
GAZİ ÜNİVERSİTESİ
Gazi
1926

TÜRKÇE KONUŞMA TANIMA SİSTEMLERİ İÇİN DERİN ÖĞRENME TABANLI MODELLERİN GELİŞTİRİLMESİ

Saadin OYUCU

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

EKİM 2020

Saadin OYUCU tarafından hazırlanan "TÜRKÇE KONUŞMA TANIMA SİSTEMLERİ İÇİN DERİN ÖĞRENME TABANLI MODELLERİN GELİŞTİRİLMESİ" adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi Bilgisayar Mühendisliği Ana Bilim Dalında DOKTORA TEZİ olarak kabul edilmiştir.

Danışman: Doç. Dr. Hüseyin POLAT

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi ……..……..

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum.

Başkan: Prof. Dr. Recep DEMİRCİ

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi ...……

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum.

Üye: Prof. Dr. Ertan ONUR

Bilgisayar Mühendisliği Ana Bilim Dalı, Orta Doğu Teknik Üniversitesi ...……

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum.

Üye: Prof. Dr. Aysun COŞKUN

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi .....……

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum.

Üye: Prof. Dr. İlyas ÇANKAYA

Elektrik-Elektronik Mühendisliği, Ankara Yıldırım Beyazıt Üniversitesi ...……

Bu tezin, kapsam ve kalite olarak Doktora Tezi olduğunu onaylıyorum.

Tez Savunma Tarihi: 20/10/2020

Jüri tarafından kabul edilen bu çalışmanın Doktora Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum

……..…….

Prof. Dr. Cevriye GENCER

Fen Bilimleri Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

……..
Saadin OYUCU
20/10/2020

TÜRKÇE KONUŞMA TANIMA SİSTEMLERİ İÇİN DERİN ÖĞRENME TABANLI MODELLERİN GELİŞTİRİLMESİ

(Doktora Tezi)

Saadin OYUCU

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Ekim 2020

ÖZET

Kelime Hata Oranı (KHO) düşük Otomatik Konuşma Tanıma (OKT) sistemlerinde, büyük miktarda konuşma ve bu konuşmalar ile eşleştirilmiş metin veri kümesine ihtiyaç duyulmaktadır. Bu nedenle çalışma kapsamında Türkçe OKT veri kümesi hazırlamaya yönelik farklı bir yaklaşım sunulmuştur. Sunulan yaklaşımda üç farklı yöntem kullanılmıştır. İlk yöntemde, işitme güçlüğü çeken kişiler için hazırlanan altyazı belgeleri filmlerden elde edilen konuşma bilgisi ile eşleştirilmiştir. İkinci yöntemde, veriler bir mobil uygulama aracılığıyla gerçek kullanıcılardan elde edilmiştir. Üçüncü yöntemde ise transfer öğrenme yaklaşımı kullanılmıştır. Elde edilen veriler gerçek kullanıcıların onayına sunulmuştur. Türkçe OKT sistemi için gerekli Akustik Model (AM), Dil Modeli (DM) ve Okunuş Sözlüğü (OS) hazırlanan veri kümesi kullanılarak geliştirilmiştir. Yapay sinir ağı, Gauss Karışım Modeli ve Saklı Markov Modeli tabanlı akustik modellerin ilk konuşma tanıma sonuçları verilmiştir. Ayrıca OKT sistemlerinin başarımını düşürecek akustik bilgilerin ortadan kaldırılması için konuşma içerisinde geçen sessizliklerin kaldırılması ve konuşmaların parçalara ayrılması gerçekleştirilmiştir. OS'nin oluşturulmasındaki sesbirim kuralları belirlenmiştir. Günlük konuşma içerisinde sıklıkla kullanılan yabancı kelimeler ve Türkçede birden fazla okunuşa sahip olan kelimelerin farklı okunuşları OS'ye eklenmiştir. OKT için iyi dizayn edilmiş bir DM'nin AM ile birlikte kullanılması KHO'yu düşürmektedir. Bu nedenle çalışmada, Türkçe OKT'nin KHO başarımını arttırmak için cümle düzeyinde bir DM iyileştirme yöntemi önerilmiştir. Sonuç olarak, Türkçe için literatürdeki yetersiz kaynak durumu telafi edilmiştir. Ayrıca, AM, DM ve OS gerçekleştirilen iyileştirmeler ile KHO düşük ve geniş kelime dağarcığına sahip bir Türkçe OKT sistemi geliştirilmiştir. Geliştirilen OKT sistemine erişimi kolaylaştırmak için web servis tabanlı bir platform hazırlanmıştır. Kullanıcıların platforma erişimi, platform ile birlikte hazırlanan web arayüzü üzerinden gerçekleştirilmiştir. Ayrıca geliştirilen uygulama programlama arayüzleri sayesinde farklı uygulama ve servislerin platforma erişimi sağlanmıştır. Böylelikle mobil cihazlarda ve nesnelerin interneti ekosisteminde sorunsuz çalışabilen geniş kelime dağarcığına sahip bir Türkçe OKT platformu geliştirilmiştir.

Bilim Kodu	:	92432
Anahtar Kelimeler	:	Türkçe Konuşma Tanıma, Türkçe Dil Modelleme, Akustik Modelleme, Okunuş Sözlüğü, Yapay Zeka, Derin Öğrenme
Sayfa Adedi	:	111
Danışman	:	Doç. Dr. Hüseyin POLAT

DEVELOPMENT OF DEEP LEARNING BASED MODELS FOR TURKISH SPEECH RECOGNITION

(Ph. D. Thesis)

Saadin OYUCU

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

October 2020

ABSTRACT

Automatic Speech Recognition (ASR) systems with low Word Error Rate (KHO) need a large amount of speech and a data set of text matched with these speeches. For this reason, a different approach to preparing a Turkish ASR data set is presented in the scope of the study. Three different methods were used in the proposed process. In the first method, subtitle documents prepared for people with hearing difficulties were matched with movies' speech information. In the second method, data were obtained from real users via a mobile application. In the third method, the transfer learning approach was used. The obtained data were submitted to the approval of real users. The Acoustic Model (AM), Language Model (LM) and lexicon required for the Turkish ASR system were developed using the prepared data set. The first speech recognition results of different acoustic models based are given. Also, to eliminate acoustic information that would reduce the performance of ASR systems, silences in the speech were removed and speeches were divided into parts. Also, the phoneme rules in the creation of the lexicon have been determined. Foreign words that are frequently used in daily speech and different readings of words that have more than one pronunciation in Turkish were added to the lexicon. Using LM together with AM in ASR systems decreases WER. Therefore, in the study, a sentence-level LM improvement method is proposed to increase the performance of Turkish ASR's WER. As a result, the low resource situation stated in the literature for Turkish has been compensated. Also, with the improvements made on AM, LM and lexicon, a Turkish ASR system with low WER and large vocabulary has been developed. A web service-based platform has been prepared to facilitate access to the developed ASR system. Users were provided with access to the ASR system via the web interface designed with the platform. Also, different applications access to the platform has been provided through the application programming interface. Thus, a Turkish ASR platform with a large vocabulary has been developed that can work smoothly on mobile devices and the Internet of Things ecosystem.

Science Code	:	92432
Key Words	:	Turkish Speech Recognition, Turkish Language Modelling, Lexicon, Acoustic Modelling, Artificial Intelligence, Deep Learning
Page Number	:	111
Supervisor	:	Assoc. Prof. Dr. Huseyin POLAT

TEŞEKKÜR

Lisans ve yüksek lisans çalışmalarımda olduğu gibi bu tez çalışmasında da tez konusunun belirlenmesi, tezin planlanması, tez çalışmalarının yürütülmesi ayrıca bana vermiş olduğu tavsiye ve cesaretten dolayı değerli danışman hocam Doç. Dr. Hüseyin POLAT'a saygı ve şükranlarımı sunuyorum. Birlikte çalışma fırsatı bulduğum ve ihtiyacım olduğu her an bilgi ve fikirlerini benden esirgemeyen ve tez izleme jürimde yer alan değerli hocam Prof. Dr. Recep DEMİRCİ'ye ve bu tez çalışmasının ilerleme aşamasındaki görüşleri, önerileri ve araştırmamızı çeşitli açılardan genişletmemizi teşvik eden yorum ve katkılarından dolayı değerli hocam Prof. Dr. Ertan ONUR'a saygılarımı sunuyor ve çok teşekkür ediyorum. Ayrıca ses, konuşma, dil bilimi ve akademik bilgisinden yararlandığım değerli hocam Prof. Dr. Hayri SEVER'e bana sürekli verdiği ümit ve desteklerden dolayı saygılarımı sunuyor ve teşekkür ediyorum. Hayatım boyunca bana sabır ve anlayış gösteren, maddi ve manevi her zaman yanımda olarak yükümü hafifleten ve her daim bana moral, motivasyon kaynağı olan sevgili eşim Yurdanur OYUCU'ya, oğlum Mehmet Han'a, kızlarım Hacer Yuşa ve Merve Asel'e, her zaman yanımda olduklarını hissettiren ve bu günlere gelmem de büyük emeği olan başta annem Güler OYUCU ve ağabeyim Necati OYUCU olmak üzere aileme, eğitimin ve bilimin önemini her konuşmasında dile getiren ayrıca maddi ve manevi desteğini sürekli esirgemeyen kayınpederim Mehmet KESKİNOĞLU'na gönülden sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	**Açıklamalar**
KHz	Kilohertz
ms	Milisaniye

Kısaltmalar	**Açıklamalar**
AFD	Ayrık Fourier Dönüşümü (Discrete Fourier Transform)
AGKM	Altuzay Gauss Karışım Modeli (Subspace Gaussian Mixture Model)
AKD	Ayrık Kosinüs Dönüşümü (Discrete Cosine Transform)
AM	Akustik Model (Acoustic Model)
ASDA	Ağırlıklı Sonlu Durum Alıcıları (Weighted Finitestate Acceptors)
BB	Bağlam Bağımlı (Context Dependency)
CHO	Cümle Hata Oranı (Sentence Error Rate)
DAA	Doğrusal Ayrımcı Analizi (Linear Discriminant Analysis)
DM	Dil Modeli (Language Model)
DÖ	Derin Öğrenme (Deep Learning)
DÖK	Doğrusal Öngörülü Kodlama (Linear Predictive Coding)
DSA	Derin Sinir Ağı (Deep Neural Network)
DVM	Destek Vektör Makinesi (Support Vector Machine)
DZA	Dinamik Zaman Atlatılması (Dynamic Time Warping)
EBA	En Büyüklenme Algoritması (Expectation-Maximization)

EBS	En Büyük Sonsal (Maximum A Posteriori)
GİB	Grafik İşlem Birimi (Graphic Processing Unit)
GKM	Gauss Karışım Modeli (Gaussian Mixture Model)
GTÜ	Geçitli Tekrarlanan Üniteler (Gated Recurrent Unit)
HFD	Hızlı Fourier Dönüşümü (Fast Fourier Transform)
HMİD	Hiper Metin İşaretleme Dili (Hypertext Markup Language)
HMTP	Hiper Metin Transfer Protokolü (Hyper-Text Transfer Protocol)
İBSA	İleri Beslemeli Sinir Ağları (Feed Forward Neural Networks)
JNB	Javascript Nesne Belirtimi (Javascript Object Notation)
KAHO	Karakter Hata Oranı (Char Error Rate)
KHO	Kelime Hata Oranı (Word Error Rate)
MBAS	Microsoft Bilişsel Araç Seti (Microsoft Cognitive Toolkit)
MFKK	Mel Frekans Kepstral Katsayısı (Mel Frequency Cepstral Coefficient)
MİB	Merkezi İşlem Birimi (Central Processing Unit)
MO	Maksimum Olabilirlik (Maximum Likelihood)
MODD	Maksimum Olasılık Doğrusal Dönüşüm (Maximum Likelihood Linear Transform)
ODTÜ	Orta Doğu Teknik Üniversitesi (Middle East Technical University)
OKT	Otomatik Konuşma Tanıma (Automatic Speech Recognition)
OS	Okunuş Sözlüğü (Lexicon)
SAK	SMM Araç Seti (Hidden Markov Model Toolkit)
SDA	Sonlu Durum Ağı (Finite-State Network)
SDD	Sonlu Durum Dönüştürücü (Finite-State Transducer)
SGAPA	Savunma Gelişmiş Araştırma Projeleri Ajansı (Defense Advanced Research Projects Agency)
SGİ	Stokastik Gradyan İniş (Stochastic Gradient Descent)

SMM	Saklı Markov Modeli (Hidden Markov Model)
SRIDM	SRI Dil Modelleme Araç Seti (SRI Language Modeling Toolkit)
TBMM	Türkiye Büyük Millet Meclisi (Grand National Assembly of Turkey)
TDT	Temsili Durum Transferi (Representational State Transfer)
TSA	Tekrarlayan Sinir Ağları (Recurrent Neural Networks)
UKSB	Uzun Kısa Süreli Bellek (Long Short Term Memory)
UPA	Uygulama Programlama Arayüzü (Application Programming Interface)
YSA	Yapay Sinir Ağı (Artificial Neural Network)
ZBS	Zamansal Bağlantılı Sınıflandırma (Connectionist Temporal Classification)
WSTTOKTP	Web Servis Tabanlı Türkçe OKT Platformu (Web Service Based Turkish OKT Platform)

1. GİRİŞ

İnsanların birbirleri ile anlaşabilmesi, duygu ve düşüncelerini başkalarına aktarabilmesi için iletişim kurmaları gerekir. İlk insanların birbirleri ve diğer canlılar ile gerçekleştirdiği iletişimin içgüdüsel olarak sağlandığı bilinmektedir. İnsanlığın geçirdiği evrimin iletişim ile bağlantılı olduğu da ileri sürülmektedir. Zaman içerisinde sürekli olarak gelişen ve değişen iletişim kavramı insanların yaşamını biçimlendirmeye devam etmektedir. İletişim, toplumların daha çağdaş olmaları yolundaki etkisini sürdürmekte ve sürdürdüğü bu etkiyi giderek arttırmaktadır.

Zaman içerisinde iletişim yöntemleri gelişim ve değişim göstermiştir. Bu gelişim sayesinde içgüdüsel olarak sağlanan iletişim sözlü olarak sağlanmıştır. Daha sonraları ise sadece sözlü olarak değil yazılı iletişim de kurulabilmiştir. Özellikle matbaanın icadı ile yazılı iletişim kurma yöntemi yaygınlaşmış ve kitleler ile rahatlıkla iletişim kurulabilmiştir. İletişim kurma yöntemindeki bu değişiklik haber verme, okuma, yönlendirme gibi farklı kavramların ve yayıncılık, editörlük gibi birçok meslek dalının ortaya çıkmasını sağlamıştır. Yazılı iletişimin yaygınlaşmasından sonra özellikle teknoloji araçlarının gelişmesi ve kullanımının artması iletişim alanında büyük değişiklikler sağlamıştır. Bu değişikliklerin başında mesafe tanımaksızın kurulabilen iletişim ağları gelmektedir. Bu ağlar özellikle radyo, televizyon, bilgisayar, internet, tablet ve akıllı cep telefonlarının yaygınlaşması ile etkisini daha fazla göstermiştir [1].

Teknolojinin gelişmesi ile ortaya çıkan yeni iletişim kanalları temelde insanların birbirleri ile iletişimin kurmasını desteklemektedir. Ancak bu iletişim kanalları zaman içerisinde yapısal olarak değişmiş ve insan-makine etkileşimi kavramı ortaya çıkmıştır. İnsan-makine etkileşimi sayesinde artık sadece insanlar arasında değil insanların bilgisayar veya diğer elektronik cihazlar ile iletişim kurabilmesi sağlanmıştır. Erken evrelerde basit tuşlu sistemler ile gerçekleştirilen insan-makine veya elektronik cihaz etkileşimi günümüzde yerini dokunmatik sistemlere bırakmıştır. Dokunmatik ekranlar, kullanım kolaylığı sağlamaktadır. Ancak bu kolay kullanım insanların hareketlerini az da olsa kısıtlamaktadır. Bu nedenle insanlar, ana dillerini kullanarak başka insanlar arasında kurduğu iletişime benzer şekilde makine veya elektronik cihazlar ile iletişim kurmayı istemektedir. Makine veya elektronik

cihazlar ile iletişim kurmayı gerektiren durumlarda iletişimin doğal konuşma ile sağlanması diğer yöntemlere göre daha kolay ve hızlıdır. İnsanlar konuşurken ağız hariç başka hiçbir vücut uzvunu kullanmadan iletişim kurabilmektedir. Bu nedenle doğal konuşma ile gerçekleştirilen bir iletişim insanlara esneklik sağlamaktadır. Konuşmanın sağladığı kullanım kolaylığı, hız ve esneklik sayesinde makine veya elektronik cihazların doğal konuşma ile kontrol edilmesi kolaylaşmış ve zaman içerisinde yaygınlaşmıştır.

Çeviksoy ve Boran'a (1999) göre "Bir dilin imkânlarını kullanarak, olay, durum, bilgi, istek, fikir, duygu, düşünce ve hayalleri sesle (sözlü olarak) anlatmaya konuşma denir." [2]. Çeviksoy ve Boran'ın konuşmanın tanımında belirttiği gibi bir konuşma belirli bir bilgiyi içermektedir. Bu bilgi karşı tarafa sözel olarak iletişim kanalı ile aktarılmaktadır. Doğal konuşmaların duyulabilmesinde kullanılan iletişim kanalı ses dalgalarıdır. İnsan kulağı kendisine gelen ses dalgalarını yakalar ve onları insan beyninin yorumlayabileceği bilgilere dönüştürür. Ses dalgası içerisinde müzik, konuşma, gürültü veya daha farklı ses bilgileri yer alabilmektedir. İnsan beyni bu sesleri ayırt edebilmekte ve konuşma içerisindeki bilgiyi anlamlandırabilmektedir.

Ses, yavaş veya hızlı titreşen basınç dalgaları ile oluşmaktadır. Yavaş titreşimler alçak, hızlı titreşimler ise yüksek tonları oluşturmaktadır. Ses dalgaları elektronik iletişim kanalları içerisinde farklı bir ortama iletilebilmektedir. Dolayısıyla ses içerisindeki konuşma bilgisi elektronik iletişim kanalları sayesinde makine veya elektronik cihazlara rahatlıkla aktarılabilir. Ancak makine veya elektronik cihazlara aktarılan ses bilgisinin makine veya elektronik cihazların anlayabileceği bir yapıya dönüştürülmesi gerekir. Bu dönüşüm kısaca insanlar tarafından konuşulan ifadelerin makine veya elektronik cihazlar tarafından okunabilir metne dönüştürülmesini temsil etmektedir.

Prakoso, Ferdiana ve Hartanto'ya (2017) göre insanlar tarafından konuşulan ifadeleri bilgisayar tarafından okunabilir metne dönüştüren sistemlere Otomatik Konuşma Tanıma (OKT, Automatic Speech Recognition) denilmektedir [3]. Çakır'a (2017) göre "konuşma tanıma sistemi, kullanıcının belirli kurallar ile oluşturulan ve kurallarının bilgisayar tarafından bilindiği, birtakım sesli ifadeleri, bilgisayar tarafından anlaşılabilecek formata dönüştürme işlemidir" [4]. Bilgisayarlar veya elektronik cihazlar tarafından anlaşılabilecek

formata dönüştürülen konuşma bilgisi ile birçok yeni uygulama geliştirilmiştir. OKT sistemleri sesli komut uygulamaları, akıllı ev sistemleri, güvenlik sistemleri, eğitim sistemleri, çağrı merkezleri, konferans merkezleri, otomatik konuşma raporlama, dikte yazılımları vb. birçok alanda ve uygulamada kullanılmaktadır. OKT sistemlerinin kullanıldığı alanlar Çizelge 1.1'de detaylı olarak sunulmuştur.

Çizelge 1.1. OKT sistemlerinin kullanıldığı alanlar ve uygulamalar

Kullanım Alanı	Kullanım Şekli	Girdi	Çıktı
İletişim alanı	Sesli komutlar ile kişi aramalarının yapılmasında	Konuşma sinyali	Konuşulan kelimeler
İletişim alanı	Telefon mesajlarının konuşarak yazılmasında	Konuşma sinyali	Konuşulan kelimeler
Eğitim alanı	Dil eğitiminde kelimelerin doğru okunuşunun öğretilmesinde	Konuşma sinyali	Konuşulan kelimeler
Elektronik	Elektronik eşyalara konuşarak komutlar verilmesinde	Konuşma sinyali	Konuşulan kelimeler
Savunma	Savunma alanında kullanılan araçların ses ile kontrol edilmesinde	Konuşma sinyali	Konuşulan kelimeler
Robotik	Robotların konuşarak idare edilmesinde	Konuşma sinyali	Konuşulan kelimeler
Sağlık alanı	Sağlık raporlarının konuşarak metne aktarılmasında	Konuşma sinyali	Konuşulan kelimeler
Çağrı merkezleri	Telefon görüşmelerinin otomatik yazıya aktarılmasında	Konuşma sinyali	Konuşulan kelimeler
Otomotiv	Araç aksesuarlarının konuşarak komuta edilmesinde	Konuşma sinyali	Konuşulan kelimeler
Adli	Mahkeme ifade ve kararlarının otomatik olarak metne aktarılmasında	Konuşma sinyali	Konuşulan kelimeler
İstihbarat	Telefon kayıtlarının otomatik metne aktarılmasında	Konuşma sinyali	Konuşulan kelimeler

Konuşma tanıma sistemleri sadece sosyal, kültürel ve ekonomik alanda değil aynı zamanda adli, istihbari alanlarda da yaygın olarak kullanılmaktadır [5]. İstihbari alanda şüpheli

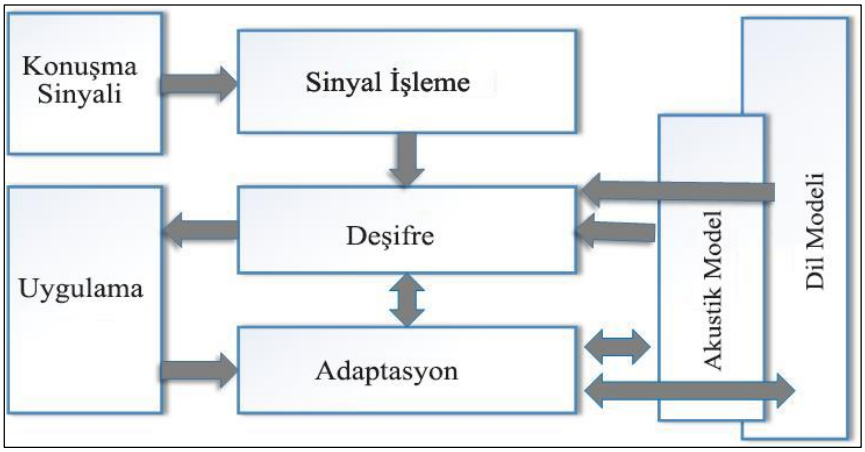
şahısların konuşmalarının OKT sistemleri ile metne aktarılması ve bu metinler üzerinde suç analizi çalışmaları yapılmaktadır. Yine istihbarat amacıyla terör gruplarına ait karasal, uydu, kablo ve internet üzerinden yapılan radyo ve televizyon yayınlarında geçen konuşmalar OKT sistemleri ile metin haline getirilmektedir. Bu metinler üzerinde kelime yakalama veya büyük veri analizi yapılabilmektedir.

Kullanım alanı her geçen gün artan OKT sistemlerinden beklenilen performans, insan kulağının konuşmaları anlayabildiği seviyeye gelmektir. OKT sistemlerinin genel yapısı karmaşıktır ve geliştirme sürecinde farklı disiplinlerin bir arada kullanılmasını gerektirir. Bu nedenle, OKT üzerinde yoğun çalışmalar yapılsa da henüz istenilen başarım seviyesine ulaşılamamıştır. Ancak bilgisayar teknolojisinin ilerlemesi ve hesaplama birimlerinin performansının arttırılması ile birlikte yaygınlaşan örüntü tanıma, işaret işleme, doğal dil işleme, görüntü işleme, metin madenciliği ve bilgi edinim teorileri alanlarında elde edilen kazanımlar OKT sistemlerinin başarımının artmasında önemli bir etken olmuştur. Özellikle hızlı, yüksek kapasiteli merkezi işlemci, grafik işlemci ve bellek teknolojilerinin performanslarının sürekli artması OKT sistemleri için gerekli olan karmaşık ve güçlü hesaplama altyapısı gereksiniminin karşılanmasını sağlamıştır. Bu teknolojik ilerlemeler ile günümüzde OKT sistemleri, karmaşık ve çok katmanlı modeller üzerine inşa edilip başarım oranları arttırılmıştır.

Klasik bir OKT mimarisinde Sinyal İşleme (Feature Extraction), Deşifre (Decoder), Akustik Model (AM, Acoustic Model) ve Dil Modeli (DM, Language Model) gibi önemli bileşenler bulunmaktadır (Şekil 1.1). Bu bileşenlerin birlikte kullanımı OKT sistemlerinin karmaşıklığını arttırmasına rağmen başarım oranları üzerinde büyük bir etkiye sahiptir [6]. Şekil 1.1'de belirtilen Uygulama bileşeni OKT sisteminin hangi alan için geliştirildiğini göstermektedir. Adaptasyon ise uygulama alanına göre yapılacak adaptasyon işlemleri için gereklidir.

OKT sistemin ilk giriş noktası olan sinyal işleme aşamasında, ses sinyalinden özellik çıkarım işlemi gerçekleştirilmektedir [7]. Özellik çıkarım işlemi bir konuşmayı farklı diğer konuşmalardan ayırmak için gerçekleştirilen önemli bir adımdır. Konuşma eyleminin doğası gereği her konuşmanın içerinde farklı bireysel özellikler bulunmaktadır [8]. Bu bireysel

özellikler, Mel Frekans Kepstral Katsayıları (MFKK, Mel Frequency Cepstral Coefficient) veya Doğrusal Öngörülü Kodlama (DÖK, Linear Predictive Coding) gibi farklı özellik çıkarım teknikleriyle elde edilmektedir.



Şekil 1.1. OKT sisteminin genel mimarisi

OKT'nin diğer bileşenlerinden biri olan Deşifre, sinyal işleme aşamasında elde edilen öznitelik vektörlerini AM kullanarak sesbirim ve DM kullanarak kelime dizilerine dönüştürmektedir. Fonem bir dildeki her bir harfin karşılığı olarak ifade edilmektedir. Deşifre işleminde kullanılan AM ve DM bileşenlerinin her biri farklı bir yapıdadır. Bu bileşenler için gerekli modeller oluşturulduktan sonra örnek bir veri kümesi üzerinden eğitim işleminin gerçekleştirilmesi gerekir. AM bileşeni ile akustik ortam bilgisi, fonetik, mikrofon, konuşmacı ve çevre değişkenliği, konuşmacılar arasındaki cinsiyet ve lehçe farklılıkları hakkında detaylı bilgiler elde edilmektedir [9]. OKT sistemine giriş olarak verilen konuşma sinyali AM sayesinde bir sesbirim çıktısı üretmektedir. Fonem çıktısı Okunuş Sözlüğü (OS, Lexicon) adı verilen bir kelime ve kelimelerin sesbirim okunuşlarını içeren bir sözlüğe gönderilmektedir. OS sesbirim dizisine karşılık hangi kelimenin gelmesi gerektiğine karar vermektedir. Ardından kelime çıktısı bir DM ile karşılaştırılmaktadır [10].

OKT için gerekli olan DM, bir sistem dizisinde hangi kelimelerin bir kelime dizisi oluşturabileceğinin, hangi kelimelerin birlikte ortaya çıkabileceğinin olasılığını vermektedir [11]. AM çıktısının DM ile karşılaştırılması işleminde DM üzerinde bir arama işlemi gerçekleştirilmektedir [12]. OKT sistemlerinde kullanılacak arama algoritmasının karmaşıklığı, DM'in getirdiği kısıtlamalarla ve belirlenen arama alanıyla doğrudan ilgilidir.

Sonlu durum dilbilgileri ve n-gramlar dâhil olmak üzere farklı dil modellerinin etkisi deşifre sonucunu doğrudan etkilemektedir.

OKT sistemleri için gerekli AM, DM ve OS'yi oluşturabilmek için genellikle insan gücü ile hazırlanmış konuşma ve birebir eşleştirilmiş metin verileri kullanılmaktadır. Bu verilerin AM ve DM'lerin eğitim işleminde kullanılabilmesi için bazı özelliklere sahip olması beklenmektedir. Verilerin birebir eşleştirilmiş konuşma ve metin karşılıklarının olması, birden fazla konuşmacı ve farklı cinsiyete sahip konuşmacılardan kayıtların alınması gerekmektedir. OKT için gerekli veri kümesinin hazırlanması farklı diller için farklı zaman, maliyet ve zorlukları beraberinde getirmektedir. Dünya genelinde konuşulan birçok dil bulunmaktadır. Bu dillerin morfolojisi birbirinden farklıdır. Aynı şekilde farklı diller için geliştirilen OKT sistemlerinin de doğası gereği farklı olması beklenmektedir.

