir ses birimidir. Örneğin; İngilizcede yaklaşık 40 adet fonem bulunmaktadır. Fonemler daha sonra bunları yaygın olarak bilinen cümleler, sözcükler ve ifadelerle karşılaştıran bir matematiksel model yoluyla bir ağ üzerinde çalıştırılır. Ortaya çıkan sonuç metin olarak ya da ses dosyasına en yakın versiyona bağlı olarak bir bilgisayar dosyası formatında sunulur. Konuşmadan metne dönüştürme, genellikle sözlü

iletişimi yazılı forma aktarmak amacıyla kullanılan bir işlemdir. Bu işlem bir dizi farklı amaç için kullanılabilir ve birçok alanda çalışmalar yapılmıştır. Bazı kullanım alanları ve örnekleri aşağıda yer almaktadır:

Transkripsiyon: Araştırma, röportaj, seminer veya ders notları gibi konuşmaların yazılı kayıtlarının oluşturulmasında yer almaktadır. Örneğin bir üniversite öğrencisi derse katılırken not almak yerine dersin ses kaydını transkribe edebilir.

Dijital İşleme ve Analiz: Konuşma metinleri, bilgisayarlar tarafından işlenebilir hale getirilmektedir. Konuşma metinlerinin STT yöntemiyle metin formatına dönüştürülmesi duygu analizi, konuşma tanıma ve diğer yapay zekâ uygulamaları için önemlidir.

Engelli Bireyler için Erişilebilirlik: İşitme veya görme engelli bireyler için konuşma ve sesli içeriğin metne dönüştürülmesi onların bilgiye erişimini kolaylaştırır. Örneğin bir işitme engelli kişi konuşma metni olarak sunulan bir podcast'i okuyabilecektir.

Yasal ve Hukuki Belgeleme: Mahkeme kayıtları, ifadeler, röportajlar gibi hukuki belgelerin metin formatına dönüştürülmesi hukuk alanında önemlidir. STT yöntemi ile dava hazırlığı, belgeleme ve raporlama süreçleri kolaylaştırılabilir.

Eğitim ve Eğitim Materyalleri: Derslerin, konferansların veya eğitim materyallerinin transkriptlerinin oluşturulması eğitim sektöründeki önemli kullanım alanlarından. STT yöntemi, öğrencilerin ders içeriğine daha kolay erişmesini sağlar ve öğrenme sürecini destekler.

STT yöntemi ile ilgili yapılan ilgili çalışmalar aşağıda yer almaktadır:

(Thimmaraja, Y. G. ve ark., 2017) tarafından gerçekleştirilen çalışmada, İngilizce ASR sistemi oluşturmak için Kaldi konuşma tanıma araç seti kullanılarak, Dil Modeli ve Akustik Model oluşturulmuştur. Test verileri Hint eyaleti olan Karnataka'daki çiftçilerden kontrolsüz ortamlarda toplanmış ve ASR modellerinin geliştirilmesi için kullanılmıştır. Toplanan konuşma verileri hazır bir araç kullanılarak metne dönüştürülmüştür. Geliştirilen Otomatik Konuşma Tanıma sistemlerinin,

çiftçilere tarımsal emtia fiyatlarına ve hava durumu bilgilerine zamanında erişim sağlayacak sistemlerde kullanılabileceği çalışmada belirtilmiştir.

(Baniardalan, F. ve ark., 2017) tarafından gerçekleştirilen çalışmada, düşük bant konuşma işlemi ve akustik modelleme yöntemlerinden faydalanılarak Mel düşük bant konuşma tanıma işlemi önerilmiştir. Çalışmada konuşma verilerinin yanı sıra farklı çalışmalarla karşılaştırma yapabilmek amacıyla konuşma tanıma deneyleri sunulmuştur. Gürültülü ortamlarda farklı filtrelerin kullanılmasıyla elde edilen aurora2 veri kümesi kullanılmıştır. Eğitim sürecinde daha yüksek seviyede gürültü içeren Mel düşük bantlarına daha düşük ağırlıklar tahsis edilerek gürültü davranışı eğitilebileceği ve gürültü azaltmada daha iyi performans gösterileceği belirtilmiştir.

(Sel ve ark., 2011) tarafından gerçekleştirilen "Beyin Bilgisayar Arayüzleri İçin Türkçe Metinden Konuşma Sentezleme Sistemi" adlı çalışmada, konuşma veya hareket etme yeteneği olmayan engellilere yönelik bir Türkçe metinden konuşma sentezleme sistemi geliştirilmiştir. Bu sistemde girdi olarak alınan metin anlamsal olarak incelenmemiş ve bu nedenle Türkçeye özgü ses değişimlerinin kaliteli bir şekilde okunmadığı belirtilmiştir.

(Sasithon, P. S., 2017) tarafından gerçekleştirilen bir diğer çalışmada Tayland'ın kuzeydoğusunda konuşulan Isan dili için otomatik bir rakam konuşma tanıma sistemi sunulmuştur. Konuşma özelliklerini çıkarmak için MFCC ve konuşma tanıma için Saklı Markov modelleri kullanılmıştır. Bu çalışmada yerli konuşmacılardan elde edilen bir Isan rakamı (0-999) okunuş verisi toplanmıştır. Isan rakamlarını söyleyen konuşmacılar için izole edilmiş veriler ve sürekli konuşma tanıma görevi üzerine odaklanılmıştır. Sadece rakamlar üzerine yoğunlaştığından başarı oranları yüksek olduğu tespit edilmiştir. Geniş kelime haznesine sahip bir ASR sistemi ile karşılaştırılmasının mümkün olmadığı tespit edilmiştir.

(Karamahmet, T., 2015) tarafından gerçekleştirilen "RC8660 Ses Sentezleyici ile Türkçe Metinden Konuşma Sentezleme" adlı çalışmada, yeni bir harici sözlük tanımlanmış ve doğru heceleme için gerekli kurallar bu sözlüğe entegre edilmiştir. RC 8660 kartı, çıkış sesini kontrol edebilen bir özelliğe sahiptir. RC 8660 kartında ses seviyesi, hız, karakter aralığı, yankı, ton gibi faktörler istenilen düzeyde ayarlanabilmektedir. Geliştirilen sistemin engellilere yardımcı olabilmesi için diğer sistemlere (e-kitap okuyucuları, ATM makineleri, navigasyon sistemleri vs.) de uygulanması gerektiği belirtilmiştir.

(Ho, H., Mawardi, V. C. ve ark., 2017) tarafından yapılan çalışmada, Saklı Markov modelleri kullanıcının sağladığı sesi tanımak için kullanılmıştır. Soru cevaplama sistemi, ilgili belgeleri almak için Lucene arama motorundan vektör uzay modeli kullanmıştır. Bütün deneyler Endonezce üzerine yapılmıştır. Sistemi eğiten konuşma verileri tek bir kaynaktan alınmıştır. Sisteme giren toplam konuşma verileri 30 kelimelik kompozisyon olup her kelime 30 kez tekrarlanarak oluşturulmuştur. Konuşma eğitim verilerinin miktarının artırılması gerektiği belirtilmiştir. Ayrıca belge üzerinden alınacak cevabı görmek için sözdizimsel ve semantik yöntemlerin kullanılması sonuçları iyileştirebileceği belirtilmiştir.

(Aşlıyan, R. ve ark., 2018) tarafından yazılan "Türkçe Metinler İçin Hece Tabanlı Konuşma Sentezleme Sistemi" adlı makalede, metni konuşmaya çeviren bir sistem tasarlanmıştır. Geliştirilen sistemde en küçük ses birimi olarak heceler kullanılmıştır. Test edilen sistem ortalama bir puan almıştır. Seslendirilen metnin iyi anlaşıldığı ancak vurgu ve tonlama konularında iyileştirmeler yapılması gerektiği vurgulanmıştır.

(Oflazoglu, Ç. ve ark., 2011) tarafından gerçekleştirilen "Türkçe Duygusal Konuşma Veri tabanı" çalışmasında, 55 adet Türkçe dilinde filminden çıkartılan 5304 konuşma sinyali ve bu sinyallerin metinsel içerikleri yer almaktadır. Konuşma sinyalleri mutlu, nötr, şaşkın, üzgün, korku gibi duygulardan oluşmaktadır. Veri setleri, aktivasyon, baskınlık ve değerlik olmak üzere üç boyutlu duygu uzayından etiketlenerek oluşturulmuştur.

(Syadida, A. Q. ve ark., 2017) tarafından gerçekleştirilen bu çalışmada, açık kaynaklı Sphinx4 araç seti kullanılarak sürekli ASR için Endonezce üzerine bir uygulama geliştirilmiştir. Çalışmada 407 cümleden oluşan bir metin derlemi ve 4070 konuşma dosyası kullanılmıştır. Veri oluşturmak için 10 Endonezyalı konuşmacıdan veri toplanmıştır. Ancak çalışmada kullanılan eğitim verisi sürekli konuşma tanıma işleminde yetersiz kalmıştır, bu nedenle verilerin çoğaltılması gerekmektedir.

(Güldalı, K., 2009) tarafından yapılan "Türkçe Metin Seslendirme" çalışmasında, ses birimi olarak heceyi kullanan ve eklemeli sentezleme tekniği kullanan bir sistem geliştirilmiştir. Girdi metni alınarak boşluklar ve satır başlarından arındırılmıştır. Ardından kelimeler tespit edilip hecelerine ayrılmış ve bu heceler liste yapısında tutulmuştur. Her hece için veri tabanından ilgili ses dosyası getirilmiş ve eklemeli sentezleme algoritmasıyla seslendirme işlemi gerçekleştirilmiştir. Yapılan testlerde anlaşılabilirlik açısından ortalamanın üzerinde sonuçlar elde edilmiştir.

(Arık, G., 2011), “Görme Engelliler İçin Bilgisayar Kullanımının Etkinleştirilmesi, Erişilebilirlik ve Bir Türkçe Hece Tabanlı Konuşma Sentezleme Sisteminin Geliştirilmesi” isimli çalışmada, eklemeli sentezleme yöntemi kullanarak hece tabanlı bir metinden konuşma sentezleme sistemi geliştirmiştir. Yapılan testlerde kalitenin ortalamanın üzerinde olduğu saptanmıştır. Kullanıcı üzerinde yapılan testlerde ise ses kalitesinin daha iyi olması gerektiği, seslerin diksiyonu düzgün bir konuşmacı tarafından seslendirilmesi gerektiği ve vurgusal anlamda iyileştirmeler yapılması gerektiği sonuçları elde edilmiştir.

(Tekindal, B. ve ark., 2012) tarafından yapılan "Görme Engelliler İçin Türkçe Metinden Konuşma Sentezleme Yazılımı Geliştirilmesi" çalışmada eklemeli sentezleme yöntemini kullanan bir metin okuyucu yazılımı geliştirilmiştir. Bu yazılım elektronik ortamda bulunan bir dokümanı insan sesine yakın bir sesle konuşmaya çevirebilmektedir. Görme kaybı olan 10 kişi üzerinde yapılan testlerde vurgunun daha fazla iyileştirilmesi gerektiği ve aynı zamanda hece ses dosyalarının daha kaliteli bir ses kayıt ortamında yapılması gerektiği belirlenmiştir.

(Tong, R. ve ark., 2017) tarafından gerçekleştirilen çalışmada çocukların konuşmasını tanımak için yetişkinlerden elde edilen veriler kullanılmıştır. Önceden eğitilmiş yetişkin modeline akustik model uyarlaması yapılmıştır. Sistemin eğitimi için Mandarin yetişkin konuşma eğitim kümesi kullanılmıştır. Dört farklı cep telefonu ile 975 kişiden yaklaşık 360 bin farklı konuşma kaydı alınmıştır. Çocukların konuşmalarının işlenmesi büyük ölçekli çocuk konuşma verilerinin bulunmaması nedeniyle yetişkinler için konuşma tanıma görevinden daha zorlayıcı olmuştur.

(Polur, P.D. ve ark., 2001) yaptıkları uygulamada, Geri Yayılımlı Çok Katmanlı Algılayıcı ile İngilizce "yes (evet)" ve "no (hayır)" kelimelerinin tanınması üzerine odaklanmışlardır. “Evet” ve “hayır” kelimelerinin özniteliklerini çıkarmak için 13 adet MFCC kullanılmıştır. Sistem iki farklı kadın konuşmacının ses kayıtlarıyla test edilmiştir. Test sonuçlarına göre evet kelimesi için %75 ve hayır kelimesi için %90 oranında başarı elde edilmiştir.

(Tora ve ark., 2017) tarafından gerçekleştirilen "Implementation of Turkish Text-To-Speech Synthesis On A Voice Synthesizer Card With Prosodic Features" çalışmada, Türkçe metin okuma sentezi bir ses sentezleyici kart ile gerçekleştirilmiştir. Tasarlanan sistem daha anlaşılır ve doğal konuşma üretebilmek için prozodik özellikler uygulayarak zenginleştirilmiştir. Prozodik özellikler arasında perde, hız, hacim, gecikme ve ton bulunmaktadır. Anlaşılabilirlik testlerinde 5 üzerinden 3,29

puan elde edilmiştir. Geliştirilen sistem, konuşma teknolojisi ile ilgilenen araştırmacılar için bir test ortamı olarak kullanılabileceği belirtilmiştir.

(Azam, S.M. ve ark, 2007) tarafından gerçekleştirilen uygulamada ise Urdu dilinde seslendirilen rakamların tanınması amaçlanmıştır. Bu çalışmada Geri Yayılımlı ÇKA yapısı kullanılmıştır. Bu çalışmada her bir kelime için 39 defa MFCC hesaplanmıştır. Çalışmada bir konuşmacının 100 defa seslendirmesi oluşturulan bir veri tabanı kullanılmıştır. Bu çalışmada başarı oranının %98 olduğu belirtilmektedir.

(Yusuf, S.M. ve ark., 2021) tarafından yayımlanan çalışmada, konuşmadan duygu tanıma sistemlerindeki veri azlığı ve çevresel gürültüyü önemli sorunlar olarak belirtilmiştir. Bu sorunları çözmek için yeni birçok pencereli spektrogram modeli önerilmiştir. IEMOCAP adını verdikleri veri setini ikiye bölerek kullanmışlardır. İlk veri seti AHSN olarak isimlendirilmiş olup öfkeli, mutlu, üzgün ve nötr duygusal konuşma örneklerinden oluşmaktadır. İkinci veri seti AESN olarak isimlendirilmiş olup öfke, heyecan, üzgün ve nötr duygusal konuşma örneklerinden oluşmaktadır. Temiz konuşma örnekleri ile 15, 20 ve 40 desibellik gürültülü konuşma örnekleri üzerinde deneyler gerçekleştirilmiştir. İlk veri setinde Ağırlıklı Ortalama Duyarlılık değeri temiz konuşma örneklerinde %65,61, 40 desibel gürültülü örneklerde %65,90 olarak bulunmuştur. İkinci veri setinde ise Ağırlıklı Ortalama Duyarlılık değeri temiz konuşma örneklerinde %62,99, 40 desibel gürültülü örneklerde %62,81 olarak tespit edilmiştir.

(Erdem, E.C. ve ark, 2015) tarafından yayınlanan çalışmada, farklı dillerdeki filmlerden ve televizyon programlarından görsel-işitsel yüz video klipleri çıkarabilen yarı otomatik bir yöntem önerilmiştir. İlgili video kayıtlarındaki tepe karelerden elde edilen resimler ile “BAUM-2i” adını verdikleri bir veri seti oluşturulmuştur. Kayıtlar 5 ile 73 yaş aralığında 118 kadın ve 168 erkek olmak üzere toplam 286 aktörden elde edilmiştir. Türkçe ve İngilizce olmak üzere toplam 1047 kayıt bulunmaktadır. Bu kayıtlardan 248'i mutluluk, 173'ü öfke, 51'i öğrenme, 152'si sürpriz, 68'i korku, 49'u aşışılama ve 169'u nötr olmak üzere sekiz farklı duygudan oluşmaktadır.

Sesten metne çevirmenin hali hazırda kullanıldığı uygulamalar bulunmaktadır. Kullanılan uygulamalar şu şekildedir;

- **Transkriptor:** Otomatik olarak toplantıyı, videoyu ve ses kaydını yazıya çeviren bir uygulamadır.

- **Google Keyboard:** Google tarafından Android ve İOS cihazlar için programlanmış olan Keyboard, bir sanal klavye uygulaması olarak sunulmuş olsa da 100'den fazla dil aktif olarak ses kaydından yazıya çevrilebilmektedir.

- **Dictation.io:** Farklı dillerde kayıtları yazıya çevrilebilen Dictation.io, dönüştürülen metni kullanıcının cihazına kaydetmektedir.

- **Speechy:** Speechy, 80'den fazla dil seçeneğine sahip olan uygulama hem tablet ve bilgisayarda hem akıllı telefonlarda da kullanılmaktadır.