Türkçe zengin morfolojiye sahip bir dildir. Türkçenin üretken morfolojisi, pek çok eşsiz kelime formunu beraberinde getirmektedir. Fince, Estonca ve Çekçe gibi diğer morfolojik açıdan zengin diller gibi Türkçe'nin de işlenmesi zordur. Ayrıca Türkçe geniş kelime dağarcığı sahip bir dildir. Geniş kelime dağarcığı, konuşma tanıma için gerekli tanınabilir kelime kümesinin boyutunu arttırmaktadır. Konuşma tanıma için gerekli tanınabilir kelime kümesinin boyutunun artması ile bazı sorunlar ortaya çıkmaktadır. Bu sorunlardan ilki eğitim işleminde kullanılacak veri kümesinin çok fazla konuşma ve metin verisi içermesi gerekliliğidir. AM için gerekli akustik bilgi büyük boyuttaki eğitim verisinden elde edilebilir. Diğer bir sorun ise sondan eklemeli, uzun cümleler kurulabilen ve üretken dil yapısının modellenmesidir.

Bu tezde yapılan çalışmalar temel olarak geniş kelime dağarcığına sahip Türkçe OKT için gerekli AM ve DM'nin geliştirilmesini ayrıca okunuş sözlüğünün oluşturulmasını kapsamaktadır. Geniş kelime dağarcığına sahip Türkçe OKT sisteminin geliştirilmesinde karşılaşılan iki temel sorun ele alınmıştır. Bu sorunlardan ilki geniş kelime dağarcığına sahip bir Türkçe veri kümesinin mevcut olmamasıdır. Bu nedenle OKT sistemlerinin geliştirilmesinde kullanılacak geniş kelime dağarcığına sahip bir Türkçe veri kümesi hazırlanmıştır. Tez kapsamında hazırlanan veri kümesi üç farklı yöntemle elde edilmiştir. İlk yöntemde uzun metrajlı filmler ve bu filmlere ait alt yazı belgeleri, ikinci yöntemde bir

mobil uygulama ve üçüncü yöntemde ise transfer öğrenme yaklaşımı kullanılmıştır. Oluşturulan Türkçe veri kümesi gerçek kullanıcıların onayına sunulmuştur. Bu sayede daha doğru veriler elde edilmiştir. Yeni elde edilen Türkçe veri kümesi ve mevcut veri kümeleri ayrı ayrı kullanılarak farklı Türkçe OKT sistemleri geliştirilmiştir. Geliştirilen Türkçe OKT sistemlerine ait konuşma tanıma performansları üzerinde gerçekleştirilen deneyler, sonuçların karşılaştırılabilir olduğunu ve veri kümesinin OKT üzerindeki etkisini göstermiştir.

Geniş kelime dağarcığına sahip Türkçe OKT geliştirilmesindeki bir diğer sorun ise ortam ve konuşmacı değişikliğinden etkilenmeyen bir AM'nin geliştirilmesi ve DM ile Türkçenin modellenmesidir. Türkçe OKT için gerekli AM için belirli zaman sinyali çerçevesindeki sesbirime ait sonsal olasılık hesaplanmaktadır. Akustik modelleme Gauss Karışım Modeli (GKM, Gaussian Mixture Model), Saklı Markov Modeli (SMM, Hidden Markov Model) veya Yapay Sinir Ağı (YSA, Artificial Neural Network) ile oluşturulabilmektedir. Bu çalışmada öncelikle mevcut ve yeni hazırlanan veri kümeleri kullanılarak GKM-SMM tabanlı bir Türkçe OKT sistemi geliştirilmiştir. Daha sonra Derin Sinir Ağı (DSA, Deep Neural Network) kullanılarak bir AM geliştirilmiştir. Geliştirilen AM etkisini daha net ortaya koyabilmek için OKT sistemine girdi olarak verilen konuşma dosyasındaki sessizlik bilgisi ortadan kaldırılmıştır. Ayrıca uzun bağımlılıkları önlemek için AM girişindeki konuşma dosyası parçalara ayrılmıştır. Bu yöntem kullanılarak elde edilen konuşma parçaları ayrı birer girdi olarak sıralı şekilde OKT sistemine verilmiştir. Ayrı ayrı girdi olarak verilen konuşma parçalarına karşılık gelen metin verisi birleştirilerek sunulmuştur. Yapılan deneylerde bu yaklaşımın OKT sisteminin başarımını arttırdığı görülmüştür. Ancak bazı durumlarda OKT başarımına olumsuz etki ettiği de belirlenmiştir. Başarıma olumsuz etki etmesinin en büyük nedeni ise çok kısa konuşma parçalarından istenilen konuşma özellik bilgilerinin elde edilememesidir.

Bu tez çalışmasında hazırlanan AM, DM ile desteklenmiştir. Ayrıca bir OS, OKT sistemine eklenmiştir. Türkçe OKT sistemi için gerekli olan DM cümle düzeyinde ele alınmıştır. Öncelikle N-Gram tabanlı ve daha sonra YSA temelli bir DM oluşturulmuştur. DM'nin en iyi sonuç veren listeleri iyileştirilmiştir. Bu iyileştirme için farklı bir yöntem önerilmiştir. Önerilen yöntem skip-gram ve n-gram özelliklerini kullanmaktadır. OKT için gerekli bütün bileşenlerde farklı materyal ve metotlar kullanılmıştır. OKT'nin her aşamasında iyileştirme

yapılmıştır. Geniş kelime dağarcığına sahip bir Türkçe OKT sistemi geliştirildikten sonra hazırlanan web servis tabanlı bir platform ile kullanıma sunulmuştur.

Tez çalışmasının geri kalanı şu şekilde düzenlenmiştir. Konu ile ilgili çalışmalar bölüm 2'de sunulmuştur. Bölüm 3'te OKT sistemlerinin genel yapısı, konuşma tanıma sistemlerinde kullanılan yöntemler, modeller ve çalışma kapsamında kullanılan materyaller ile ilgili bilgi verilmiştir. Bölüm 4'te tez çalışması için gerekli olan veri kümesinin hazırlanması ve doğrulanması açıklanmıştır. Bölüm 5'de geniş kelime dağarcığına sahip Türkçe OKT sisteminin ve web servis tabanlı Türkçe OKT platformunun geliştirilmesine yer verilmiştir. Ayrıca sistem geliştirilirken gerçekleştirilen deneyler ve deneyler sonucunda elde edilen bulgular bu bölümde açıklanmıştır. Son olarak ise çalışma kapsamında elde edilen sonuçlar verilerek gelecek çalışmalar için önerilerde bulunulmuştur.

2. İLGİLİ ÇALIŞMALAR

Konuşma işleminin mekanik veya elektronik cihazlar tarafından algılanması ve insan-makine etkileşimi gerektiren görevlerin sadece konuşma ile yapılabilme arzusu OKT alanındaki araştırmaların 1950 yılından itibaren büyük ilgi görmesini sağlamıştır. Konuşmanın istatistiksel ve YSA tabanlı modellemesindeki büyük ilerlemeler OKT sistemlerinin başarımını arttırmıştır. Artan başarımları sayesinde çağrı merkezi ve otomatik kontrol sistemleri gibi insan-makine etkileşimini gerektiren görevlerde yaygın olarak kullanılmıştır. Bu bölümde, teknolojik ve bilimsel bir derinlik sağlamak için geçmişten günümüze OKT sistemleri ile ilgili bilgi verilmiştir.

OKT sistemlerinin geliştirilmesine yönelik ilk girişimler, akustik sesbirim ile ilgili temel fikirlerin kullanıldığı 1950'li yıllara dayanmaktadır. Bu yıllarda araştırılan konuşma tanıma sistemlerinin çoğunda her bir ifadenin ses bölgesindeki spektral genlik frekansları kullanılmıştır. Bu frekanslar analog bir filtre bankası ve mantık devrelerinin çıkış sinyallerinden elde edilmiştir. 1952'de Bell Laboratuvarında sadece tek bir konuşmacının sesli harflerinin tanınması, 1956'da RCA Laboratuvarında ise tek bir konuşmacının on adet tek heceli kelimesi tanınmaya çalışılmıştır [13]. 1959'da İngiltere'deki Üniversite Koleji'nde dört sesli harf ve dokuz ünsüz harfi tanımak için bir fonetik tanıyıcı geliştirilmiştir [14]. Bu çalışma, OKT sistemlerinde sesbirim düzeyinde istatistiksel sözdiziminin ilk kullanımını göstermiştir. 1959'da Massachusetts Teknoloji Enstitüsü Lincoln Laboratuvarında konuşmacıdan bağımsız ancak belirli bir kalıba uygun on sesli harf tanıyabilen bir sistem geliştirilmiştir [15].

1960'larda bilgisayarlar hala yeterince gelişmiş hesaplama yeteneğine sahip olmadığı için konuşma tanıma amacıyla özel amaçlı donanımlar geliştirilmiştir. Japonya'daki Radio Laboratuvarında sesli harfleri tanıyan bir donanım[16] ve Kyoto Üniversitesinde 1962 yılında sesbirim tanıyan bir donanım[17] geliştirilmiştir. Konuşma tanımanın en önemli zorluklarından biri konuşma olaylarında zaman ölçeklerinin eşit olmamasıdır. 1960'lı yıllarda Martin ve RCA Laboratuvarındaki arkadaşları konuşma başlangıç ve bitişlerini güvenilir bir şekilde tespit etmeye dayanan temel bir zaman normalizasyon yöntemi

sunmuşlardır [18]. Ardından ilk konuşma tanıma şirketlerinden biri olan Threshold Technology, Brian Martin tarafından kurulmuştur.

1960 ve 1970'lerin sonlarına doğru konuşma tanıma alanında önemli ve başarılı çalışmalar gerçekleştirilmiştir. 1968 yılında Sovyetler Birliği'nde Vintsyuk, zaman serisi analizi olarak da bilinen Dinamik Zaman Atlatılması (DZA, Dynamic Time Warping) yöntemini sunmuştur [19]. DZA değişebilir iki zamansal dizi arasındaki benzerliği ölçmek için kullanılabilen bir yöntemdir. Dinamik zaman analizi yaklaşımlarının konuşma tanıma görevinde kullanılması aynı tarihlerde yaygınlaşmıştır. Japonya'daki NEC Laboratuvarlarında Sakoe ve Chiba, tekdüzelik sorununu çözmek için dinamik bir programlama tekniği kullanmaya başlamıştır [20]. 1970'lerin sonlarında Viterbi dâhil olmak üzere birçok dinamik programlama yaklaşımı OKT alanında kullanılmıştır [20]. Viterbi, en olası gizli durum dizisini bulmak için dinamik bir programlama algoritması olarak bilinen ve Viterbi yolu adı verilen bir yaklaşımdır. Dinamik programlama yaklaşımındaki bu gelişmelerin ardından bir şablona yerleştirilmiş ve gürültü gibi dış etkilerden izole edilmiş kelime tanıma yaklaşımı yaygınlaşmıştır.

İzole edilmiş kelime ya da ayrık söyleniş tanıma alanı Sovyetler Birliği ve Japonya'daki temel çalışmalara dayanmaktadır. Sovyetler Birliği'nde Velichko ve Zagoruyko, konuşma tanımada örüntü tanıma fikirlerinin kullanılması gerektiğini savunmuştur [21]. Itakura, Bell laboratuvarlarında DÖK spektral parametrelerine dayanan uygun bir mesafe ölçüsü sayesinde DÖK'ün OKT sistemlerinde nasıl uygulanabileceğini göstermiştir [22]. Sürekli konuşma tanıma alanında ise 1960'ların sonunda Carnegie Mellon Üniversitesi'nde Reddy, sesbirimlerin dinamik takibini kullanarak sürekli konuşma tanıma alanında öncü araştırmalar yapmıştır [23]. Diğer yandan AT&T Bell Labs, konuşmacıdan bağımsız konuşma tanıma sistemleri geliştirmeye yönelik bir dizi deney başlatmıştır [24]. Bu hedefe ulaşmak için geniş bir kullanıcı kümesinden farklı kelimelerin tüm varyasyonlarını temsil eden veri kümeleri ihtiyacı ortaya çıkmıştır. Bu ihtiyacı karşılayabilmek için ilk önemli adımı Japonya atmıştır [25]. Ardından TIMIT ve LibriSpeech adında İngilizce konuşma-metin veri kümeleri geliştirilmiştir. TIMIT ve LibriSpeech, farklı cinsiyet ve lehçelerden gelen Amerikan İngilizcesi konuşanların fonetik ve dil bilimsel olarak yazıya dökülmüş bir konuşma veri kümesidir.

Japonya ve Sovyetler Birliği'ndeki gelişmelerden sonra Amerika'da birçok gelişmiş sistem ve teknolojinin çıkış noktası olan Savunma Gelişmiş Araştırma Projeleri Ajansı (SGAPA, Defense Advanced Research Projects Agency) tarafından bir konuşma tanıma projesi finanse edilmiştir [26]. SGAPA tarafından finanse edilen bu konuşma tanıma projesinde tanıyıcı tarafından dikkate alınan alternatiflerin sayısını önemli ölçüde azaltmak için semantik bilgiler kullanılmıştır. 1.011 kelimelik bir kelime dağarcığını kullanarak kabul edilebilir doğrulukta konuşmayı tanıyabildiği gösterilmiştir. Bu proje, hesaplamayı azaltmak ve en yakın eşleşen dizeyi verimli bir şekilde belirlemek için Sonlu Durum Ağından (SDA, Finite-State Network) yararlanan ilk sistemdir.

Konuşmacıdan bağımsız OKT sistemleri üzerine çalışmalar devam ederken akıcı bir şekilde birbirine bağlı bir kelime dizisinin tanınabilmesi sorunu ortaya çıkmıştır. Akıcı kelimeleri tanıyan bir sistem oluşturma sorunu 1980'lerde araştırmaların odak noktası olmuştur. Ardışık kelimelerin sıralı bir modelini eşleştirmeye dayanan çok çeşitli algoritmalar geliştirilmiştir. Bu algoritmaların başında Sakoe'nin iki seviyeli dinamik programlama yaklaşımı, Bridle ve Brown'un tek geçiş yöntemi [27], Myers ve Rabiner'ın Bell Laboratuvarında geliştirdikleri seviye oluşturma yaklaşımı [28], Lee ve Rabiner'ın Bell Laboratuvarlarında geliştirdikleri kare senkronize seviye oluşturma yaklaşımı [29] gelmektedir. Bu en iyi eşleştirme yöntemlerinin her birinin uygulama alanına göre avantajları mevcuttur. Ancak ardışık kelimelerin sıralı bir modelini eşleştirmeye dayanan yaklaşımların performans verimliliği istenilen seviyeye ulaşamamıştır.

1980'lerde konuşma tanıma araştırmaları sezgisel şablon tabanlı yaklaşımdan istatistiksel modelleme çerçevesine doğru geçiş evresini yaşamıştır. Günümüzde dahi en pratik konuşma tanıma sistemleri 1980'lerde geliştirilen istatistiksel yöntemlere dayanmaktadır. Bu istatistiksel yöntemlere 1990'larda önemli iyileştirmeler yapılmıştır. İyileştirmelerin başında SMM gelmektedir. SMM 1980'lerde geliştirilen istatistiksel modellemeye dayanan anahtar teknolojilerden biridir [30,31].

SMM, gözlemlenebilir olmayan ancak bir dizi gözlem üreten ve başka bir rastlantısal süreç yoluyla gözlemlenebilen bir süreci ifade etmektedir. SMM başta IBM ve SGAPA olmak üzere birçok araştırma laboratuvarında sıklıkla kullanılmaya başlanmıştır. SMM tabanlı

yaklaşımların kullanılmasının ardından dil modellemesi üzerine çalışmalar gerçekleştirilmiştir. IBM'in OKT için birincil odağı, istatistiksel tabanlı dil kuralları ile temsil edilen bir dil model yapısının geliştirilmesi olmuştur. 1985 yılında N kelime dizisinin ortaya çıkma olasılığını tanımlayan N-gram modeli tanıtılmıştır [22]. Aynı yıllarda sinir ağlarının konuşma tanımaya uygulama fikri yeniden gündeme gelmiştir. Aslen psikolog olan Rosenblatt yapay bir nöronun eğitilebilir olduğunu göstermiştir [32]. Böylelikle SMM durumlarının YSA üzerinde inşa edilebilmesinin önü açılmıştır. Nöron eğitiminin ardından bir ya da iki gizli katmanı olan sinir ağlarını eğitmek için geri yayılma ile bağlantılı yaklaşımlar geliştirilmiştir [33]. Rumelhart ve arkadaşları derin sinir ağlarını eğitmek için geri yayılım kullanmış ve geri yayılma algoritmasının popülerleşmesini sağlamıştır. Yapay sinir ağlarının eğitilebilme problemleri giderildikten sonra OKT sistemlerini geliştirebilmek için farklı yaklaşımlar sunulmuştur. Ancak sinir ağları için gerekli hesaplama alt yapısının uygun olmaması 1990'lı yıllarda istatistiksel temelli yaklaşımların sinir ağlarına göre daha popüler hale gelmesini sağlamıştır.

1990'larda diğer bir konuşma tanıma sorunu olan ses verisi için dağılım tahminini gerektiren örüntü tanıma sorunu ele alınmıştır. Örüntü tanıma sorunu deneysel olarak öğrenilen tanıma hatasının en aza indirilmesini içeren en uygun hale getirme problemine dönüştürülmüştür [34]. 2000'li yılların başında ise kendiliğinden konuşma tanıma problemi ele alınmıştır. Her ne kadar bir metnin okunarak seslendirilmesi en gelişmiş konuşma tanıma teknolojisi araçları kullanılarak %95'ten daha yüksek doğrulukla tanınabilse de kendiliğinden konuşma için bu değer önemli ölçüde azalmaktadır. Konuşma tanıma uygulamalarının genişletilmesi, büyük ölçüde kendiliğinden konuşma için tanıma performansının artırılmasına bağlıdır. Kendiliğinden konuşma tanıma performansını artırmak için çeşitli projeler gerçekleştirilmiştir. Japonya'da 5 yıllık ulusal bir veri kümesi elde etme projesi yürütülmüştür [35]. Kendiliğinden konuşma veri kümesi olan 700 saatlik konuşmaya karşılık gelen yaklaşık 7 milyon kelimeden oluşan metin veri kümesi elde edilmiştir. Geliştirilen bu veri kümesi üzerinde yeni teknikler uygulanmıştır. Bu yeni teknikler arasında akustik modelleme, cümle sınırı tespiti, telaffuz modellemesi, akustik ve dil modeli uyarlaması yer almaktadır [25].

Kendiliğinden konuşmanın tanınması çalışmalarının ardından konuşma tanıma alanındaki diğer bir sorun olan konuşma sağlamlığı ile ilgili çalışmalar da gerçekleştirilmiştir. Konuşma

tanıma sistemlerinin sağlamlığını daha da arttırmak için söylem doğrulaması ve güven önlemleri yoğun bir şekilde kullanılmıştır [36]. Güven önlemi, sistemin kullanıcılardan aldığı bilgiye göre uygun bir yanıt sağlamak için referans kılavuz kullanmasını gerektirmektedir. Anlamsal olarak önemli parçaları tespit etmek ve kendiliğinden konuşulan ifadelerde alakasız kısımları reddetmek için tespit tabanlı bir yaklaşım önerilmiştir [37]. Bu birleşik tanıma ve doğrulama stratejisi özellikle kötü seslendirilmiş konuşma ifadeleri için iyi sonuçlar vermiştir. Akustik ortamlardan elde edilen bilgileri daha iyi modellemek ve gürültü sağlamlığını arttırmak için akustik model üzerine farklı yaklaşımlar sunulmuştur. Geleneksel SMM'lerden daha gelişmiş akustik modeller oluşturmak için dinamik Bayes ağı geliştirilmiştir [38].

1980'li yıllardan bu yana özellikle 2012 yılına kadar konuşmaların tanınması için geliştirilen modellerde SMM kullanılmıştır. Yakın geçmişte SMM'ler ile GKM'lerin birlikte kullanılması ise başarılı sonuçlar alınmıştır. GKM'ler ile akustik özellikler ve sesbirimler arasındaki ilişki modellenirken SMM'ler ile konuşmanın sıralı yapısı elde edilmiştir. 1980'lerin sonlarında ve 90'ların başında ise sinir ağı temelli konuşma tanıma modelleri geliştirilmiştir. Bu sinir ağı temelli modellerin ilk yıllarında sinir ağı tabanlı konuşma tanımanın başarımı GKM-SMM tabanlı sistemlerin performansıyla benzerlik göstermiştir. Robinson ve Fallside 1991 yılında TIMIT veri kümesi üzerinde sinir ağı yöntemini uygulamış ve %26 sesbirim hata oranı elde etmişlerdir [39]. Akustik özelliklerin sesbirimlere ilişkilendirilmesinde çok daha büyük modellerde yani daha derin modellerde ve çok büyük veri kümeleriyle sinir ağları eğitildiğinde, sinir ağlarının GKM'lerden daha iyi sonuçlar verdiği görülmüştür [4,40]. 2009'dan başlayarak RBM'ler, sabit boyutlu bir giriş penceresinde spektral akustik temsillerin alınmasında ve SMM durumlarının koşullu olasılıklarının tahmin edilmesinde kullanılmıştır [41]. TIMIT veri kümesindeki ilk sonuçlar, bu tür derin ağların eğitilmesinin aslında TIMIT üzerindeki SMM tanıma oranını önemli ölçüde artırmaya yardımcı olduğunu göstermiştir. Sesbirim hata oranını %26'dan yaklaşık %23'e düşürmüştür [42].

Sesbirimlerin tanınmasındaki başarım oranlarının artmasından sonra sesbirim dizilerinin tanınmasıyla ilgili çalışmalar gerçekleştirilmiştir. Bu çalışmalar gerçekleştirilirken IBM, Bell ve NEC gibi önemli konuşma tanıma araştırma gruplarının birçoğu akademik araştırmacılarla işbirliği içinde Derin Öğrenmeyi (DÖ, Deep Learning) keşfetmeye

başlamıştır. DÖ, bilgisayarın daha basit kavramlardan daha karmaşık kavramlar oluşturmasını sağlamıştır. DÖ, geleneksel makine öğreniminden daha fazla miktarda öğrenilmiş fonksiyon bileşimini veya öğrenilmiş kavramları içeren modellerin incelenmesi olarak güvenli bir şekilde kabul edilmiştir. DÖ başlıklı kitabın yazarlarına göre makine öğrenimi; karmaşık, gerçek dünyadaki ortamlarda çalışabilen yapay zekâ sistemleri oluşturmak için tek geçerli yaklaşımdır. DÖ ise dünyayı kavramların ve temsillerin iç içe bir hiyerarşisi olarak temsil etmeyi öğrenerek, her bir kavramı daha basit kavramlarla ilişkili olarak tanımlanmasıdır. Ayrıca daha soyut temsillerle büyük güç ve esneklik sağlayan özel bir tür makine öğrenimi olduğu belirtilmiştir [43].

DÖ tabanlı yaklaşımlar özellikle görüntü işleme, konuşma tanıma, dil modelleme, ayrıştırma, konuşma çevirisi, araç otomasyonu ve benzeri birçok araştırma alanında yüksek başarım göstermiştir. DÖ yaklaşımlarının yüksek başarımının nedeni büyük miktarda veri kümesi ile eğitildiğinde çok karmaşık ve birleşik yapıları bulabilme ve farklı öğrenme yetenekleridir [44]. DÖ yaklaşımları ile birden fazla katmana sahip derin yapıdaki modeller, OKT için gerekli olan modelleri geliştirmek için kullanılmıştır [45]. Fohr ve arkadaşları derin sinir ağlarını SMM ile birleştirmiş ve farklı bir mimari sunmuştur [45]. Bu mimari kurgu, haber yayınlarından oluşan bir veri kümesi ile eğitilmiştir. Elde edilen başarım oranları tek başına kullanılan SMM veya derin olmayan yapay sinir ağlarına göre çok daha yüksektir.

DÖ kavramı ortaya çıktıktan sonra derin ağların mimarisinde kullanılmak üzere daha büyük etiketlenmiş veri kümelerine olan ihtiyaç arttırmıştır. Sürekli büyüyen eğitim veri kümesi boyutu konuşma tanıma ile ilgili konferanslarda DÖ'ye hızlı bir geçiş sunmuştur [39,46-48]. DÖ, konuşma tanımaya yönelik araştırmaların çoğundaki yenilikler sunmuştur ve halen bu yenilik dalgası devam etmektedir. Bu yeniliklerden biri tam bağlantılı ileri geri sinir ağlarının kullanılmasının yerine evrişimsel sinir ağlarının kullanılmasıdır [49,50]. Bu kullanımda, giriş bilgisi uzun bir vektör olarak değil, bir zaman eksenine ve spektral bileşenlerin frekansına karşılık gelmektedir.

DÖ yaklaşımları çoğunlukla İngilizce, Çince ve İspanyolca gibi kaynak bakımından zengin diller için sıklıkla uygulanmıştır. Türkçe üzerine gerçekleştirilen OKT araştırmaları

Türkçe'nin sondan eklemeli bir yapıya sahip olması nedeni ile yetersiz kalmaktadır. OKT sistemlerinde DÖ tabanlı yaklaşımlar her ne kadar başarım oranlarını arttırsa da hala istenilen seviyeye ulaşılamamıştır. Özellikle geniş kelime dağarcığına sahip dillerde başarım oranları daha da düşmektedir. Bu nedenle geniş dağarcık problemini çözmek amacıyla heceleme veya gövde sonlarındaki ekleri ortadan kaldırarak çözümleme gibi yaklaşımlar önerilmiştir [51].

Konuşma tanıma görevinde öncelikle sesbirimlerin, sesli ve sessiz harflerin tanınması gerçekleştirilmiştir. Ancak konuşma sağlamlığı ve kendiliğinden konuşma gibi sorunlar ortaya çıkmış ve çözümler aranmıştır. İstatistiksel yaklaşımların yanı sıra sinir ağı yaklaşımlarının kullanılması yaygınlaşmıştır. Derin sinir ağlarının başarımı ve DÖ yaklaşımının konuşma tanımaya uygulanması ile OKT sistemlerinin başarımı arttırılmıştır. Ancak daha derin ve karmaşık modellerin OKT sistemlerinde kullanılması bazı sorunları ortaya çıkarmıştır. Bu sorunların başında yüksek hesaplama ihtiyacının karşılanamaması gelmektedir. Merkezi İşlem Birimlerinin (MİB, Central Processing Unit) bu konudaki yetersizliği derin ve karmaşık sinir ağlarının eğitimi işleminin Grafik İşlem Birimlerine (GİB, Graphic Processing Unit) aktarılması ile giderilmiştir. Raina ve arkadaşlarının 2009 yılında gerçekleştirdiği bir çalışmada MİB yerine GİB kullanmanın işlem hızının 70 kat daha hızlı olabileceği gösterilmiştir [52]. Böylelikle OKT sistemleri için gerekli olan hesaplama ihtiyacı sağlanmış ve çok katmanlı mimarilerin geliştirilmesine olanak sağlanmıştır.

Önceki çalışmalar istatistiksel yaklaşımlardan ve DÖ tabanlı yaklaşımlara kadar detaylı olarak incelenmiştir. DÖ tabanlı yaklaşımların yaygınlaşması ile gerçek hayattaki problemler daha iyi modellenmeye başlamış ve çözüm aranmıştır. Diğer yandan geliştirilen OKT araç setleri OKT alanında gerçekleştirilen çalışmalar için kolaylıklar sağlamıştır. Örneğin SMM Araç Seti (SAK, Hidden Markov Model Toolkit) kullanılarak konuşma bozukluğu olan ve Malayca konuşan çocuklar üzerine bir çalışma gerçekleştirilmiştir [53]. Malay-Polinezya dilleri, ayrıca Avustronezya dilleri olarak da bilinen batı alt familyasına ait fonetik bir dildir. Bu çalışmada veri hazırlama görevleri, dilbilgisi ve okunuş sözlüğü SAK araç seti ile geliştirilmiştir. Çalışma kapsamında konuşma engelli bireylerin konuşma anlaşılırlığını ölçmede OKT uygulamalarının potansiyeli araştırılmıştır. Farklı bir çalışmada ise çocukların konuşmalarını tanımak için yetişkinlerden elde edilen veriler kullanılmıştır [54]. Sistemin eğitimi için Mandarin yetişkin konuşma eğitim kümesi "King-ASR-118

mobile speech corpus”dan alınan veriler kullanılmıştır. Çocuk konuşmalarının işlenmesi, büyük ölçekli çocuk konuşma verilerinin bulunmaması nedeniyle yetişkinler için konuşma tanıma görevinden daha zorlayıcıdır. Diğer yandan Lin ve arkadaşlarının boğaz mikrofonu kullanan kişiler üzerine uyguladıkları OKT sistemi dikkat çekicidir [55]. Geleneksel bir akustik mikrofon sinyaline göre boğaz mikrofonunun işlenmesi zorlu bir görevdir. Bunun gibi zorlu görevlerin üstesinden gelebilmek için gerçekleştirilen ilk çalışmalar, zaman gecikmeli çok katmanlı yapay sinir ağlarının küçük harf grupları arasında ayrımcılık sağlayabileceğini göstermiştir. Konuşma tanıma sağlamlığını arttırmaya yönelik bir çalışmada ise düşük bant konuşma işlemi ve akustik modelleme yönteminin özelliklerinden faydalanılmıştır [56]. Çalışmada Gürültülü ortamlarda farklı filtrelerin kullanılması ile elde edilen aurora2 veri kümesi kullanılmıştır.