- **Just Press Record, SpeechTexter, Voice Notes** vb. programlar konuşma tanıma yöntemini kullanmak isteyen kişiler erişebileceği uygulamalardır (Gökhan, K., 2022).



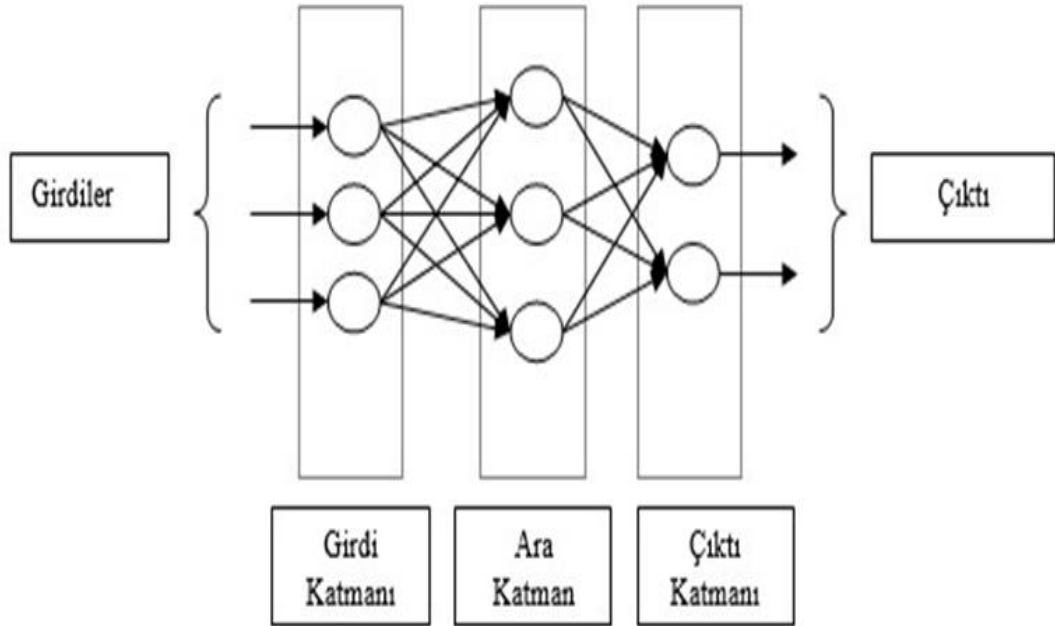
3. MATERYAL VE YÖNTEM

3.1. Yapay Sinir Ağları

Yapay sinir ağı, verileri insan beyninden esinlenerek bilgisayarlara işlemeyi öğreten bir yapay zekâ yöntemidir. Yapay Sinir Ağları, yapay zekâ uygulamalarında kullanılmak üzere hayvanların beyinlerinin bilgiyi işleme biçiminden esinlenilerek geliştirilen hesaplama sistemleridir (Ocak, Mahir E., 2020). Konuşma tanıma ve sesleri yazıya çevirme gibi görevlerde yapay sinir ağı kullanarak karmaşık ses sinyallerini analiz edilebilir ve elde edilen analizler sonucunda metinsel çıktılar elde edilebilmektedir. Bu yaklaşım geliştirilen sistemlerde kullanıcılardan biri olan konuşmacıların farklı dilleri, aksan farkı, hatalı söz veya kelime kullanımı gibi sorunları aşmak için etkili bir çözüm sunmaktadır.

İnsanların beyinlerinde birbirlerine sinapslarla (sinir hücreleri arasındaki iletişim noktaları) bağlı çok sayıda nöron (sinir hücresi) bulunur. Benzer biçimde yapay sinir ağı da yapay sinapslarla birbirine bağlanmış yapay nöronlardan oluşur. Yapay sinir ağındaki her bir nöron başka nöronlar tarafından kendisine iletilen bilgileri işler ve bulduğu sonucu başka nöronlara iletir (Ocak, Mahir E., 2020).

Aşağıdaki resimde Yapay Sinir Ağlarının katmanları olan Girdi, Ara ve Çıktı katmanı ve Yapay Sinir Ağı çalışma mantığı yer almaktadır.



Şekil 3.1. Yapay Sinir Ağlarının çalışma mantığı

Yapay Sinir Ağları üç ana katmanda incelenir;

Giriş Katmanı,

Ara (Gizliler) Katmanı,

Çıkış Katmanı

Bilgiler ağa girdi katmanından iletilir. Ara katmanlarda işlenerek oradan çıktı katmanına gönderirler. Bilgi işlemeden kasıt ağa gelen bilgilerin ağırlık değerleri kullanılarak çıktıya dönüştürülmesidir. Ağırlık girdiler için doğru çıktıları üretebilmesi için ağırlıkların doğru değerleri olması gerekmektedir.

Yapay Sinir Ağları insan beyninin en temel özelliği olan öğrenme fonksiyonunu gerçekleştiren bilgisayar sistemleridir. Bu ağlar birbirine bağlı yapay sinir hücrelerinden oluşmaktadır. Her bağlantının bir ağırlık değeri vardır. Yapay sinir ağlarının sahip olduğu bilgi bu ağırlık değerlerinde sahip olup ağa yayılmıştır (Öztemel, E. ,2003). CNN, RNN ve Derin Dil Modelleri (Deep Language Models) gibi yapay sinir yapıları, konuşma sinyallerini segmentlere ayırma, özellik çıkarımı yapma, dil modelleri oluşturma ve elde edilen sonuçları metinsel şekle dönüştürme gibi adımlarda etkili olabilmektedir (Hingston, G. ve ark, 2012). Yapay Sinir Ağları temelli konuşma tanıma ve sesleri yazıya çevirme sistemlerinin performansını artırmak için optimizasyon teknikleri de kullanılmaktadır. Veri artırma yöntemlerinin kullanılması ve ağ yapılarının ayarlanması, örüntü tanıma algoritmalarının optimize edilmesi gibi durumlar ile optimizasyon yöntemleri sistemin daha iyi genelleme yapmasını daha az hata ile doğru sonuçlar bulmasını sağlayabilmektedir.

Yapay sinir ağları temelli konuşma tanıma ve sesleri yazıya çevirme teknikleri birçok alanda uygulama potansiyeline sahiptir. Örneğin; çağrı merkezi sistemlerinde müşteri hizmetleri için otomatik konuşma tanıma ve metinleştirmenin kullanılması müşteri taleplerinin daha hızlı ve etkili bir şekilde yönetilmesini sağlamaktadır. Bunun yanı sıra konuşma tanıma ve sesleri yazıya çevirme teknolojisi, Google Sesli Asistan veya Apple Siri ses sistemi gibi birçok farklı uygulama alanına sahiptir. Çeşitli alanlarda uygulanmasıyla beraber yapay sinir ağları eğitilerek günümüzde hayatı kolaylaştıran birçok ürün ve proje hayata geçirilebilmektedir. Yapay sinir ağları, beyindeki sinir hücreleri ve sinir ağı yapılarını taklit eden bir modeldir ve karmaşık veri işleme görevlerinde etkili bir şekilde kullanılmaktadır. Konuşma tanıma için yapay sinir ağları ses verilerini analiz etmek, ses özelliklerini çıkarmak ve bu özelliklerin metinsel karşılıklarını belirlemek amacıyla kullanılan bir yöntemdir.

Yapay sinir ağıları, konuşma tanıma sistemleri transkript oluşturma, metin tabanlı arama ve dil modellemesi gibi birçok alanda kullanılmaktadır. İnsanlar yaklaşık 3000 yıl önce veri işleme çalışmalarını mağaraya resim çizerek bizlere o dönemden veriler hakkında bilgiler vererek başlamıştır ve günümüzde tüm verileri küçük bir depolama cihazlarına veya çiplerle veri işleme süreçleri devam etmektedir. Konuşma tanıma, insan sesini algılayarak onu metin ve komutlara dönüştüren önemli bir yapay zekâ uygulamasıdır. Yapay sinir ağıları, konuşma tanıma sistemlerinde etkili bir şekilde kullanılan ve çeşitli görevler için başarılı sonuçlar veren güçlü bir tekniktir. Yapay Sinir Ağları, bilgi işleme süreçlerinde doğrusal olmama, öğrenme, genelleme, uyarlanabilirlik yeteneklerine sahiptir.

3.2. Konuşmadan Metne Yöntemi

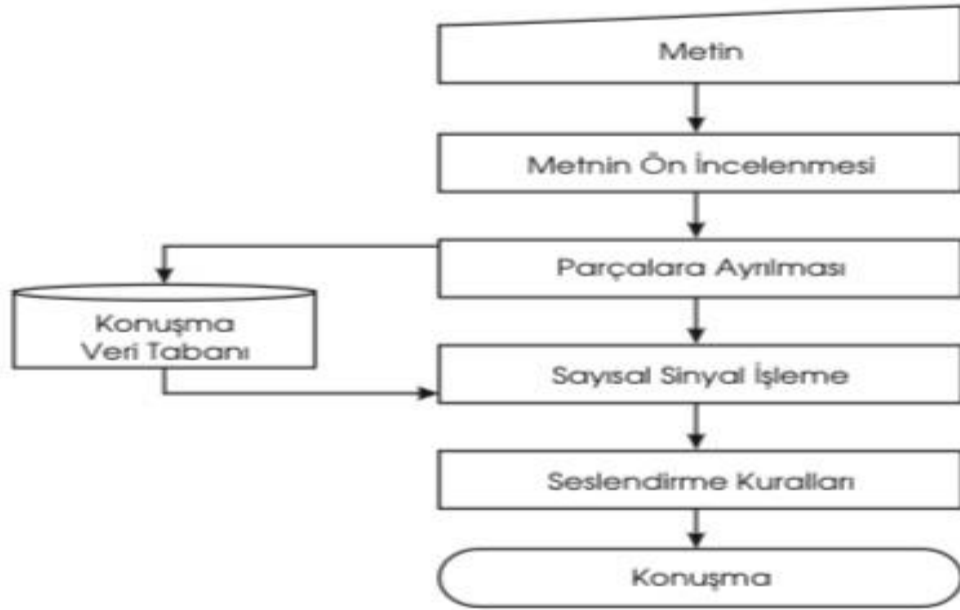
STT yöntemi konuşulan dili tanımlamak ve işlemek için kullanır. Konuşmayı sentezleme süreç iki adımdan oluşmaktadır. Bu adımlar sırasıyla, doğal dil işleme ve dijital sinyal işlemedir. Sesi yazıya çevirme teknolojisi sesi dinleyerek ve elde ettiği değerlerden metne dönüştürerek çalışır. Yöntem konuşulan sözcüklerden ses sinyallerini sıralamak ve bu sinyalleri karakterler kullanarak metne çevirmek için dil algoritmasını kullanır. Bir kişinin söylediği sözler bir dizi titreşim üretir. Yöntem bu titreşimleri alır ve bir analogdan dijital dönüştürücü kullanarak bunları dijital bir dile dönüştürür. Bu dönüştürücü, ses dosyasındaki sesleri alır ve dalgaları ayrıntılı olarak ölçer ve karşılık gelen sesleri ayırt etmek için bu veriler filtrelenir. Sesler daha sonra binlerce saniyeye bölünür ve fonemlerle (belirli bir dilde bir kelimeyi diğerlerinden ayıran ses birimleri) eşleştirilir. Bir sonraki adımda fonemler onları bilinen kelimeler ve cümlelerle karşılaştıran matematiksel bir model kullanarak ağdan geçirilir. Ardından metin, sesin en olası versiyonuna göre sunulmaktadır (Bulutistan, 2023).

3.3. Konuşma Tanıma Yönteminin Çalışma Mantığı

Konuşma tanıma yöntemi, projede kullanılan STT yönteminin ters aşamasıdır. Bir video/ses dosyasında bulunan tüm gürültüler, STT yöntemin kullanımıyla metne çevrilebilmektedir. SR yönteminde ise yazılı metin sese çevrilmektedir. SR yönteminin kütüphanesi kodlama aşamasında STT yönteminin kullanımı için projede yer almıştır. Bu yöntemin ilk aşaması yapı analizidir. Yapı analizi adımı yazılı metin işlenmektedir. Cümlelerin ve kelimelerin oluşturduğu gürültülerin nerede başladığı ve nerede bittiği belirlenmektedir. Burası noktalama işaretleri ve tarihlerin yazılış

formatının da işlendiği bölümdür. İkinci aşama metin ön işleme aşamasıdır. Bu aşamada dilin özel yapısı incelenir ve işlenir. Bu adımda metinde yer alan kısaltmalar konuşma diline adapte edilir. Üçüncü aşama yazılı metinden ses birimine çevrilmesi aşamasıdır. Burada her bir kelime ses birimlerine çevrilmeye başlar. Örneğin ‘mekân’ sözcüğü ‘m ek â n’ olarak birimlere ayrılır. Dördüncü aşama ölçü aşamasıdır. Bu aşamada cümle yapısı, kelime vurguları, duraklatacak yerler, cümleler için en uygun ölçülerin bulunması gerçekleşir. Beşinci aşama dalga üretimi aşamasıdır. Bu aşama son adımdır. Burada fonem ve ölçü bilgisi dikkate alınarak bir birleştirme gerçekleştirilir. Bu birleştirme birden fazla yöntem ile yapılır. Bu yöntemlerden biri kayıtlı insan sesi ile birleştirme, diğeri ise sinyal işleme tekniğidir. En yaygın kullanılan sinyal işleme tekniğidir. Bu teknik ses birimleri sinyaller halinde işlenir, birbirlerine uyumlu sinyaller birleştirilir (G., Şevval ,2022).

Tüm bu aşamalar Şekil 3.2.’de yer almaktadır.



Şekil 3.2. Speech Recognition (Konusma Tanıma) Yönteminin çalışma mantığı (G., Şevval, 2022)

SR yönteminin ters aşaması STT yöntemi verileri metne çevirmektedir.

STT, genel olarak iki ana aşamaya ayrılabilir.

- **Ses Ön İşleme:** Bu aşamada sesli girdinin gürültü ve diğer bozulmalardan temizlenmesi gerekir.

- **Kelime tanıma:** Bu aşamada sesli girdideki kelimeler, bir kelime darağacı ve bir olasılık modeli kullanılarak belirlenir.

STT, Doğal Dil İşleme (NLP) alanındaki en zor tekniklerinden biridir. Bunun nedeni sesli girdinin doğası gereği belirsiz ve karmaşık olmasıdır. Sesli girdide aynı ses, farklı kelimelere karşılık gelebilir. Örneğin; “cat” ve “hat” kelimeleri aynı şekilde söylenebilir. Ayrıca sesli girdi, gürültü, aksanlar ve diğer faktörler nedeniyle bozulabilir. Video platformlarında kullanıcıların oluşturduğu videolarda telaffuz şekilleri oluşturulan videonun dili ve o dile ait lehçelerden dolayı bazı kelimeler aynı anlama gelebilmektedir. STT teknikleri son yıllarda önemli gelişmeler kaydetmiştir. Bu ilerlemeler daha güçlü bilgisayar donanımları ve daha gelişmiş algoritmalar sayesinde olmuştur. STT, insan konuşmasını bir analogdan dijital bir forma çeviren yapay zekâ ile geliştirilmiş bir yöntemdir. Ayrıca toplanan verilerin dijital formu bir metin formatına dönüştürülür. Konuşma teknolojisi, şu adlarla da anılır;

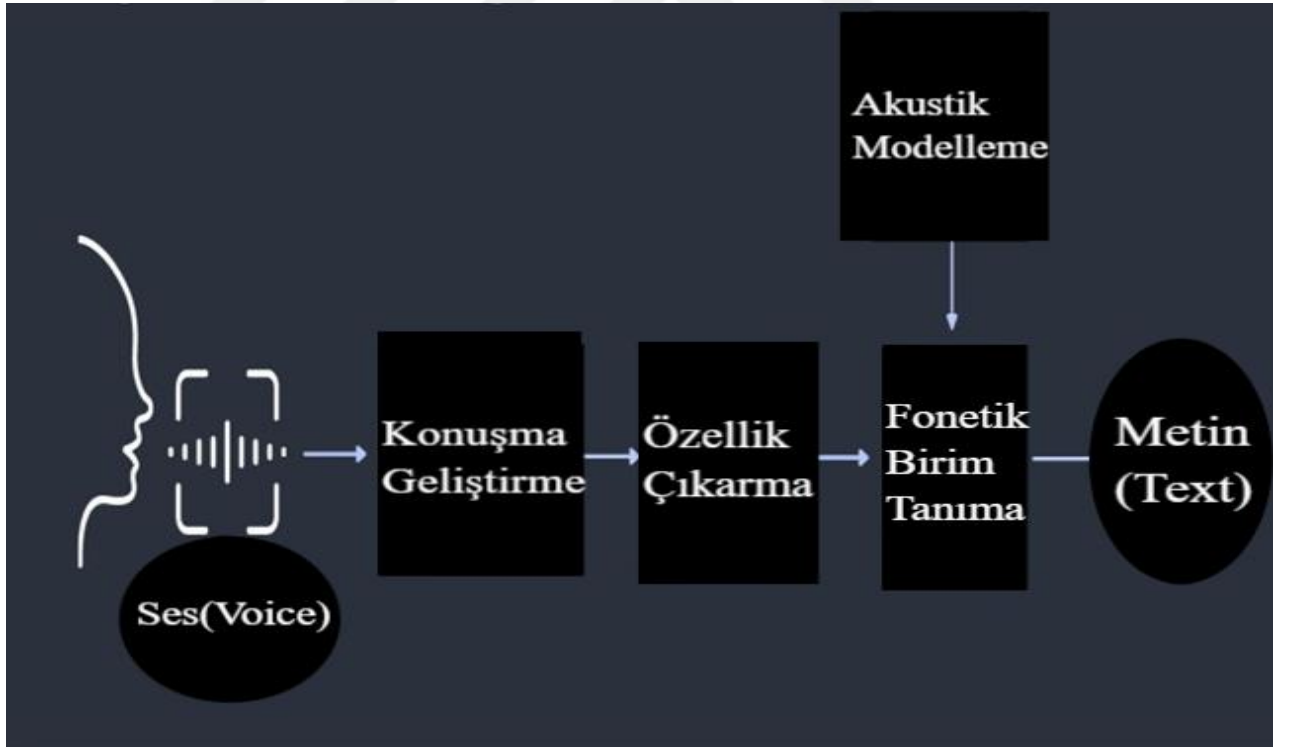
- Otomatik Konuşma Tanıma (ASR),
- Konuşma tanıma,
- Bilgisayar konuşma tanıma,
- Ses transkripsiyon,
- Ekran okuma

Sesten metne çeviri yazılımının çalışması karmaşıktır ve birden çok adımın uygulanmasını içerir. Süreç aşağıdaki gibidir;

Başlangıçta bir analogdan dijitale dönüştürücü kullanan bir bilgisayar programı, titreşimleri işitsel sinyallerden ayırt etmek için sağlanan verilere dilsel algoritmalar uygular. Daha sonra ses dalgaları ölçülerek ilgili sesler filtrelenir. Ayrıca sesler yüzdeler veya binde saniyeler halinde dağıtılır, daha sonra bölümlere ayrılır ve fonemlerle eşleştirilir. Fonemler ayrıca mevcut verileri iyi bilen kelimeler, cümleler ve ifadelerle karşılaştırmak için matematiksel bir modelden geçirilir. Çıktı bir metin veya bilgisayar tabanlı ses dosyasındadır (Shaip, 2022). Yöntem, program oluştururken test aşamasında tespit edilen geliştirmeler sonucu IT sektörünün temel amacı olan doğruluk oranı artırımı ve amaca uygun çalışılabilirlik prensibe uygun bir şekilde geliştirilmiştir. Yöntem geliştirilmeden denendiğinde video paylaşım platformlarında bulunan videolara 120 saniyelik (2 dakikalık) çevirebildiği daha sonra hata verdiği tespit edilmiştir. Yöntem, bölümlendirme yöntemi ve frekans değerlerine göre sözcük / saniye eşleştirmesini daha doğru yapabilmesi için geliştirilmiştir. Gürültü oldukça frekans değerleri çevrilip kaydedilmektedir. Gürültü olmadığında o saniyeden itibaren chunk (yığın)’da o andaki bölüm(section) kapanmaktadır. Yeni bir gürültü duyulduğunda, bir sonraki bölüm açılmaktadır. Yöntem frekans değerlerine göre ayrılmadığı zaman bölme yönteminde

bir kelimenin baştan tekrar kaydedilebildiğini hem veri tabanında sütun fazlalığı hem de sözcük/saniye eşleştirmesinde hatalar olabileceği tespit edilerek bölme yöntemi ve frekans/gürültü seviyelerine göre değerler veri tabanına kaydedilmiş ve doğruluk oranı artırılmıştır. Video süresi boyunca oluşan bölümler (sections), bir araya gelerek chunk (yığın)’ı oluşturmaktadır. Oluşan yığınlar bölümleme yönteminin çalışma prensibini temsil etmektedir. Konuşma tanıma teknolojisinin birçok faydası olabildiği gibi aksaklıklarda yaşanabilmektedir. Aksan ve lehçeler üzerinde kontrol kazanma, konuşulan cümlelerin bağlamını anlama, giriş kalitesini yükseltmek için arka plan seslerinin ayarlanması, etkin işleme için kodu farklı dillere çevirme, video dosyaları durumunda konuşmada kullanılan görsel ipuçlarının analiz edilmesi durumlarından oluşabilmektedir. Konuşma tanıma ile ilgili en büyük zorluk çıktısını %100 doğru bir şekilde oluşturmaktır.

Şekil 3.3.’de STT yönteminin çalışma mantığı yer almaktadır. Ses veya gürültü, Konuşma Geliştirme yöntemiyle sesli ifadeyi iyileştirmek ve elde edilen verileri optimize etmek için gerçekleştirilmektedir.



Şekil 3.3. STT yöntemi çalışma mantığı

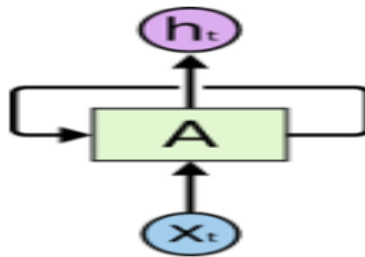
3.4. Derin Öğrenme Mimarileri

Derin öğrenme yöntemleri de konuşma tanıma alanında büyük bir ilerleme sağlamıştır. (Hinton, G. ve ark, 2012) tarafından, derin inanç ağı kullanarak akustik modelleme konusunda çalışmalarını sunulmuştur. Derin öğrenme mimarileri kullanılarak elde edilen sonuçlar, RNN ve LSTM Ağlarının konuşma tanıma performansını artırmada önemli bir rol oynadığını göstermektedir. Microsoft gibi büyük teknoloji şirketleri de derin öğrenme tekniklerini konuşma araştırmalarında aktif olarak kullanmaktadır.

3.4.1. Tekrarlayan Sinir Ağları (RNN)

RNN, model olarak doğal öğrenme yöntemlerinden tecrübe ile öğrenmeyi temel almaktadır. İnsanlar her adımda öğrenmeye yeniden başlamazlar her adımda eski tecrübelerinden yararlanarak öğrenmeye devam ederler. Ancak geleneksel yapay sinir ağlarında insanlarda bulunan bu tecrübe ile anlamlandırma özneteliği bulunmaz ve bu onların en büyük eksikliğidir (Ateş, E. 2019). RNN, girdi verilerini ve önceki adımlarda hesaplaması yapılmış bilgileri kullanarak bir zaman verisi veya sıralı veri içindeki desenleri ve ilişkileri modellemek için kullanılır. RNN’de her bir adım işleminin bir önceki çıktıya bağlandığı çalışma mimarisine dayandırılmıştır. Her birime ait bilgilerin kullanılabilmesi özellikle görüntü üzerinde gerçekleştirilen işlemlerde yüksek doğruluk ortaya koymaktadır (Somuncu, E. ve ark., 2022).

Şekil 3.4.’de basit tekil tekrarlayan sinir ağ görüntülenmektedir. A ismi verilen dikdörtgen bir yapay sinir ağındaki düğümü temsil etmektedir. Ağın girdi değeri X_t ’dir. Yapay sinir ağının çıktı değeri h_t ’dir. Düğümünden değerlendirme sonucu çıkan değer yine kendisine dönerek döngüsel bir eğitim modeli oluşturmaktadır. Bu döngü ile geçmiş verilerde kullanılabildiğinden yeni bilgi ve eski bilgi harmanlanarak bir sınıflandırma yapılabilmektedir.

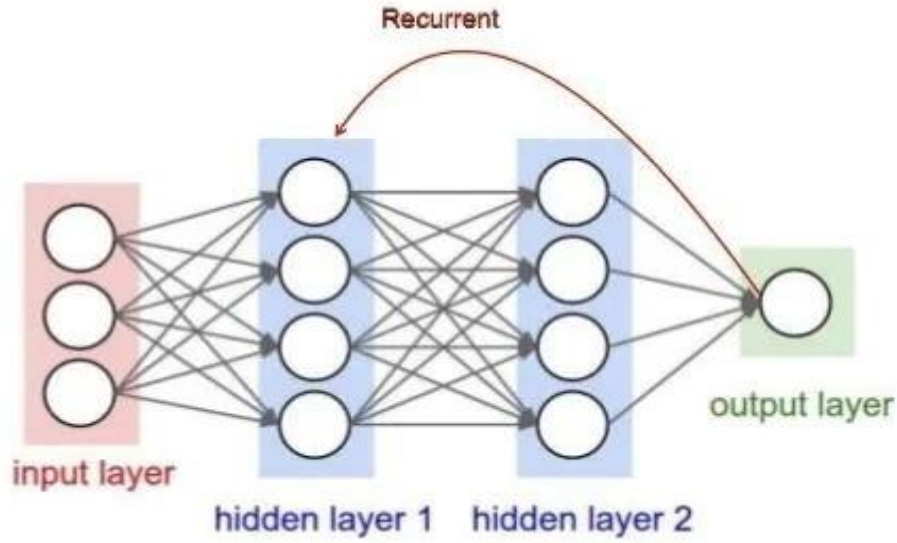


Şekil 3.4. Tekil tekrarlayan sinir ağ gösterimi (Olah, 2015)

İlk olarak Basit Tekrarlayan Ağ (Simple Recurrent Network-SRN), Jeff Elman tarafından tasarlanmıştır. Elman’ın cümle yapısı simülasyonunda kullanılan her bir

kelime için gizli kalıpların üzerinde ortalama örüntü kümeleme sonucunda isim ve fiil kategorileri temiz şekilde ayrılmıştır. Bunun yanında isimler arasında canlı-cansız ayrımları hatta insan-hayvan, hayvanlar arasında avcı-yırtıcı gibi kümeler de ayrılmıştır (Şeker, A. ve ark., 2017). Sıralı veri analizi ve işlemesi için tasarlanmış bir yapay sinir ağı türüdür. Metin analizi, zaman serileri tahmini, konuşma tanıma, el yazısı tanıma ve diğer sıralı veri işleme görevlerinde başarıyla kullanılmaktadır.

Şekil 3.5.'de Tekrarlayan Sinir Ağlarının yapısı görünmektedir.



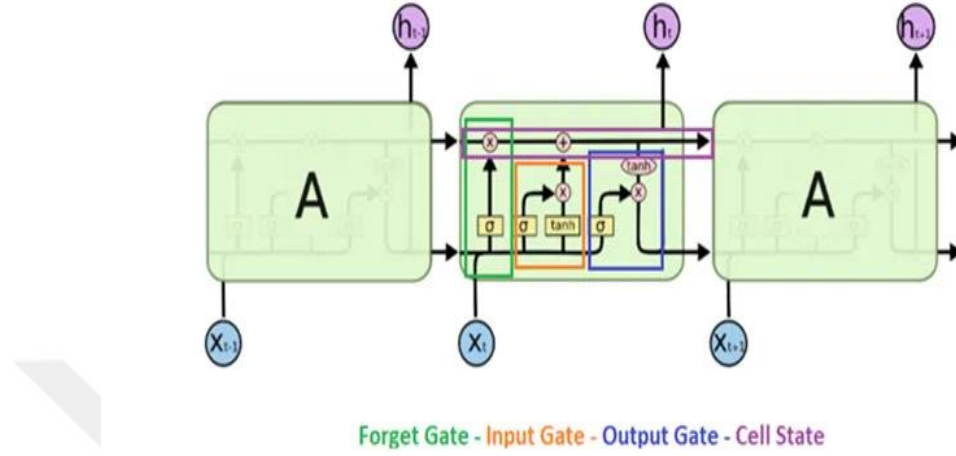
Şekil 3.5. Tekrarlayan Sinir Ağları yapısı

3.4.2. Uzun Kısa Vadeli Hafıza Ağları (LSTM Ağları)

RNN mimarilerinde önceki bilgi kullanımına dayalı bir yaklaşım vardır. Örneğin; “Ağaç toprakta yetişir” cümlesinde “toprak” kelimesini tahmin etmek kolaydır fakat bağlamlar arası boşluk arttığında RNN modelin geçmişten gelen bir bilgiyi kullanması oldukça zordur (Şeker, A. ve ark., 2017). Özellikle Doğal Dil İşleme (NLP), metin analizi, konuşma tanıma, zaman serisi tahmini ve diğer sıralı veri işleme görevlerinde büyük bir başarı elde etmiştir.

Uzun Kısa Vadeli Bellek (LSTM), özellikle zaman serileri verileri gibi uzun aralıklar boyunca bağımlılıkları modellemek için tasarlanmış bir yapay sinir ağı mimarisidir. LSTM’ler geleneksel RNN temel sorunlarından biri olan uzun vadeli bağımlılıkları ele alma yeteneği ile ortaya çıkar. LSTM, geçit (gate) olarak adlandırılan yapılar tarafından düzenlenmiştir. Geçitler, isteğe bağlı olarak bilginin iletilmesini sağlayan bir yoldur. Standart ileri beslemeli sinir ağlarının aksine LSTM’nin geri bildirim bağlantıları vardır.

Şekil 3.6.'de de görüldüğü gibi bir LSTM hücresinde veriyi güncellemek veya unutmak amacıyla sigmoid ve tanh fonksiyonlarından yararlanır (T., Doruk, 2022).

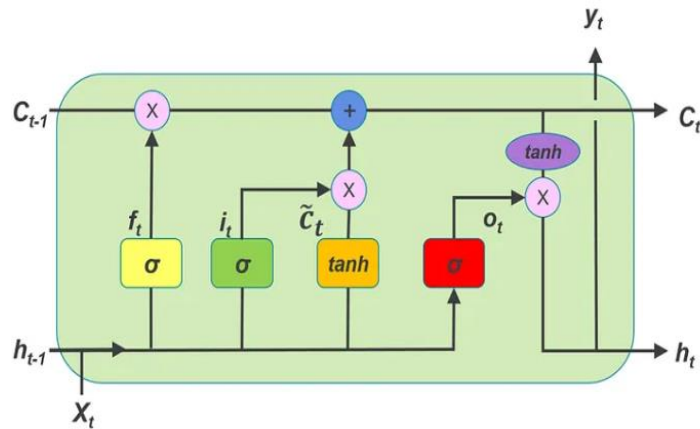


Şekil 3.6. LSTM ağırları (Doruk, T., 2022)

Sigmoid Fonksiyonu: Doğrusal olmayan bir aktivasyon fonksiyonudur. Sigmoid fonksiyonu verileri 0 ve 1 arasına sıkıştırır. Fonksiyon çıktısı 0 olursa bilgi unutulmuş kabul edilir. Benzer şekilde 1 ile sonuçlanırsa bilgi korunur.

Tanh fonksiyonu: Tanh fonksiyonu sigmoid fonksiyonu gibi doğrusal olmayan bir fonksiyondur ancak tanh fonksiyonun aralığı sigmoid'den farklı olarak $(-1,1)$ 'dir (T., Doruk, 2022).

Bir LSTM hücresi Şekil 3.7.'de de görüldüğü gibi unutma kapısı, giriş kapısı ve çıkış kapısından oluşmaktadır.



Şekil 3.7. Unutma Kapısı, Giriş Kapısı, Çıkış Kapısı (T., Doruk, 2022)

- **Unutma Kapısı (Forget Gate):** Unutma kapısı, hücre durumundan hangi bilgilerin atılacağını belirler. Bu sayede gereksiz veya uzun vadeli bağlantı kurulmuş bilgileri unutulabilmektedir.