Literatürde OKT sistemlerinin çok farklı diller üzerine inşa edildiği görülmüştür. Ayrıca farklı dillerdeki konuşma tanıma görevleri için farklı ortam (gürültü veya diğer akustik parametreler) değerlendirmeleri için yapılan çalışmalar da mevcuttur. Galic ve arkadaşlarının gerçekleştirmiş olduğu kısık sesli konuşmaların tanınması çalışması [57], Shahnawazuddin ve arkadaşlarının gürültülü ortamlarda çocukların konuşmalarının tanınması için yaptıkları çalışma [58] örnek olarak verilebilir. Ayrıca Google tarafından geliştirilen görsel ve işitsel bilgilerin aynı anda kullanıldığı bir projede gürültülü ortamlardaki konuşmaların daha yüksek başarım oranları ile metne aktarılması çalışılmıştır [59]. OKT işlemi konuşma-metin verileri ile birebir eşleştirilen görsel anlatım ile desteklenmiştir. Bu çalışma da ana sorun gürültülü ortamda (üst üste binen karşılıklı konuşmalar, alkış veya yüksek sesle konuşarak diğer konuşmaları bastırmak) geçen konuşmayı tanımlayabilmektir. Bu sorunu çözebilmek için gerçek hayatta olduğu gibi sadece bir konuşmacıya yoğunlaşarak ve konuşmacının yüz hareketleri ele alınarak bir OKT sistemi geliştirilmiştir. Böylelikle gerçek hayatta olduğu gibi gürültülü bir ortamda olunsa dahi hem görsel iletişim hem de duyusal iletişim ile OKT sistemlerinin başarımı arttırılmaya çalışılmıştır [59].

OKT sistemlerinin başarımını artırmaya dayalı çalışmalar yapılırken sadece yetersiz veri kaynağı değil yetersiz donanım kaynağı da bir sorun olarak karşımıza çıkmaktadır. OKT sistemlerinin daha derin ve daha karmaşık yapılar ile gerçek hayattaki problemleri modelleyebilmesi ve hesaplama yoğunluğunun GİB üzerine aktarılması ile yetersiz donanım

sorunu nispeten giderilmiştir. OKT sistemlerinin gerçek hayatta kullanımını yaygınlaştırmak için çevrimiçi OKT sistemleri geliştirilmeye başlanmıştır [60,61]. Ancak OKT sistemlerinin çevrimiçi kullanılması yüksek hesaplama ihtiyacını beraberinde getirmektedir. Çevrimiçi kullanımda OKT için gerekli olan donanım kaynağı genellikle uzak ortamlarda bulunan güçlü sunucular ile sağlanmıştır. Bu nedenle çevrimiçi OKT sistemleri için sunduğu yüksek hesaplama alt yapısı sayesinde bulut tabanlı OKT uygulamaları geliştirilmiş ve yaygınlaşmıştır.

OKT sistemlerinin yaygınlaşması ve ekonomik kazanımların ortaya çıkması ile bu alana yapılan yatırımlar artmıştır. Amazon firması "Amazon Triscribe", Microsoft firması "Azure Bing Speech", Google firması "Cloud Speech-to-Text", IBM ise "Watson Speech-to-Text" projesini başlatmıştır [62]. Bu kapsamda bulut platformu üzerinden sunulan uygulama programlama arayüzleri sayesinde farklı uygulamalar geliştirilmiştir. Özellikle Google'un sunduğu "Ok Google", Microsoft'un sunduğu "Cortana", IBM'in sunduğu "Watson" ve Apple'ın sunduğu "Siri" kişisel asistan uygulamaları en iyi örneklerdendir. Ayrıca bulut tabanlı OKT servis sağlayıcıları sundukları kullanıcı programlama arayüzleri yardımı ile OKT sistemlerinin kullanımı yaygınlaşmıştır. Ancak kullanıcılar bulut tabanlı OKT sistemlerinin gizliliğinden ve güvenliğinden tam anlamı ile emin değildirler. Bu nedenle çevrimdışı OKT sistemlerin yetersiz donanım kaynağına rağmen başarımlarını arttıracak yaklaşımlar gerçekleştirilmelidir [63].

OKT sistemleri üzerinden sadece konuşma bilgisinin metne dönüştürülmesi değil farklı çalışmalarda gerçekleştirilmiştir. Örneğin, konuşma duygusunun tanımlanması [64,65], negatif etki ve saldırganlığın otomatik olarak tanımlanmasını sağlayan konuşma analizinin yapılması [66] ve cinsiyet belirlenmesi [67] gibi çalışmalar da mevcuttur. Ayrıca aksan tanıma çalışmaları da OKT sistemlerinin başarımını artırmada önemli rol oynayacağı gibi aynı zamanda konuşmacı hakkında detaylı bilgiler vermektedir [68]. Aksan tanıma çalışmalarında dil bilimsel bir yaklaşım izlenmektedir. Bir dile ait bir kelimenin okunuş biçiminde sergilenen morfolojik yaklaşım o konuşmacının konuştuğu aksan hakkında bilgiyi içermektedir [46]. Sagha ve arkadaşlarının yaptığı bir çalışmada ise konuşma bilgisi üzerinden yaş bilgisi, cinsiyet bilgisi ve konuşma bilgisinin OKT sistemleri üzerindeki performansı araştırılmıştır [69].

Konuşma bilgisi üzerinden yaş ve cinsiyetin tespiti OKT sistemlerinin başarımı arttıran önemli faktörlerdendir [70]. Cinsiyet ve yaşa göre akustik bilgiler değişmektedir. Ön işlem olarak cinsiyetin belirlenmesi ve akustik bilginin belirlenen cinsiyete özel üretilen modele girdi olarak verilmesi başarımı arttırmaktadır. Cinsiyet ve yaş bilgisi ile başarım oranları arttırılsa da gerçek hayatta sürekli olarak kullanılan bir durumun modellenmesi sorunu ortaya çıkmıştır. Bu sorun gerçek hayatta insanların birbirleriyle konuşurken kullandığı çok dilli iletişim sorunudur [71,72]. Konuşmanın anlaşılması üzerine yapılan çalışmalar genellikle hem görsel hem de işitsel bilgilere dayandırılmıştır. Özellikle konuşma bilgisi karmaşık olduğunda veya gürültülü bir ortamda iletişim gerçekleştiğinde, görsel bilgiden faydalanılması konuşma tanıma başarımını arttırmıştır. Konuşma tanımada görsel yüz bilgisinin, özellikle dudak bilgisinin kullanımı araştırılmış ve sonuçlar her iki bilgi türünün kullanılmasının daha iyi tanıma performansı sağladığını göstermiştir [59].

Sonuç olarak daha önceleri sesbirim ve harf tanıma ile başlayan ardından sesbirim dizilerinin tanınması ile gelişim yoluna devam eden konuşma tanım üzerine birçok farklı çalışma yapılmıştır. Fonem dizilerinin tanınabilmesi ile kelimelerin tanınması ve istatistiksel dil modellerinin geliştirilmesi ile cümlelerin tanınabilmesi sağlanmıştır. İstatistiksel yaklaşımların yerine YSA ve DÖ tabanlı yaklaşımların geliştirilmesi ile konuşma tanıma alanında gerçek dünya problemleri çözülmeye başlamıştır. Konuşmsa bilgisi üzerinden duygu tanıma, sinirlilik durumu algılama, kısık sesli kişiler için OKT, gürültülü ortamlar için OKT gibi birçok çalışmada klasik SMM modelleri ile birlikte derin sinir ağı yapılarının kullanıldığı görülmüştür [40,73,74]. En iyi performans gösteren değerleri seçmek için bir dizi ağ mimarisi ve öğrenme parametresinin hazırlanması gerektiği belirtilmiştir. Yapılan birçok çalışma araştırmalara açık veri setleri üzerine gerçekleştirilmiştir [72,73,75]. Ayrıca gürültü sağlamlığı, konuşmacı bağımsızlığı, kendiliğinden konuşma ve çok dilli konuşma problemleri üzerine araştırmalar yapılmıştır.

Konuşma tanıma alanı ile ilgili birçok çalışma yapılsa da OKT başarımının tam olarak istenilen doğruluk ve hesaplama seviyesine geldiği söylenememektedir. Özellikle Türkçe üzerine yapılan çalışmalarda veri kümesi boyutlarının çok küçük olması Türkçe OKT sistemlerinin başarımını düşürmektedir [76,77]. Ayrıca mevcut yöntemlerin Türkçe OKT üzerindeki performansının değerlendirilmesini zorlaştırmaktadır. Bu nedenle öncelikle Türkçe üzerine geniş kelime dağarcığına sahip bir veri kümesi hazırlanmalı ve mevcut

yöntemlerin Türkçe OKT üzerine etkisi araştırılmalıdır. Geniş kelime dağarcığına sahip Türkçe DM, okunuş sözlüğü ve AM birlikte kullanılarak bir OKT sistemi geliştirilmeli ve eğitim veri kümesi sürekli genişletilmelidir.

3. MATERYAL VE METOTLAR

Bu bölümde öncelikle konuşma bilgisinin doğal üretimine yer verilmiştir. İnsanlar tarafından üretilen konuşma bilgisinin OKT sistemleri ile tanınabilmesi için gerekli yapılar açıklanmıştır. OKT sisteminin genel yapısı, konuşma tanıma sistemlerinde kullanılan yöntemler, modeller ve materyaller ile ilgili bilgiler verilmiştir. Konuşma tanıma sistemlerinin genel mimarisi, bileşenleri ve kullanılan algoritmalar incelenmiştir. Konuşma tanıma sistemlerinde doğruluk oranı ölçümü için kullanılan metrikler belirlenmiştir.

3.1. Konuşma Bilgisi Üretimi

İnsanlar tarafından üretilen konuşma bilgisi akciğer, ses telleri ve ağız tarafından oluşturulmaktadır. Öncelikle akciğer, solunum sayesinde konuşma için gerekli enerjiyi üretmektedir. Ses telleri bu enerjiyi duyulabilir sese çevirme görevini üstlenmektedir. Bu işleme kısaca ses üretimi denilmektedir. Ses üretimi aşamasında gırtlak kullanılmaktadır. Ağız ve burundaki boşluklar ses telleri tarafından üretilen bilgiyi insan kulağı tarafından anlaşılabilir konuşmaya dönüştürmektedir. Bu işleme genellikle söyleyiş veya boğumlanma işlemi denilmektedir. Anatomik yapısından dolayı her insanın söyleyiş biçimi diğerlerinden farklıdır. Ancak konuşma bilgisinin üretim kaynağı benzerdir.

Konuşma bilgisinin kaynağı, akciğerlerden dışarı atılan hava olarak bilinmektedir. Solunum sırasında ses telleri hava akışını serbest bırakmak için glottis adı verilen bir boşluk oluşturur. Bu boşluk içerisinde çok az duyulabilen veya hiç duyulamayan bir ses ortaya çıkmaktadır. Sesler çıkarılırken ses telleri glottis'i kapatarak akciğerlerin yüksek basınçla dolmasına neden olmaktadır. Bu basınç yeterince yüksek olduğunda ses telleri birbirinden ayrılmaktadır. Ses telleri elastik ve aerodinamik yapısından dolayı geri yayılım, tekrar basınç oluşturma ve ses tellerinin tekrar açılması gibi görevleri yerine getirmektedir. Ses üretimi (fonasyon) olarak bilenen bu periyodik süreç ötümlü seslerin üretilmesini sağlamaktadır. Bu periyodik süreçteki salınımın temel frekansı, ses tellerinin uzunluğu ve kütlesi ile belirlenirken ses tellerinin gerginliği ile kontrol edilmektedir. Erkeklerde temel frekans 80-200 Hz arasındayken, kadınlarda 150-300 Hz arasındadır. Gırtlaktan üretilen ses ise gırtlağın üzerindeki ses yolunda bulunan zar (yutak içine doğru burun boşluğu bağlantısını açar ve

kapatır), diş, dil, dudak ve çene pozisyonları üzerinde oynanarak değişime uğratılabilir. Bu organların pozisyonuna bağlı olarak farklı tınlama frekansları oluşmaktadır. Bu tınlama frekansları Formant olarak adlandırılmaktadır. Formant konuşmanın ve konuşmacının anlaşılabilir olması için gereklidir [78]. Konuşma analizinde Formant kavramı, gırtlaktaki tınlaşım olarak bilinmektedir. Ayrıca ses sinyalini oluşturan belirgin frekans bileşenleri de Formant olarak tanımlanmaktadır. Harflerin tanınması için önemli özniteliklerden birisi de Formant bilgisidir. En düşük frekansta bulunan Formant F1, ikinci seviyede bulunan Formant F2 ve üçüncü seviyede bulunan Formant ise F3 ile ifade edilmektedir.

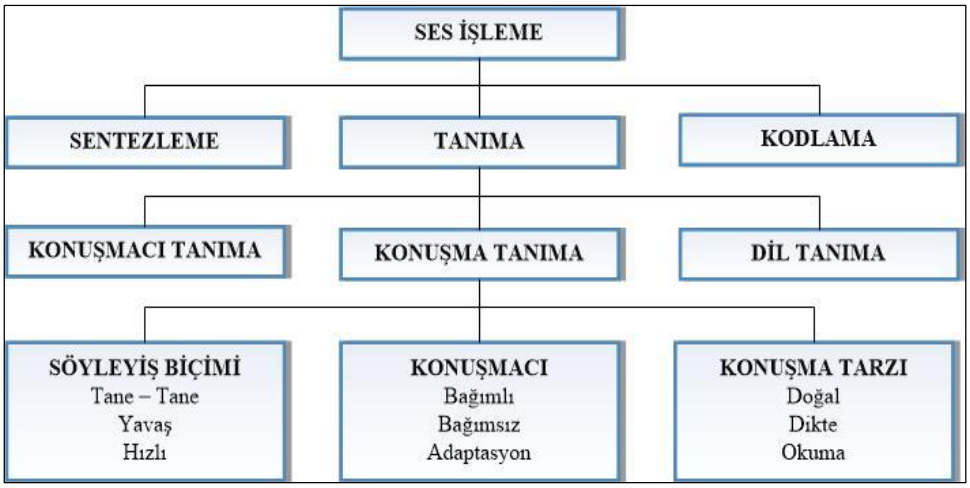
İnsanlar, seslerin üretimi için belirtilen durumları kombine ederek teorik olarak sonsuz sayıda farklı ses üretebilir. Ancak her dil sınırlı sayıda fonetik yapıya sahiptir. Fonetik bir dildeki tanımlanabilen, fiziksel olarak ölçülebilen konuşma seslerini incelemektedir. Fonetik, sesleri söylenişinden akustik özelliklerine ve beyinde algılanmasına kadar olan üç özelleşmiş bölüme ayırmaktadır.

Birinci bölüm seslerin üretilişi sırasında dudak, dil, diş, çene ve ses telleri gibi konuşma organlarının pozisyonu, hareketi ve şeklinin incelenmesidir. Bu incelemeye bağlı olarak seslerin sınıflandırılması ile ilgilenmektedir. İkinci bölüm konuşma ile üretilen ses dalgalarının frekans ve enerji gibi ölçülebilir fiziksel özelliklerini incelemektedir. Bu özellikler kişiden kişiye değişebilmektedir. Üçüncü bölüm konuşma seslerinin insanda algılanma ve tanınmasını incelemektedir. Genlik-zaman gösteriminde sinyalin periyodik olup olmamasına bakılarak çıkarılan sesin ötümlü veya ötümsüz olduğu tespit edilebilmektedir. Ancak bu gösterimden sesbirimlerin (fonem) kimliği belirlenemez. Bunun için dalga formundan frekans alanına geçiş yapmak gerekir. Konuşma tanıma sistemlerinde bu dönüşümü yapma işlemine öznitelik çıkarma işlemi denilmektedir.

Bir konuşmayı farklı diğer konuşmalardan ayırt etmek için kullanılan özellik çıkarımı, konuşma tanımanın en önemli adımlarındandır. Konuşma eyleminin doğası gereği her konuşmanın, konuşma bilgisi içerisine yerleşmiş bireysel özellikler bulunmaktadır [8]. Bu Bireysel özellikler, farklı öznitelik çıkarım teknikleriyle elde edilmektedir. OKT sistemleri için önerilen ve başarılı bir şekilde kullanılan birçok öznitelik çıkarım tekniği mevcuttur. Ancak bu çalışmada güncel olarak sık kullanılan öznitelik çıkarım teknikleri açıklanmıştır.

3.2. Otomatik Konuşma Tanıma Sistemleri

OKT sistemleri konuşmacıya bağımlılık, sözlük boyutu, konuşma biçimi ve konuşma türü açısından çeşitli sınıflarda incelenmektedir [79]. Şekil 3.1'de genel olarak ses işlemedeki süreç ve OKT sistemlerinin sınıflandırılması verilmiştir.

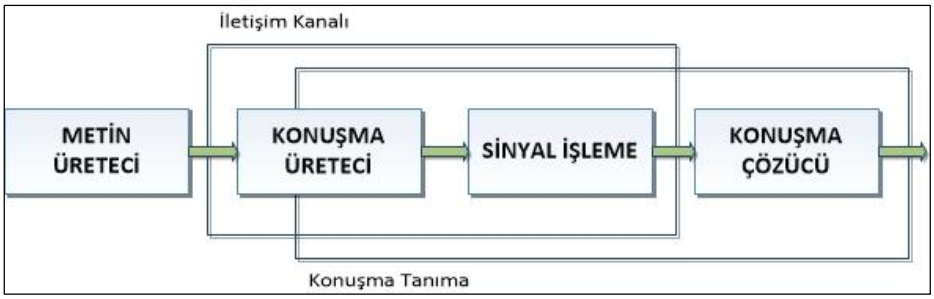


Şekil 3.1. Ses işlemindeki süreç ve OKT sistemlerinin sınıflandırılması

OKT sisteminin görevi konuşma sinyalini birtakım algoritma ve modelleri kullanarak işleyip metine dönüştürmektir. Konuşma tanıma işlemi için öncelikle metin üreteci aracılığıyla kaynak kelime dizisi hazırlanmaktadır (Şekil 3.2). Kaynak kelime dizisi sisteme giriş olarak verilen konuşma bilgisinden üretilmektedir. Farklı format veya kod yapısındaki konuşma bilgisi ilk giriş aşamasında istenilen diziye dönüştürülmektedir. Kaynak kelime dizisi konuşmanın ses sinyali dalga formunu ve konuşma sinyali bileşenini üretmek için bir iletişim kanalından geçirilmektedir. Son olarak konuşma kod çözücü sayesinde akustik sinyalin, orijinal kelime dizisine karşılık gelen en ideal dizi haline getirilmesi sağlanmaktadır [80].

Klasik bir konuşma tanıma sisteminin temel görevi konuşma içerisinde geçen ifadelerin tahmin edilmesi olarak tanımlanmaktadır. Ancak tahmin işleminin yerine getirilebilmesi için öncelikle konuşma bilgisinden o konuşmaya ait özelliklerin çıkarılması gerekmektedir [7]. Bu aşamada konuşma sinyallerinden akustik özellik dizileri elde edilmektedir. Ardından bu özellik dizilerinden konuşmaya ait sesbirimlerin olasılıkları hesaplanmaktadır.

Hesaplanan sesbirim olasılıkları sesbirimlerini tahmin etmeye çalışan bir sınıflandırıcıya iletilmektedir. Tahmin görevinde iki temel modelleme kullanılmaktadır. Bu modeller konuşma çözücüsü içerisinde yer alan akustik ve dil modellemedir. Sesbirim olasılıkları, akustik modeli temsil eden SMM ve DM ile birlikte konuşmayı çözmek için kullanılmaktadır. Kelimeler ve cümleler DM yardımı ile tahmin edilmektedir.



Şekil 3.2. Konuşma tanıma sistemi için kaynak-kanal modeli

AM içerisinde akustik ortam bilgisi, fonetik, mikrofon ve çevre değişkenliği, konuşmacılar arasındaki cinsiyet ve lehçe farklılıkları hakkında detaylı bilgiler yer almaktadır [9]. Dil modeli ise hangi kelimelerin bir araya gelerek olası bir kelime dizisi oluşturduğunu, hangi kelimelerin birlikte ortaya çıkacağını vermektedir. OKT sistemlerinde kullanılan akustik modelleme ve dil modelleme işleminin bölünmesi, istatistiksel konuşma tanıma sistemlerinde Eş. 3.1'deki gibi tanımlanmıştır [81].

$$\mathrm{W} = \arg\max \mathrm{P}(\mathrm{W} \mid \mathrm{X}) = \arg\max \frac{\mathrm{P}(\mathrm{W})\mathrm{P}(\mathrm{X}|\mathrm{W})}{\mathrm{P}(\mathrm{X})} \tag{3.1}$$

Eş. 3.1, akustik sesbirim veya öznitelik vektör dizisi $X = X_1, X_2, \ldots X_n$ durumu için konuşma tanıma sisteminde karşılık gelen kelime dizisini vermektedir. Eş. 3.1'de ifade edilen maksimum sonsal (posterior) olasılık $P(W|X)$ değerine sahiptir. Eşitliğin maksimizasyonu sabit X gözlemiyle gerçekleştirildiğinden payın maksimize edilmesine doğrudan bağlıdır. $P(W)$ ve $P(X|W)$ bileşenleri dil modellemesi ve akustik modelleme tarafından hesaplanan olasılık büyüklüklerini oluşturmaktadır. Büyük kelime dağarcığına sahip OKT sistemlerinde çok sayıda kelime olduğu için bir kelimeyi bir sesbirim dizisine parçalamak gerekebilir. Bu işleme kısaca okunuş sözlüğü modellemesi denir [82]. $P(X|W)$ fonetik modelleme ile yakından ilişkilidir. $P(X|W)$, konuşmacı varyasyonlarını, telaffuz varyasyonlarını, çevresel

varyasyonlar ve içeriğe bağlı fonetik varyasyonları dikkate almaktadır. OKT sistemleri, (X giriş) konuşma sinyalini kelime ile eşleştirmek için en iyi eşleşen kelime dizisini (W) bulmaktadır. Bu noktada gerçekleştirilen deşifre işlemi, basit bir örüntü tanıma probleminden daha fazlası olarak göze çarpmaktadır. Buradaki zorluk geniş kelime dağarcığı olan bir dilde OKT işlemini gerçekleştirmek için neredeyse sonsuz sayıda kelime kalıbına bakılmasının gerekliliğidir.

3.2.1. Konuşma özelliklerinin çıkarılması

Konuşma eylemindeki farklı bireysel nitelikler OKT sistemlerinde kullanılan farklı özellik çıkarım teknikleriyle elde edilmektedir. Özellik çıkarımının temel amacı uzun ve düzensiz bir yapıda olan konuşma verilerindeki konuşma içeriğini tanıyabilmektir. Bu işlem sonrasında elde edilen bilgiler bir sonraki aşama olan konuşma sınıflandırma işlemine girdi olarak verilmektedir. OKT sistemlerinde sık kullanılan MFKK ve DÖK özellik çıkarım teknikleri aşağıda açıklanmıştır.

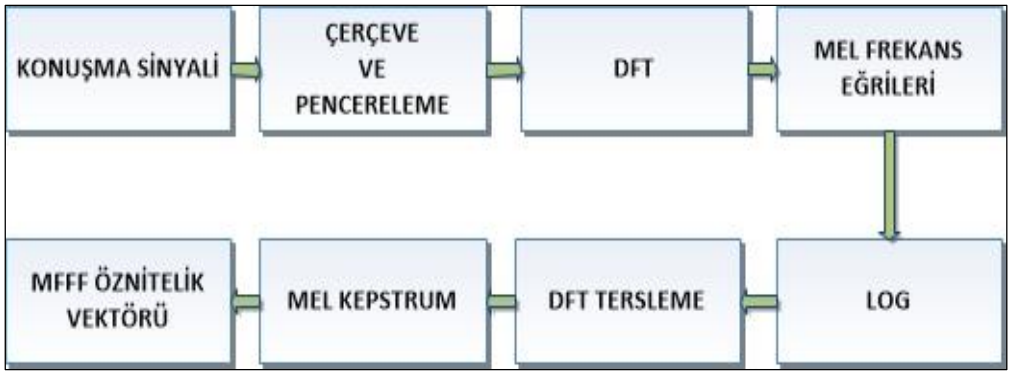
Mel frekanslı kepstrum katsayıları

Konuşma özelliklerini elde etmek için kullanılan MFKK, temel olarak ses sinyallerini insanların nasıl algıladığı prensibine dayanmaktadır. İnsan kulağı, frekansı 1 KHz'den düşük olan sesleri doğrusal, daha yüksek değere sahip olanları ise logaritmik olarak algılamaktadır. Bu nedenle frekans bantları MFKK'de logaritmik olarak konumlandırılmıştır. MFKK hesaplamanın tekniği kısa vadeli analize dayanmaktadır. Her çerçeveden bir MFKK vektörü Eş. 3.2 kullanılarak hesaplanmaktadır [81].

$$\mathrm{Mel(f)} = 2595 * \log10(1 + \frac{\mathrm{f}}{700}) \tag{3.2}$$

Bir konuşma örneğinin Kepstral katsayılarını çıkarmak için öncelikle konuşma sinyali süreksizliğinin en aza indirilmesi gerekmektedir. Bu işlem için konuşma sinyaline Hamming penceresi uygulanmaktadır. Mel filtre bankasını oluşturmak için Ayrık Fourier Dönüşümü (AFD, Discrete Fourier Transform) kullanılmaktadır. Mel frekans eğrisine göre

filtrelerin genişliği değişmektedir. Bu değişim sayesinde merkez frekans etrafında yer alan kritik bant üzerindeki öznitelikler hesaplanmaktadır. Son olarak ise elde edilen katsayılar ters yönlü Fourier dönüşüm katsayılarının hesaplanması için kullanılmaktadır [83]. MFKK öznitelik çıkarım işleminde yer alan adımlar Şekil 3.3'te gösterilmiştir.

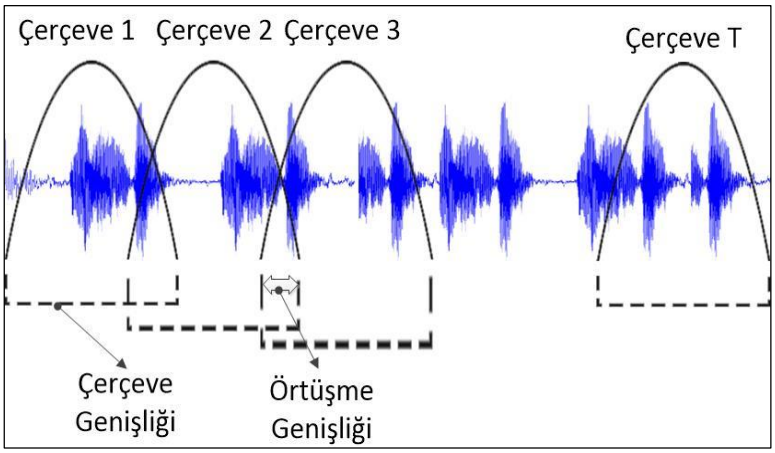


Şekil 3.3. MFKK özellik çıkarım işlem adımları [46]

Şekil 3.3'te gösterildiği gibi konuşma sinyali öncelikle çerçeveleme işlemine tabi tutulmaktadır. Bu işlemin amacı sadece küçük bir zaman aralığında sabit kalan konuşma sinyalini sabit kaldığı aralıklarda parçalara ayırmaktır. Yapılan araştırmalarda en etkili çerçeve aralığının 20-30 ms arasında olduğu belirtilmektedir. Konuşma sinyali uzunluğuna göre birden fazla çerçeve ile temsil edilmektedir. Çerçevelere ayrılan konuşma sinyalinde olabilecek kayıpları önlemek için örtüşme (overlapping) işlemi gerçekleştirilmektedir. Bu işlemde her çerçeve kendinden önceki çerçevenin belirli bir zaman aralığını kapsamaktadır. Bu işlem bir çerçeveden diğerine geçişin yumuşak olmasını sağlamaktadır. Çerçeveleme işleminden hemen sonra pencereleme işlemi gerçekleştirilmektedir. Bu işlemin amacı her bir çerçevenin başındaki ve sonundaki süreksizliği önlemektir. Pencereleme çeşitli filtreler ile gerçekleştirilmektedir. Bu filtrelerden Hamming, Barlett, Kaiser Mel Cepstrum ve Blackman en çok kullanılan filtrelerdir. Çerçeveleme, örtüşme ve pencereleme işlemleri şekil 3.4'te detayı olarak verilmiştir.