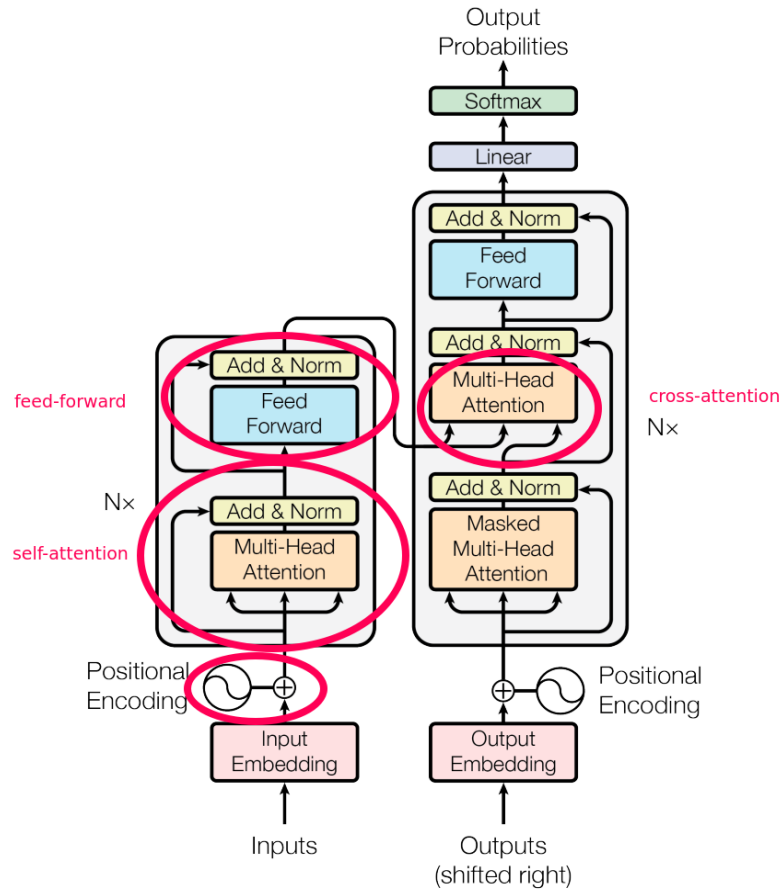
- **Giriş Kapısı (Input Gate):** Hücre durumuna ne kadar bilgi ekleyeceğini kontrol eder. Bu sayede gereksiz veya uzun vadeli bağlantı kurulmuş bilgileri unutulabilmektedir.

- **Çıkış Kapısı:** Hücre durumunu temel alarak ağı ne kadar bilgi üreteceğini belirlemektedir.

3.4.3. Dikkat Mekanizması (Self Attention Mechanism)

Attention (dikkat) kelime anlamı olarak bir veya birkaç şeye odaklanırken geri kalan şeylerin ihmal edilmesi anlamına gelir. Bir sinir ağı insan beyni hareketlerinin sadeleştirmiş şekilde çalıştırılması üzerine temellendirilmiştir. Attention mekanizması da sinir ağına benzer şekilde benzer olanlara odaklanarak geri kalanların filtrelenmesi yani yok sayılması temeline dayanmaktadır (Atıcı, A., 2021). Self-Attention mekanizması, özellikle “Transformer” adlı bir temel öğrenme modelinin temel bir bileşeni olarak öne çıkmıştır. Transformer modeli, özellikle doğal dil işleme görevlerinde (metin çevirisi, konuşmadan metne, duygu analizi, dil modelleme vb.) büyük başarılı sonuçlar elde etmiştir ve bu başarının ardında Self-Attention mekanizması önemli bir rol oynamaktadır. Self-Attention mekanizması bir giriş dizisindeki her bir ögenin diğer tüm ögelerle olan ilişkilerini dikkate alan bir yapıdır. Bu her bir ögenin diğer ögelerle olan bağlantılarını daha iyi anlamasını ve modelin genel bağlamı daha etkili bir şekilde kavramasını sağlamaktadır. Self Attention mekanizması, metin işlemede kullanılan bir mimaridir. Bu mekanizma, metin içindeki düzenlilikleri ve bağlamları daha iyi anlamaya yaramaktadır. Self-Attention, metin içindeki her bir dizinin metin içindeki diğer tüm dizinlerle nasıl etkileşimde olduğunu hesaplar. Bu etkileşimler, metnin anlamının daha iyi anlaşılmasına yardımcı olmaktadır (Deniz, Y., 2023).

Şekil 3.8.’de bir transformer’in self attention mekanizması yer almaktadır. Şekil 3.8.’de görüldüğü gibi kodlayıcı modülü mavi renkte, kod çözücü ise yeşil renktedir. Ayrıca hem kodlayıcı ve kod çözücü modellerinin Multi-Head Attention adı verilen bir katman kullandığını da görebilmekteyiz.



Şekil 3.8. Bir Transformer'ın Self Attention Mekanizması(Vaswani, A. ve ark., 2023)

3.5. Fourier Dönüşümü

Fourier Dönüşümü, bir fonksiyonun zaman veya uzay alanındaki temsilini frekans alanındaki temsiliyle değiştiren önemli bir matematiksel işlemidir. Bu dönüşüm, orijinal fonksiyonun frekans bileşenlerini açığa çıkarır ve böylece bir sinyalin frekans bileşenlerini analiz etmek için kullanılır. Temel olarak bir fonksiyonun Fourier dönüşümü, o fonksiyonun bir serisinin toplamı olarak ifade edilir. Özellikle karmaşık sinyallerin işlenmesinde ve analizinde oldukça önemli olan bu dönüşüm ses işleme alanında da yaygın olarak kullanılmaktadır.

Fourier dönüşümü genellikle aşağıdaki formülle ifade edilir:

$$F(\omega) = \int_{-\infty}^{\infty} f(t) \cdot e^{-j\omega t} dt$$

$F(\omega)$, fonksiyonun Fourier dönüşümünü temsil eder.

$f(t)$, orijinal fonksiyondur.

ω , açısal frekanstır.

J , kompleks sayılardır.

T , zaman değişkenidir.

STT yönteminin kullanımının yaygınlaşmasıyla birlikte karmaşık kelimelerin doğru bir şekilde ayırt edilmesi önemlidir. Bu yöntemin etkinliğini artırmak için özellikle frekans alanında doğru analizler yapılmalıdır. Bu bağlamda Fourier Dönüşümü önemli bir yöntem olarak öne çıkmaktadır. Projede, STT yöntemi kullanılırken karşılaşılan sözcük tekrarı ve zaman farklılığı gibi problemleri çözmek için Fourier Dönüşümü önemli bir rol oynamıştır. Bu dönüşüm ses sinyallerinin frekans bileşenlerini analiz ederek, gürültüyü ve farklı ses özelliklerini belirlememize yardımcı olur. Bu analiz sonucunda gürültülerin frekans değerleri belirlenmiş ve bu bilgi bölüme izin verilen süre içinde doğru bir şekilde eklenmiştir.

Fourier Dönüşümü, karmaşık bir matematiksel işlem olmasına rağmen ses işleme gibi alanlarda sıklıkla kullanılan bir tekniktir. Projede bu dönüşüm sayesinde geliştirilen STT yöntemi, sözcük/saniye uyumunu daha başarılı bir şekilde sağlamış ve doğruluk oranını arttırmıştır. Ayrıca Fourier Dönüşümü videolardaki ses sinyallerini analiz ederek, belirli bir kelimenin veya ifadenin hangi saniyelerde geçtiğini belirlememize yardımcı olur. Böylece kullanıcılar, aradıkları içeriğe hızlı ve etkili bir şekilde erişebilmektedirler.

Projede MFCC algoritması ile Fourier Dönüşümü kullanılarak elde edilen frekans bilgileri STT yönteminin doğruluğunu ve verimliliğini artırmaya yönelik olarak kullanılmıştır. Özellikle gürültüyü ve farklı ses özelliklerini belirlemek için frekans analizleri gerçekleştirilmiştir. Bu analizler sonucunda sözcük/saniye uyumu ve doğruluk oranı artırılmıştır. Fourier Dönüşümü, STT yöntemiyle yapılan bu çalışmada temel bir rol oynamış ve karmaşık ses verilerinin analiz edilmesinde ve işlenmesinde büyük önem taşımıştır.

3.6. Mel Frekans Kepstral Sayıları Algoritması

MFCC algoritması, ilk kez Davis ve Mermelstein tarafından 1980 yılında duyurulmuştur. Duyurulduktan sonra MFCC, en çok tercih edilen ses işleme algoritmalarından biri haline gelmiştir (Davis, S.B., 1980). MFCC kullanılarak yapılan çalışmalarda yüksek tanıma oranları elde edildiği tespit edilmiştir.

MFCC algoritmasının blok diyagramı Şekil 3.9.'da gösterilmiştir. Konuşma girişi 10000 Hz'den yüksek bir frekansta örneklendirilmelidir. Bu frekans eşliğinin

seçilmesinin sebebi düşük frekanslarda ve çözünürlükte yapılan örneklemede yeterli nokta örneklenemeyeceği için eğimin değiştiği yerlerde sapmalar olacak ve gerçek ses işaretinden farklı bir işaret örneklenmiş olacaktır. Bu örneklenen işaretler, 5 kHz'ye kadar bütün frekansları yakalayacaktır. 5 kHz, insan sesinin aşağı yukarı bütün enerji seviyelerini içermektedir (Demirci, D.M., 2005).

MFCC, insan kulaklarının davranışlarını taklit ederken aynı zamanda ses dalga yapısından çok daha az etkilenir. Bu özelliği sayesinde, MFCC değişimlerden daha az etkilenir ve daha kararlı sonuçlar elde edilir.



Şekil 3.9. Mel Frekans Kepstral Sayıları algoritmasının diyagramı

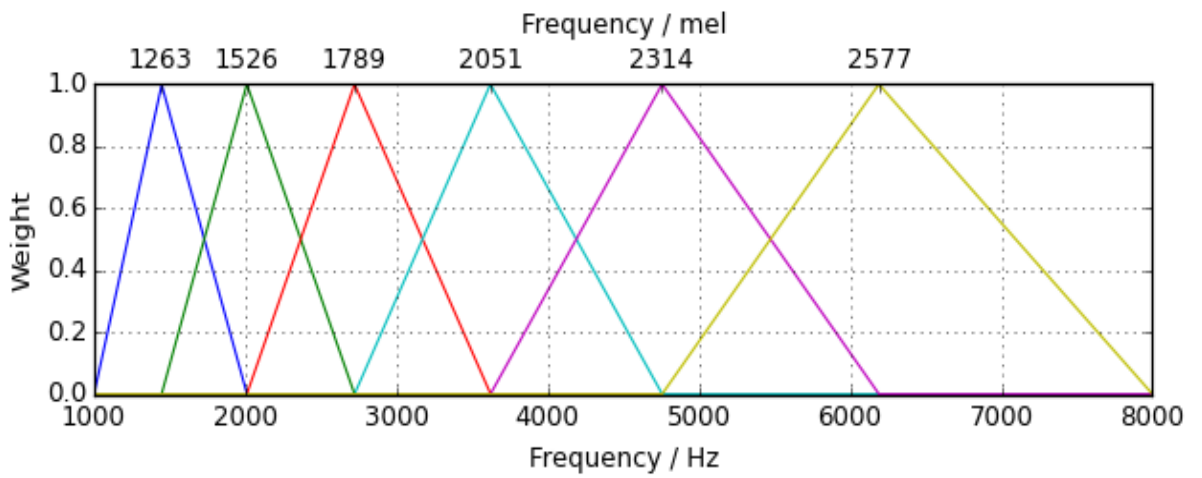
MFCC adımlarından sırasıyla;

Çerçeveleme adımında, ses sinyali küçük zaman aralıklarına bölünür. Bu zaman aralıklarına çerçeve denir. Bu adımda konuşma sürekli bir işaret, N adet örnekten oluşan çerçevelere ve komşu M örnekten oluşan çerçevelere bölünür ($M < N$). Örneğin; ilk çerçeve N örnekten oluşurken, ikinci çerçeve ilk çerçeveden M örnek sonra başlar ve ilk çerçevenin $N-M$ örneğine kadar üzerine binmektedir. Bu adım, konuşma sinyalini daha küçük ve işlenebilir parçalara bölmektedir. Bu adım ses sinyalini işlemek için önemlidir çünkü bu sayede konuşma sinyali daha küçük parçalara ayrılarak daha rahat analiz edilebilir. Bu parçalara ayrılma işlemi sinyali zamansal olarak böler ve her bir çerçeveyi ayrı ayrı işleyerek konuşma tanıma, ses sentezi, ses sınıflandırma ve diğer ses işleme görevlerini gerçekleştirmek için kullanılabilir. Bu çerçeveler genellikle belirli bir süre aralığında alınır ve sinyalin zamansal özelliklerini daha iyi temsil etmeye yardımcı olmaktadır. Her çerçeve, frekans bileşenlerine ayrılmak üzere Fourier dönüşümüne tabi tutulmaktadır.

Pencereleme aşamasında ise her çerçeve pencereleme işlemine tabi tutulmaktadır. Bu işlem, çerçevelerin başı ve sonundaki süresiz bölümleri minimize etmeyi amaçlamaktadır. Pencereleme çerçevelerin başı ve sonundaki bilgi içermeyen bölümleri kırparak spektral bozulmayı önlemektedir (Demirci, D.M., 2005).

Mel frekansına çevirme adımı insanın frekans içeriklerini algılamasının lineer olmadığı psikofiziksel çalışmalarla ortaya konmuştur. Bu nedenle ses frekansları mel ölçeğine dönüştürülür. Mel frekansı 1000 Hz'in altında lineer, 1000 Hz'in üstünde ise logaritmik bir dağılıma sahiptir. Mel ölçeği insanın duyma eşiğinin 1000 Hz'lik bir ses için 40 dB üstünde olduğu bir noktadan başlayarak tanımlanmaktadır.

Mel ölçeğinde, belirli bir mel frekansı aralığına sahip olan filtreler kullanılarak ses spektrumu bölünür. Bu filtre bankası, insanın duyma yeteneğini daha doğru bir şekilde temsil eden bir frekans spektrumunu yansıtmak için tasarlanmıştır. Bu işlem sesin temsilini insan algısına daha uygun hale getirmektedir



Şekil 3.10. Mel Filtre Bankı

Mel Filtre bankı, üçgen bant geçiren ve sabit aralıklarda bant genişliğine sahip olan bir mel frekansı aralığına bağlı bir bankadır. Bu bankanın yayılışı ve bant genişliği sabit bir mel frekansı aralığına göre belirlenir. Filtre bankası frekans alanında uygulandığı için işaretin FFT ile frekans alanına dönüştürülmüş hali üzerine uygulanır. Her bir filtre, frekans alanında bir histogram kutusu oluşturur ve bu filtre bankasının birbirine binen kutularından oluşan bir seti temsil eder.

Kepstrum aşamasında logaritması alınmış mel spektrumunun tekrar zaman alanına çevrilmesi gerekmektedir. Bu işlem sonucunda elde edilen veriye MFCC denir. Her bir çerçevenin analizinde kepsral gösterim işaretin bölgesel spektral özelliklerini güzel bir şekilde temsil eder. Mel Spektrum Katsayıları ve onların logaritması gerçek sayılar olduğu için, bu verileri zaman alanına çevirmek için DCT kullanılmaktadır.

Programda STT yönteminin geliştirilmesinde büyük bir veri seti kullanılmıştır. Test aşamalarında tespit edilen geliştirmeler sonucunda projede MFCC algoritması kullanılmıştır. Sesten metne çevrim sürecinde veri önce ses özelliklerine dönüştürülür. Bu özellikler genellikle MFCC algoritması kullanılarak çıkarılmaktadır. İlgili özellikler, sesin frekans bileşenlerini temsil eden sayısal değerlerdir.

3.7. Google Speech to Text yöntemi nasıl çalışır?

Google Speech-to-Text yönteminin çalışma mantığı derin öğrenme temellidir. Yöntemin çalışma mantığı;

Google, büyük bir veri seti üzerinde eğitilmiş konuşma tanıma sistemine sahiptir. Veri önce ses özelliklerine dönüştürülür. Bu özellikler MFCC gibi teknikler kullanılarak çıkarılır. İlgili özellikler sesin frekans bileşenlerini temsil eden sayısal değerlerdir. Derin öğrenme modeli genellikle özel olarak tasarlanmış bir sinir ağı içerir. Örneğin; Rekürsif Sinir Ağları (RNN), uzun kısa vadeli bellek (LSTM) veya dikkat mekanizması (self attention mechanism) gibi yapılar kullanılabilir. Bu mimari, ses özelliklerini metne dönüştürme görevini öğrenmektedir. Eğitim sürecinde model veri seti üzerinde iteratif olarak çalıştırılır. Ses özellikleri ve eğitim verisi arasındaki ilişkileri anlamak için model parametreleri ayarlanır. Bu süreçte model dil yapılarını, konuşma desenlerini ve aksanları öğrenir. Kullanıcı konuştuğunda ses verisi önce özelliklere dönüştürülür. Bu ses özellikleri daha sonra eğitilmiş model tarafından analiz edilir. Model ses özelliklerini kullanarak metni oluşturmaya çalışır. Model konuşma girişini analiz ettikten sonra bir metin çıktısı üretir. Bu metin, kullanıcının söylediklerini temsil eder. Google'ın bu alandaki hizmetleri genellikle çok büyük ölçekli ve geniş veri setleri üzerinde eğitildiği için genel olarak yüksek doğruluk seviyelerine sahiptir.

3.8.Bölümleme Yöntemi

Bölümleme yöntemi, program geliştirme aşamasında kullanılan ST yönteminin izin verdiği 120 saniyelik süreç dahilinde geliştirilmiştir. STT yöntemi sestten metne çevrim sürecinde video süresi fark etmeksizin 120 saniye süresince destek verebiliyorken, 120 saniyeden büyük videolarda yöntem kodlama aşamasında hata vermektedir. Programın temel amacı olan uzun süreli izleme sürelerinden tasarruf sağlayarak kullanıcının istediği zamana erişmesi olduğundan kullanıcıların programı uzun süreli videoları kullanması için bölümleme yöntemi geliştirilmiştir.

İlk aşamada videolar 120 saniyelik bölümlere ayrılmıştır. Test aşamalarında videolar 120 saniyelik bölümlere ayrıldığında yeni bölüm oluşturulurken kelime tekrarları yapıldığı ve veri tabanına aynı değerde kelime/sözcük kayıtları olduğu tespit edilmiştir. Ayrıca bu bölme yönteminde videolarda herkesin sesi ve konuşma hızına göre ayarlama yapılmasına rağmen hatalar alınabileceği tespit edilmiştir.

Elde edilen sonuca göre bölümleme yöntemi için video ayrıştırma işlemi 0,7 saniye üzerindeki sessizlikler arasında video bölümü yaparak cümle cümle kayıt satırları oluşturulmuştur. Gürültü oldukça frekans değerleri çevrilip kaydedilmektedir. Gürültü olmadığında o saniyeden itibaren chunk (yığın)'da o andaki bölüm (section) kapanmaktadır. Yeni bir gürültü duyulduğunda, bir sonraki bölüm açılmaktadır. Yöntem, frekans değerlerine göre ayrılmadığı zaman bölme yönteminde bir kelimenin baştan tekrar kaydedilebildiğini hem veri tabanında sütun fazlalığı hem de sözcük/saniye eşleştirmesinde hatalar olabileceği tespit edilerek bölme yöntemi ve MFCC algoritması ile Fourier Dönüşümü kullanılarak frekans/gürültü seviyelerine göre değerler veri tabanına kaydedilmiş ve doğruluk oranı artırılmıştır. Video süresi boyunca oluşan bölümler (sections), bir araya gelerek chunk (yığın)'ı oluşturmaktadır. Oluşan yığınlar bölümleme yönteminin çalışma prensibini temsil etmektedir.

4. ARAŞTIRMA SONUÇLARI VE TARTIŞMA

4.1. STT Yöntemi ve Geliştirilme Süreci

Sesi yazıya çevirme teknolojisi, sesi dinleyerek ve elde ettiği değerlerden metne dönüştürerek çalışır. Kullanılan veri önce ses özelliklerine dönüştürülmektedir. Bu özellikler MFCC tekniği kullanılarak çıkartılmıştır. Bu projede kullanılan STT yöntemi elde ettiği değerlerde bazen konuşma hataları lehçe farklılıklarından dolayı hatalı çevrimler yapabilmektedir. Ayrıca yöntemin yaklaşık olarak videonun ilk 120 saniyesinde doğru çalıştığı ancak sonrasında hata verebileceği gözlemlenmiştir. Bu tespitlerin ardından özellikle bazı videoların uzun süreli olabileceği örneğin; bir videonun süresinin 300 dakika veya daha uzun olabileceği ihtimali düşünülerek, yöntemin sorunsuz çalışabileceği saniye aralıklarını belirlemek için bir döngüye alınmıştır. Program kullanıldıkça uzun süreli videolarda STT yönteminin izin verdiği süre olan 120 saniyede her bir yenilemede bazı sözcükler arasında süre dolduğundan yeni bölümde o saniyeden tekrar başladığı tespit edilmiştir. Kullanılan STT yönteminin uzun süreli videolarda bölümlene yöntemi kullanıldığı zaman izin verilen süre bittikten sonra tekrar yeni bölme oluşturduğunda sözcük/saniye eşleştirmesi hatası olabileceği tespit edilmiştir. MFCC algoritması ile Fourier yönteminin kullanılmasıyla gürültüye göre elde edilen frekans değerleri ve bu değerlere göre kullanılan STT yönteminde kullanılan bölümlene yöntemiyle beraber hem doğru saniye/sözcük eşleştirmesi yapıldığı hem de sözcük değerlerinin videolardaki saniyelere göre daha doğru çevrildiği tespit edilmiştir. Yöntemin projeye uygun bir şekilde başarılı entegrasi ve ihtiyaç dahilinde özelliklerle geliştirilmesiyle beraber aşağıda deneyler aşamasında da görüleceği gibi YouTube otomatik çeviri sistemine göre daha yüksek doğruluk oranlarında metin çevrimi yaptığı görülmektedir, bu başarının uygulanan bölümlene yöntemi ve MFCC algoritmasıyla Fourier dönüşümünün uygulanmasından kaynaklanmaktadır.

4.2. Uygulamanın Karşılaştırılması

YouTube, Google'a ait bir çevrimiçi video paylaşım ve sosyal medya platformudur. Kullanıcılar oluşturulan videolara YouTube'un otomatik çeviri sistemi ile altyazı ekleyebilmektedirler. YouTube'un otomatik çeviri sistemi ve geliştirilen program arasındaki doğruluk oranları için 40 adet video Çizelge 4.1.'de karşılaştırılmıştır.

Çizelge 4.1. YouTube ve Geliştirilen Program karşılaştırması

Video	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6	Video 7	Video 8	Video 9	Video 10
YouTube	%87,5	%82,8	%90	%80	%75	%93,06	%90,83	%97,5	%92,85	%96,41
Program	%97,5	%90	%90	%93,3	%93,18	%92,65	%94,16	%95	%96,41	%96,41

Video	Video 11	Video 12	Video 13	Video 14	Video 15	Video 16	Video 17	Video 18	Video 19	Video 20
YouTube	%95,88	%89,58	%95,83	%91,66	%93,75	%100	%85,93	%90	%84,37	%85,71
Program	%97,64	%95,83	%97,61	%97,22	%100	%100	%95,31	%95	%93,75	%96,42

Video	Video 21	Video 22	Video 23	Video 24	Video 25	Video 26	Video 27	Video 28	Video 29	Video 30
YouTube	%89,28	%92,30	%82,50	%93,33	%95,45	%100	%96,67	%97,27	%100	%97,91
Program	%82,14	%100	%92,50	%100	%93,63	%100	%93,33	%97,27	%100	%95,83

Video	Video 31	Video 32	Video 33	Video 34	Video 35	Video 36	Video 37	Video 38	Video 39	Video 40
YouTube	%95,45	%97,65	%94,79	%99,29	%95,55	%92,85	%100	%96,15	%96,87	%90
Program	%93,63	%96,09	%97,91	%96,45	%96,66	%92,85	%96,66	%96,15	%100	%96,66

Video 1:

Geliştirilen program: 40 kelimenin 39'unu doğru çevirdi (%97,5 doğruluk).

YouTube otomatik çeviri: 40 kelimenin 35'ini doğru çevirdi (%87,5 doğruluk).

Video 2:

Geliştirilen program: 70 kelimenin 63'ünü doğru çevirdi (%90 doğruluk).

YouTube otomatik çeviri: 70 kelimenin 58'ini doğru çevirdi (%82,8 doğruluk).

Video 3:

Her iki sistemde 100 kelimenin 90'ını doğru çevirdi (%90 doğruluk).

Video 4:

Geliştirilen program: 75 kelimenin 70'ini doğru çevirdi (%93,3 doğruluk).

YouTube otomatik çeviri: 75 kelimenin 60'ını doğru çevirdi (%80 doğruluk).

Video 5:

Geliştirilen program: 44 kelimenin 41'ini doğru çevirdi (%93,18 doğruluk).

YouTube otomatik çeviri: 44 kelimenin 33'ünü doğru çevirdi (%75 doğruluk).

Video 6:

Geliştirilen program: 245 kelimenin 227'sini doğru çevirdi (%92,65 doğruluk).

YouTube otomatik çeviri: 245 kelimenin 228'ini doğru çevirdi (%93,06 doğruluk).

Video 7:

Geliştirilen program: 120 kelimenin 113'ünü doğru çevirdi (%94,16 doğruluk).

YouTube otomatik çeviri: 120 kelimenin 109'unu doğru çevirdi (%90,83 doğruluk).

Video 8:

Geliştirilen program: 80 kelimenin 76'sını doğru çevirdi (%95 doğruluk).

YouTube otomatik çeviri: 80 kelimenin 78'ini doğru çevirdi (%97,5 doğruluk).

Video 9:

Geliştirilen program: 28 kelimenin 27'sini doğru çevirdi (%96,41 doğruluk).

YouTube otomatik çeviri: 28 kelimenin 26'sını doğru çevirdi (%92,85 doğruluk).

Video 10:

Her iki sistemde 28 kelimenin 27'sini doğru çevirdi (%96,41 doğruluk).

Video 11:

Geliştirilen program: 170 kelimenin 166'sını doğru çevirdi (%97,64 doğruluk).

YouTube otomatik çeviri: 170 kelimenin 163'ünü doğru çevirdi (%95,88 doğruluk).

Video 12:

Geliştirilen program: 48 kelimenin 46'sını doğru çevirdi (%95,83 doğruluk).

YouTube otomatik çeviri: 48 kelimenin 43'ünü doğru çevirdi (%89,58 doğruluk).

Video 13:

Geliştirilen program: 48 kelimenin 47'sini doğru çevirdi (%97,61 doğruluk).

YouTube otomatik çeviri: 48 kelimenin 46'sını doğru çevirdi (%95,83 doğruluk).

Video 14:

Geliştirilen program: 36 kelimenin 35'ini doğru çevirdi (%97,22 doğruluk).

YouTube otomatik çeviri: 36 kelimenin 33'ünü doğru çevirdi (%91,66 doğruluk).

Video 15:

Geliştirilen program: 32 kelimenin tamamını doğru çevirdi (%100 doğruluk).

YouTube otomatik çeviri: 32 kelimenin 30'unu doğru çevirdi (%93,75 doğruluk).

Video 16:

Her iki sistemde 100 kelimenin tamamını doğru çevirdi (%100 doğruluk).

Video 17:

Geliştirilen program: 64 kelimenin 61'ini doğru çevirdi (%95,31 doğruluk).

YouTube otomatik çeviri: 64 kelimenin 55'ini doğru çevirdi (%85,93 doğruluk).

Video 18:

Geliştirilen program: 40 kelimenin 38'ini doğru çevirdi (%95 doğruluk).

YouTube otomatik çeviri: 40 kelimenin 36'sını doğru çevirdi (%90 doğruluk).

Video 19:

Geliştirilen program: 32 kelimenin 30'unu doğru çevirdi (%93,75 doğruluk).

YouTube otomatik çeviri: 32 kelimenin 27'sini doğru çevirdi (%84,37 doğruluk).

Video 20:

Geliştirilen program: 28 kelimenin 27'sini doğru çevirdi (%96,42 doğruluk).

YouTube otomatik çeviri: 28 kelimenin 24'ünü doğru çevirdi (%85,71 doğruluk).

Video 21:

Geliştirilen program: 44 kelimenin 23'ünü doğru çevirdi (%82,14 doğruluk).

YouTube otomatik çeviri: 44 kelimenin 25'ini doğru çevirdi (%89,28 doğruluk).

Video 22:

Geliştirilen program: 26 kelimenin tamamını doğru çevirdi (%100 doğruluk).

YouTube otomatik çeviri: 26 kelimenin 24'ünü doğru çevirdi (%92,30 doğruluk).

Video 23:

Geliştirilen program: 40 kelimenin 37'sini doğru çevirdi (%92,50 doğruluk).

YouTube otomatik çeviri: 40 kelimenin 33'ünü doğru çevirdi (%82,50 doğruluk).

Video 24:

Geliştirilen program: 30 kelimenin tamamını doğru çevirdi (%100 doğruluk).

YouTube otomatik çeviri: 30 kelimenin 28'ini doğru çevirdi (%93,33 doğruluk).

Video 25:

Geliştirilen program: 110 kelimenin 103'ünü doğru çevirdi (%93,63 doğruluk).

YouTube otomatik çeviri: 110 kelimenin 105'ini doğru çevirdi (%95,45 doğruluk).

Video 26:

Her iki sistemde 33 kelimenin hepsini doğru çevirdi (%100 doğruluk).

Video 27:

Geliştirilen program: 30 kelimenin 28'ini doğru çevirdi (%93,33 doğruluk).

YouTube otomatik çeviri: 30 kelimenin 29'unu doğru çevirdi (%96,67 doğruluk).

Video 28:

Her iki sistemde 44 kelimenin 43'ünü doğru çevirdi (%97,27 doğruluk)

Video 29:

Her iki sistemde 40 kelimenin hepsini doğru çevirdi (%100 doğruluk).

Video 30:

Geliştirilen program: 48 kelimenin 46 âdetini doğru çevirdi (%95,83 doğruluk).

YouTube otomatik çeviri: 48 kelimenin 47'sini doğru çevirdi (%97,91 doğruluk).

Video 31:

Geliştirilen program: 110 kelimenin 103'ünü doğru çevirdi (%93,63 doğruluk).

YouTube otomatik çeviri: 110 kelimenin 105'ini doğru çevirdi (%95,45 doğruluk).

Video 32:

Geliştirilen program: 128 kelimenin 123'ünü doğru çevirdi (%96,09 doğruluk).

YouTube otomatik çeviri: 128 kelimenin 125'ini doğru çevirdi (%97,65 doğruluk).

Video 33:

Geliştirilen program: 96 kelimenin 94'ünü doğru çevirdi (%97,91 doğruluk).

YouTube otomatik çeviri: 96 kelimenin 91'ini doğru çevirdi (%94,79 doğruluk).

Video 34:

Geliştirilen program: 141 kelimenin 136'sını doğru çevirdi (%96,45 doğruluk).

YouTube otomatik çeviri: 141 kelimenin 140'ını doğru çevirdi (%99,29 doğruluk).

Video 35:

Geliştirilen program: 90 kelimenin 87'sini doğru çevirdi (%96,66 doğruluk).

YouTube otomatik çeviri: 90 kelimenin 86'sını doğru çevirdi (%95,55 doğruluk).

Video 36:

Her iki sistemde 70 kelimenin 65 adetini doğru çevirdi (%92,85 doğruluk).

Video 37:

Geliştirilen program: 30 kelimenin 29'unu doğru çevirdi (%96,66 doğruluk).

YouTube otomatik çeviri: 30 kelimenin tamamını doğru çevirdi (%100 doğruluk).

Video 38:

Her iki sistemde kelimenin 50'sini doğru çevirdi (%96,15 doğruluk).

Video 39:

Geliştirilen program: 64 kelimenin tamamını doğru çevirdi (%100 doğruluk).

YouTube otomatik çeviri: 64 kelimenin 62'sini doğru çevirdi (%96,87 doğruluk).

Video 40:

Geliştirilen program: 30 kelimenin 29'unu doğru çevirdi (%96,66 doğruluk).

YouTube otomatik çeviri: 30 kelimenin 27'sini doğru çevirdi (%90 doğruluk).

YouTube'un otomatik çeviri sistemi ve geliştirilen program arasındaki doğruluk oranları için 40 adet video Çizelge 4.2.'de karşılaştırılmıştır.

Çizelge 4.2. YouTube ve Geliştirilen Program karşılaştırması

Video	Video 41	Video 42	Video 43	Video 44	Video 45	Video 46	Video 47	Video 48	Video 49	Video 50
YouTube	%95	%96,66	%94,42	%98,38	%94,44	%95,18	%96,92	%85,33	%88	%90
Program	%90	%96,66	%92,85	%98,82	%93,33	%97,27	%96,15	%94,66	%96	%96
Video	Video 51	Video 52	Video 53	Video 54	Video 55	Video 56	Video 57	Video 58	Video 59	Video 60
YouTube	%98,46	%98,88	%97,10	%97,27	%98,66	%94,54	%97,89	%98,85	%97,59	%98,18
Program	%100	%100	%96,53	%99,09	%98,66	%96,36	%98,94	%97,14	%96,98	%94,54
Video	Video 61	Video 62	Video 63	Video 64	Video 65	Video 66	Video 67	Video 68	Video 69	Video 70
YouTube	%88,88	%92,94	%95	%96	%100	%96	%96,42	%96,55	%98,46	%98,87
Program	%100	%96,47	%97,5	%96	%99,23	%92	%97,85	%96,55	%98,46	%97,75
Video	Video 71	Video 72	Video 73	Video 74	Video 75	Video 76	Video 77	Video 78	Video 79	Video 80
YouTube	%100	%100	%100	%99,42	%99,31	%96,92	%97,66	%95,94	%100	%100
Program	%100	%96,34	%100	%100	%96,55	%92,30	%96,42	%91,89	%100	%100

Video 41:

Geliştirilen program: 40 kelimenin 36'sını doğru çevirdi (%90 doğruluk).