MFKK özellik çıkarma işleminin ilk adım olan çerçeveleme, örtüşme ve pencereleme işlemi tamamlandıktan sonra zaman bölgesinden frekans bölgesine geçiş işlemi gerçekleştirilmektedir. Her çerçeveye Hızlı Fourier Dönüşümü (HFD, Fast Fourier Transform) uygulanarak zaman bölgesinden frekans bölgesine geçiş sağlanmaktadır. HFD

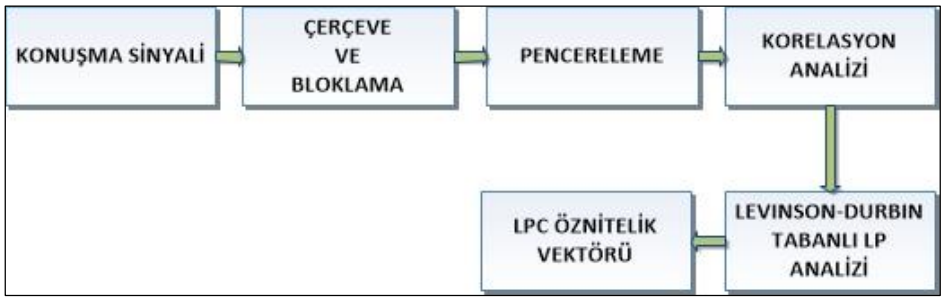
işleminin hemen ardından Mel frekans eğrileri elde edilmektedir. Mel ölçeği sayesinde Mel frekans dönüşümleri elde edilmektedir. Mel ölçeği algısal ve logaritmik olarak artan filtre öbeklerini kullanmaktadır. Bu filtre öbeği %50 oranında birbiri üstüne örtüşen, üçgen bant geçiren filtre kullanan, aralıkları ve bant genişliği sabit Mel frekans aralığına bağlı bir öbektir. Bu öbeğe 10 Hz'den 1000 Hz'e kadar doğrusal olarak genişleyen 10 filtre yerleştirilmektedir. Frekans her iki katına çıktığında logaritmik olarak 5 filtre bu frekans aralığına yerleştirilmekte ve kepstrum katsayıları hesaplama işlemine geçilmektedir. Ardından ses sinyali örneğinin frekans bileşen genliklerinin logaritması alınmaktadır. Son olarak elde edilen değer HFD tersleme işlemine tabi tutulmaktadır. Bu işlemler sonucunda ses sinyalinin kepstral değerleri elde edilmektedir [84].



Şekil 3.4. Çerçeveleme, örtüşme ve pencereleme işlemleri

Doğrusal öngörülü kodlama

En güçlü sinyal analiz tekniklerinden biri olan DÖK konuşmanın temel parametrelerini tahmin etmek için kullanılmaktadır [85]. DÖK'ün arkasındaki temel fikir bir konuşma özniteliğinin geçmiş konuşma örneklerinin doğrusal bir kombinasyonu olarak tahmin edilebilmesidir. Gerçek konuşma örnekleri ile tahmin edilen değerler arasındaki kareler arası farkların en aza indirilmesiyle, benzersiz bir parametre ya da bağımsız katsayılar elde edilmektedir. Bu katsayılar konuşmanın DÖK'ü için temel oluşturmaktadır. Analiz işlemi zaman içindeki konuşmanın doğrusal tahmin modelini hesaplama yeteneğini sağlamaktadır [86]. DÖK konuşma öznitelik çıkarma işleminde yer alan adımlar Şekil 3.5'te gösterilmiştir. DÖK kapsamındaki öznitelik çıkarma işlemleri bu adımlara göre gerçekleştirilmektedir.



Şekil 3.5. DÖK özellik çıkarım işlem adımları

Şekil 3.5'te gösterildiği gibi konuşma sinyalinin ilk giriş aşamasında MFKK'de olduğu gibi çerçeveleme ve pencereleme işlemi gerçekleştirilmektedir. Ancak DÖK'de tek bir çerçeve değil birden fazla çerçevenin bir araya gelerek oluşturduğu bloklar pencereleme işleminde kullanılmaktadır. Pencerelenmiş sinyalin her bir çerçevesine oto korelasyon analizi uygulanmaktadır. DÖK analizi, her bir çerçeveye ait $p + 1$ oto korelasyonundan DÖK parametre kümesi hesaplanmaktadır. Oto korelasyon analizinden DÖK analizine geçişte Levinson-Durbin metodu gibi farklı yöntemler kullanılmaktadır. Levinson-Durbin algoritmasındaki amaç doğrusal öngörü filtresi ile ilgili öngörü hata değişmesinin yinelemeli olarak bulunmasıdır. DÖK analizine geçişte oto korelasyon analizinin yerine kovaryans analizi de yapılmaktadır. Ancak uygulamalarda yaygın olarak kullanılan oto korelasyon analizi tercih edilmektedir.

3.2.2. Akustik modelleme

OKT sistemlerinin doğruluğunu arttırmada konuşmacı ve ortam farklılıkları önemli rol oynamaktadır. Bir konuşma sinyalinin akustik modellenmesi genel olarak özellik vektör dizileri üzerinden istatistiksel bilgilerin oluşturulması işlemi olarak tanımlanmaktadır. Akustik modelleme OKT'de gürültüye karşı dayanıklılığın arttırılmasında, konuşmanın öznitelik vektörlerini yeniden şekillendirmek için sesbirim sınıflandırıcıdan gelen geri bildirim bilgisinin kullanımını da içermektedir [87].

Akustik modeldeki sınıflandırma işlemi iki farklı yaklaşımla ele alınmaktadır. İlk yaklaşım SMM ve GKM'nin kullanıldığı istatistiksel yaklaşımlardır. Diğer yaklaşım ise YSA ve Destek Vektör Makinelerinin (DVM, Support Vector Machine) kullanıldığı yöntemlerdir. Akustik modellemede ilk olarak belirli zaman sinyali çerçevesindeki sesbirime ait sonsal

olasılık hesaplanmaktadır. Sesbirim sırası SMM üçlü durum yapısına benzer şekilde modellenebilmektedir. Genel olarak sesbirimlerin hizalanması normal Gauss dağılımını kullanarak elde edilmektedir [88]. SMM'de, durumlar arasındaki geçişlerde Markov varsayımları yapılıp bir durumdan başka duruma geçişte sadece bir önceki durum göz önünde bulundurulmaktadır. Bu kısıtlama uzun kelime bağımlılığı olan dizilerin modellenmesini zorlaştırmaktadır. Ayrıca, SMM'deki durumların yayım olasılığı da birbirinden bağımsızdır. SMM temelde bir zaman diliminin durumunu bir kez değiştiren sonlu durum makinesi olarak düşünülebilir. Çok sayıda SMM değerlendirme yöntemi, durum geçiş olasılıklarının ve SMM'nin her bir durumundaki yayılma olasılık yoğunluklarını tahmin etmek için geliştirilmiştir.

YSA tabanlı akustik modelde sesbirim sonsal olasılığı her bir çerçeve için bağımsızdır [74]. Bu bağımsızlık kelimede bulunan sesbirimlerin birbirinden bağımsız olması anlamına gelmektedir. YSA tabanlı akustik modelde temel olarak ses özniteliklerinin her birine sesbirim etiketi tahsis edilmektedir. Temelde her SMM durumunun gözlem olasılığı Gauss karışımları kullanılarak hesaplanmaktadır. Ancak YSA veya DSA tabanlı akustik modelde her SMM durumunun gözlem olasılığı, YSA kullanılarak hesaplanmaktadır.

3.2.3. Dil modelleme

OKT için gerekli olan dil modeli, hangi kelimelerin birlikte kullanılması ile olası bir kelime dizisi oluşturduğunu, hangi kelimelerin birlikte ortaya çıkacağını vermektedir [11]. AM çıktısının dil modeli ile karşılaştırılması işleminde DM üzerinde bir arama işlemi gerçekleştirilmektedir [12]. OKT sistemlerinde kullanılacak arama algoritmasının karmaşıklığı, DM'nin getirdiği kısıtlamalarla ve belirlenen arama alanıyla doğrudan ilgilidir. Sonlu durum dilbilgileri ve n-gramlar dâhil olmak üzere farklı dil modellerinin etkisi deşifre sonucunu doğrudan etkilemektedir. SMM tabanlı konuşma tanıma sistemlerinin başarısı AM'den daha çok DM'ye bağlıdır. YSA tabanlı OKT sistemlerinde DM üzerindeki yükün bir kısmı AM'ye aktarılmıştır. Ancak gürbüz bir DM hala OKT başarımını doğrudan etkilemektedir. Dolayısıyla, gürbüz bir dil modelinin geliştirilmesi Kelime Hata Oranı (KHO, Word Error Rate) oranı düşük bir OKT sistemi elde etmek için gereklidir.

Türkçe OKT üzerine son yıllarda birçok çalışma gerçekleştirilmiştir [89-92]. Ancak DM üzerinde yapılan araştırmaların çoğu DM boyutunun OKT üzerindeki etkisine yoğunlaşmıştır [93-97]. Elde edilen sonuçlar özellikle Türkçe gibi sondan eklemeli diller için büyük boyutlu DM'lerin gerekliliğini ortaya koymuştur. DM'ler üzerinden farklı olasılıkların ortaya çıkarılması çeşitli zorluklar içermektedir. DM üzerinden elde edilebilecek olasılıkların ortaya çıkarılması için farklı teknikler kullanılmaktadır. DM hazırlamak için genel olarak n-gram tabanlı modelleme yöntemi kullanılmaktadır [98].

N-gram tabanlı DM'lerde Markov varsayımları gerçekleştirilmektedir. Genel olarak 2 ile 4 arası geçmişteki kelime sırası dikkate alınarak olasılıklar hesaplanmaktadır. Ancak n-gram tabanlı DM, uzun bir cümledeki kelimelerin dizilişini modelleyememektedir. Çünkü n-gram'ler ile kısıtlı bir geçmişe bakılmaktadır. Bu sorunu giderebilmek için YSA temelli DM'ler geliştirilmiştir [98]. YSA tabanlı dil modelleri Markov varsayımında bulunmadığı için kelimelerdeki uzun bağımlılıklar modellenebilmektedir [99-101]. YSA tabanlı DM'ler genellikle n-gram tabanlı DM'lerden daha iyi performans göstermektedir [102]. YSA tabanlı dil modellerinin en büyük avantajı, kelime dizilim olasılıklarını sürekli bir uzayda tahmin etmesidir.

YSA tabanlı DM'lerin performansını arttırmak için dropout [103], Bayes başarım iyileştirmesi [104] ve tekrarlayan sinir ağı [105] gibi bazı teknikler önerilmiştir. Bu tekniklerin tümü Türkçe OKT uygulamalarında sınırlı performans iyileştirmesi sağlamıştır. Bunun temel nedeni Türkçe'nin üretken dil yapısıdır [106]. Üretken dil yapısı ile çok fazla uzun cümleler kurulabilmektedir. Bu nedenle bir sonraki kelime tahminini doğru yapabilmek için çok fazla geçmişe dönük kelime analiz edilmelidir.

Geçmişe dönük çok fazla kelimenin analiz edilmesi hesaplama karmaşıklığını beraberinde getirmektedir. Bu nedenle dil modelleri için kullanılacak yöntemler temelde iki konuya dikkat etmelidir. Bunlardan ilki geçmişe dönük çok daha fazla kelimeden yararlanabilmektir. İkincisi ise hesaplama karmaşıklığının azaltılmasıdır. Bu çalışmada Türkçe OKT sistemleri için gerekli olan DM cümle düzeyinde ele alınmıştır ve DM'in N-best (en iyi kelime birleşimleri) listeleri iyileştirilmiştir. Tez kapsamında istatistiksel ve YSA tabanlı DM yaklaşımları üzerinde deneyler gerçekleştirilmiştir.

3.2.4. Okunuş sözlüğü

Okunuş sözlüğü içerisinde konuşma tanıma sisteminin tanıması gereken kelimeler yer almaktadır. OKT sisteminin kullanım alanına göre sözlükteki kelime sayısı değişmektedir. Büyük sözlüklü OKT sistemlerinde yüzbinlerce kelimenin farklı okunuşları yer almaktadır [107]. Türkçe geniş kelime dağarcığına sahip bir dildir. Bu nedenle okunuş sözlüğündeki kelime sayısı oldukça fazla olmalıdır. Okunuş sözlüğünde yer almayan kelime sayılarının azaltılması ve gereksiz kelimelerin sözlükte bulundurulmaması önem arz etmektedir. Çizelge 3.1'de bazı kelimelerin farklı okunuşları verilmiştir.

Çizelge 3.1. Örnek kelime ve okunuş biçimleri

Kelime	Okunuş 1	Okunuş 2
Merhaba	M E R H A B A	M E R A B A
Nasılsınız	N A S I L S I N I Z	N A S S I N I Z
TTNET	T E T E N E T	T İ T İ N E T
Mobile	M O B İ L	M O B A Y L
ADSL	A D E S E L E	E Y D İ E S E L

Okunuş sözlüğündeki tüm kelimelerin okunuşu fonetik semboller ile belirlenmektedir. Türkçe, yazıldığı gibi okunan bir dil olduğu için okunuş sözlüğü hazırlama süreci daha kolay olup kelimedeki harfler fonetik sembol olarak da kullanılabilmektedir. Aynı kelime farklı şekillerde okunabildiği için okunuş sözlüğünde bulunan her kelimenin çeşitli okunuş şekilleri fonetik olarak belirtilmelidir. Ayrıca, yabancı kelimeler ve kısaltmalar gibi kelimelerin de yazılışı ile okunuşları farklı olabileceği için bütün farklı okunuşlar sözlükte bulunmalıdır. Genel bağlam için hazırlanan kelimeler genelde büyük metin derleminden elde edilen tekil kelime listesinden oluşmaktadır. Kelimeler dil bilim uzmanları tarafından incelenip farklı okunuş ve yazılışları ile sözlüğe dâhil edilmektedir. Okunuş sözlüğüne eklenen kelimelerde imla kurallarının doğru olması, geniş kelime dağarcığına sahip olması ve güncel konuşma dilinde kullanılan kelimeleri içermesi gerekmektedir.

3.2.5. Deşifre

Deşifre sırasındaki amaç elde edilen X sinyaline ait en olası kelime sırasını bulmaktır. Deşifre işleminde ilk adım olarak akustik model yardımı ile sesbirim dizisi elde edilmektedir. OKT sisteminde akustik model, giriş öznitelik vektörleri sırasına göre en iyi konumlanan bir sesbirim dizisini vermektedir. Elde edilen sesbirim dizisi okunuş sözlüğü yardımı ile kelime formuna dönüştürülmektedir. Kelimeler ise dil modeli yardımıyla en uygun kelime dizisi haline getirilmektedir [108].

Deşifre işleminde kullanılan modeller arasında bir bağ bulunmaktadır. Bu nedenle doğru veri ile eğitilmiş akustik ve dil modelleri ile birlikte deşifre işleminin süreci genellikle bir arama işlemi olarak adlandırılabilir [10]. OKT sistemlerinde kullanılacak arama algoritmasının karmaşıklığı, dil modellemenin getirdiği kısıtlamalarla belirlenen arama alanıyla yüksek oranda ilişkilidir. Deşifre işlemi konuşmayı soldan sağa bir süreç olarak işaret eden bilgileri ve zamana bağlı bir süreç tarafından sağlanan verimliliği içermektedir. Ancak gerçek uygulamalarda akustik modelin tek başına kullanılması kelime hata oranını oldukça yükseltmektedir. İstatistiksel tabanlı konuşma tanıma sistemlerinin başarısı akustik model yerine daha çok dil modeline bağımlı kalmaktadır. Dolayısıyla, oldukça güçlü bir dil modelinin eğitilmesi doğruluk oranı yüksek bir konuşma tanıma sistemi elde etmek için zorunlu hale gelmektedir. Deşifre işlemlerinde kullanılan Zamansal Bağlantılı Sınıflandırma (ZBS, Connectionist Temporal Classification) algoritması, sesbirim hizalama çıktısındaki bir alt sembolü ile X dizisi arasında bağımsızlık varsayımını temel almaktadır [109]. Girdi ile çıktı arasında derin bir yapı olsa bile girdi ile çıktı arasında güçlü bir Markov varsayımı bulunmaktadır. Bir çerçeve üzerindeki tahmin ile komşusundaki çıktı arasında koşullu bağımsızlık mevcuttur. Bu nedenle ZBS algoritması çıktısındaki sembollerin öğrenemeyip deşifre aşamasında güçlü bir dil modeline ihtiyaç duymaktadır.

3.2.6. Başarım metrikleri

OKT sistemlerinin başarımını değerlendirmek için genelde KHO metriği kullanılmaktadır. KHO, hipotez ve referans kelimeler arasındaki farklılıkları değerlendirmektedir. KHO hesaplaması Eş. 3.3'de gösterilmiştir.

$$\text{KHO} = \frac{D + S + I}{\text{N}} \times 100 \tag{3.3}$$

Eş. 3.3'de gösterilen N, referans kelimedeki toplam sembol sayısını vermektedir. D, hipotezdeki kelimeye göre silinmiş sembollerin sayısını, S değiştirilen sembollerin sayısını I ise ilave sembollerin sayısını temsil etmektedir. KHO başarım metriğinin benzeri olarak Cümle Hata Oranı (CHO, Sentence Error Rate) ve Karakter Hata Oranı (KAHO, Char Error Rate) kullanılmaktadır. CHO işleminde referans cümle ile hipotez cümle karşılaştırılmaktadır. KAHO metriğinde ise karakterler kendi arasında karşılaştırılmaktadır. En çok kullanılan başarım metriği KHO olarak bilinmektedir.

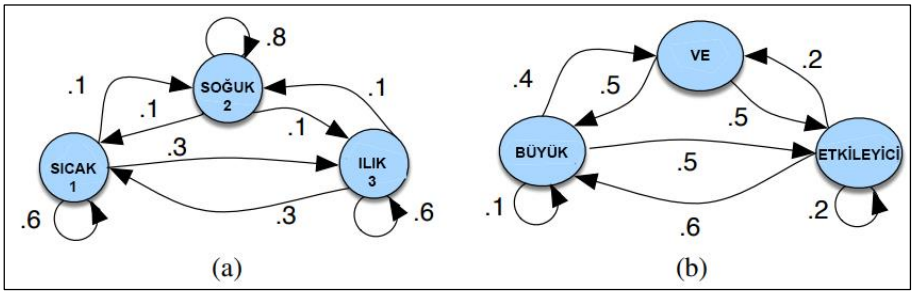
3.3. Saklı Markov Modeli

Saklı Markov Modeli, akustik ve dil modelleme de sıklıkla kullanılan bir yöntemdir. Konuşma tanıma probleminin çözümü için etiketlenmiş veri gerekmektedir. Etiketlenmiş bir veri kümesi içerisinde bulunan kelimeler SMM ile modellenebilir. Bu bölümde, SMM'nin çalışma prensibi verilmiş ve SMM için kullanılan algoritmalar açıklanmıştır.

SMM, durumlar arasındaki Markov zincirini güçlendirmeye dayanmaktadır. Bir Markov zinciri, durumların olasılıkları hakkında bilgi vermektedir. Zincirler, belirli bir kümeden değerler alabilen rastgele değişkenler ile temsil edilmektedir. Bu kümeler kelimeler veya etiketler gibi herhangi bir veriyi temsil eden simgeler olabilmektedir. Bir Markov zinciri gelecek değeri tahmin ederken mevcut durumun ne olduğuna dair çok güçlü bir varsayım yapmaktadır. Markov zincirleri arasında nasıl bağlantı sağlanacağı, geçmişe dönük kaç durumun göz önünde bulundurulacağı ile ilgilidir. Ancak genellikle mevcut durum ve sonraki durum göz önünde bulundurulmaktadır. Şekil 3.6'da hava durumunu ve bir kelimeden sonra hangi kelimenin gelebileceğini tahmin etmek için hazırlanan bir SMM gösterilmiştir.

Şekil 3.6'da hava durumu (a) ve bir kelime dizisi (b) için durumlar ve durumlar arası geçişler verilmiştir. Hava durumu için sıcak, soğuk ve ılık durumlarının olduğu bir dizi hava olayına olasılık atamak için gerçekleştirilen Markov zinciri gösterilmiştir. Modelin inşası için bir

başlangıç dağılımı (π) gereklidir. Hava durumu tahmini için π = [0.1, 0.7, 0.2] seçimi durum 2'de (soğuk) başlama olasılığı 0,7, durum 1'de (sıcak) başlama olasılığı 0,1 anlamına gelmektedir. Markov varsayımının geleceği tahmin ederken geçmişin bir önemi yoktur, sadece şimdiki zamana bakılmaktadır. Durumlar arası geçişler olasılık temelli gerçekleşmektedir. Belirli bir durumdan çıkan bağlantıların değerleri 1'e eşit olmalıdır. Şekil 3.6'da verilen ikinci şekilde $w_1 \ldots w_n$ kelime dizisine olasılık atamak için bir Markov zinciri gösterilmiştir. Bu Markov zinciri temel olarak her bir kenar $P(w_i|w_j)$ olasılığını ifade eden bigram dil modelini temsil etmektedir.



Şekil 3.6. Hava durumu (a) ve kelime tahmin (b) durumları [110]

Bir Markov zincirinde ilgilenilen durumlar çoğu zaman gizlidir. Normal durumda bir metindeki konuşma parçası etiketleri gözlemlenememektedir. Bunun yerine kelimeleri görmekte ve etiketleri kelime dizisinden çıkarma işlemi gerçekleştirilmektedir. SMM, hem gözlemlenen olaylar hakkında hem de olasılıklardaki etkili faktörler olarak düşünülen olaylar hakkında bilgi edinmemize izin vermektedir. Bir SMM, $Q = q_1 q_2 \ldots q_n$ belirtiminde olduğu gibi bir dizideki n adet durum ile belirtilmektedir. Ardından n adet durum için bir geçiş olasılığı matrisi oluşturulmaktadır. Standart bir SMM yapısı aşağıdaki bileşenlerden oluşmaktadır:

- Durum Sayısı: X saklı sürecinin Markov zinciri olması nedeniyle, sistemin durum uzayı kesiklidir. Standart SMM'de durum uzayı sonludur. "M" durum uzayındaki toplam durum sayısını temsil etmektedir. M tane saklı durum içeren bir sistem için, M-durumlu SMM'ler incelenmelidir. Burada $S = \{S_1, S_2, \ldots, S_M\}$ kesikli sonlu durum kümesi olsun, bu durumda standart SMM'de durumlar birbirinden bağımsızdır. Durumların sayısı, SMM'in boyutunu göstermektedir.

- Durum Geçiş Olasılıkları Dağılımı: S durum kümesi üzerinde tanımlanan $P = [P_{ij}]_{MxM}$ matrisi ile gösterilmektedir. P_{ij} elemanı, sistemin S_i durumundan S_j durumuna geçme olasılığıdır. Durumlar arasındaki geçiş olasılıkları bütün durumların bir sonraki durumda nerede olduğunun hesaplanması şeklinde ifade edilmektedir. Standart SMM'de durum geçiş olasılıkları zaman içerisinde değişmez ve bu olasılıklar gözlemlerden bağımsızdır. Gözlemler koşullu olasılıkları temsil etmektedir. Örneğin dondurma çok yendiğine göre hava sıcak olmalıdır. Bu modele verilen bir gözlem dizisinden durumların sonucu elde edilmektedir.

- Gözlemler Kümesi: Gözlem süreci kesikli ya da sürekli değerler alabilmektedir. Y'nin X'e koşullu dağılımı, tek bir (kesikli ya da sürekli) parametrik aileye aittir. Gözlemler kümesi gözlem sürecinde elde edilen veri kümesini temsil etmektedir. Burada koşullu olasılıkları elde edebilmek için Poisson, Binom, Normal, Gamma, Beta vb. dağılım aileleri kullanılmaktadır. İlgilenilen dağılım ailesinin parametreleri, X saklı süreci tarafından belirlenmektedir. Bu durumda, Y'nin dağılımı, belirlenen parametrik ailenin ilişkili karma dağılımına göre belirlenmektedir.

- Gözlem Olasılıkları: Sistemin durumu bilindiğinde, ortaya çıkan gözlem diğer gözlemlerden bağımsızdır. Gözlem olasılıkları, $B = [b_i(y)]_{mxM}$ matrisi ile gösterilir. Gözlem sürecinin tanım kümesine göre M değeri tanımlanmaktadır. bi(y) değeri, sistem Si durumunda iken, y gözleminin ortaya çıkma olasılığıdır. Sürekli bir dağılım ailesi ele alındığında, bi(y) bir olasılık yoğunluk değeri olarak tanımlanır.

- Başlangıç Durum Dağılımı: Başlangıç durum dağılımı M adet durumun belirtildiği bir olasılık vektörü ile gösterilmektedir. Her bir durum için başlangıç olasılıkları $\pi_i = P(X_1 = S_i)$ şeklinde ifade edilmektedir. Başlangıç durumunu belirten eleman, başlangıç anında sistemin Si durumunda bulunma olasılığını temsil etmektedir.

SMM'in model bileşenleri dikkate alındında SMM'i tanımlayan parametreler; başlangıç durum dağılımı, gözlem olasılıkları ve durum geçiş olasılıkları dağılımıdır. Konuşma tanıma işleminde SMM durumları sesbirimlerini temsil etmektedir. Durum geçiş olasılıkları ise sesbirimlerinden hangi kelimelerin üretildiğini belirtmektedir. Diğer yandan başlangıç

durum dağılımı ise ilk öznitelik vektörünün girişinden itibaren hesaplanarak elde edilmektedir.

3.4. Gauss Karışım Modeli

GKM parametrik bir olasılık fonksiyonunu Gauss bileşenlerinin yoğunlukları toplamı olarak ifade eden bir modeldir. Konuşma tanıma görevinde SMM ile birlikte kullanılmasıyla başarılı sonuçlar elde edilmiştir. GKM parametreleri bir veri kümesinde tekrarlanan beklentinin En Büyüklenmesi Algoritması (EBA, Expectation-Maximization) veya En Büyük Sonsal (EBS, Maximum A Posteriori) kestirimi ile bulunabilmektedir. GKM, M adet Gauss yoğunluğunun ağırlıklı toplamı olarak Eş. 3.4'te verilmiştir. Eş. 3.4'te verilen x, çok boyutlu rastsal vektörü temsil etmektedir.

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i) \tag{3.4}$$

Eş. 3.4 vektör öznitelik türüne göre değişebilir ancak konuşma tanımada konuşmaya ait öznitelik vektörleri kullanılmaktadır. Eşitliğin sağ tarafı Gauss karışım ağırlıkları ve Gauss yoğunluk bileşenlerini temsil etmektedir. Her yoğunluk bileşeni Gauss rastlantı fonksiyonu formülü ile elde edilmektedir. Gauss rastlantı fonksiyonu ortalama vektörü, kovaryans matrisini belirtmektedir. Gauss karışım modeli parametrelerinin tanımlanabilmesi için kovaryans matrislerinin ve Gauss karışım model ağırlıklarının tüm yoğunluk bileşenlerinden elde edilmesi gerekmektedir. Bu parametreler Eş. 3.5'de verilmiştir. Kovaryans matrisleri Eş. 3.5'te toplam işareti ile belirtilmiştir. Bu matrisler genellikle diyagonal formda bulunmaktadır.

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \qquad i = 1, \dots, M \tag{3.5}$$

Bir GKM model yapılanması (bileşen sayısı, tam diyagonal kovaryans matrisleri ve bağlama parametresi gibi) GKM parametrelerinin tahminine ve veri kümesi büyüklüğüne göre tanımlanmaktadır. GKM'nin geliştirilmesinde kullanılacak uygulama alanı göz önünde

bulundurulmalıdır. Ayrıca özniteliklerin istatistiki açıdan bağımsız olmamaları durumunda bile kovaryans matrislerinin gerekli olmasına dikkat edilmelidir.

Modelin geliştirilmesinde kullanılacak veri kümesi vektörlerine göre GKM yapılandırılmasında, GKM parametrelerinin veri kümesi vektörlerinin dağılımı ile uyuşacak şekilde kestirimi amaçlanmaktadır. GKM parametrelerinin kestirimi için bazı yaklaşımlar mevcuttur [111]. Günümüzde kullanılan en popüler ve başarılı yaklaşım Maksimum Olabilirlik (MO, Maximum Likelihood) kestirimidir. MO kestirim yönteminde amaç verilen eğitim verisindeki GKM olabilirliğini en büyük hale getirecek model parametrelerinin bulunmasıdır. Fakat parametrelerin doğrusal olmaması GKM olabilirliğini en büyük hale getirecek model parametrelerinin bulunmasını zorlaştırmaktadır.