YouTube otomatik çeviri: 40 kelimenin 38'ini doğru çevirdi (%95 doğruluk).

Video 42:

Her iki sistemde 30 kelimenin 29'unu doğru çevirdi (%96,66 doğruluk).

Video 43

Geliştirilen program: 70 kelimenin 64'ünü doğru çevirdi (%92,85 doğruluk).

YouTube otomatik çeviri: 70 kelimenin 65'ini doğru çevirdi (%94,42 doğruluk).

Video 44:

Geliştirilen program: 186 kelimenin 184'ünü doğru çevirdi (%98,92 doğruluk).

YouTube otomatik çeviri: 186 kelimenin 183'ünü doğru çevirdi (%98,38 doğruluk).

Video 45:

Geliştirilen program: 90 kelimenin 84'ünü doğru çevirdi (%93,33 doğruluk).

YouTube otomatik çeviri: 90 kelimenin 85'ini doğru çevirdi (%94,44 doğruluk).

Video 46:

Geliştirilen program: 110 kelimenin 107'sini doğru çevirdi (%97,27 doğruluk).

YouTube otomatik çeviri: 110 kelimenin 108'ini doğru çevirdi (%95,18 doğruluk).

Video 47:

Geliştirilen program: 130 kelimenin 125'ini doğru çevirdi (%96,15 doğruluk).

YouTube otomatik çeviri: 130 kelimenin 126'sını doğru çevirdi (%96,92 doğruluk).

Video 48:

Geliştirilen program: 75 kelimenin 71'ini doğru çevirdi (%94,66 doğruluk).

YouTube otomatik çeviri: 75 kelimenin 61'ini doğru çevirdi (%85,33 doğruluk).

Video 49:

Geliştirilen program: 25 kelimenin 24'ünü doğru çevirdi (%96 doğruluk).

YouTube otomatik çeviri: 25 kelimenin 22'sini doğru çevirdi (%88 doğruluk).

Video 50:

Geliştirilen program: 100 kelimenin 96'sını doğru çevirdi (%96 doğruluk).

YouTube otomatik çeviri: 100 kelimenin 90'ını doğru çevirdi (%90 doğruluk).

Video 51:

Geliştirilen program: 65 kelimenin 64'ünü doğru çevirdi (%98,46 doğruluk).

YouTube otomatik çeviri: 65 kelimenin tamamını doğru çevirdi (%100 doğruluk).

Video 52:

Geliştirilen program: 180 kelimenin 180'ini doğru çevirdi (%100 doğruluk).

YouTube otomatik çeviri: 180 kelimenin 179 tanesini doğru çevirdi (%98,88 doğruluk).

Video 53:

Geliştirilen program: 173 kelimenin 167'sini doğru çevirdi (%96,53 doğruluk).

YouTube otomatik çeviri: 173 kelimenin 168'ini doğru çevirdi (%97,10 doğruluk).

Video 54:

Geliştirilen program: 110 kelimenin 109'unu doğru çevirdi (%99,09 doğruluk).

YouTube otomatik çeviri: 110 kelimenin 107'sini doğru çevirdi (%97,27 doğruluk).

Video 55:

Her iki sistemde 75 kelimenin 74'ünü doğru çevirdi (%98,66 doğruluk).

Video 56:

Geliştirilen uygulama: 55 kelimenin 53'ünü doğru çevirdi (%96,36 doğruluk).

YouTube otomatik çeviri: 55 kelimenin 52'sini doğru çevirdi (%94,54 doğruluk).

Video 57:

Geliştirilen program: 190 kelimenin 188'ini doğru çevirdi (%98,94 doğruluk).

YouTube otomatik çeviri: 190 kelimenin 186'sını doğru çevirdi (%97,89 doğruluk).

Video 58:

Geliştirilen program: 175 kelimenin 170'sini doğru çevirdi (%97,14 doğruluk).

YouTube otomatik çeviri: 175 kelimenin 173'ünü doğru çevirdi (%98,85 doğruluk).

Video 59:

Geliştirilen uygulama: 166 kelimenin 161'ini doğru çevirdi (%96,98 doğruluk).

YouTube otomatik çeviri: 166 kelimenin 162'sini doğru çevirdi (%97,59 doğruluk).

Video 60:

Geliştirilen program: 55 kelimenin 52'sini doğru çevirdi (%94,54 doğruluk).

YouTube otomatik çeviri: 55 kelimenin 54'ünü doğru çevirdi (%98,18 doğruluk).

Video 61:

Geliştirilen program: 9 kelimenin tamamını doğru çevirdi (%100 doğruluk).

YouTube otomatik çeviri: 9 kelimenin 8'ini doğru çevirdi (%88,88 doğruluk).

Video 62:

Geliştirilen program: 85 kelimenin 82'sini doğru çevirdi (%96,47 doğruluk).

YouTube otomatik çeviri: 85 kelimenin 79'unu doğru çevirdi (%92,94 doğruluk).

Video 63:

Geliştirilen program: 80 kelimenin 78'ini doğru çevirdi (%97,5 doğruluk).

YouTube otomatik çeviri: 80 kelimenin 76'sını doğru çevirdi (%95 doğruluk).

Video 64:

Her iki sistemde 100 kelimenin 96'sını doğru çevirdi (%96 doğruluk).

Video 65:

Geliştirilen program: 130 kelimenin 129'unu doğru çevirdi (%99,23 doğruluk).

YouTube otomatik çeviri: 130 kelimenin tamamını doğru çevirdi (%100 doğruluk).

Video 66:

Geliştirilen program: 50 kelimenin 46'sını doğru çevirdi (%92 doğruluk).

YouTube otomatik çeviri: 50 kelimenin 48'ini doğru çevirdi (%96 doğruluk).

Video 67:

Geliştirilen program: 140 kelimenin 137'sini doğru çevirdi (%97,85 doğruluk).

YouTube otomatik çeviri: 140 kelimenin 135'ini doğru çevirdi (%96,42 doğruluk).

Video 68:

Her iki sistemde 145 kelimenin 140 adetini doğru çevirdi (%96,55 doğruluk)

Video 69:

Her iki sistemde 65 kelimenin 64'ünü doğru çevirdi (%98,46 doğruluk).

Video 70:

Geliştirilen program: 89 kelimenin 87'sini doğru çevirdi (%97,75 doğruluk).

YouTube otomatik çeviri: 89 kelimenin 88'ini doğru çevirdi (%98,87 doğruluk).

Video 71:

Her iki sistemde 65 kelimenin tamamını doğru çevirdi (%100 doğruluk).

Video 72:

Geliştirilen program: 82 kelimenin 79'unu doğru çevirdi (%96,34 doğruluk).

YouTube otomatik çeviri: 82 kelimenin tamamını doğru çevirdi (%100 doğruluk).

Video 73:

Her iki sistemde 55 kelimenin tamamını doğru çevirdi (%100 doğruluk)

Video 74:

Geliştirilen program: 175 kelimenin tamamını doğru çevirdi (%100 doğruluk).

YouTube otomatik çeviri: 175 kelimenin 174'ünü doğru çevirdi (%99,42 doğruluk).

Video 75:

Geliştirilen program: 145 kelimenin 140'ını doğru çevirdi (%96,55 doğruluk).

YouTube otomatik çeviri: 145 kelimenin 144'ünü doğru çevirdi (%99,31 doğruluk).

Video 76:

Geliştirilen program: 65 kelimenin 60'ını doğru çevirdi (%92,30 doğruluk).

YouTube otomatik çeviri: 65 kelimenin 63'ünü doğru çevirdi (%96,92 doğruluk).

Video 77:

Geliştirilen program: 84 kelimenin 81'ini doğru çevirdi (%96,42 doğruluk).

YouTube otomatik çeviri: 84 kelimenin 82 tanesini doğru çevirdi (%97,66 doğruluk)

Video 78:

Geliştirilen program: 74 kelimenin 68'ini doğru çevirdi (%91,89 doğruluk).

YouTube otomatik çeviri: 74 kelimenin 71'ini doğru çevirdi (%95,94 doğruluk).

Video 79:

Her iki sistemde 99 kelimenin tamamını doğru çevirdi (%100 doğruluk).

Video 80:

Her iki sistemde 63 kelimenin tamamını doğru çevirdi (%100 doğruluk).

YouTube'un otomatik çeviri sistemi ve geliştirilen program arasındaki doğruluk oranları için 20 adet video Çizelge 4.3.'de karşılaştırılmıştır.

Çizelge 4.3. YouTube ve Geliştirilen Program karşılaştırması

Video	Video 81	Video 82	Video 83	Video 84	Video 85	Video 86	Video 87	Video 88	Video 89	Video 90
YouTube	%97,39	%94,44	%99,17	%96,87	%100	%98,14	%97,69	%100	%98,63	%100
Program	%98,26	%94,44	%100	%96,87	%100	%98,76	%99,23	%100	%98,63	%96,66

Video	Video 91	Video 92	Video 93	Video 94	Video 95	Video 96	Video 97	Video 98	Video 99	Video 100
YouTube	%98,30	%98,48	%99,48	%94,67	%99,34	%99,32	%100	%100	%99,86	%99,32
Program	%96,29	%99,24	%95,89	%97,04	%100	%98,64	%98,57	%97,87	%99,80	%100

Video 81:

Geliştirilen program: 115 kelimenin 113'ünü doğru çevirdi (%98,26 doğruluk).

YouTube otomatik çeviri: 115 kelimenin 112'sini doğru çevirdi (%97,39 doğruluk).

Video 82:

Her iki sistemde 90 kelimenin 85'ini doğru çevirdi (%94,44 doğruluk).

Video 83:

Geliştirilen program: 121 kelimenin tamamını doğru çevirdi (%100 doğruluk).

YouTube otomatik çeviri: 121 kelimenin 120'sini doğru çevirdi (%99,17 doğruluk).

Video 84:

Her iki sistemde 160 kelimenin 155 tanesini doğru çevirdi (%96,87 doğruluk).

Video 85:

Her iki sistemde 121 kelimenin tamamını doğru çevirdi (%100 doğruluk).

Video 86:

Geliştirilen program: 162 kelimenin 160'ını doğru çevirdi (%98,76 doğruluk).

YouTube otomatik çeviri: 162 kelimenin 159'unu doğru çevirdi (%98,14 doğruluk).

Video 87:

Geliştirilen program: 130 kelimenin 129'unu doğru çevirdi (%99,23 doğruluk).

YouTube otomatik çeviri: 130 kelimenin 127'sini doğru çevirdi (%97,69 doğruluk).

Video 88:

Her iki sistemde 158 kelimenin tamamını doğru çevirdi (%100 doğruluk)

Video 89:

Her iki sistemde 146 kelimenin 144'ünü doğru çevirdi (%98,63 doğruluk).

Video 90:

Geliştirilen program: 120 kelimenin 116'sını doğru çevirdi (%96,66 doğruluk).

YouTube otomatik çeviri: 120 kelimenin tamamını doğru çevirdi (%100 doğruluk).

Video 91:

Geliştirilen program: 135 kelimenin 130'unu doğru çevirdi (%96,29 doğruluk).

YouTube otomatik çeviri: 135 kelimenin 133'ünü doğru çevirdi (%98,30 doğruluk).

Video 92:

Geliştirilen program: 132 kelimenin 131'ini doğru çevirdi (%99,24 doğruluk).

YouTube otomatik çeviri: 132 kelimenin 130'unu doğru çevirdi (%98,48 doğruluk).

Video 93:

Geliştirilen program: 195 kelimenin 188'ini doğru çevirdi (%95,89 doğruluk).

YouTube otomatik çeviri: 195 kelimenin 194'ünü doğru çevirdi (%99,48 doğruluk).

Video 94:

Geliştirilen program: 676 kelimenin 656'sını doğru çevirdi (%97,04 doğruluk).

YouTube otomatik çeviri: 676 kelimenin 640'ını doğru çevirdi (%94,67 doğruluk).

Video 95:

Geliştirilen program: 153 kelimenin 152'sini doğru çevirdi (%99,34 doğruluk).

YouTube otomatik çeviri: 153 kelimenin tamamını doğru çevirdi (%100 doğruluk).

Video 96:

Geliştirilen program: 148 kelimenin 146'sını doğru çevirdi (%98,64 doğruluk).

YouTube otomatik çeviri: 148 kelimenin 147'sini doğru çevirdi (%99,32 doğruluk).

Video 97:

Geliştirilen program: 140 kelimenin 138'ini doğru çevirdi (%98,57 doğruluk).

YouTube otomatik çeviri: 140 kelimenin tamamını doğru çevirdi (%100 doğruluk).

Video 98:

Geliştirilen program: 141 kelimenin 138'ini doğru çevirdi (%97,87 doğruluk).

YouTube otomatik çeviri: 141 kelimenin tamamını doğru çevirdi (%100 doğruluk).

Video 99:

Geliştirilen program: 1500 kelimenin 1497'sini doğru çevirdi (%99,80 doğruluk).

YouTube otomatik çeviri: 1500 kelimenin 1498'ini doğru çevirdi (%99,86 doğruluk).

Video 100:

Geliştirilen program: 148 kelimenin tamamını doğru çevirdi (%100 doğruluk).

YouTube otomatik çeviri: 148 kelimenin 147'sini doğru çevirdi (%99,32 doğruluk).

100 videodan oluşan video karşılaştırmasında geliştirilen program 45 videoda YouTube'a göre daha fazla çevrim oranı yakalamıştır. YouTube ise 33 videoda geliştirilen programa göre daha fazla çevrim oranı yakalamıştır. 22 videoda ise geliştirilen program ile YouTube otomatik çeviri sistemi aynı oranı yakalamıştır. Geliştirilen program, YouTube otomatik çeviri sistemine göre yüz videoluk deney sürecinde daha fazla videoyu doğru çevirmiştir.

YouTube'un otomatik çeviri sistemi ve geliştirilen program arasındaki doğruluk oranlarının ortalaması Çizelge 4.4'te karşılaştırılmıştır. Çizelge 4.5.'de görüleceği gibi geliştirilen program, YouTube otomatik çeviri sistemine göre ortalama olarak daha fazla doğruluk oranı yakalamıştır.

Çizelge 4.4. YouTube ve Geliştirilen Program karşılaştırması

Video	100 adet Video'nun Ortalaması
YouTube	%95,33
Program	%96,66

Deney süreci uzun ve kısa videolardan seçilen çeşitli veri setleri üzerinde gerçekleştirilmiştir. Her iki sistem de Türkçe ve İngilizce içeriğe sahip videoların ses ve gürültü düzeylerini karşılaştırmak amacıyla kullanılmıştır. Elde edilen sonuçlara göre geliştirilen programın YouTube otomatik altyazı sistemine kıyasla daha yüksek bir doğruluk oranı (%96,66'a karşı %95,33) sergilemesi dikkat çekicidir. Çizelge 4.4. ve Çizelge 4.5.'de de görülebileceği gibi geliştirilen program hem doğruluk oranı bakımından hem de doğru video çevrimi sayısı bakımından YouTube otomatik çeviri sistemine göre başarı yakalamıştır. Bu olumlu performans farklılıklarının temel nedenleri arasında birkaç etken bulunabilir. İlk olarak, geliştirilen programın daha spesifik bir dil modeli içermesi ve sadece konuşmadan metne çevrime odaklanması önemli bir etkendir. Bu, programın dilin bağlamını daha iyi anlamasına ve bu doğrultuda daha doğru çeviriler yapmasına olanak tanır. Ayrıca geliştirilen programın eğitildiği veri setinin çeşitliliği ve büyüklüğü modelin genel dil yeteneklerini artırabilir. Daha fazla veri programın dilin farklı varyasyonlarını ve konuşma modellerini daha etkili bir şekilde kavramasına katkı sağlayabilir. Geliştirilen programın, YouTube'un otomatik çeviri sistemine göre daha yüksek doğru oranında başarı yakalaması STT yöntemine uygulanan MFCC algoritması ile Fourier Dönüşümü ve Bölümleme yönteminden kaynaklanmaktadır. Sonuç olarak geliştirilen programın daha yüksek doğruluk oranlarına sahip olmasının kullanılan STT yönteminin deneyler aşamasında tespit edilen geliştirmelere göre entegre edilmesinden kaynaklanması kaynaklanmaktadır. Bu faktörler programın çeşitli videolardaki konuşmaları daha etkili bir şekilde anlamasını ve çevirmesini sağlamış olabilir.

4.3. Uygulama Arayüzü

Aşağıda anlatılacak uygulama kılavuzu kullanıcılara kolay ve anlaşılabilir bir arayüz sunan, hedefe uygun bir şekilde çalışan "Aranan Kelimenin Saniyesine Götüren" yazılımı hakkında detaylı bilgiler içermektedir.

Bu uygulama video platformlarında bulunan videoların indirilmesini, indirilen videonun ses dosyasından metne çevrilmesini ve metne çevrilen videodaki her kelimenin geçtiği saniyenin kaydedilmesini ve bunların bir ID'ye atanmasını sağlar. Ayrıca hem ID hem de kelime olarak sözcüklerin geçtiği saniyeleri gösterir. Kullanıcı, bu arama sonucuna tıkladığında kendi medya oynatıcısından videonun o saniyeden itibaren başladığı bir programla çalışır.

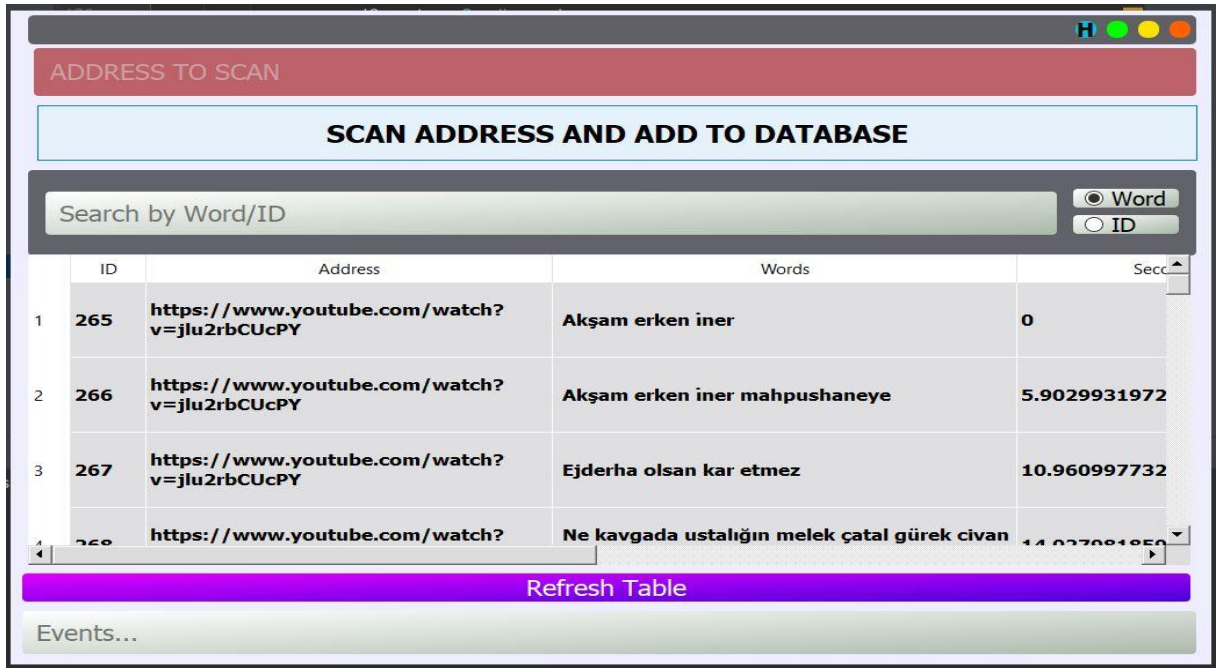
Uygulama, içerik üreticileri ve izleyiciler için kolaylık sağlayacak şekilde tasarlanmıştır. Bu projede Python programlama dili kullanılmış ve kodlamalar PyCharm Community Edition programı ile gerçekleştirilmiştir. Veri setleri, Python kütüphanesi olan sqlite3 kullanılarak kayıt altına alınmıştır ve bu kütüphane ile veri setleri işlenmektedir. Projede STT yöntemi Fourier Dönüşümü, MFCC tekniği ve Bölümleme Yöntemi kullanılarak geliştirilerek uygulanmıştır.

Video indirme süresi kullanıcının internet sağlayıcısının hızına bağlı olarak değişebilmektedir. Örneğin; standart bir Wifi ağındaki bir dakikalık bir video için indirme süresi yaklaşık 30 saniye sürmektedir. Beş dakikalık bir video için video indirme süresi 75 saniye, 20 dakikalık bir video için ise yaklaşık 150 saniye sürmektedir.

İndirilen videolardaki seslerin kelimelere çevrim süresi video uzunluğuna bağlı olarak değişmektedir. Örneğin bir dakikalık video indikten sonra sözcüklerin metne çevrilmesi yaklaşık olarak 30 saniye sürmektedir. 5 dakikalık bir video için seslerin metne çevrimi 60 saniye, 20 dakika ile 60 dakika arasındaki videoların kelime çevrimi yaklaşık olarak üç ila beş dakika arasında olmaktadır.

Uygulamanın asıl amacı, zaman tasarrufu sağlamaktır. Fiber internet hızlarının sağlandığı bölgelerde özellikle bir saat üstü videolar için kullanıcıya üç ila beş dakika arasında hızlı bir çözüm sunulabilir. Standart bir Wifi ağında ise tüm aşamalar bir saatten uzun videolar için dört ila yedi dakika arasında tamamlanabilecektir.

Uygulama arayüzü Şekil 4.1.'de yer almaktadır.



Şekil 4.1. Uygulama Ara Yüzü

Uygulama ara yüzü Şekil 4.1.’deki gibidir. Uygulama arayüzünde video platformundan video indirmek için adres sütünü ve bu videoyu indirmek için indirme butonu bulunmaktadır. Ayrıca kullanıcı burada bulunan “Search by Word/ID” kısmından aramak istediği kelimeyi yazıp klavyeden ENTER tuşuna bastığı anda aradığı kelimenin ID’si, video platformunda indirilen linki, aranan kelimenin geçtiği kelimeler ve saniyeler yer almaktadır. Kullanıcı ayrıca buradan linke tıkladığında o saniyeden itibaren bilgisayarın kendi media playerından video o saniyeden başlamaktadır. Bu aşamaları daha detaylı olarak anlatan görseller aşağıda anlatılmıştır.

Kullanıcı sözcüklerin geçtiği saniyeyi bulmak için videoyu indirmelidir. İndirilen video veri tabanında tutulmaktadır. Kullanıcı indirmek istediği videonun linkini kullanarak Şekil 4.2.’de ki gibi videoyu indirebilmektedir.



Şekil 4.2. Uygulama Arayüzünde Veri İndirme

Şekil 4.2.’de de anlatıldığı gibi burada kullanıcı “ADDRESS TO SCAN” alanına video linkini eklemektedir. Linkler video platformlarından alındığı için STT yönteminin indirme mantığına göre indirilmektedir. Kullanılan yöntem uygulamanın kullanımı artırmak ve doğru sözcük/saniye eşleştirilmesi yapılması maksadıyla kod kısmında bu ayarlar için geliştirilmiştir.

Video paylaşım platformundan alınan video linki Şekil 4.3.’de görülen buton ile indirilebilmektedir.



Şekil 4.3. Uygulama Arayüzünde İndirme Butonu

Şekil 4.3.’de anlatıldığı gibi Kullanıcı “SCANN ADRESS AND ADD TO DATABASE” butonuna bastığında videonun linkinin uygun olması durumunda video indirilmeye başlanmakta ve veri tabanına video eklenmektedir. Burada kullanıcı indirme işleminin tamamlandığını da görebilmektedir.

Kullanıcı indirdiği videonun başarılı bir şekilde kaydedildiğini Şekil 4.4.' de ki gibi görebilmektedir.



Şekil 4.4. Videonun başarılı bir şekilde indirilmesi

Şekil 4.4.'de görüldüğü gibi kullanıcı indirmek istediği videonun indirme durumunu burada “Video indirme işlemi tamamlandı” olarak takip edebilmektedir.

Kullanıcı videoyu indirdikten sonra sözcük araması yapmadan önce Şekil 4.5.'de görülen “Refresh Table” butonuna bastıktan sonra veri tabanını yenileyebilmektedir.



Şekil 4.5. Tablo Yenileme Ekranı

Şekil 4.5.'de bulunan Refresh Table butonu ile video indirildikten sonra veri tabanını yenileyebilmektedir.

Şekil 4.6.'de kullanıcının video indirme tuşuna bastıktan sonra Pycharm platformunda hem bölümlene sonucu MFCC algoritması ile Fourier Dönüşümü uygulanarak, frekans değerlerine göre hangi saniyeler arasında kaydedildiğini hem de tüm halini görebilmektedir. Ayrıca tüm hali txt dosyası olarak da kaydedilmektedir.

```
File 'speech.wav' already exists. Overwrite ? [y/N] Not overwriting - exiting
y
0
audio-bolunenler\chunk1.wav : 0 : Komşu aç bir kenarı eş olan açılara biz komşu açılar diyoruz örnek vermek gerekirse buradaki gibi.
17.63498866213152
audio-bolunenler\chunk2.wav : 17.63498866213152 : A diyelim b dışını diyelim olcay işin diyelim buradaki arkadaşlar.
29.382993197278914
audio-bolunenler\chunk3.wav : 29.382993197278914 : Alfa ve beta.
32.18199546485261
audio-bolunenler\chunk4.wav : 32.18199546485261 : Açıları.
34.25399892970522
audio-bolunenler\chunk5.wav : 34.25399892970522 : Komşudur.
36.522993197278915
audio-bolunenler\chunk6.wav : 36.522993197278915 : Tabii çok büyük bir kavram kargaşasına gerek yok arkadaşlar yan yana olan aç.
Komşu aç bir kenarı eş olan açılara biz komşu açılar diyoruz örnek vermek gerekirse buradaki gibi. A diyelim b dışını diyelim olcay işin diyelim buradaki arkadaşlar. Alfa ve be
```

Şekil 4.6. Uygulamanın Kod Sonucu

Kullanıcı uygulamada sözcük/saniye araması yapmak için Şekil 4.7.’de bulunan “Search by Word/ID” Kelime için Word butonuna basarak aramak istediği kelimeyi yazarak klavyeden Enter tuşuna basmaktadır. Eğer kullanıcı video ID’si ile aramak isterse aynı alandan ID seçerek video ID’sini yazarak enter tuşuna basmaktadır.

ID	Address	Words	Sec
Search by Word/ID			
☒ Word			
☐ ID			

Şekil 4.7. Sözcük/ID arama

Aranılan sözcük/ID burada Adres, video Id’si, kelimeler ve saniyeleri olarak çıkmaktadır. Kullanıcı tıkladığı linkte yanda yazılan saniyeden video açılmaktadır. Tüm bu özellikler Şekil 4.8’da yer almaktadır.

	Address	Words	Second
1	https://www.youtube.com/watch?v=exQPSW6ZNdo	Alfa ve beta	29.382993197278914
2	https://www.youtube.com/shorts/lfIJpAWmZ0k	Alfa ve beta	29.382993197278914

Şekil 4.8. Kelimelerin Geçtiği Saniye, Aranan Kelimeler, Linkleri ve Idleri



5. SONUÇLAR VE ÖNERİLER

5.1 Sonuçlar

Günümüzde birçok video paylaşım platformunda insanlar eğitim, kültür, sanat gibi birçok alanda değerli verilere ulaşabilmektedir. Bu proje ile videolar içerisinde arama yapma ve belirli kısımlara hızlıca ulaşma imkânı sağlanmıştır. Kullanıcılar indirdikleri videoların tüm sözcüklerine ulaşabilecekleri bir txt dosyası elde edebilmekte ve bu dosya üzerinden isteğe bağlı olarak notlar alabilmektedir. Bu program sayesinde yararlı ve değerli bilgilerin daha geniş bir kullanıcı kitlesine sunulması hedeflenmektedir.

Bu projede videolardaki sesleri yazıya çevirerek aranılan kelimelerin saniyesine götüren bir programın geliştirilmesi hedeflenmiştir. Bu amaçla çeşitli kaynaklardan (YouTube gibi) elde edilen video veri setleri kullanılmıştır. Veri seti farklı konuları kapsayan videoların yanı sıra çeşitli dil ve aksanlarda konuşan kişilerin içeriklerinden oluşmaktadır. Veri setleri video paylaşım platformları, eğitim kaynakları ve çeşitli dijital medya platformlarından temin edilmiştir. Veri setinin çeşitliliği, programın genel dili ve konuşma çeşitliliği üzerinde etkili olmasını sağlamıştır. Programda içinde kelime aramak istenilen videonun arayüzde linkini girilerek indirilmesi gerekmektedir. Veri seti öncelikle ses ve video bileşenlerine ayrılacak ve ses bileşeni üzerinde odaklanmaktadır. Ses bileşeni, video dosyalarından ayrıştırılarak ayrı bir veri seti olarak depolanmaktadır. MFCC algoritması ile verinin önce ses özellikleri dönüştürülmektedir. Ses verileri temizlenmekte normalleşme işlemiyle örnekleme hızı (sampling rate) belirlenmektedir. Gürültü azaltma, sinyal düzeltme ve spektral analiz gibi işlemler veri setinin kalitesini artırmak için uygulanmaktadır.

Projede kullanıcılar video platformlarından izlemek istediği bölüm için videoyu indirebilmekte ve indirilen videolardaki sözcüklerin geçtiği saniyeler veri tabanına kaydedilmektedir. Kullanıcılar aradıkları kelimenin geçtiği saniyeyi bulabilmekte ve kendi medya oynatıcılarından o saniyeden itibaren izleyebilmektedirler. Programda kullanıcı, videoyu indir butonuna tıkladığında video sisteme eklenecektir. Veri önce ses özelliklerine dönüştürülür. Bu özellikler MFCC algoritması kullanılarak çıkartılmaktadır. İlgili özellikler sesin frekans bileşenlerini temsil eden sayısal değerlerdir. Burada dosya mp4'ten wav ve mp3 formatlarına dönüştürülmektedir. Amacı ses ve görüntünün olduğu dosyadan sırasıyla wav ve mp3 formatlarıyla saniye temelinde metinleri tutmaktır. STT yöntemi kullanılarak bölümlendirme yöntemi ile yöntemin izin verdiği süre ve frekansların değerlerine göre doğru sözcük / saniye

eşleştirmesi yapılabilmesi için sesler metinlere frekans ve değerlerine göre işlenmektedir. Bölümleme yöntemi ve yöntemin izin verdiği süre, frekans değerlerine göre çevrim sonucu sözcük/saniye eşleştirmesi ile tüm veriler veri tabanına kaydedilmektedir. Kullanıcı eşleştirilmiş verileri aradığında bu yöntemler ile daha doğru sonuçlar elde edebilmesi sağlanması amacıyla tercih edilmiştir.

STT yöntemi kullanılarak yapılan ilk araştırmada yöntemin desteklediği süre sınırı nedeniyle bölümleme yöntemine başvurulmuştur. Ancak test aşamalarında karşılaşılan problemlerden biri yeni bölüm oluşturulurken sözcük tekrarı sorunu oluşmaktadır. Özellikle bazı durumlarda bölümleme yöntemindeki süre/sözcük eşleştirmesi sırasında tekrarlayan sözcüklerin ve saniye farklılıklarının tespit edilmesi gerekmektedir. Bu durumu çözmek için MFCC algoritması ile Fourier dönüşümü yöntemi kullanılarak gürültülerin frekans değerlerine göre hesaplamalar yapılmış ve gürültü oldukça bölüme izin verilen süre içinde eklenmiştir. Elde edilen sonuca göre bölümleme yöntemi için video ayrıştırma işlemi 0,7 saniye üzerindeki sessizlikler arasında video bölümü yaparak cümle cümle kayıt satırları oluşturulmuştur. Gürültü kesildiğinde ise yeni bir bölüm oluşturulmuştur. Bu yöntemle birçok veri seti üzerinde yapılan denemelerde doğruluk oranının arttığı ve sözcük/saniye uyumunun sağlandığı tespit edilmiştir. Deneyler aşamasında alınan sonuçlara göre geliştirilen program YouTube otomatik çeviri sistemine göre doğru kelime çevriminde daha fazla bir doğruluk oranı yakalamıştır. İki sisteminde aynı yöntemler üzerinden çalışmasına rağmen geliştirilen programın yüz videodan oluşan deney sürecinde hem oransal olarak fazla çevrim hem de videolardan oluşan deney verilerinde rakamsal olarak videolarda daha doğru bir kelime çevrimi ve Türkçe imla kurallarına daha uygun çevrim yaptığı tespit edilmiştir. Her iki sistemde aynı yöntem olmasına rağmen oluşan farkın geliştirilen programda bölümleme yöntemi ve MFCC algoritması ile Fourier dönüşümü yönteminin uygulanmasından kaynaklanmaktadır.

MFCC ve Fourier dönüşümü gibi ses işleme tekniklerinin bir araya getirilmesi ses verilerinin daha doğru bir şekilde analiz edilmesini ve metin formatına dönüştürülmesini sağlamaktadır. Bu teknikler sesin frekans bileşenlerini ve özelliklerini temsil eden sayısal değerlerin çıkarılmasında önemli bir rol oynamaktadır. Fourier dönüşümü, ses verilerinin frekans bileşenlerini analiz ederek gürültüyü ve farklı ses özelliklerini belirlemede yardımcı olurken, MFCC yöntemi ise insan kulağının davranışlarını taklit ederek ses verilerini temsil etmek için kullanılır.