Konuşma tanıma görevinde GKM ile SMM arasında derin bir ilişki bulunmaktadır. GKM, vektör olarak verilen bir sesbirim dizisinin olasılık dağılımını modellemektedir. Burada MO gibi farklı algoritmalar kullanılabilir. GKM bir sesbirim ile gözlemlenen olay arasındaki ilişkiyi hesaplamaktadır. Konuşma tanıma görevinde bu işlem ses sinyali ile sesbirimler arasındaki ilişki modellenmektedir. Diğer yandan SMM, sesbirim durumları arasındaki gözlemleri modellemektedir.

3.5. Yapay Sinir Ağları

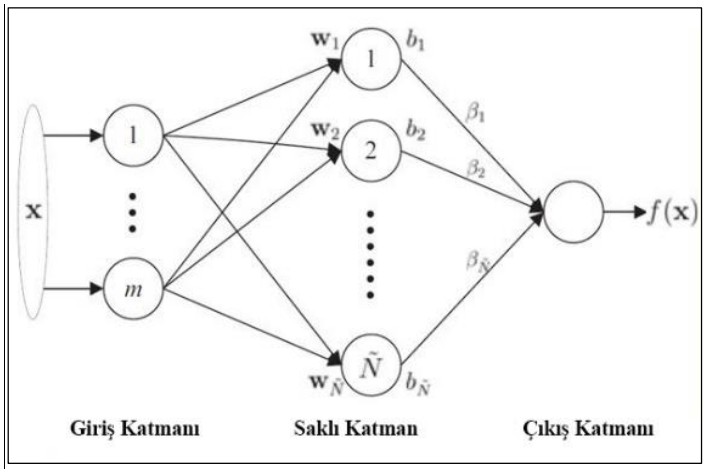
YSA'lar, biyolojik sinir ağlarından esinlenen bir makine öğrenme tekniğidir. Yapay sinir ağları doğrusal bir yapısı olmayan problemlerde d_x boyutlu bir x girdisi ile d_y boyutlu y çıktısı arasındaki ilişkiyi modellemektedir. Bu nedenle konuşma tanıma görevinde hem AM hem de DM'nin modellenmesinde sıklıkla kullanılmaktadır [112].

YSA'nın erken dönem çalışmalarında doğrusal bir sınıflandırıcı olarak kullanılması önerilmiş ve matematiksel yapısı oluşturulmuştur [32]. Matematiksel yapının oluşturulmasından sonra doğrusal olmayan sınıflandırıcıların kullanımı için yeni yöntemler geliştirilmiştir [32]. Böylelikle evrensel bir tahmin mekanizması olarak YSA'ların kullanılması yaygınlaşmıştır.

Nöronlar birbirlerine paralel olarak bağlı 3 katman halinde bir araya gelerek yapay sinir ağını oluşturmaktadır [113]. YSA yapısı temel olarak üç katmandan oluşmaktadır; girdi katmanı, gizli katman ve çıktı katmanı. Veriler yapay sinir ağına girdi katmanından iletilmektedir. Gizli katmanlarda işlenen veriler gizli katmanların ardından çıktı katmanına gönderilmektedir. Veri işlemeden kasıt ağa sunulan verilerin ağ ağırlık değerleri kullanılarak uygun çıktıya dönüştürülmesidir. Yapay sinir ağının doğru çıktıları üretebilmesi için ağ ağırlık değerlerinin doğru değerler alması gerekmektedir. Doğru ağırlık değerlerinin bulunması işlemine yapay sinir ağının eğitilmesi denilmektedir. Sinir ağının ağırlık değerleri başlangıçta rasgele atanmaktadır. Eğitim sırasında her eğitim örneği yapay sinir ağına verildiğinde ağın öğrenme kuralına göre ağırlıklar değiştirilmektedir. Geliştirilen ağ test veri kümesindeki örneklere doğru cevaplar verir ise ağ başarıyla eğitilmiş kabul edilmektedir.

3.5.1. İleri beslemeli sinir ağı

YSA'ların en temel yapısı İleri Beslemeli Sinir Ağlarıdır (İBSA, Feed Forward Neural Networks). İBSA yapı olarak hesaplama ünitelerinin farklı katmanlarda üst üste yığılması ile elde edilmektedir. Hesaplama ünitesi literatürde nöron olarak tanımlanıp kendisine gelen girdiler üzerinde birtakım matematiksel işlemler yapmaktadır. Şekil 3.7'de tek bir saklı katmandan oluşan İBSA modelinin grafiksel gösterimi verilmiştir. Bu şekilde, girdi olarak verilen bir vektör ara katmandaki nöronların matematiksel işlemlerine tabi tutulduktan sonra çıktı katmanına iletilmektedir.



Şekil 3.7. Tek katmanlı ileri beslemeli sinir ağı [60]

Sinir ağı en basit mimaride doğrusal bir katman, matris ve vektör çarpımları ve sonunda da doğrusal olmayan bir fonksiyon transferi ile ifade edilmektedir. Yapay sinir ağlarındaki en yaygın mimari genelde birden fazla katmanın birleşiminden oluşmaktadır. Bu yapı çok katmanlı algılayıcı olarak da bilinmektedir.

Sinir ağları günümüzde regresyon ve sınıflandırma problemleri için sıklıkla kullanılmaktadır. Bu tez çalışmasında, sinir ağlarını bir sesbirim sınıflandırıcısı olarak kullanılması amaçlanmıştır. Sınıflandırma probleminde, sinir ağının görevi x girdi vektörü ile etiketler i ∈ [1, . . . , I] arasındaki ilişkiyi modellemektir. Dolayısıyla, ağın çıktısı olan f(x) fonksiyonu, I boyutlu bir vektör olup fi(x) ise her bir sınıfın olasılığını göstermektedir. Sonsal olasılığı hesaplamak için ağın çıktıları üzerinde softmax fonksiyonu uygulanmaktadır. Literatürde, çapraz entropi veya kare hata belirtimleri kullanılarak sinir ağlarının yaptığı hata miktarı ölçülmektedir.

3.5.2. Geleneksel sinir ağlarının kısıtları

Geleneksel bir yapay sinir ağı, her biri gerçek değerli aktivasyon dizisi üreten nöron adı verilen ve birbirine bağlı işlemcilerden oluşmaktadır. Giriş nöronları, çevreyi algılayan algılayıcılar aracılığıyla aktive edilmektedir. Diğer nöronlar ise daha önce aktif olan nöronlardan elde edilen ağırlıklar ile aktive edilmektedir [75]. Bu çalışma mantığına sahip yapay sinir ağlarının kısıtları 5 madde ile detaylandırılmıştır. Bu maddeler şu şekildedir [114]:

- Yapay sinir ağları bir kara kutu gibi çalışmaktadır. Dolayısıyla nedensel ilişkileri tanımlayabilmek için sınırlı bir kabiliyete sahiptir. İstatistiksel modellerin birincil amacı nedensel çıkarımlar yapmaktır. Model geliştiricisi, öncelikle ağ için eğitim verilerini hazırlar, ardından ağın kendini eğitmesine ve hangi girdi değişkenlerinin önemli olduğunun belirlemesine izin vermektedir. Belli bir çıktıya, en çok katkıda bulunan değişkenlerin hangi değişkenler olduğu kolayca belirlenemez ve bir yapay sinir ağı modeli, geliştiricinin istemediği çok sayıda önemsiz öngörücü değişken içerebilmektedir. Araştırmacılar yapay sinir ağlarının iç mantığı konusundaki anlayışlarını arttırmak için aktif olarak teknikler geliştirmişlerdir. Bu tekniklerden biri,

her bir girdi değişken nöronunun birer birer çıkarıldığı sinir ağını eğitmek ve daha sonra ağ performansı üzerindeki etkisini gözlemlemektir. Diğer araştırmacılar ise çeşitli girdi parametrelerinin bağlantı ağırlıklarını incelemek için kullanılan değişkeni belirlemek için teknikler geliştirmişlerdir. Bu nedenle sinir ağlarının kara kutu özelliği azaltılmaya çalışılırken zaman kaybı bir kısıt olarak karşımıza çıkmaktadır.

- Yapay Sinir ağı modellerinin paylaşımı oldukça zor bir süreçtir. Eğitilmiş bir sinir ağı modelini farklı kullanıcıların kullanabilmeleri için eğitimli yapar sinir ağı yazılımının bir kopyasının her bir son kullanıcıya sunulması gerekmektedir. Bu durumda farklı veriler eğitim için kullanıldığında geliştiricinin kendi yazılımında bulunan matrisleri tekrar kalibre etmesi gerekir. Matrisleri tekrar optimize etme işlemi zaman alıcı bir kısıt olarak karşımıza çıkmaktadır.

- Yapay Sinir ağı modellemesi çok fazla hesaplama kaynağı gerektirmektedir. Yapay sinir ağı modeli geliştirme, çok daha fazla hesaplama zamanı gerektiren ve hesaplama açısından yoğun işlemlerin yapıldığı bir işlemler bütünüdür. Standart kişisel bilgisayarların donanımı ile bir yapay sinir ağının en düşük hata ile en iyi öğrenme durumuna yaklaşması genellikle haftalar alabilmektedir. Bu nedenle klasik özelliklere sahip bilgisayar donanımları üzerinde çalışarak işlem gören yapay sinir ağlarının modellenmesi çok zaman almaktadır.

- Yapay sinir ağı modelleri aşırı uyumluluk eğilimi gösterebilmektedir. Bir sinir ağı modelinin etkileşimleri ve doğrusal olmayan modellemeye sahip olması bazı durumlarda dezavantaj sağlamaktadır. Bu durum düşük performansa sahip bir eğitim veri kümesinde aşırı uyuma neden olabilmektedir. Genel olarak aşırı uyum üç ana yolla engellenir. Gizli düğümlerin sayısını sınırlayarak, büyük ağırlıklar için hedef fonksiyona bir ceza terimi ekleyerek veya çapraz doğrulama kullanıp eğitim veri miktarını sınırlayarak. Her üç yöntem de potansiyel olarak yararlı olmasına rağmen, geçersiz kılma yöntemini kullanarak eğitim miktarını sınırlamak, hesaplamanın yoğun olduğu en yaygın yöntemdir. Gizli düğümlerin sayısını azaltmak sonuçta ele edilen başarım oranının düşmesine neden olabilmektedir.

- Yapay sinir ağı modelinin gelişimi deneye dayalı ve disiplinler arası bir süreçtir. Elde edilen çoklu eğitim algoritmaları ve model gelişiminin deneye dayalı niteliği göz önüne

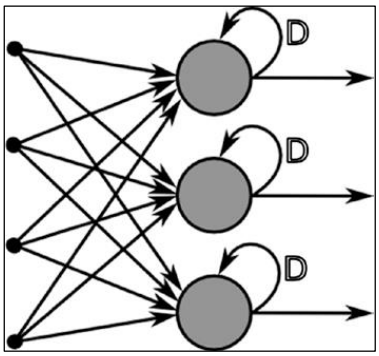
alındığında, belirli bir uygulama için en iyi çözüm veren yapay sinir ağı geliştirildiğine karar vermek zordur. Bu nedenle model geliştiricilerin öğrenme oranları, momentum terimleri ve gizli düğüm sayısı gibi eğitim parametreleri üzerine hassasiyet analizleri gerçekleştirmesi gerekmektedir. Bu işlemlerin her biri uzmanlık alanı gerektirdiğinden ve deneye dayalı olarak yapılması gerektiğinden uzun ve tecrübe gerektiren bir kısıt olarak karşımıza çıkmaktadır.

Sonuç olarak probleme uygun yapay sinir ağı yapısının belirlenmesi genellikle deneme yanılma yolu ile yapılmaktadır. Uygun yapay sinir ağının oluşturulamaması, düşük performanslı çözümlere neden olabilmektedir. Uygun ağ oluşturulduğunda ise iyi bir çözüm bulunabilir fakat yapay sinir ağı bu çözümün en iyi çözüm olduğunu garanti etmemektedir. Yapay sinir ağındaki öğrenme katsayısını, gizli katman sayısını ve gizli katmanlardaki nöron sayılarını belirlemek için genel geçer bir kural bulunmamaktadır. Bu durum iyi çözümler bulmayı zorlaştırmaktadır.

YSA parametreleri, her problem için ayrı faktörler dikkate alınarak ve tasarlayıcının tecrübesine bağlı olarak belirlenmektedir. Yapay sinir ağları sadece sayısal veriler ile çalışmaktadırlar. Sembolik ifadelerin nümerik gösterime çevrilmesi gerekmektedir. Yapay sinir ağının ne kadar veri kümesi ile eğitileceğine geliştirici karar vermektedir. Yapay sinir ağlarındaki en iyi çözüm ucu açık ve sürekli araştırılan bir konudur. Bütün bu dezavantajlarına rağmen, yapay sinir ağları kullanılarak farklı problemlere değişik şekilde çözümler üretilmekte ve uygulamalarda başarılı şekilde kullanılmaktadır [115].

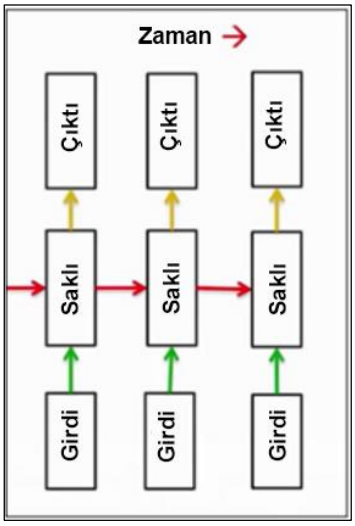
3.5.3. Tekrarlayan sinir ağları

Tekrarlayan Sinir Ağları (TSA, Recurrent Neural Networks), örüntü tanıma kapsamındaki sorunları çözmek için oluşturulan bir yaklaşımdır [116]. Temel olarak ileri beslemeli çok katmanlı algılayıcı kavramına dayalı bir yapıdır. TSA, bitişik zaman adımlarına, yayılan kenarların eklenmesiyle güçlendirilmiş sinir ağı olup, yapıya zaman kavramı getirmektedir. TSA, gizli durumları dağınık bir şekilde saklamaktadır. Bu nedenle deneyimler hakkında daha fazla bilgi sahibi içermektedir. Şekil 3.8'de basit TSA mimarisi verilmiştir.



Şekil 3.8. TSA mimarisi [116]

Diğer yapay sinir ağlarına benzer şekilde, nöronların oluşturduğu katmanlar birden fazla ağırlık ile birleştirilmektedir. Her ağırlık Şekil 3.8'de gösterildiği gibi geri yayılım algoritması ile değiştirilmektedir. Ağırlıklar, belirli bir zaman adımında aktivasyon fonksiyonlarının ağırlıklı toplamı olan bir değerlendirme istatistiğine göre değiştirilmektedir. Gerçek dünyadaki örneklerin çoğunda girdi ve çıktılar birbirinden bağımsız değildir. Örneğin, bir sonraki kelimeyi önceden tahmin etmek gerekirse, onun önündeki kelimeleri bilmek önemlidir. Şekil 3.9'da gizli durumların TSA'da nasıl bağlantılı olduğu gösterilmiştir.



Şekil 3.9. TSA'da gizli katmanların gösterimi [116]

TSA'da girdinin önceki yürütmenin çıktısı olduğu belirtilmektedir. Şekil 3.9'da gösterilen gizli katmanlar giriş ve çıkış katmanları arasında tekrarlama görevi de görmektedir. Tekrarlama görevi özellikle bir dizi gözlem sonucu oluşan durumlarda yüksek başarım

göstermektedir. Bu nedenle TSA üzerine birçok mimari yapı geliştirilmiştir. Ancak en başarılı mimari yapılar 1997'de yayınlanan iki farklı çalışmada belirtilmiştir. İlk çalışma olan Hochreiter ve Schmidhuber'in [117] Uzun Kısa Süreli Bellek (UKSB, Long Short Term Memory) üniteleri ile gizli katmandaki geleneksel düğümlerin yerini alan hesaplama birimi hücresini tanıtmaktadır. Bir YSA, bu hafıza hücreleriyle önceki tekrarlanan ağların karşılaştığı unutkanlık problemini aşabilmektedir.

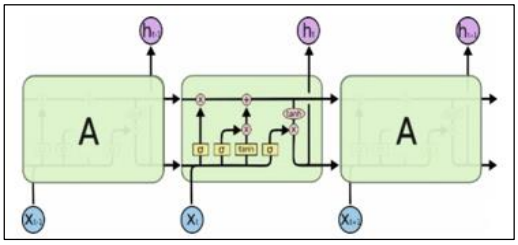
UKSB, gizli bir katmanı olan standart tekrar eden bir sinir ağına benzemekle birlikte, gizli katmandaki her sıradan düğümün yerini bir bellek hücresi almaktadır. UKSB, bellek hücresi vasıtasıyla ara bir depolama türü sunmaktadır. Schuster ve Paliwal'ın [117] yaptığı çift yönlü yinelemeli sinir ağları çalışmasında, dizinin herhangi bir noktasında çıktıyı belirlemek için hem geleceğin hem de geçmişin verdiği bilgileri kullanan bir mimari yapı sunulmuştur. Bu mimari yapı, önceki girdilerin yalnızca çıktıyı etkileyebileceği uygulamalarda veya doğal dil işleme uygulamalarında sıralama etiketi oluşturma görevlerinde başarılı bir şekilde kullanılmaktadır [117].

Tekrarlayan sinir ağlarındaki unutkanlık problemi

TSA'ların getirdiği en büyük avantaj daha önce görülen girdilerin bilgisini bir sonraki adımlara taşıyabilmektir. Bazı durumlarda bir girdi serisinde sadece birkaç adım öncenin bilgisine ihtiyaç duyulmaktadır. Örneğin, "bugün kar yağışı var ve --- çok soğuk" cümlesindeki -- kelimesini bir dil modelinin tahmin etmesi için sadece geçmişteki birkaç kelimeyi hatırlaması yeterlidir. Bu durumda TSA'lar uzun bir geçmiş ile karşı karşıya olmadıkları için bağımlılıkları iyi modelleyebilmektedir. Ancak, "ben Türkiye'de yaşıyorum ve yıllardır ... Türkçe'yi de iyi konuşabiliyorum " cümlesindeki "Türkçe'yi" kelimesinin model tarafından tahmin edilebilmesi için 5-10 kelime öncesindeki "Türkiye'de" kelimesinin model tarafından hatırlanması gerekmektedir. Dolayısıyla bu tür kullanımlarda modelin uzun bir geçmiş hakkında bilgi sahibi olması ve bu bilgileri unutmaması gerekmektedir. İlk girdi ile son girdi arasındaki mesafe uzadıkça TSA'lar geçmişte gördükleri bilgileri unutmaya başlamaktadır. Teorik olarak, TSA'lar bu tür uzun bağımlılıkları hatırlayıp unutmamaları gerekmekte, fakat gerçek kullanımlarda bu her zaman

doğru olmayıp model uzun girdilerde unutkan olmaya başlamaktadır. Bu sorunu çözmek için UKSB üniteleri önerilmiştir [118].

UKSB'ler, TSA modelinde kullanılan ve uzun bağımlılıkların model tarafından hatırlanmasına yardımcı olan ünitelerdir. Standart bir TSA'de basit bir şekilde sinir ağının çıktısı bir sonraki girdi ile birlikte aktivasyon fonksiyonuna tabi tutulmaktadır.



Şekil 3.10. TSA'de bir girdi ile katmanın önceki çıktısının kullanım durumu

Şekil 3.10'da gösterilen ilk adımda durum içerisinde hangi bilgilerin geçirilmesi gerektiğine karar verilmektedir. Bu adım unutma geçidi olan bir sigmoid katmanı ile gerçekleştirilmektedir. İkinci adımda, ünitenin durumuna hangi yeni bilgilerin eklenmesi gerektiği ile ilgili karar verilmektedir. Bir sonraki adımda da, ünite durumu güncellenmekte ve yeni durum olarak kaydedilmektedir. Son adımda da, ünitenin saklı vektörü olan h_t ve durum bilgisini içeren c_t çıktıları üretilip bir sonraki üniteye iletilmektedir. UKSB yapısı TSA'larin unutkanlık sorununu çözüp uzun dizilerdeki bilginin kaybolmamasını sağlamaktadır. UKSB'lere benzer görevi yapan farklı ünitelerde bulunmaktadır. Geçitli Tekrarlanan Üniteler (GTÜ, Gated Recurrent Unit) UKSB yapısına benzer bir yapısı olup daha az parametre içermektedir. Bu ünitelerde son zamanlarda farklı problemlerde kullanılmaktadır.

3.6. Kullanıcı Arayüzü

Kulanım alanları giderek artan OKT sistemleri kullanıcı ile etkileşime geçebilmek için bir arayüze ihtiyaç duymaktadır. Bu arayüzlerin geliştirilmesinde ve gerçek uygulamaya alınması işleminde birçok zorluk ile karşılaşılmaktadır. Fazla kaynak tüketimi ve etkileşimli arayüzlerin oluşturulamaması bu zorlukların başında gelmektedir. OKT kullanımı için web

ve masaüstü arayüzler geliştirilmektedir. Web arayüzlerinin çoğu bulut tabanlı çalışmaktadır. 2008 yılında geliştirilen web OKT dünyanın ilk bulut tabanlı konuşma tanıma motoru olarak bilinmektedir [119]. Web OKT çalışmasından sonra ortaya yeni web servis ve arayüz yaklaşımları çıkmıştır. Bu yaklaşımlar web tabanlı uygulamaların sayısının artmasına yardımcı olmuştur.

Bu tez çalışmasında, geniş kelime dağarcığına sahip Türkçe OKT'nin kolay kullanımı ve yerel sunucularda rahatlıkla çalışabilmesi için bir web servis tabanlı Türkçe OKT platformu geliştirilmiştir. Uzaktan sesli komut sistemleri, multimedya varlıklarının metne aktarılıp raporlanması vb. tüm yönetim gereksinimleri geliştirilen platform sayesinde gerçekleştirilebilmektedir.

Geliştirilen platformun ana amacı kullanıcı isteklerini mümkün olduğu kadar düşük kaynaklar ile yerine getirmektir. Diğer bir amaç ise mümkün olduğu kadar çok kullanıcıya erişebilmektir. Bu amaçla kolay kullanılabilir, ölçeklenebilir, güvenli, performansı yüksek ve katmanlı bir mimariye sahip platform geliştirilmiştir. OKT Platformunun tasarımında masaüstü veya konsol uygulaması yerine Temsili Durum Transferi (TDT, Representational State Transfer) web servis yaklaşımı tercih edilmiştir.

OKT Platformu için gerekli web servisler için Java 8 sürümü kullanılmıştır. Web servisler Apache Yazılım Vakfı tarafından geliştirilmiş açık kaynaklı bir Java uygulaması olan Apache Tomcat sunucusu üzerinde çalıştırılmıştır.

Platformun kullanıcı arayüzü, TypeScript tabanlı bir açık kaynaklı web uygulama çerçevesi olan Angular 7 kullanılarak hazırlanmıştır. Arayüz ayrıca Hiper Metin İşaretleme Dili (HMİD, Hypertext Markup Language) 5 ve açık kaynak kodlu web sayfaları veya uygulamaları geliştirmek için kullanılabilecek araçları içeren Bootstrap ile desteklenmiştir. Arayüz, Angular uygulamalarını başlatmak, geliştirmek, iskelet oluşturmak ve bakımını yapmak için kullanılan bir komut satırı arabirim aracı olan Angular komut satırı arayüzü üzerinde çalıştırılmıştır. OKT işlemi sonucunda elde edilen metin verileri Lucene kütüphanesine dayanan, Hiper Metin Transfer Protokolü (HMTP, Hyper-Text Transfer

Protocol) web arayüzü ve şema içermeyen Javascript Nesne Belirtimi (JNB, Javascript Object Notation) belgelerini kullanan dağıtılmış, çok kullanıcılı bir tam metin arama motoru olan Elasticsearch sistemine gönderilmiştir. ElasticSearch sistemine gönderilen metin verileri benzersiz olarak kimliklendirilmiştir. Bu benzersiz kimlik aynı zamanda dosya sunucusunda saklanan konuşma kayıtları içinde verilmiştir. Böylelikle metin ve konuşma kayıtları birebir eşleştirilmiştir.

4. VERİ KÜMESİNİN HAZIRLANMASI

OKT, istatistiksel örüntü sınıflamasına dayanan makine öğrenme algoritmaları ile eğitilmiş modellerle çalışmaktadır [120]. Makine öğrenme algoritmalarının eğitimi için iki ana yaklaşım kullanılmaktadır. Bu yaklaşımlar denetimsiz ve denetimli öğrenme yaklaşımlarıdır. Denetimsiz öğrenmede etiketlenmemiş verilerle kümeleme gerçekleştirilir. Denetimli öğrenme yaklaşımında ise genellikle sınıflandırma yapılmaktadır. Eğitim için etiketli verileri içeren bir eğitim veri kümesi gereklidir. OKT, SMM'ler gibi konuşma sınıflandırıcılarını eğitmek için denetimli öğrenmeyi kullanmaktadır. OKT'nin erken gelişim döneminde SMM ve GKM başarılı bir şekilde kullanılmıştır. Ancak yapay sinir ağlarının 1990'larda konuşma tanıma uygulamalarında ilk kullanımı, geleneksel GKM ve SMM teknolojisine kıyasla daha iyi performans göstermiştir [121]. YSA temelli bir yaklaşımın yüksek performans göstermesinin nedeni büyük miktarda veriyle karmaşık yapıları bulma ve öğrenme yeteneğidir [122]. Ancak YSA tabanlı yaklaşımlar, doğru sonuçlar üretmek için büyük miktarda eğitim veri kümesine ihtiyaç duymaktadır. Bu nedenle, konuşmacıdan bağımsız ve yüksek performans (yani düşük KHO) gösteren bir OKT geliştirmek için konuşmaların metinleri ile birlikte çeşitli konuşmacılardan geniş bir konuşma örneği veri kümesi hazırlanmalıdır [123].

Türkçe konuşma ve metin verilerinden oluşan yazıya dökülmüş bir konuşma topluluğu oluşturmak için yararlı ve pratik bir süreç geliştirmek önemli bir zorluk olarak görülmektedir. Yazılı konuşma verilerini toplamak için yaygın olarak kullanılan mevcut yöntemlerle ilgili en büyük sorun, sürecin hem zaman hem de bütçe açısından maliyetli olmasıdır. Ne yazık ki, tüm Türkçe tabanlı Asya dilleri, AM'leri ve DM'leri eğitmek için gerekli miktardaki veri kümesine sahip değildir. Mevcutta bulunan az miktardaki konuşma ve metin veri kümesi nedeniyle OKT uygulamaları için Türkçe düşük kaynaklı diller arasında sınıflandırılmıştır. Bu nedenle, yüksek kaynaklı dillerin veri kümelerine benzer boyutta bir Türkçe yazıya dökülmüş konuşma veri kümesi oluşturmak önem arz etmektedir.

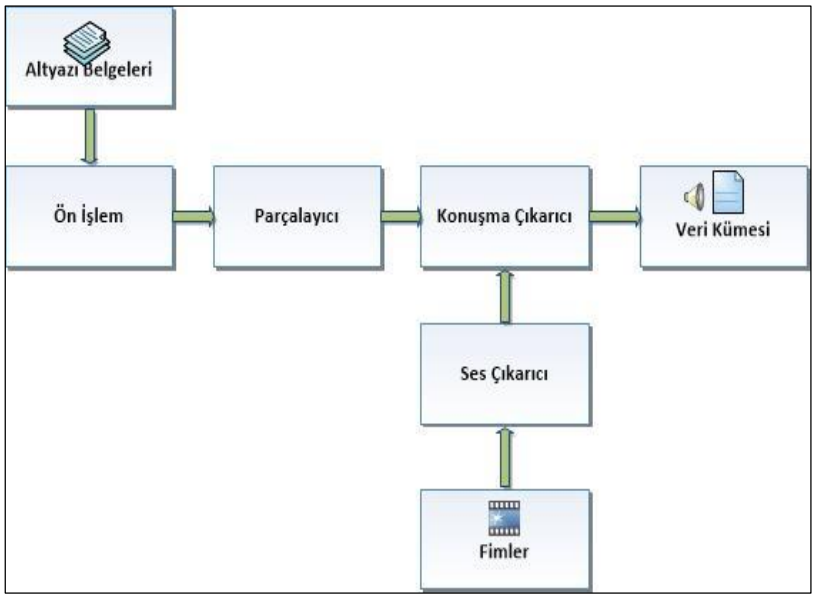
Bu çalışmada, geniş bir kelime dağarcığına sahip bir Türkçe konuşma veri kümesi hazırlanmıştır. Türkçe konuşma veri kümesi üç farklı yaklaşım ile elde edilmiştir. İlk yaklaşımda uzun metrajlı bir film kümesi, ikinci yaklaşımda bir mobil uygulama, üçüncü

yaklaşımda ise transfer öğrenme yöntemi kullanılmıştır. Bu yaklaşımlarla elde edilen Türkçe konuşma veri kümesi ayrıca gerçek kullanıcıların onayına sunulmuştur. Bu sayede daha doğru veriler elde edilmiştir.