Geliştirilen program kullanıcıların videolardaki ses verilerini metin formatına dönüştürerek aradıkları içeriğe kolayca ulaşmalarını sağlamaktadır. Ayrıca entegre edilen MFCC ve Fourier dönüşümü gibi ses işleme teknikleri ve Bölümle Yöntemi sayesinde doğruluk oranı ve performansı artırılarak daha iyi bir kullanıcı deneyimi sunulmaktadır.

5.2 Öneriler

Programın işlevselliğinin artırılması ve kullanıcı deneyiminin iyileştirilmesi amacıyla, aranan kelimenin geçtiği saniyelerdeki ses özelliklerinin doğru şekilde tespit edilmesi ve bilgisayarın kendi medya oynatıcısında bu saniyelerin öncesinde ve sonrasında gecikme olmaması için geliştirilen programın kullanılması önerilmektedir. Geliştirilen programın kullanılması YouTube'a göre kullanıcıların istedikleri kısımlara hızlıca erişebilmelerini sağlayacak ve programın kullanımını daha etkili hale getirecektir.

Elde edilen sonuca göre bölümleme yöntemi için video ayrıştırma işlemi 0,7 saniye üzerindeki sessizlikler arasında video bölümü yaparak cümle cümle kayıt satırları oluşturulmuştur. Gürültü olduğunda frekans değerleri çevrilip kaydedilmektedir. Gürültü olmadığında o saniyeden itibaren chunk (yığın)'da o andaki bölüm (section) kapanmaktadır. Yeni bir gürültü duyulduğunda, bir sonraki bölüm açılmaktadır. Deneyler sürecinde bazı videolarda 0,7 saniye üzerindeki sessizlikler için yeni bir bölüm oluşmadığı ve videoda konuşmacı söylediği cümlelerde duraksamadığı için yöntemin sessizlik algılayamadığı böylelikle çevrilen metinlerin tek bir bölüme yazılabileceği tespit edilmiştir. Bazı videolarda tüm metinlerin tek bir bölüme yazılmasının önüne geçmek için yazılımda 0,7 saniye sınırının değiştirilmesi gerekmektedir.

Programın performansını artırmak için sesin çıkışının net olduğu videoların tercih edilmesi ve veri setindeki seslerin anlaşılabilirliği için dikkat edilmesi önemlidir. Ayrıca test aşamasında tespit edilen ses gelmeyen videolar gibi sorunlu durumlarla karşılaşıldığında bu tür videoların seçilmemesi, bu videoların seçildiği takdirde STT yönteminin bu video türlerini anlayacak şekilde Fourier Dönüşümü ve MFCC algoritmasını kullanılırken geliştirilmesi beklenmektedir.

Gelecekteki çalışmalarda farklı yapay sinir ağı mimarileri ve derin öğrenme yöntemlerinin kullanılması önerilmektedir. Derin oto kodlayıcılar, dikkat mekanizması, CNN veya transformatör tabanlı modeller gibi yeni yaklaşımların kullanılması konuşma tanıma performansını artırabilir. Bu farklı mimarilerin karşılaştırmalı analizleri ve optimizasyon stratejileriyle en etkili yapay sinir ağı yapısının belirlenmesi sağlanabilir.

Daha geniş ve çeşitli veri kümelerinin kullanılması da önemlidir. Farklı dil ve aksanlardan, farklı kaynaklardan elde edilen ses verileri modelin genelleme yeteneğini artırabilir. Ayrıca çeşitli gürültü ve bozulma koşullarında kaydedilen ses verilerinin kullanılması modelin gerçek dünya uygulamalarında daha iyi performans göstermesini sağlayabilir.

Eğitim sürecinin optimize edilmesi ve modelin performansının değerlendirilmesi için daha güvenilir ve kapsamlı yöntemlerin kullanılması gerekmektedir. Çapraz doğrulama, test veri setlerinin çeşitlendirilmesi ve farklı değerlendirme protokollerinin uygulanması bu alanda faydalı olabilir.

Sonuç olarak bu araştırma yapay sinir ağlarıyla STT yönteminin kullanımıyla sesleri yazıya çevirme konusunu temel almıştır. Yapay sinir ağlarının geliştirilmesi, veri kümelerinin çeşitlendirilmesi ve büyütülmesi, eğitim sürecinin iyileştirilmesi, metriklerin ve değerlendirme yöntemlerinin geliştirilmesi, uygulama ve endüstriyel kullanım alanlarında çalışmaların yapılması ve STT yönteminin belirlenen eksikliklere göre geliştirilmesi önemlidir. Bu alanda yapılan araştırma ve geliştirmeler ses tanıma ve metin dönüştürme alanında önemli ilerlemeler sağlayacaktır.

KAYNAKLAR

- Yalçın, N., 2008, Konuşma Tanıma Teorisi ve Teknikleri, Kastamonu Eğitim Dergisi, Kastamonu Üniversitesi, Kastamonu, 249-266
- Ergezer, H., Dikmen, M. ve Özdemir, E., 2003, Yapay Sinir Ağları ve Tanıma Sistemleri, PiVOLKA, 14-17
- Jurafsky, D. ve Martin, J. H., 2009, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, University of Colorado, Boulder.
- Mikolov, T., Chen, K., Corrado, G. ve Dean, J., 2013, Efficient Estimation of World Representations in Vector Space, arXiv preprint arXiv:1301.3781v3.
- Ocak Mahir, E., 2020, *Yapay Sinir Ağları Nelerdir?* [online], Bilim Genç, Tubitak, <https://bilimgenc.tubitak.gov.tr/yapay-sinir-aglari-nedir> [Ziyaret Tarihi: 10 Eylül 2023].
- Öztemel, E., 2003, Yapay Sinir Ağları, Papatya Yayıncılık, İstanbul, 44-60.
- Moralı, İ.A. ve Aygün, F.F, 2006, Çok Katmanlı Algılayıcı ve Geriye Yayılım Algoritması ile Konuşmacı Tanıma. Dokuz Eylül Üniversitesi, İzmir.
- Davis, S.B., Mermelstein, P., 1980 Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentence, IEEE Trans. on Acoustic, Speech and Signal Processing, p. 357–366
- Gandhi, P., 2019, Deep Learning Next Step : Transformes and Attention Mechanism [online], KDnuggets, <https://www.kdnuggets.com/2019/08/deep-learning-transformers-attention-mechanism.html> [Ziyaret Tarihi: 5 Aralık 2023].
- Hinton, G. E., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.R., Jaitly, N. ve Kingsbury, B., 2012, Deep neural networks for acoustic modeling in Speech Recognition: The shared views of four research groups, IEEE Signal Processing Magazine, 29(6), 82-97.
- Ahad, A., Fayyaz, A. ve Mehmood T., 2002, Speech Recognition using Multilayer Perception.Students Conference Proceedings, IEEE Volume 1 p.103-109
- Adalı, E., 2012, Doğal Dil İşleme, Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, Türkiye Bilişim Vakfı, 5, 2.
- Saijid, A. Marhon, 2006, Speaker Recognition Based On Neural Networks, The Higher Institute For Industry, Misrata, Libya.
- Delibas, A., 2008, Doğal dil işleme ile Türkçe yazım hatalarının denetlenmesi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

- Caner, M. ve Üstün, S.V., 2006, Yapay Sinir Ağları ile Konuşmacı Kimliği Tanıma Uygulaması. Pamukkale Üniversitesi Mühendislik Fakültesi Mühendislik Bilimleri Dergisi, Sayı 2, Cilt 12, s.279–284
- Sel, İ., Hanbay, D. ve Karabatak, M., 2011, Beyin Bilgisayar Arayüzleri İçin Türkçe Metinden Konuşma Sentezleme Sistemi, Fırat Üniversitesi Elektrik-Elektronik Bilgisayar Sempozyumu, Bildiri Kitabı II, 2011 Elazığ, 273-276.
- Arık, G., 2011, Görme Engelliler İçin Bilgisayar Kullanımının Etkinleştirilmesi, Erişilebilirlik ve Bir Türkçe Hece Tabanlı Konuşma Sentezleme Sisteminin Geliştirilmesi, Yüksek Lisans Tezi, Ankara, Gazi Üniversitesi Bilişim Enstitüsü.
- Tekindal, B. ve Arık, G., 2012, Görme Engelliler için Türkçe Metinden Konuşma Sentezleme Yazılımı Geliştirilmesi, Bilişim Teknolojileri Dergisi, 5 (3), 9-18
- G.,Şevval, 2022, Text to Speech nedir? [online], Hosting , <https://www.hosting.com.tr/blog/text-to-speech> [Ziyaret Tarihi: 20 Mayıs 2023].
- Şeker, A., Diri, B. ve Balık, Hasan H., 2017, Derin Öğrenme Yöntemleri ve Uygulamaları hakkında bir inceleme, Gazi Mühendislik Bilimleri Dergisi, 3(3):47-64.
- Yusuf, S. M., Adedokun, E. A., Muazu, M. B., Umoh, I. J., ve Ibrahim, A. A., 2021, RMWSaug: Robust Multi-window Spectrogram Augmentation Approach for Deep Learning based Speech Emotion Recognition, In 2021 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-6), IEEE.
- Demirci, M.D., 2005, Bilgisayar Destekli Ses Tanıma Sistemi Tasarımı, İstanbul Üniversitesi Fen Bilimleri Enstitüsü, Bilimsel Dergi Makalesi, İstanbul.
- Shaip, 2022, Speech-to-Text Algoritması nedir ve otomatik konuşma’da nasıl çalışır? [online], Shaip, <https://tr.shaip.com/blog/automatic-speech-recognition-a-sr/> [Ziyaret Tarihi: 24 Haziran 2023].
- Chollet, F., 2021, Deep Learning with Pyhton, Simon and Schuster, Shalter Island
- Baniardalan, F., & Akbari, A. ,2017, A weighted denoising auto-encoder applied to mel sub-bands for robust speech recognition, Iranian Conference on Intelligent Systems and Signal Processing, 38- 42.
- NewlabsTurkey, 2022, Sesten Metine Yararlı Bazı Uygulamalar [online], NewLabsTurkey, <https://www.newslabturkey.org/2022/10/12/sesten-metine-bazi-yararli-uygulamalar/> [Ziyaret Tarihi: 24 Nisan 2023].
- Aygün, F. F. ve Moralı, İ. A., 2007, Çok Katmanlı Algılayıcı ve Geriye Yayılım Algoritması ile Konuşmacı Ayırt Etme, Akademik Bilişim '07-IX. Akademik Bilişim Konferansı Bildirileri 31 Ocak-2 Şubat 2007 Dumlupınar Üniversitesi, Kütahya, 57-62

- Dede, G. F. ve Sazlı, M. H., 2008, Yapay Sinir Ağları ile Konuşma Tanıma, Ankara Üniversitesi Fen Bilimleri Enstitüsü Elektronik Mühendisliği Anabilim Dalı
- Ateş, E., 2019, Derin Öğrenme ile Sesli Komut Tanıma, Eskişehir Osmangazi Üniversitesi Elektronik Mühendisliği Anabilim Dalı.
- Güldalı, K., 2009, Türkçe Metin Seslendirme, Yüksek Lisans Tezi, İstanbul, İTÜ Fen Bilimleri Enstitüsü.
- Doruk, T., 2022, Uzun-Kısa Vadeli Bellek(LTSM) Nedir ve Nasıl çalışır?, [online], <https://medium.com/teknopar-akademi/uzun-k%C4%B1sa-s%C3%BCreli-bellek-lstm-nedir-ve-nas%C4%B1l-%C3%A7al%C4%B1%C5%9F%C4%B1-cd152009422b> [Ziyaret Tarihi: 22 Mart 2023].
- Oğuzlar, A., 2003, Veri Ön İşleme, Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 21, ss.67-76.
- Süleyha, B., 2007, Yapay Sinir Ağları Kullanarak Konuşmacı Tanıma, T.C. Uludağ Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi, Bursa.
- Azam, S.M., Mansoor, Z.A., Mughal, M.S. ve Mohsin, S, 2007, Urdu spoken digits recognition using classified MFCC and backpropagation neural network., Computer Graphics, Imaging and Visualization Conference.
- Syadida, A. Q., Ignatius, D. R., Setiadi, M., ve Setyono, A, 2017, Sphinx4 for indonesian continuous speech recognition system, International Seminar on Application for Technology of Information and Communication, 264-267.
- Natro, 2023, Doğal Dil İşleme (NLP) nedir? [online], <https://www.natro.com/blog/dogal-dil-isleme-nlp-nedir/> [Ziyaret Tarihi: 10 Ekim 2023].
- Somuncu, E. ve Atasoy, N.A., 2022, Evrişimli tekrarlayan sinir ağı ile metin görüntüleri üzerinde karakter tanıma uygulaması gerçekleştirilmesi, Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, Ankara, 37:1, 17-27
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoerit, J., Lones, L., Gomez, A.N., Kaiser, L. ve Polosukhin, I., 2023, Attention Is All You Need, arXiv:1706.03762v7 [cs.CL]
- Atcılı, A., Attention Mekanizması [online], Medium, <https://medium.com/machine-learning-t%C3%BCrkiye/attention-mekanizmasi-6d3566c0518e> [Ziyaret Tarihi: 5 Şubat 2023].
- Thimmaraja, Y. G., ve Jayanna, H. S., 2017, Creating language and acoustic models using kaldi to build an automatic speech recognition system for Kannada

- language. IEEE International Conference Recent Trends Electronic Information Communication Technology, 161-165.
- Polur, P.D., Zhou, R., Yang, J., Adnani, F. ve Hobson, R.S, 2001, Isolated speech recognition using artificial neural networks, 23rd Annual EMBS Conference.
- Şeker, S.E., 2015, Doğal Dil İşleme (Natural Language Processing), YBS Ansiklopedi, İstanbul, 2:4, 1
- Deniz, Y., 2023, Self-Attention, LinkedIn, Mekanizması[online], <https://www.linkedin.com/pulse/self-attention-mekanizmas%C4%B1-yusuf-deni%CC%87z-/?originalSubdomain=tr> s [Ziyaret Tarihi: 10 Haziran 2023].
- Bengio, Y., Lamblin, P., Popovici, D., ve Larochelle, H., 2007, Greedy layer-wise training of deep networks. Advances in Neural Information Processing Systems, 19(1), 153-162.
- Morandi, A., 2012, A fast learning algorithm for deep belief nets geoffrey. Monthly Notices of the Royal Astronomical Society, 425(3), 2069-2082.
- Tong, R., Wang, L., ve Ma, B., 2017, Transfer learning for children's speech recognition, International Conference on Asian Language Processing, 36-39
- Tora, H., Uslu, İ. B. Ve Karamahmet, T., 2017, Implementation Of Turkish Text-To-Speech Synthesis On A Voice Synthesizer Card With Prosodic Features, Anadolu Üniversitesi Bilim ve Teknoloji Dergisi A- Uygulamalı Bilimler ve Mühendislik, 18 (3), 584- 594.
- Oflazoglu, Ç., ve Yildirim, S. , 2011, Turkish emotional speech database, In 2011 IEEE 19th Signal Processing and Communications Applications Conference (SIU) (pp. 1153-1156), IEEE.
- Bulutistan, 2023, Sesi Yazıya Çevirme Neden Önemlidir? [online], Bulutistan, <https://bulutistan.com/blog/sesi-yaziya-cevirme-nedir-neden-onemlidir-openaiin-yeni-konusmayi-anlama-ve-metne-cevirme-sistemi-whisper/> [Ziyaret Tarihi: 24 Ağustos 2023].
- Sasithon, P. S., 2017, Isarn digit speech recognition using HMM, International Conference Informatic Technology, 1-5, [42].
- Ho, H., Mawardi, V. C. ve Dharmawan, A. B., 2017, Question answering system with hidden markov model speech recognition, International Conference on Science in Information Technology, 257-262.
- Eroglu Erdem, C., Turan, C. ve Aydin, Z., 2015, BAUM-2: a multilingual audio-visual affective face database, Multimedia tools and applications, 74(18), 7429-7459.

Uzun, E., 2007, Görme Engelliler İçin Basılı Doküman Yorumlama ve Seslendirme Sisteminin Gerçekleştirilmesi, Yüksek Lisans Tezi, Trabzon KTÜ Fen Bilimleri Enstitüsü.

Şentürk, T. ve Adalı, E., 2016, Türkçe Metin Seslendirme, Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 4 (1).

Aşlıyan, R. ve Günel, K., 2008, Türkçe Metinler için Hece Tabanlı Konuşma Sentezleme Sistemi, Akademik Bilişim, 2008.

Karamahmet, T., 2015, Implementation Of Turkish Text To Speech Synthesis, Yüksek Lisans Tezi, Ankara, Atılım Üniversitesi.