4.1. Filmlerin ve Altyazı Belgelerinin Kullanılması

Bu çalışmada, Türkçe filmlerden toplanan konuşma ve bu konuşmalara karşılık gelen zamana göre hizalanmış altyazı belgeleri kullanılmıştır. Önerilen yöntemi kullanarak, herhangi bir dilde büyük miktarda film ve bu filmlerin altyazılarının erişilebilir olması şartıyla büyük miktarlarda veri kümesi hazırlamak mümkündür. Şekil 4.1 filmlerden veri toplama işleminin ana bileşenlerinin bir blok diyagramını göstermektedir.

Şekil 4.1'de gösterilen ön işlem bileşeni, altyazı belgelerini otomatik hizalama için optimize etmektedir. Her altyazı belgesi, OKT sisteminin eğitiminde kullanılabilecek bir biçime dönüştürülür. Altyazı belgesindeki her konuşma bölümünün metin karşılığı, zaman aralığı etiketinin hemen altındadır. Her bölüm için hazırlanan metinlerdeki noktalama işaretleri kaldırılmıştır. Ayrıca sayı ve sembollerin telaffuzu yazılı olarak sunulmuştur. Altyazı belgelerindeki metinler bu işlemlerden sonra parçalayıcı bileşeni için hazırdır.



Şekil 4.1. Filmlerden Türkçe konuşma veri kümesi toplama süreci

Parçalayıcı bileşeni, konuşmaya karşılık gelen metni altyazı belgesinden ayıklamak için kullanılır. Bu bileşende, her konuşma bölümü altyazı belgesinden başka bir belgeye kopyalanmıştır. Ardından farklı bir metin dosyası olarak kaydedilmiştir. Bu işlem sırasında metin dosyasına verilen kimlik, konuşma dosyasına verilecek şekilde hafızada tutulmuştur. Altyazı belgelerinden elde edilen metin birimlerine karşılık gelen konuşma dosyaları, ses çıkarıcı kullanılarak oluşturulmuştur.

Ayrıştırma ve filmlerden konuşma çıkarma işleminden önce, ses çıkarıcı bileşeni yardımı ile film videolarından ses bilgisi elde edilmiştir. Konuşma çıkarıcı bileşeni sayesinde, parçalayıcı bileşeni tarafından elde edilen metin bilgisine karşılık gelen konuşma bilgisi elde edilmiştir. Konuşma bilgilerinin zaman aralığı, altyazı belgesindeki her konuşma bölümünün başında yer almaktadır. Konuşma çıkarıcı işlemi bu zaman aralığı bilgisine göre gerçekleştirilmiştir. Parçalayıcının çıkışında kaydedilen her bir metin dosyasına sağlanan kimlik, konuşma çıkarıcı bileşenin çıkışındaki ses dosyasına verilmiştir. Böylece, veri kümesi için gerekli konuşma ve birebir metin eşleştirmesi sağlanmıştır.

Birebir metin karşılığı olan bir Türkçe konuşma veri kümesi oluşturmak için toplam 150 Türkçe film belirlenmiştir. Filmler, özetleri ve türleri kontrol edilerek belirlenmiştir. Bununla birlikte, toplanan filmlerin tamamı veri kümesi oluşturulması için kullanılmamıştır. Bu filmlerin bazılarının konuşma verilerinin düşük konuşma-gürültü oranı veya çok sayıda konuşmacı nedeniyle uygun olmadığı görülmüştür. Ön inceleme sürecinden sonra 150 Türkçe filmden ancak 120 tanesi konuşma veri kümesi oluşturmak için seçilmiştir. Genellikle, bir konuşma veri kümesinin saat cinsinden uzunluğu, AM'yi eğitmede ne kadar yararlı olabileceğinin bir göstergesi olarak kabul edilmektedir. Bu çalışmada 90 saatlik Türkçe konuşma veri kümesi altyazılı Türkçe filmlerinden çıkarılmıştır. Çıkarılan konuşmanın metin karşılığı her filmle ilişkili önceden hazırlanmış altyazılardan elde edilmiştir. Çizelge 4.1, seçilen filmlerden çıkarılan konuşma verilerinin istatistiklerini özetlemektedir.

Çizelge 4.1'de gösterildiği gibi, film başına ortalama 45,0 dakikalık konuşma ve metin önerilen yöntem kullanılarak elde edilmiştir. Çizelge 4.1 ayrıca film başına kelime sayısını ve benzersiz kelime sayısını (kelime büyüklüğü) göstermektedir.

Çizelge 4. 1. Filmlerden elde edilen veri kümesi ile ilgili istatistikler

Film No.	Kelime Sayısı	Benzersiz Kelime Sayısı	Uzunluk (dakika)
Film_1	4,056	1,308	55.0
Film_2	2,306	900	62.0
…	…	…	…
Film_120	5,440	1,605	44.0
Ort.	4,842	1,229	45.0

4.1.1. Altyazıların ön işlenmesi

Ön işlem bileşeninin amacı her altyazıyı veya veriyi konuşma tanıma geliştirme araçları tarafından kullanılabilecek bir biçime dönüştürmektir. Şekil 4.2'de Türkçe film altyazı belgesinin bir örneği verilmiştir.

```
1
00:01:31,540 --> 00:01:32,256
Evet burası çok güzel !!

2
00:01:33,860 --> 00:01:38,058
Peki siz ne zaman geliyorsunuz???

3
00:01:49,060 --> 00:01:54,373
[MUSIC]

4
00:01:55,420 --> 00:02:02,019
Aslında ben arabayı alıp gitmek istiyordum.

5
00:02:02,380 --> 00:02:02,892
<i>Buyurun gidelim.</i>
```

Şekil 4.2. Türkçe film altyazı belgesi örneği

Şekil 4.2'de gösterildiği gibi konuşmanın her bir bölümü (zaman aralığı) numaralandırılmıştır. Birbirini izleyen bölümler arasına bir çizgi yerleştirilmiştir. Her konuşma bölümünün transkripsiyonuna karşılık gelen zaman aralığı metin etiketlerinin hemen altında yer almaktadır.

Türkçe' de, özellikle yazım, kodlama ve biçimlendirme hatalarına eğilimli olan harfler sıklıkla kullanılır. Örneğin; gereksiz yeni satırlar, yanlış yazılmış kelimeler, metin kodlama hataları ve değişken sayı biçimleri Türkçe altyazılarda en sık gözlenen tutarsızlıklardır. Ek olarak, altyazı belgeleri değişen Unicode standartlarına sahip farklı metin editörlerinde hazırlanmaktadır. Bu farklılıklar bazı Türkçe karakterlerin yanlış görüntülenmesine neden olmaktadır. Örneğin, Türkçe karakterler Ü, Ç, Ö, İ, Ş ve Ğ farklı metin editörlerinde farklı kodlama standartlarına sahiptir. Bu nedenle bu çalışmada tutarlılık ve doğruluk için tüm altyazı belgeleri tek bir Unicode standardında yeniden kodlanmıştır (UTF-8). Ayrıca ön işlem adımında, alfa sayısal olmayan karakterler metin dosyalarından çıkarılmıştır. Bir altyazı belgesindeki tüm sayılar resmi kelime okunuş formlarına dönüştürülmüştür. Bu işlem ile AM eğitimi sırasında sayıların fonetik temsili gerçekleştirilmiştir. Örneğin 86 rakamı kendisine karşılık gelen yazılı form "seksen altı" ya dönüştürülmüştür.

4.1.2. Altyazı belgelerinde yanlış yazılan kelimeler

Yanlış yazılmış kelimeler, AM ve DM eğitimi için kullanılan istatistiksel dağılımları olumsuz etkilemektedir. Bu nedenle ön işlem sırasında yazım hataları tespit edilmiş ve düzeltilmiştir. Altyazı belgelerinde sıklıkla karşılaştığımız üç farklı yazım hatası kategorisi oluşmuştur. Bunlar;

Ekleme hataları: Kelimeye bir veya daha fazla harfin eklenmesidir. Örneğin, yanlış yazılmış "merrhabaa" kelimesinde iki harf ekleme hatası vardır ve doğru form "merhaba" olmalıdır.

Silme hataları: Kelimenin doğru biçiminde bir veya daha fazla harfin eksik olmasıdır. Örneğin, "arbala" kelimesinde iki harf silinmiştir. Doğru form "arabalar" olmalıdır.

Değiştirme hataları: Orijinal kelimede bir veya daha fazla harfin diğer harflerle değiştirilmesidir. Örneğin, "çanda" kelimesinde bir değiştirme hatası vardır ve doğru form "çanta" olmalıdır.

Bu hataların bir kısmı bir Türkçe yazım denetimi çerçevesi kullanılarak otomatik olarak giderilmiştir. Yanlış yazılan kelimeleri düzeltmek için Zemberek aracı kullanılmıştır [123]. Başta Türkçe olmak üzere Türkçe kökenli diller için tasarlanan Zemberek, bir dizi açık kaynaklı doğal dil işleme kütüphanesi ve aracıdır. Çizelge 4.2, altyazılardan elde edilen veri kümesi istatistiklerini göstermektedir.

Çizelge 4.2. Altyazı belgelerinden elde edilen veri kümesi istatistikleri

Veri Kümesi Adı	Toplam Kelime Sayısı	Sözlük Boyutu	Kayıp (%)
Film veri kümesi	670,546	100,128	5,2

Çizelge 4.2'de gösterildiği gibi toplam kelimelerin % 5,2'si yanlış yazılmıştır. Yanlış yazılan kelimeler konuşma tanıma için yanlış sesbirim dizeleri vereceğinden düzeltilmiştir. Bu işlem veri kümesinin doğruluğunu artırarak sonuçta konuşma tanıma için gerekli veri kümesinin kalitesini arttırmıştır.

4.1.3. Filmlerin konuşmalara ayrılması

Ön işlem adımları tamamlandıktan sonra zamana göre ayarlanmış altyazı metinleri, konuşma verilerinin bölümlere ayrılmasında kullanılmıştır. İlk olarak her filmden çıkarılan ses WAV formatında ayrı bir dosya olarak kaydedilmiştir. Konuşma tanıma için genel tercih olduğundan örnekleme hızı ve örnekleme boyutu 16 kHz ve 16 bit olarak ayarlanmıştır.

Konuşma bölümleri ve altyazı belgeleri ile eşleştirilen dosyalar arasındaki yanlış hizalamaları tanımlamak ve düzeltmek için çıkarılan tüm ses dosyalarının altyazıları ile birlikte kontrol edilmiştir. Ses dosyasındaki bir konuşma bölümünün başlangıç ve bitiş zamanları altyazı belgesindeki başlangıç ve bitiş zamanlarıyla eşleşmediğinde yanlış hizalama meydana gelmektedir. Örneğin ses dosyasının 120. saniyesinde bir cümle başlamakta ve 150. saniyede sona ermektedir. Ancak altyazı belgesinde bu cümle 125 ve 155. saniyeler arasında konuşulmaktadır. Yanlış hizalamalar, altyazılardan rastgele örnekler seçerek kontrol edilmiştir. Başlangıç zamanını kullanan ve hata oranının belirli bir eşiğin

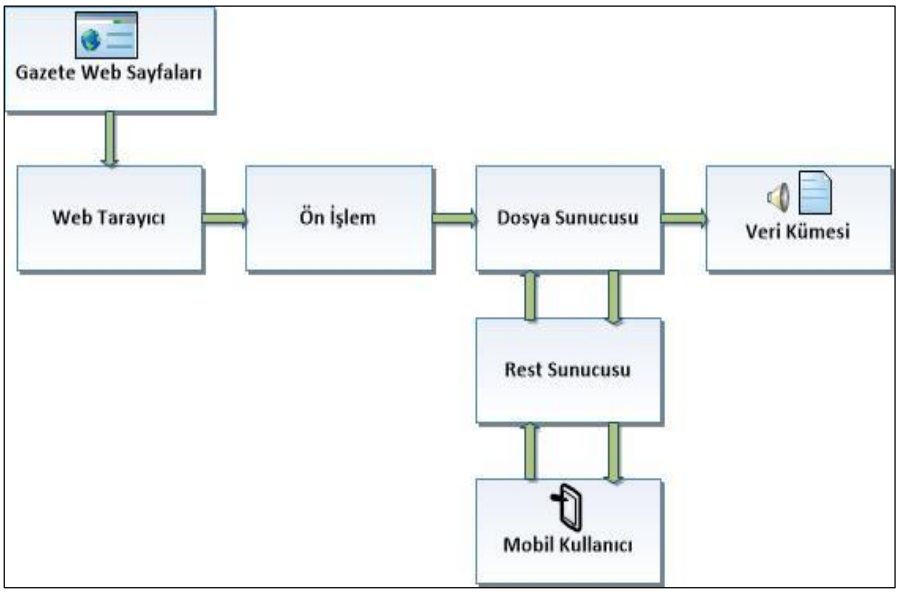
üstünde olup olmadığını kontrol eden bir konuşma tanıma testi uygulanmıştır. Bu testi geçemeyen filmler, yanlış hizalanmaya neden olan zaman kaymasını belirlemek için gerçek kullanıcılar ile bir dinleme testi yapılarak dikkatle incelenmiştir. Zaman kayması film boyunca tutarlı ise altyazı zamanı ek açıklamasının düzeltilmesi gerçekleştirilmiştir. Eğer zaman kayması tutarlı değil ise film güvenilmez olarak işaretlenmiş ve veri kümesinden çıkarılmıştır.

4.2. Mobil Uygulama ile Veri Kümesinin Elde Edilmesi

Türkçe, sondan eklemeli bir dildir ve kelime dağarcığı diğer dillere göre çok geniştir. Bu nedenle sadece filmler içerisinde geçen konuşmalardan elde edilen bir veri kümesi ile geniş kelime dağarcığına sahip Türkçe OKT geliştirmek mümkün değildir. Bu çalışmada, veri kümesinin boyutunu artırmak için sadece filmler kullanılmamıştır, bununla beraber farklı alanlarda (dergi, spor, teknoloji, gündem, vb.) konuşma ve metin verilerini elde etmek için bir mobil uygulama geliştirilmiştir. Şekil 4.3, mobil uygulama üzerinden veri toplama işleminin ana bileşenlerinin bir blok diyagramını göstermektedir.

Yeni yayınlanan gazete makaleleri bir web tarayıcısı yardımıyla elde edilmiştir. Web tarayıcı bileşeni, yayınlanmış Türkçe gazetelerin web sayfalarına belirli zamanlarda erişim sağlamıştır. Bu süreç günde bir kez ile sınırlandırılmıştır. Belirlenen gazete makaleleri spor, magazin ve genel makaleler olarak sınıflandırılmıştır. Ön işlem bileşeni, gazetelerden elde edilen metinleri birer cümle olacak şekilde parçalamaktadır. Bu aşamada kelime kaybının olmamasına dikkat edilmiştir. Ön işlem bileşeninde tek tek cümleler tanımlanmış ve dosya sunucusuna aktarılmıştır.

TDT sunucusunda çalışan bir web servis Uygulama Programlama Arayüzü (UPA, Application Programming Interface) önceden hazırlanmış metinleri bir mobil kullanıcıya göndermiştir [124]. Mobil kullanıcı bileşeni, metni kullanıcıya göstermekte ve kullanıcıdan metni seslendirmesini istemektedir. Oluşturulan ses kayıt dosyası tanımlanarak dosya sunucusuna aktarılmıştır. Böylece kullanıcıların seslendirdiği metin ve ses dosyaları eşleştirilerek saklanmıştır. Bu çalışmada geliştirilen ve gerçek kullanıcılar tarafından kullanılan mobil uygulamanın ekran görüntüsü Şekil 4.4'te verilmiştir.



Şekil 4. 3. Mobil uygulama ile veri kümesinin elde edilmesi

Şekil 4.4 Mobil uygulama arayüzü

Şekil 4.4'te gösterilen mobil uygulama ile kullanıcılara gerekli metin bilgisi sağlanmıştır. Ancak kullanıcıdan seslendirme işlemine başlamadan önce bazı bilgileri girmesi istenmiştir.

Bu bilgiler cinsiyet, yaş ve isim bilgisidir. Veri kümesi hazırlanırken cinsiyet dağılımının orantılı olması için bu bilgiler gereklidir. Ayrıca bu çalışmada mobil uygulama kullanıcılarının veri kümesine katkısını ölçmek için bir web arayüzü geliştirilmiştir. Mobil kullanıcıların ana dili Türkçe'dir. Mobil kullanıcı istatistiklerini sağlayan web arayüzü Şekil 4.5'te verilmiştir.

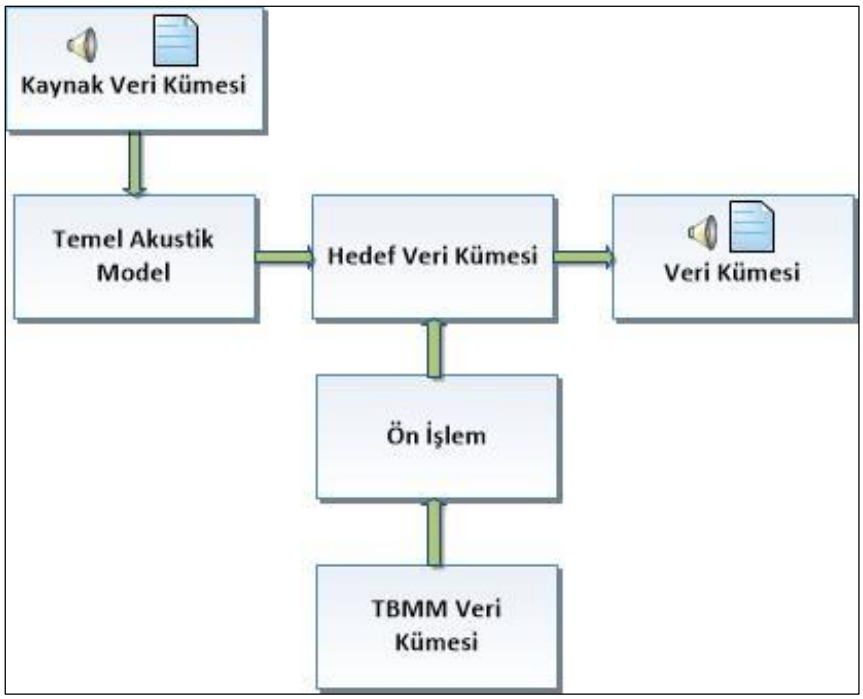
Şekil 4.5. Mobil kullanıcı istatistikleri için web arayüzü

İstatistikler kimlerin kaç oturum açtığı, en son ne zaman veri gönderdiği ve ne kadar veri işlediği gibi mobil uygulama tarafından sağlanan bilgiler ile elde edilmiştir. Bu bilgiler geliştirilen web arayüzü üzerinden takip edilmiştir. Kullanıcıların ortalama 55 dakikalık ses kaydı oluşturduğu tespit edilmiştir. Ayrıca mobil kullanıcıların yaşları 19-52 arasında ve yaş ortalaması 26 olarak belirlenmiştir. Geliştirilen mobil uygulama toplamda 130 gerçek kişi tarafından kullanılmıştır (50 erkek ve 80 kadın). Bu yöntemle toplam 120 saatlik konuşma veri kümesi elde edilmiştir.

4.3. Transfer Öğrenme Yaklaşımı ile Veri Kümesinin Elde Edilmesi

Transfer öğrenimi, bir modelden diğerine bilgi aktarmak için makine öğrenmesi sürecinde kullanılan standart bir yöntemdir [97]. Transfer öğreniminin amacı mevcut kaynak modeli, sınırlı hedef verileri olan yeni bir göreve uyarlamaktır. Bu çalışmada Türkiye Büyük Millet Meclisi (TBMM, Grand National Assembly of Turkey) oturum kayıtları üzerinde transfer öğrenimi gerçekleştirilmiştir.

Transfer öğrenimi için GKM-SMM tabanlı temel düzeyde bir Türkçe OKT sistemi geliştirilmiştir. AM eğitimi açık kaynaklı Kaldi konuşma tanıma araç seti ile gerçekleştirilmiştir. Kaynak AM'i eğitmek için kullanılan veri kümesi film altyazılarından ve mobil uygulamalardan elde edilmiştir. Kaynak AM için gerekli veri kümesi toplam 210 saatlik konuşma verisinden oluşmaktadır. Kaynak DM eğitimi için, Çizelge 4.2'de verilen istatistiklere sahip bir metin veri kümesi kullanılmıştır. Bu veri kümesi kullanılarak, 3-gram tabanlı bir dil modeli geliştirilmiştir. Elde edilen Türkçe OKT, TBMM'nin konuşma verilerine hedef veri olarak uygulanmıştır. Şekil 4.6, transfer öğrenimi ile veri toplama işleminde kullanılan ana bileşenlerin bir blok diyagramını göstermektedir.



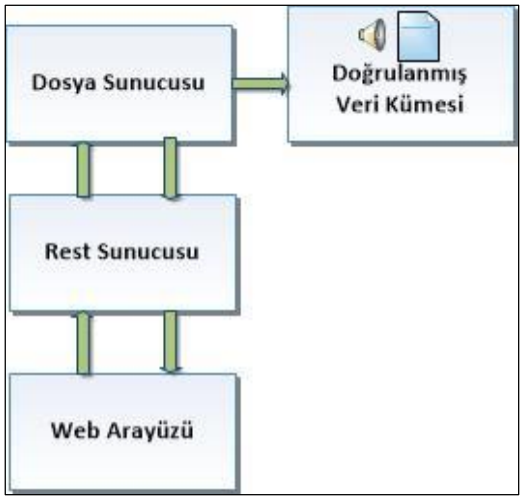
Şekil 4.6. Transfer öğrenme yaklaşımının kullanımı

TBMM veri kümesi iki farklı dosya içermektedir. Bu dosyalar TBMM oturum video kayıtları ve bu kayıtların metin raporlarıdır. TBMM oturum raporları ve ilgili videolar gerçek kullanıcılar tarafından daha önce eşleştirilmiştir. Bu eşleştirme ön işlem süreçlerini kolaylaştırmıştır. Ön işlem bileşeninde, öncelikle video görüntülerinden ses dosyaları çıkarılmıştır. Ayrılan ses dosyaları 15-25 saniye aralıklar ile parçalara ayrılmıştır. Bu ses dosyaları önceden kaynak veri kümesi ile eğitilmiş OKT sistemine girdi olarak verilmiştir. OKT sisteminin çıktısında elde edilen metinler TBMM raporlarında aranmış ve ilgili

metinler bulunmuştur. Transfer öğrenme yöntemi ile toplamda 250 saatlik konuşma veri kümesi elde edilmiştir.

4.4. Veri Kümesi Doğrulama İşlemleri

Film altyazı belgeleri, mobil uygulama ve transfer öğrenimi yaklaşımı kullanılarak TBMM oturum kayıtları üzerinden elde edilen veri kümesi gerçek kullanıcılar tarafından doğrulanmıştır. Doğrulama işlem için bir web arayüzü geliştirilmiştir. Şekil 4.7, veri kümesi doğrulama işleminin ana bileşenlerini göstermektedir.



Şekil 4.7. Veri kümesi doğrulama işlemleri

Dosya sunucusundaki ses-metin eşlemeleri tek tek alınarak bir TDT web servisi yardımıyla web arayüzüne aktarılmıştır. Gerçek kullanıcılar, web arayüzü üzerinden ses-metin eşleşmelerini kontrol etmiştir. Bu kontrol sürecinde kullanıcılar kendilerine gönderilen ses verilerini dinlemiştir. Ayrıca ses dosyasıyla birlikte alınan metin bilgilerini de kontrol etmiştir. Doğrulama işlemi için oluşturulan web arayüzü Şekil 4.8'de verilmiştir.

Şekil 4.8'de görüldüğü gibi kullanıcılar, ses-metin eşlemesini kontrol edebilmekte ve doğrulayabilmektedir. Kullanıcılar kendilerine verilen metni düzeltebilmekte ve dosya sunucusuna geri gönderebilmektedir. Ayrıca ses verilerinde gürültü, üst üste binme veya

belirsizlik var ise bu verileri belirtilen duruma göre etiketleyebilmektedir. Doğrulama işlemi sonucunda gerçek kullanıcılar tarafından kontrol edilmiş bir veri kümesi elde edilmiştir.

Şekil 4.8. Veri kümesi doğrulama web arayüzü

4.5. Test Veri Kümesinin Oluşturulması

Literatürdeki birçok çalışmanın alana özgü (hukuk, medikal, spor vb.) çalışmalardan oluştuğu gözlenmiştir. Alana özgü çalışmalarda konuşulan kelime dağarcığı belirli bir kümeye aittir. Ancak kendiliğinden (spontane) konuşmaları belirli bir kümeye ait bilgiler ile gerçekleştirmek mümkün değildir. Bu nedenle bu çalışma kapsamında geniş kelime dağarcığına sahip bir veri kümesi hazırlanmıştır.

Gerçek hayatta konuşmalar farklı konulardan oluşmaktadır. Örneğin spor, siyaset veya bilim bu konulardan birkaçını temsil etmektedir. Bu çalışmada geliştirilen geniş kelime dağarcığına sahip Türkçe OKT sistemini test edebilmek için farklı alanlardan oluşan bir test veri kümesi hazırlanmıştır. Test veri kümesi için belirlenen alanlar aşağıda listelenmiştir.

- Siyaset
- Bilim
- Ekonomi
- Spor
- Kültür
- Magazin
- Sanat

- Hukuk
- Teknoloji
- Sağlık

Belirtilen alanlar dikkate alınarak hazırlanan test veri kümesinin detayları Çizelge 4.3'te verilmiştir. Test veri kümesi için gerekli veriler Youtube üzerinden elde edilmiştir. Öncelikle video kayıtlarının alanı belirlenmiştir. Ardından video içerisindeki konuşmalar gerçek kullanıcılar tarafından metne aktarılmıştır.

Çizelge 4.3. Test veri kümesine ait bilgiler

Konuşmacı	Adet	Kelime Sayısı	Süre (Dakika)
Farklı Erkek Konuşmacı Sayısı	175	1575	142.91
Farklı Kadın Konuşmacı Sayısı	47	423	38.38
Toplam	222	1998	181.29

Çizelge 4.3'de detayları verilen test veri kümesi bu çalışmada geliştirilen geniş kelime dağarcığına sahip Türkçe OKT sisteminin başarımını değerlendirmede kullanılmıştır. Ayrıca bu test veri kümesi bulut üzerinden hizmet veren Google Türkçe OKT sistemi üzerinde de kullanılmıştır. Böylelikle Google ile karşılaştırmalı deneysel sonuçlar için bir zemin oluşturulmuştur.

5. SİSTEMİN GELİŞTİRİLMESİ VE DENEYSEL SONUÇLAR

Bu bölümde, geniş bir kelime dağarcığına sahip veri kümesi kullanılarak Türkçe OKT sisteminin geliştirilmesi ayrıntılı olarak açıklanmıştır. OKT sistemi geliştirilirken Türkçe'yi daha iyi modelleyebilmek için AM ve DM üzerinde iyileştirmeler yapılmıştır. Ayrıca geniş kelime dağarcığına sahip OS ile sistem desteklenmiştir. OKT sistemine eklenen ön işlem bileşeni ile sessizlik bilgisinin ortadan kaldırılması sağlanmıştır. Geliştirilen sistemin her aşamasında deneysel sonuçlara yer verilmiştir. Deneylerde veri kümesi boyutunun önemini ortaya koyabilmek için mevcut Boğaziçi ve Orta Doğu Teknik Üniversitesi (ODTÜ, Middle East Technical University) veri kümesi ile karşılaştırmalı deneyler gerçekleştirilmiştir. Klasik GKM-SMM yaklaşımı ve DSA temelli modeller kullanarak Türkçe OKT sisteminin geliştirilmesi ayrıntılı olarak verilmiştir.

5.1. Konuşma Tanıma için Gerekli Veri Kümesi

Türkçe üzerine gerçekleştirilen veri kümesi çalışmaları yakın geçmişi işaret etmektedir. İlk Türkçe konuşma veri kümesi 2006 yılında ODTÜ tarafından hazırlanmıştır. Yaklaşık 500 dakikalık konuşma ve konuşma dosyaları ile eşleştirilmiş metin verisini içermektedir [17]. 2012 yılında Boğaziçi Üniversitesi tarafından bir başka Türkçe veri kümesi hazırlanmıştır. Boğaziçi Üniversitesi tarafından hazırlanan veri kümesi yaklaşık 130 saatlik konuşma verisi içermektedir. Ancak deneylerimizde bu veri kümesinin boyutunun daha düşük olduğu sonucuna varılmıştır. Her iki veri kümesi içerisinde geçen konuşmalar ve konuşmacılar farklılık göstermektedir. Her iki veri kümesi de 16 bit çözünürlük ve 16 kHz örnekleme frekansına sahiptir.

Bu çalışmada, gerçekleştirilen deneylerde farklı bit çözünürlükleri ve farklı örnekleme frekansına sahip konuşma dosyaları da 16 bit çözünürlük ve 16 kHz örnekleme frekansına sabitlenmiştir. Böylece, daha doğru sonuçlar elde edebilmek için akustik özellikler birbirlerine benzetilmiştir. Örnekleme frekansının konuşma tanıma deneylerindeki başarımını gösterebilmek için farklı örnekleme frekansına sahip dosyalar üzerinde deneyler gerçekleştirilmiştir. Böylelikle farklı örnekleme frekanslarının konuşma tanıma performansı üzerindeki etkisi araştırılmıştır.

5.2. Geliştirme Parametreleri

Ses dosyalarından konuşmaya ait özniteliklerin elde edilmesi ve geliştirilen modellere bilgi girişinin sağlanması için öznitelik çıkarma işlemi gerçekleştirilmiştir. Öznitelik çıkarma işlemi için MFKK kullanılmıştır. Öznitelik çıkarma işlemleri için kullanılan Mel frekans ölçeği, insan kulağının duyabileceği frekanstaki değişimin algısını net olarak gösterebilmektedir. 1000 Hz'e kadar seslerin algılanması doğrusal iken frekans arttıkça değişimin algısı logaritmik hale gelmektedir. MFKK hesaplaması aşağıda belirtilen adımlarda gerçekleştirilmiştir.

Ön vurgulama: Ön vurgulama bir sinyalin yüksek frekans bileşenlerini daha baskın hale getirmek için kullanılmaktadır. OKT uygulamalarında MFKK çıkarmak için bir ön işlem olarak gerçekleştirilmektedir. Ön sentezleme olarak da bilinen bu işlemdeki amaç ses sinyalini daha yüksek frekanslara yükseltmek için bir filtre uygulamaktır. Uygulanan bu filtre yüksek örnekleme frekanslarının daha küçük genliklere sahip olması nedeniyle frekans spektrumunu dengeleme görevine yardımcı olmaktadır. Bu işlem aynı zamanda, HFD işlemi uygulandığında ortaya çıkan sayısal problemleri önlemek ve sinyal-gürültü oranını iyileştirmek için de kullanılmaktadır. Ön vurgulama bir sinyale $y(t) = x(t) - ax(t-1)$ şeklinde uygulanır. Tez çalışmasında ön vurgulama işlemi için filtre α katsayısı 0,97 olarak belirlenmiştir.

Çerçeveleme: Sinyali kısa süreli çerçevelere bölmek için ön vurgunun ardından, çerçeveleme işlemi gerçekleştirilmektedir. Bu adım, ses sinyalindeki frekansların zamanla değişmesi nedeniyle yapılmaktadır. Çerçeveler genellikle 20 ms - 40 ms arasında periyodlara bölünmekte ve ardışık çerçeveler arasında % 50 oranında örtüşme oranı kullanılmaktadır. Bu çalışmada öncelikle her bir konuşma dosyasındaki çerçeve sayısı hesaplanmıştır (her seferinde 10 ms kaydırılan 25 ms'lik çerçeveler elde edilmiştir). Çerçevelere ayrılan konuşma sinyalinde olabilecek kayıpları önlemek için 10 ms'lik örtüşme işlemi gerçekleştirilmiştir. Bu işlem bir çerçeveden diğerine geçişin yumuşak olmasını sağlamıştır.

Pencereleme: Ses sinyaline uygulanan çerçeveleme işleminden sonra pencereleme işlemi gerçekleştirilmektedir. Pencereleme işlemindeki amaç çerçeveleme işleminden kaynaklı

spektral bozulmaları önlemek ve ses sinyalinde oluşabilecek süreksizlikleri gidermektir. Yaygın olarak kullanılan birçok farklı pencereleme fonksiyonu mevcuttur. Bu çalışmada Hamming pencere fonksiyonu kullanılmıştır. Hamming penceresinin uygulanmasının nedeni HFD tarafından sinyalin sabit olduğu varsayımını önlemek ve spektral sızıntıyı azaltmaktır.

Hızlı fourier dönüşümü: N adet örnekten oluşan ses sinyalinin zaman alanından frekans alanına dönüştürmek amacıyla HFD işlemi gerçekleştirilir. Pencerelenen sinyalin genlik spektrumu HFD ile hesaplanmaktadır. Bu çalışmada Frekans spektrumunu hesaplamak için her karede bir N noktası için HFD gerçekleştirilmiştir. N değeri, 256 olarak belirlenmiştir.

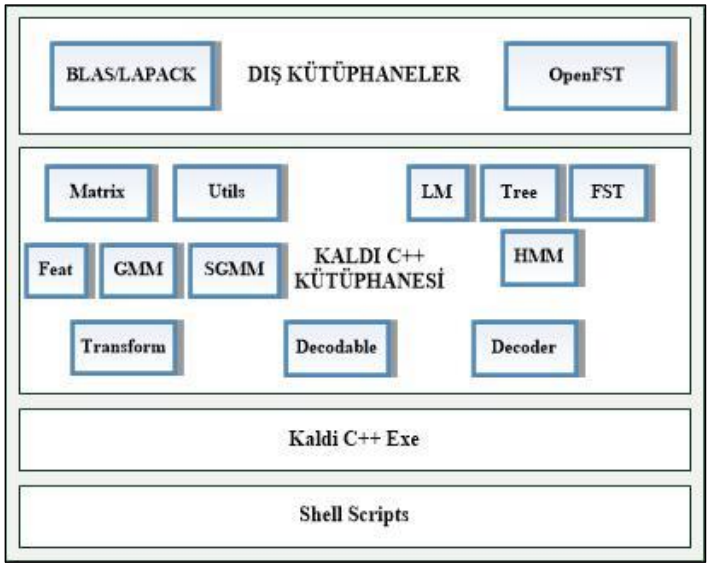
Mel filtre bankası: HFD ile genlik spektrumu hesaplanan ses sinyali, Mel ölçeklerinde yerleştirilmiş üçgen filtre bankasından geçirilmektedir. Mel ölçekli filtreler, 1 kHz'e kadar doğrusal, 1 kHz'den daha yüksek frekanslarda logaritmik aralıklarla değişmektedir. Bu çalışmada, 27 adet Mel ölçekte yerleştirilmiş üçgen filtre bankası kullanılmıştır.

MFKK özellikleri: Ses sinyalinin filtre bankası çıkışında elde edilen işaretin logaritması alınarak özniteliklerin dinamik değişimlere karşı daha az hassas olması sağlanmaktadır. Logaritmik filtre çıkışlarına yapılan Ayrık Kosinüs Dönüşümü (AKD, Discrete Cosine Transform) ile frekans alanından tekrar zaman alanına geçilerek MFKK katsayıları elde edilmiştir.

5.3. Akustik Model

Bu çalışmada geliştirilen OKT sistemi ilk olarak klasik GKM-SMM tabanında daha sonra da DSA tabanlı olarak geliştirilmiştir. Bu iki farklı OKT modeli aynı veri kümeleri kullanılarak geliştirilmiştir. Akustik modelin eğitimi sırasında dil modellemesi, zaman kazanmak için sürekli olarak değiştirilmemiştir. Bu durum geliştirme aşamasında zaman kazandırmıştır. DM, GKM ve DSA için kullanılan istatistiksel metin veri kümesi kullanılarak üretilmiştir. Ek olarak, sözlük boyutu da sabit tutulmuştur. Böylelikle akustik modelin etkisi daha net ortaya çıkmıştır.

OKT sistemi geliştirilirken Kaldi araç setinden faydalanılmıştır. Kaldi, C ++ ile yazılmış ve "Apache License v2.0" altında lisanslanan konuşma tanıma uygulamaları için kullanılan açık kaynaklı araç setidir [125]. Şekil 5.1'de Kaldi araç setinin blok diyagramı verilmiştir.



Şekil 5.1. Kaldi araç seti [125]

Kaldi araç seti, temel olarak iki harici kütüphaneye bağlıdır. Bunlardan ilki sonlu durum çerçevesi için kullanılan "OpenFst", diğeri ise sayısal cebir kütüphanesidir. Sayısal cebir kütüphanesi "BLAS" ve "LAPACK" olarak ikiye ayrılmıştır. Kaldi kütüphane bileşenleri, her biri harici kütüphanelerden sadece birine bağlı olacak şekilde konumlandırılmıştır. Kütüphane işlevlerine erişim C++ dilinde yazılmış kod parçacıkları ile sağlanmaktadır. Kaldi'de hazırlanan kod parçacıkları ve kütüphaneler konuşma tanıma sistemini oluşturmak ve çalıştırmak için betik dili tarafından çağrılmaktadır.

Bu çalışmada Kaldi araç setinin 5.0 sürümü kullanılarak bir Türkçe OKT sistemi geliştirilmiştir. Klasik GKM-SMM tabanlı OKT sistemindeki AM, Gauss karışım modelleri kullanılarak geliştirilmiştir. Sesbirimlere dayanarak temel GKM modelleri oluşturulmuştur. Daha sonra eğitim veri kümesindeki metinlerin sesbirimlere kodlanması gerçekleştirilmiştir. Sesbirim modellerinin birleştirildiği bir eğitim örneğinden oluşturulacak klasik GKM modeli, eğitim örneğindeki metnin kelime tabanlı analizi ve okunuş sözlüğündeki kelimelere karşılık gelen sesbirim analizi ile belirlenmiştir. Tüm sesbirimler için GKM model

parametreleri çok sayıda veri kümesi örneği üzerinde eğitilmiştir. Bu çalışmada AM eğitimi için hazırlanan veri kümesi, model parametrelerini tahmin etmek için yeterince büyüktür.

AM'ler SMM, GKM ve DSA kullanılarak birçok şekilde üretilebilir. Bu çalışmada Altuzay Gauss Karışım Modeli (AGKM, Subspace Gaussian Mixture Model) tabanlı akustik modeller kullanılmıştır. AGKM, güçlü konuşmacı adaptasyonu sayesinde GKM akustik modellemesi için etkili bir yöntemdir. AGKM temel veri kümesi dağılımı olarak Gauss karışımlarını kullanmaktadır. AGKM ayrıca ortalama, kovaryans ve karışım ağırlıkları gibi duruma bağlı GKM parametrelerini de içermektedir. Belirtilen bağımlı parametreler küçük boyutlu altuzay için hesaplanmıştır. Küçük boyutlu bir altuzay oluşturmak için Gauss bileşenlerinin ortalaması alınmıştır. Konuşmacı adaptasyon eğitimi için Maksimum Olasılık Doğrusal Dönüşüm (MODD, Maximum Likelihood Linear Transform) ve Doğrusal Ayrımcı Analizi (DAA, Linear Discriminant Analysis) kullanılmıştır. Bir DAA-MLLT modeli oluşturmak için DAA özelliklerinin üstüne MODD uygulanmıştır. Daha sonra DAA-MODD modellerinin üzerinde konuşmacı adaptasyon eğitimi gerçekleştirilmiştir. GKM modellerimizde toplam 3500 Bağlam Bağımlı (BB, Context Dependency) triphone durumu bulunmaktadır. Ayrıca, durum başına ortalama 14 Gauss bileşenine sahiptir. Her triphone yapısı üç BB durumu ile modellenmiştir.

OKT sistemi ilk olarak klasik GKM-SMM tabanlı olarak geliştirilmiştir. GKM modellemesinin ardından DSA tabanlı bir AM geliştirilmiştir. Sinir ağlarını kullanan OKT sistemi, klasik GKM tabanlı OKT sisteminden farklıdır. Klasik GKM-SMM kullanan OKT sisteminde her SMM durumunun gözlemlenme olasılığı GKM kullanılarak hesaplanmıştır. Fakat DSA tabanlı AM'de, her bir SMM durumunun gözlemlenme olasılığı bir DSA kullanılarak hesaplanmıştır. DSA tabanlı AM'de, ses özellikleri alınarak bu özelliklere bir sesbirim etiketi atanmıştır. DSA'yı eğitmek için öncelikle geleneksel GKM-SMM sistemi tarafından üretilen sesbirim-ses hizalamaları elde edilmiştir. Bu nedenle DSA akustik modellemesi bir GKM-SMM'yi eğitmek için kullanılan MFKK özelliklerine ve GKM-SMM tarafından üretilen karar ağacına bağlı kalmıştır.

Bu çalışmada, DSA için gerekli akustik özellik vektörleri ilk katmana yerleştirilmiştir. DSA, akustik her çerçeveye bir sesbirim etiketi atayacak şekilde tasarlanmıştır. DSA tabanlı AM,

DSA'ya girdi olarak MFKK vektörleri kullanılarak geliştirilmiştir. MFKK'deki her bir çerçeve başına sesbirim durumu olasılıkları (SMM durumları) tahmin edilmiştir. Ardından tahmin edilen sesbirim durumu arka planlarını SMM durumu ile karşılaştırmak için Stokastik Gradyan İniş (SGİ, Stochastic Gradient Descent) metodu uygulanmıştır. Bu işlemin ardından DSA'nın ağırlıkları ayarlanmıştır. DSA'nın eğitimini hızlandırmak için gradyan, mini parça (mini batch) olarak bilinen küçük bir eğitim verisi üzerinden hesaplanmıştır. Her mini parça hesaplanmasından sonra ağırlıklar güncellenmiştir. DSA ağında her biri 500 nöron içeren beş gizli katman kullanılmıştır. GİB, DSA tabanlı modeli, MİB ise GKM tabanlı modeli eğitmek için kullanılmıştır. Bu iki işlem arasındaki fark eğitim hızını etkilemektedir. GİB'deki deneyler ve eğitim NVIDIA TITAN X grafik kartında gerçekleştirilmiştir. Öğrenme katsayısı 0.0000025 ve tekrar sayısı 6 olarak belirlenmiştir. Nöron aktivasyon fonksiyonu olarak hiperbolik tanjant kullanılmıştır.

5.4. Dil Modeli

DM'lerin görevi eğitim veri kümesinden elde edilen istatistiksel bilgi ile kelimelerin olasılıksal dizilimini bulmaktır ($w_1^n = w_1 \dots w_n$). Her bir kelimenin koşullu olasılığı istatistiksel olarak Eş. 5.1'deki gibi hesaplanmaktadır [126].

$$\mathrm{P}(w_1^n) = \mathrm{P}(w_1).\mathrm{P}(w_2|w_1) \dots \mathrm{P}(w_n|w_1^{n-1}) = \prod_{i=1}^{n} P(w_i|w_1^{i-1}) \tag{5.1}$$

Koşullu olasılıklar, uzun bağımlılık gerektiren durumlarda güvenilir sonuçlar vermemektedir. Bu nedenle bir kelime dizisinin gerçek olasılığına w_1^{i-1} dizisi yerine $n-1$ ifadesinin w_{i+n-1}^{i-1} tanımına bağlı olacak şekilde yaklaşılmıştır. İstatistiksel dil modelinde kelimeler geçmişteki kelime grubunu dikkate almaktadır. Geçmişteki kaç kelimenin dikkate alındığı n-gram olarak ile ifade edilmektedir. N-gram olasılıkları, veri kümesi içerisindeki belirli bir n-gram oluşumunu hesaplamaktadır. Elde edilen değer aynı $n-1$ kelimesi dizisi ile başlayan tüm n-gramların sayısına bölünerek tahmin edilmektedir (Eş. 5.2).

$$P(w_n|w_{n-1}, w_{n-2}, \dots w_{n-N+1}) = \frac{C(w_{n-N+1}, \dots w_{n-1}, w_n)}{\sum_{w_i} C(w_{n-N+1}, \dots w_{n-1}, w_i}} \tag{5.2}$$

N-gram dil modellemesindeki temel problem veri kümesi boyutunun küçük olması durumunda OKT başarımının düşmesidir. Eğer eğitim veri kümesi küçük ise, birçok kelime dizisine çok küçük olasılıklar atanabilmektedir. Deşifre aşamasında küçük olasılıklar atanmış kelime dizilerinin ortadan kaldırılmasını önlemek için genellikle küçük olasılıklar yapay olarak arttırılmaktadır.

Her dilin kendisine ait bir doğası ve yapısı vardır. Ayrıca her dilin üretkenlik yapısı diğerlerinden farklıdır. Bu nedenle farklı her dil için farklı DM oluşturulmalıdır. Fakat OKT için çoğunlukla İngilizce DM üzerine çalışmalar yapılmıştır. İngilizce'de yer alan kelimelerin morfolojik biçimlerinin sözlüğünü oluşturmak teorikte ve pratikte mümkündür. Ancak bu işlem Türkçe gibi sondan eklemeli diller için neredeyse imkânsızdır [127]. Türk Dil Kurumu tarafından yayınlanan Türkçe sözlük yaklaşık 70.000 kelime içermektedir. Oxford İngilizce Sözlüğündeki kelimelerin sayısı teknik ve bilimsel terimler hariç yaklaşık 500.000'dir [128]. Teknik ve bilimsel terimlerle bu sayı bir milyona yaklaşmaktadır. Almanca'nın sözlük boyutu 185.000 ve Fransızca'nın 100.000 civarında kelime haznesine sahip olduğu bilinmektedir [128]. Bu rakamlar kelime çekimleri ve türevleri, isimler, teknik ve bilimsel terimleri içermemektedir. Dilin morfolojik verimliliğinden kaynaklanan kelime artışı dil modelini doğrudan etkilemektedir. Bu durumda, yetersiz veri kümesi ile istatistiksel DM oluşturulduğunda OKT'nin başarı oranı düşecektir.

Yetersiz veri kümesi dezavantajını giderebilmek için literatürde skip-gram kullanımı önerilmiştir [129]. Mikolov ve arkadaşlarının sunduğu skip-gram modeli bir hedef kelimeyi çevreleyen kelimeleri tahmin etmede en iyi kelime temsillerini bulmayı amaçlamaktadır. Matematiksel olarak Eş. 5.3'deki gibi gösterilmiştir.

$$\frac{1}{T}\sum_{t=1}^{T}\Big(\sum_{-c\leq j\leq c, j\neq 0}\log P\left(w_t+j|w_t\right)\Big) \tag{5.3}$$

Eş. 5.3'teki $w_1, w_2, \ldots, w_t$, eğitim veri kümesinde kelimeleri, c ise hedef kelimenin etrafındaki çerçevenin boyutunu ifade etmektedir. w_t'nin temsili ile tahmin edilecek içerik kelimelerinin kümesi hesaplanmaktadır [129]. Ancak n-gram ve skip-gram tabanlı DM'leri uzun cümlelerdeki kelimelerin dizilişi modelleyememektedir. Çünkü n-gram'ler ile kısıtlı

bir geçmişe, skip-gram'ler ile belirli bir atlama değerinden sonraki kelimeye bakılmaktadır. Uzun cümlelerdeki kelimelerin dizilişini modellemek için YSA temelli DM'ler geliştirilmiştir [98].

YSA tabanlı dil modelleri Markov varsayımında bulunmadığı için kelimelerdeki uzun bağımlılıklar modellenebilmektedir. Dil modellerinin eğitimi için farklı YSA modelleri kullanılmıştır. Dil modeli ilk olarak İleri Beslemeli Sinir Ağı (İBSA, Feed Forward Neural Network) mimarisinde sınanmıştır. İBSA'da, tüm katmanlar birbirine bağlıdır ve bir katmandaki yapay nöronun çıkışı bir sonraki katmandaki bir nöronun giriş birimi haline gelir. Girdi katmanı girdiyi sinir ağına vererek dış ortamla iletişim kurar. Veriler, girdi katmanından gizli katmanlara gönderilir ve gizli katmandaki veriler üzerinde işlemler uygulandıktan sonra çıktı katmanına gönderilir. İBSA sadece bir çıkış nöronuna sahip olmayabilir aynı zamanda birden fazla çıkış nöronuna da sahip olabilir. Çıktı katmanındaki nöron sayısı, sinir ağının tipi ile doğrudan ilişkili olmalıdır [130]. İBSA'nın eğitimi, en iyi tahmini yapmak için katmanlar arasındaki bağlantı ağırlıklarının değerlerini belirlemektir. Başlangıçta, ağırlık değerleri rastgele atanmıştır. İBSA daha sonra kullanılan öğrenme algoritmasının kurallarına göre ağırlık değerlerini değiştirir.

DM, farklı bir YSA yapısı olan TSA üzerinde de sınanmıştır. TSA, son yıllarda konuşma tanıma, dil modelleme ve görüntü analizi gibi birçok uygulama alanında başarılı sonuçlar vermiştir [131]. TSA'daki temel fikir önceki bilgilerin ağın sonraki adımlarında kullanılabilmesidir. Bu yapı genel olarak geçmişe bağımlı çözülebilecek problemler üzerine uygulanmaktadır. Bazen bir girdi serisinde sadece birkaç adım öncenin bilgisine ihtiyacımız olabilir. Örneğin, "dünyada korona salgını var --- çok ölümcül" cümlesindeki kelimeyi bir dil modelinin tahmin etmesi için sadece geçmişteki birkaç kelimeyi hatırlaması yeterlidir. Bu durumda TSA'lar uzun bir geçmiş ile karşı karşıya olmadıkları için bağımlılıkları iyi modelleyebilmektedir. Ancak, "ben uzak doğuya seyahat etmeyi sevdiğimden … ve Çin gibi birçok ülkeye gittiğimden koronavirüs hastası olabilirim" cümlesindeki "koronavirüs" kelimesinin model tarafından tahmin edilebilmesi için 9-10 kelime öncesindeki "uzak doğu" kelimelerinin model tarafından hatırlanması gerekir. Dolayısıyla bu tür kullanımlarda modelin uzun bir geçmiş hakkında bilgi sahibi olup ve bu bilgileri unutmaması gerekir.

İlk girdi ile son girdi arasındaki mesafe uzadıkça TSA'lar geçmişte gördükleri bilgileri unutmaya başlar. Teorik olarak, TSA'ların bu tür uzun bağımlılıkları hatırlayıp unutmamaları gerekir. Fakat gerçek kullanımda bu her zaman doğru olmayıp model uzun girdilerde unutkan olmaya başlamaktadır. Bu sorunu çözmek için UKSB üniteleri önerilmiştir [130]. UKSB, uzun vadeli bağımlılıkları hatırlamaya ve dolayısıyla bağlam farkındalığına sahip sinir ağları elde etmek için kullanılır. Literatürdeki birçok çalışmada UKSB ünitelerinin TSA ile birlikte kullanılmasının başarımı arttırdığı belirtilmiştir [132-134]. Bu nedenle DM için önerilen yöntem UKSB yapısına sahip bir TSA ile geliştirilmiştir.

5.4.1. Skip-gram ve n-gram'lerin birlikte kullanımı

Bu çalışmada, skip-gram özelliği, uzak bağlam kelimelerinin sayısını hesaplamak için önerilmiştir. Örneğin "teknoloji öğrenmek en büyük hedeftir" cümlesinde "hedef" kelimesi cümle sonu olarak belirlenmiştir. Bu durumda 1,2,3-skip gram işlemi uygulandığında "en büyük" ifadesi skip edilen kelimeleri belirtir. Karmaşıklığı azaltmak için skip değeri 1 seçilmiştir. Temel olarak belirli bir k-skip değerinden sonra n-gram işlemi gerçekleştirilmiştir. Önerilen bu yöntem Türkçe OKT sistemine uygulanmıştır. Ayrıca veri kümesi değişiminin önerilen yöntem üzerindeki etkisi gösterilmiştir. Gerçekleştirilen deneylerde düşük dereceli modellerde araya ekleme, yüksek dereceli modellerde ise yetersiz veri kümesi sorunu ile karşılaşılmıştır.

Önerilen yöntem modellenirken cümle sonuna ve başına bir işaretçi eklenmiştir. Bu işaretçiler cümle başlangıç ve bitişini temsil etmektedir. Bu nedenle kelime sayısı $n + 2$ olacak şekilde işlem yapılmıştır. Temelde iki farklı yaklaşım ile model üretilmiştir. Birinci yaklaşımda n sıralı dizilimlerden elde edilen bir olasılık hesaplanmıştır. İkinci yaklaşımda ise skip işleminden sonra bir olasılık değeri hesaplanmıştır.

DM Eş. 5.4 ve 5.5 dikkate alınarak hazırlanmış ve Türkçe OKT sistemine uygulanmıştır. Eş. 5.4 ve 5.5, n kelimeden oluşan bir cümleyi temsil etmektedir. Cümlenin başına ve sonuna işaretçi eklendiğinden dolayı denklemlerde sınır $k \geq 2$ olarak belirlenmiştir. Denklem 5'te $S(x)$ ifadesi x'in düzenli DM skorunu (sentence), $P_k(x)$ k-inci değerden türetilmiş S olasılığını belirtmektedir. Denklem 5.5'deki $i + k(x)$ ise i-skip-(i+k)-gram modelinden

türetilmiş olasılığı belirtmektedir. Eş. 5.4 ve 5.5 birleştirilerek logaritması alınmış ve Eş. 5.6 elde edilmiştir.

$$S(x) = P_k(x).\frac{P_{k+1}(x)}{h_k^0(x)} \cdot \frac{P_{k+2}(x)}{h_{k+1}^1(x)} \dots \cdot \frac{P_{2k-1}(x)}{h_{2k-2}^{k-2}(x)} \tag{5.4}$$

$$S(x) = P_k(x).\prod_{i=0}^{k-2} \frac{P_{i+k+1}(x)}{h_{i+k}^i(x)} \tag{5.5}$$

$$\log S(x) = \sum_{i=0}^{k-2} \left(\log P_{i+k+1}(x) - \log h_{i+k}^i(x)\right) + \log P_k(x) \tag{5.6}$$

Eş. 5.6, log-domaininde enterpolasyonlu bir iyileştirme algoritması olarak kabul edilebilir. Geleneksel enterpolasyonlu iyileştirme algoritmalarından farklı olarak daha yüksek dereceli Markov varsayımları ve skip-gram özellikleri kullanılmıştır. Ayrıca log-domainine cümle düzeyinde bir enterpolasyon uygulanmıştır. Önerilen yaklaşımın amacı uzun bağımlılıkları modelleyebilmektir. Türkçe'de uzun ve değişik kalıplar içerisinde cümleler üretilebilir. Bu nedenle skip-gram işlemi uzun bağımlılıkları modellemek için yardımcı edebilir. Temel varsayımımız ise önerilen yöntemin n-gram modeline göre daha güvenilir olduğudur. Bu varsayımı açıklamak için Türkçe yapısını incelemek gerekir. Türkçe doğası gereği bazı özelliklere sahiptir. Bunlar;

- Türkçe sondan eklemeli bir dildir. (kök - $ek_1, ek_2, \dots, ek_n$)
- Sunum (kök) - da - ki - ler
- Çekoslavakya (kök) -lı -laş -tır -a -ma -dık -lar - ı -mız -dan
- Türkçe serbest kelime düzen ve dizilimine sahip bir dildir. (Altı çizili olan kelime hedef kelimeyi, eğik yazılmış olan kelime ise geçmiş bilgiyi temsil etmektedir.)
- Ben çocuğa *kitabı* <u>verdim</u>
- P(W) = P(ben)P(çocuğa | ben) …. P(<u>verdim</u> | *kitabı*)
- Çocuğa kitabı *ben* <u>verdim</u>
- P(W) = P(çocuğa)P(kitabı | çocuğa) …. P(<u>verdim</u> | *ben*)
- Ben kitabı *çocuğa* <u>verdim</u>
- P(W) = P(ben)P(kitabı | ben) …. P(<u>verdim</u> | *çocuğa*)

- Çocuğa ben kitabı verdim
- P(W) = P(ben)P(ben | çocuğa) …. P(verdim | *kitabı*)

Belirtilen özellikler Türkçe'nin koşullu olasılıklar altında net bir kestirim yapılabilmesini zorlaştırmaktadır. Dil modeli için en büyük olabilirlik kestirimi yapan bir yöntem kullandığımızı varsayalım. Bu durumda farklı kelime kombinasyonları ve Türkçenin belirtilen özelliklerinden dolayı en iyi n hipotezi doğru sonuçlar vermeyebilir. En iyi n hipotezine göre;

- Futbol maçı gitti 0,55 (1)
- *Futbol maçına gitti 0,53 (2)*
- Futbol maçı sordu 0,16 (3)

Elde edilen en iyi sonuç 0,55'lik değer ile ilk sıradaki sonuç iken, gerçekte doğru sonuç ikinci sıradadır. Ancak koşullu olasılıklar değiştirilebilir. Denklem 5.2, 5.5 ve 5.6 koşullu olasılıkların değiştirilebildiğini göstermektedir. Burada $P(w_2|w_1)$ koşullu olasılığı $P(w_2|w_1).P(w_2|\langle/x\rangle w_1)/P(w_2|\langle/x\rangle)$ ile değiştirilmiştir. Kısaca DM içerisinde geçmişin modellenmesi $\langle/x\rangle$ işlemine bağlı kılınmıştır. Ancak $P(X|w_1)$ durumlarında w_2 yerine X durumunun gelmesi eğitim corpusunda X durumunun w_2 durumundan daha sık görülmesi ile açıklanabilir. $P(w_2|w_1)$ durumunun olasılığı $P(X|w_1)$ durumunun olasılığından daha yüksektir. Türkçe doğası gereği uzun bağımlılıklara izin vermektedir. Bu durumda $P(w_2|\langle/x\rangle w_1)$ işleminin daha büyük koşullu olasılık sonuçlarını vereceği unutulmamalıdır. Teorik olarak daha uzun bağımlılıklar doğru kelimelerle ilgili daha belirgin anlamsal bilgiler sağlamaktadır. Bu nedenle uzun bağımlılıkların modellenmesi önemlidir.

Önerilen yöntemin gerçek bir Türkçe OKT sistemine uygulanması bazı zorluklar içermektedir. Bu zorluklardan biri hesaplama işleminin diğer yaklaşımlara göre daha fazla olmasıdır. Özellikle eğitim ve test aşamasında k değerinin büyük seçilmesi $2k-1$ modelinin bir $k+tn$ modeli düzenlemesini gerektirir. Bu nedenle, önerilen yöntemi temsil etmek için tek bir model kullanılmıştır.

Eş. 5.7'de X operatörü, bir kelimeyi n. konumdan alıp çıkarmayı hedeflemektedir. Örneğin $X_3 w_1^4$ ifadesi sonuçta $w_1 w_2 w_4$ çıkışını verecektir. Bu sayede n-best listelerinin yeniden değerlendirilmesine ilişkin sonuçlar elde edilmiştir. Önerilen yöntem istatistiksel (n-gram) ve TSA tabanlı dil modellerine uygulanmıştır.

$$S(x) = \prod_{j=1}^{n+1} P(w_j | w_{j-k+1}^{j-1}) \tag{5.7}$$

Temel DM, n-gram (n=2,3,4,5) olarak hazırlanmıştır. Modelin geliştirilmesinde SRI Dil Modelleme Araç Seti (SRIDM, SRI Language Modeling Toolkit) kullanılmıştır [135]. SRIDM öncelikle konuşma tanıma, istatistiksel etiketleme ve makine çevirisi için kullanılan istatistiksel DM'leri oluşturmak ve uygulamak için geliştirilen bir araç setidir. YSA temelli DM'ler ise Microsoft Bilişsel Araç Seti (MBAS, Microsoft Cognitive Toolkit) ile geliştirilmiştir [136]. MBAS Microsoft Research tarafından geliştirilen DÖ çerçevesidir. MBAS, sinir ağlarını yönlendirilmiş bir grafik aracılığıyla bir dizi hesaplama adımı olarak tanımlamaktadır.

5.5. Okunuş Sözlüğü

Okunuş sözlüğü ile harfler belirli bir sesbirim ile temsil edilmektedir. Bu çalışmada kullanılan veri kümesindeki harflerin sesbirim karşılıkları yazılmıştır. Bu işlemi gerçekleştirebilmek için sesbirim kural tablosu oluşturulmuştur. Çizelge 5.1'de harflere karşılık gelen sesbirimler verilmiştir. Çizelge 5.1 bir kural tablosu olarak tanımlanabilir. Veri kümesindeki bütün kelimeler öncelikle harflere parçalanmış ardından bu harflere karşılık gelen sesbirimler yazılmıştır. Bu işlem için küçük bir kod parçacığı yazılmış ve Kaldi araç setine eklenmiştir. Ancak tek başına bu işlem geniş kelime dağarcığına sahip Türkçe için yeterli değildir. Türkçe de bazı kelimelerin farklı okunuşları mevcuttur. Ayrıca kısaltmalar ve yabancı kelimelerde farklı okunuşa sahip olabilir. Bu nedenle okunuş sözlüğüne bütün okuma biçimlerinin verilmesi başarımı arttıracaktır.

Bu çalışmada bütün okunuş sözlüğü gerçek kullanıcılar tarafından kontrol edilmiş ve farklı okunuşlara sahip kelimeler veya kısaltmalar sözlüğe eklenmiştir. Türkçe kelimelerin

bulunmasında Türk Dil Kurumu sözlüğünden yararlanılmıştır. Öncelikle her kelime aşağıda belirtilen gruplara ayrılmıştır.

- Okunuş sözlüğüne eklenmesin
- Yazıldığı gibi okunan kelimeler
- Farklı okunan ve hecelenebilen kelimeler (TRT gibi)
- Farklı okunan ve tamamen farklı yazılan kelimeler (AİHM, AIDS gibi)

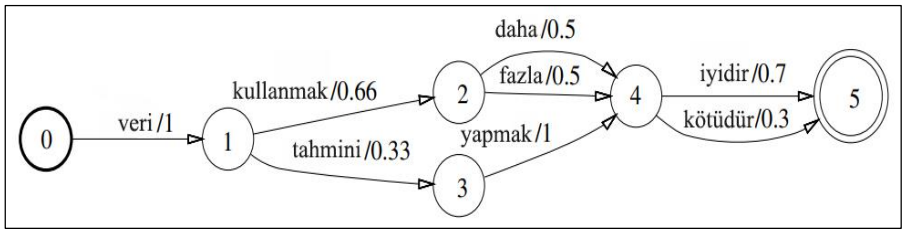
Çizelge 5.1. Harflere karşılık gelen sesbirimlerin gösterimi

‘harf’ :’sesbirim’	‘harf’ :’sesbirim’
'a': 'AA',	'o': 'O',
'b': 'B',	u'ö': 'AX',
'c': 'JH',	'p': 'P',
u'ç': 'CH',	'r': 'R',
'd': 'D',	's': 'S',
'e': 'EH',	u'ş': 'SH',
'f': 'F',	't': 'T',
'g': 'G',	'u': 'U',
u'ğ': ':',	u'ü': 'UH',
'h': 'HH',	'v': 'V',
u'ı': 'IY',	'y': 'Y',
u'i': 'IX',	'z': 'Z',
'j': 'ZH',	'x': 'KS',
'k': 'K',	'w': 'VI',
'l': 'L',	'q': 'KU',
'm': 'M',	u'â': 'AE'

Okunuş sözlüğüne eklenen kelimeler bu kurallar dikkate alınarak eklenmiştir. Örneğin AA Anadolu Ajansının kısaltmasıdır. Ancak hiçbir yerde AA diye bir okunuş ile Anadolu Ajansından bahsedilmemektedir. Yazıldığı gibi okunan bir kısaltma örneği ABD olabilir. ABD Anabilim Dalının kısaltmasıdır. Diğer yandan TRT yazıldığı gibi okunur ancak AIDS ve benzeri kısaltmalar tamamen okunuşları ve yazılışları farklı kısaltmalardır. Sabit kurallar ile oluşturulan kural tablosu bu yazım şekillerini yanlış modelleyecektir. Bu nedenle okunuş sözlüğü gerçek kullanıcılar tarafından kontrol edilerek düzenlenmiştir.

5.6. Deşifre

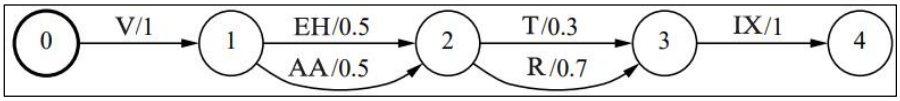
Bu çalışmada Kaldi araç seti OKT sisteminin temel yapı taşlarını oluşturmada kullanılmıştır. Ancak AM, DM ve okunuş sözlüğü gibi bütün bileşenlerde iyileştirmeler yapılmıştır. Kaldi, kod çözme işleminde ağırlıklı Sonlu Durum Dönüştürücülerinin (SDD, Finite-State Transducer) oluşturulması, birleştirilmesi, optimize edilmesi ve aranması için OpenFST kütüphanesini kullanmaktadır.



Şekil 5.2. Sonlu durum dil modeli yapısı

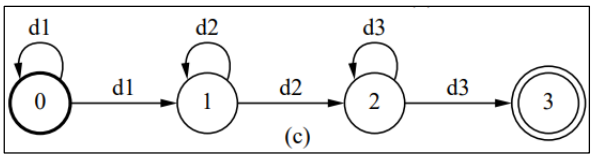
SDD, durum geçişleri hem giriş hem de çıkış sembolleri ile etiketlenmiş sonlu bir otomattır. Bu nedenle dönüştürücü boyunca bir giriş sembolü dizisinden veya dizeden bir çıkış dizesine eşlenmesi kodlanır [137]. SDD, giriş ve çıkış simgelerine ek olarak geçişlere ağırlık eklemektedir. Ağırlıklar, bir giriş dizesini bir çıkış dizesine eşlemenin toplam ağırlığını hesaplamak için yollar boyunca biriken olasılıkları, süreleri veya cezaları kodlayabilmektedir. Dolayısıyla SDD'ler, konuşma işlemede yaygın olan olasılıklı sonlu durum modellerini temsil etmek için ideal bir seçimdir. Şekil 5.2, 5.3 ve 5.4 OKT'de kullanılan ağırlıklı otomata örnekleri verilmiştir. Şekil 5.2, OKT'deki sonlu durum dil modeli yapısını göstermektedir.

Şekil 5.2'de gösterildiği gibi kelime dizeleri her yol boyunca kelimelere ve olasılıklara karşılık gelen geçiş olasılıklarının çarpımı ile belirtilmiştir. Şekil 5.3, dil modelinde kullanılan bir kelimenin, olası durumlarını vermektedir.



Şekil 5.3. Veri kelimesinin olası okunuş durumu

Her olası okunuş, bir yol boyunca sesbirim dizelerinden oluşmaktadır. Sesbirim kural tablosu okunuş sözlüğünde verilen kurallara göre kodlanmıştır. Geçiş olasılığının ağırlıkları veri kümesindeki bilgilere göre değişmektedir. DM ve AM için gerekli SDD'ler oluşturulduktan sonra bir sesbirim için soldan sağa, üç dağıtıma sahip SMM yapısı kodlanmıştır. Etiketler bir yol boyunca belirtilen sesbirim için akustik dağılımların olması gereken dizelerini belirtmektedir (Şekil 5.4).



Şekil 5.4. SMM durumu

Şekil 5.4'te verilen otomata durum, başlangıç durumu, son durum ve durumlar arasında bir dizi geçiş işlemlerinden oluşmaktadır. Her geçişin bir kaynak durumu, bir hedef durumu, bir etiketi ve ağırlığı mevcuttur. Bu tür otomatalar, Ağırlıklı Sonlu Durum Alıcıları (ASDA, Weighted Finitestate Acceptors) olarak bilinmektedir. Kabul edilen her dizeye bir ağırlık atanmaktadır. Ağırlık atama işlemine son diziler de dâhildir.

Konuşma tanıma mimarilerinde genellikle deşifre işlemine otomatları birleştirme ve optimize etme görevi verilmiştir. Kod çözücü, okunuş sözlüğündeki kelime okunuş biçimlerini bulur ve bunları dilbilgisi ile değiştirir. Fonetik ağaç gösterimleri bu noktada yolları optimize etmek ve büyük kelime dağarcığı tanıma görevinde arama verimliliğini arttırmak için kullanılmaktadır. SDD'lerin kullanımı ve SMM durumlarına dönüştürülme

işlemi tamamlandıktan sonra artık deşifre işlemi için bir yol haritası çıkarabiliriz. Deşifre işlemi için genel görünüm HCLG = H o C o L o G şeklindedir ve HCLG ayrıntıları Çizelge 5.2'de verilmiştir.

Çizelge 5.2. Deşifre işlemi yol haritası

Sembol	Dönüştürücü	Giriş Sırası	Çıkış Sırası
G	Dil Modelini Kodlayan Alıcı	Kelimeler	Kelimeler
L	Okunuş Sözlüğü	Fonemler	Kelime
C	Bağlam Bağımlılık	Bağlamlara Bağlı Fonemler	Fonemler
H	SMM	SMM Durumları	Bağlam Bağımlı Fonemler

Çizelge 5.2, OKT için gerekli yol haritasını vermektedir. G, dilbilgisini veya dil modelini temsil eden bir alıcıdır. Bu alıcı durumunda giriş ve çıkış sembolleri aynıdır. L, okunuş sözlüğünü temsil etmektedir. Çıktısı sesbirimlere karşılık gelen kelimelerdir ve girdi olarak sesbirimleri kullanmaktadır. C, bağlam bağımlılığını temsil etmektedir. SDD yapısı açıklanırken gerekli detaylar verilmiştir. Bu alıcının çıkış sembolleri sesbirimlerdir ve giriş sembolleri bağlama bağlı sesbirimlerdir. Kısaca N adet sesbirimin pencerelerini temsil etmektedir. Bağlam bağımlılığı, bağlam genişliği ve merkez konumu ile elde edilmektedir. H, SMM tanımlarını içermektedir. Çıkış sembolleri bağlam bağımlı sesbirimleri temsil etmektedir. Giriş sembolleri sıfırdan başlayarak olasılık dağıtım işlevi için bir dizin olarak kullanılan bir sayıdır ve diğer bilgileri kodlayan geçiş kimlikleridir.

5.7. Deneysel Sonuçlar

Deşifre işleminde konuşma tanıma işlemi için bir yol haritası sunulmuştur. Bu yol haritasındaki bütün adımlarda kullanılan modeller verilmiştir. Bu çalışma kapsamında gerekli modellerin her birinde ve okunuş sözlüğünde iyileştirmeler yapılmıştır. Deneylerde

elde edilen sonuçlar KHO olarak ifade edilmiştir. Çalışmada temel olarak iki yaklaşım karşılaştırılmıştır. Bu yaklaşımlardan ilki klasik GKM-SMM diğeri ise DSA yapısındaki OKT sistemidir. Sadece yaklaşımlar değil aynı zamanda Türkçe geniş kelime dağarcığını modelleyebilmek için DÖ temelli çok katmanlı sinir ağlarının eğitiminde kullanılan verilerin çokluğu ve çeşitliliği ele alınmıştır.

Geliştirilen iki farklı OKT sistemi benzer sözlük ve dil modellerini kullanacak şekilde deneyler gerçekleştirilmiştir. İlk olarak AM'nin etkisini ortaya koyabilmek için sadece AM'ler değiştirilmiştir. Gerçekleştirilen deneylerde geniş kelime dağarcığına sahip Türkçe OKT'nin modellenmesinde DSA, GKM tabanlı sistemden daha iyi performans göstermiştir. Bu sonuç, Türkçe OKT sistemlerinde akustik modellemenin DSA ile daha iyi sınıflandırıldığını göstermektedir. DSA'nın daha iyi performans göstermesinin nedeni bitişik pencerelerdeki mevcut bağlam bilgilerini daha iyi yakalayabilmesidir. Bu çalışmada elde edilen veri kümesi HS veri kümesi olarak adlandırılmıştır. HS veri kümesi ve mevcut diğer veri kümelerinin GKM-SMM'e dayalı geniş kelime dağarcığına sahip Türkçe OKT sistemine etkisi Çizelge 5.3'te verilmiştir. Çizelge 5.3'te verilen OKT sisteminde kullanılan dil modeli istatistiksel tabanlı olup 3-gram olacak şekilde geliştirilmiştir.

Çizelge 5.3'te farklı örneklem frekanslarında model gelişimi incelenmiştir. Örneklem frekansı azaldığında performans oranı düşmüştür. Deneyler sonucunda modelin geliştirilmesinde kullanılan örneklem frekansı ve test sürecinde kullanılan örneklem frekansındaki fark performansı olumsuz etkilemiştir. Bu nedenle test prosedürleri aynı örnekleme frekansına dayanmaktadır. Bu durumun nedeni bazı akustik bilgilerin frekans değişiminden etkilenmesidir.

Her bir deney aşamasında belirtilen veri kümesinin 1/3'ü test amaçlı ayrılmıştır. ODTÜ veri kümesi, yapılan testlerde en başarısız KHO oranına sahiptir. Doğrulanmış HS veri kümesi ise OKT için en başarılı KHO oranını vermiştir. Veri kümelerinin detayları incelendiğinde ODTÜ veri kümesi içerisinde birçok anlamsız örnek olduğu görülmüştür. Bu anlamsız ve aynı zamanda çok kısa bilgiler OKT sisteminin performansını olumsuz yönde etkilemiştir. Boğaziçi veri kümesi örneğine baktığımızda ise içeriğin net olmasına rağmen yeterli akustik çeşitlilik olmadığı sonucuna varılmıştır.

Çizelge 5.3. GKM tabanlı Türkçe OKT sistemi

Eğitim Veri Kümesi Adı	Test Veri Kümesi Adı	KHO (%)	Örneklem Frekansı
ODTÜ Veri Kümesi	ODTÜ Veri Kümesi	70,7	16 KHZ
Boğaziçi Veri Kümesi	Boğaziçi Veri Kümesi	27,7	16 KHZ
Doğrulanmamış HS Veri Kümesi	Doğrulanmamış HS Veri Kümesi	55,2	16 KHZ
Doğrulanmış HS Veri Kümesi	Doğrulanmış HS Veri Kümesi	24,7	16 KHZ
ODTÜ Veri Kümesi	ODTÜ Veri Kümesi	71,2	8 KHZ
Boğaziçi Veri Kümesi	Boğaziçi Veri Kümesi	28,9	8 KHZ
Doğrulanmamış HS Veri Kümesi	Doğrulanmamış HS Veri Kümesi	58,1	8 KHZ
Doğrulanmış HS Veri Kümesi	Doğrulanmış HS Veri Kümesi	25,8	8 KHZ

Türkçe geniş kelime dağarcığına sahip bir OKT sisteminin performansı sadece AM'den elde edilen sesbirim bilgileri ile değil aynı zamanda DM tarafından elde edilen kelime içeriğinin temsiline de bağlıdır. Bu nedenle metin veri kümesinin hazırlanmasına ek olarak geliştirilen her model farklı n-gram değerlerine sahip olacak şekilde deneyler gerçekleştirilmiştir. Deneylerde istatistiksel dil modeli 2,3,4 ve 5-gram olarak geliştirilmiştir. Ancak en başarılı sonuçlar 3 ve 4-gram değerlerinde alınmıştır. Çizelge 5.4'de farklı n-gram değerlerinin farklı veri kümesi ile geliştirilmiş Türkçe OKT sistemi üzerindeki etkisi gösterilmiştir.

Çizelge 5.4'te verilen sonuçlar, HS veri kümesinin fonetik içeriğinin dengeli ve farklı dil modelleri için iyi performans sergilediğini göstermektedir. Daha büyük dil modelleri için KHO oranındaki düşüş Türkçenin doğası ile açıklanmaktadır. Dil modeli eğitimi için daha büyük metinlerin kullanılmasının daha iyi sonuçlara yol açtığı bilinmektedir.

DSA'ya dayalı OKT sisteminin deneylerinde veri kümesinin boyutu çok önemlidir. Ancak sadece veri kümesi değil sistemi etkileyen farklı parametrelerin olduğu da bilinmektedir. Bir

DSA'nın derinliği performans üzerinde önemli bir etkiye sahiptir. Ancak farklı gizli katmanlardan oluşan büyük modellerin hem eğitim hem de kod çözme işlemi uzun sürmektedir. Bu nedenle, yüksek bir model boyutu seçmek yerine uygun miktarda gizli katmanın seçilmesi önerilmektedir. Teorik olarak, DSA'ların basit sinir ağlarından ziyade daha karmaşık fonksiyonları modelleyebilmesi gerekir. Ancak, uygulamalarda modellerin derinlik ve karmaşıklık kullanımını optimize etmek bir problem olarak görülmektedir.

Çizelge 5.4. Farklı n-gram değerlerinin Türkçe OKT sistemine etkisi

Eğitim Veri Kümesi Adı	N-Gram	KHO (%)
ODTÜ Veri Kümesi	3-Gram	70,7
ODTÜ Veri Kümesi	4-Gram	72,4
Boğaziçi Veri Kümesi	3-Gram	27,7
Boğaziçi Veri Kümesi	4-Gram	28,2
Doğrulanmamış HS Veri Kümesi	3-Gram	55,2
Doğrulanmamış HS Veri Kümesi	4-Gram	58,9
Doğrulanmış HS Veri Kümesi	3-Gram	24,7
Doğrulanmış HS Veri Kümesi	4-Gram	26,1

Gizli katman sayısına ve gizli katman boyutuna ek olarak, önemli bir parametre olan öğrenme oranının sistemin performansı üzerinde etkisi olmaktadır. Öğrenme oranının seçiminde veri kümesi boyutu önemlidir. Büyük veri kümesi için düşük bir öğrenme oranı daha uzun bir eğitim süresi gerektirecektir. DSA için diğer önemli bir parametre mini parça boyutudur. Mini parça boyutu, 128, 256 veya 512 gibi aralıklarla seçilmelidir. Büyük bir boyutun seçilmesi, özellikle GİB kullanıldığında matris çarpımında kullanılan optimizasyonlar ile etkileşime girdiği için faydalı olarak kabul edilmektedir. MİB kullanımı durumunda, büyük bir mini parça boyutu kararsızlığa neden olmaktadır. Bu nedenlerden dolayı, mini parça boyutu çok parçacıklı MİB tabanlı eğitim için 128 ve GİB tabanlı eğitim için 512 olarak ayarlanmıştır. Bu çalışmada elde edilen veri kümesi ve mevcut veri kümeleri kullanılarak geliştirilen DSA tabanlı OKT sisteminin performansı Çizelge 5.5'de verilmiştir. Çizelgedeki test işlemlerinde 16 kHz örnekleme frekansı kullanılmıştır.

Çizelge 5.5. DSA tabanlı OKT sistemi

Eğitim Veri Kümesi Adı	Test Veri Kümesi Adı	KHO (%)	Örneklem Frekansı
ODTÜ Veri Kümesi	ODTÜ Veri Kümesi	64,5	16 KHZ
Boğaziçi Veri Kümesi	Boğaziçi Veri Kümesi	22,6	16 KHZ
Doğrulanmamış HS Veri Kümesi	Doğrulanmamış HS Veri Kümesi	49,2	16 KHZ
Doğrulanmış HS Veri Kümesi	Doğrulanmış HS Veri Kümesi	18,7	16 KHZ

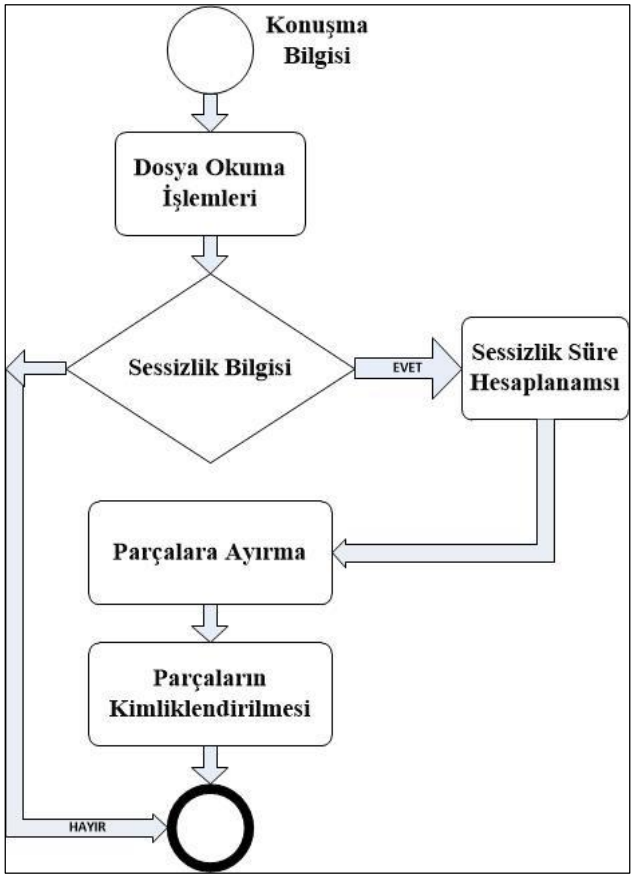
Çizelge 5.5'de sunulan sonuçlar mevcut veri kümelerinin, HS veri kümesi ile karşılaştırılabilir olduğunu göstermektedir. Bu nedenle çalışma kapsamında literatürdeki en geniş ve ayrıntılı bir Türkçe konuşma tanıma veri kümesinin hazırlandığı ve bu veri kümesinin diğer mevcut veri kümelerine göre daha başarılı sonuçlar verdiği görülmüştür.

Elde edilen veriler ile deneyler gerçekleştirildiğinde çok katmanlı derin sinir ağı yaklaşımlarının klasik GKM-SMM tabanlı yaklaşımlardan daha etkili görülmüştür. Ancak bu durum geniş kelime dağarcığına sahip Türkçe OKT için tek başına yeterli değildir. AM, geliştirilirken HS veri kümesi ile çok fazla örnek sesbirimin sisteme eklenmesi sağlanmıştır. Bu durum başarılı bir AM'nin oluşturulmasını desteklemiştir. Bilindiği gibi Türkçe sondan eklemeli ve üretken bir yapıya sahiptir. HS veri kümesi ile olabildiğinde kelime çeşitliliği sağlanmış ve AM geliştirilmiştir.

Geliştirilen AM farklı kelimelere ait sesbirimlerin modellenmesini sağlamıştır. Ancak deneylerde uzun bağımlılıkları içeren bir konuşmada performansın düştüğü görülmüştür. Bu nedenle AM farklı bir yaklaşım ile ele alınmıştır. Konuşma içerisindeki uzun bağımlılıkları engelleyebilmek için ön işlem bileşeni eklenmiştir. Ön işlem bileşeni için akış şeması Şekil 5.5'te verilmiştir.

Farklı çalışmalarda konuşma etkinliği, sessizlik bilgisi veya gürültü tanıma için birçok yaklaşım ve değerlendirme çerçevesi sunulmuştur. Fakat bu alan ile ilgili farklı çözümler problemi kapsamlı bir yaklaşım ile ele alsa da istenilen başarım sağlanamamıştır.

Gürültülere ve dış etkilere karşı daha sağlam ve başarımı yüksek bir Türkçe OKT sistemi için konuşma etkinliğini tanıma uygulamalarının geliştirilmesi gerekmektedir. Bu nedenle çalışma kapsamında sessizlik içeren bir konuşma içerisinde sesbirimlerin birbiri arasında bağımsız olmasının sağlanması ve sadece konuşma içeren kısımların AM'ye girdi olarak verilmesi üzerine bir yöntem sunulmuştur. Ayrıca konuşmayı parçalara ayırarak uzun bağımlılıkların önüne geçilmesi amaçlanmıştır.

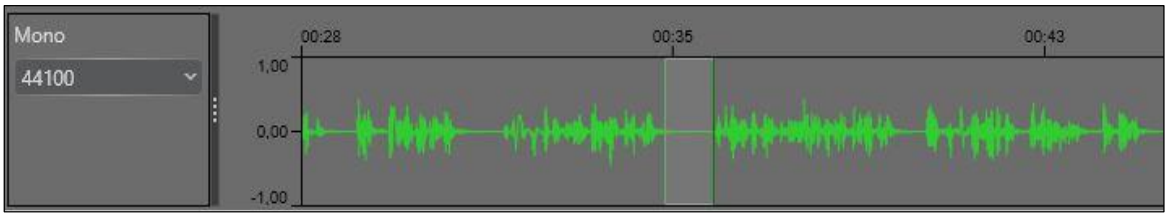


Şekil 5.5. Ön işlem bileşeni için akış şeması

Konuşma sinyalinin ele alındığı ilk andan itibaren öncelikle konuşmanın başlaması beklenmiştir. Geliştirilen sistem konuşma başladığı anda aktif hale gelmektedir. Sistemin aktif hale geldiği an parçalama işleminin başladığı zaman başlangıcıdır. Konuşma içerisinde bir sessizlik ortamı ile karşılaşıncaya kadar konuşma parçasının sürekliliği devam etmektedir. Sessizlik bilgisi ile karşılaşıldığı an ise konuşma parçasının bitiş noktasıdır. Geliştirilen sistemde sessizlik algılama işleminin başlangıç noktası, sinyal büyüklüğünün sessizlik eşik değerinin altına düştüğü noktadır. Bitiş noktası ise sinyal büyüklüğünün sessizlik eşik değerin üstüne çıktığı nokta olarak ifade edilmiştir. Sessizlik ile karşılaşıldığı

ilk anda konuşma hemen parçalanmamıştır. Bunun nedeni ise bazı konuşmacıların sakin, aralıklı ve durağan konuşmasındaki doğal sessizliklerin gerçekte konuşmanın özelliğinde var olmasıdır. Bu nedenle sessizlik ile karşılaşıldığı ilk andan itibaren bir zaman sayacı tutulmuştur. Zaman sayacının belirli bir değerin altında kalması durumunda konuşmanın parçalanması gerçekleştirilmemiştir. Böylelikle konuşmanın sürekliliği koruma altına alınmıştır.

Belirtilen yaklaşımın ana çıkış noktası her ses veya konuşma sinyalinin belirli bir frekans bandında ilerlemesidir. Konuşma sinyali belirli bir frekansa ve genliğe sahiptir. Şekil 5.6'da bir konuşma sinyalinin zaman-genlik dalga formu gösterilmiştir.



Şekil 5.6. Konuşma dosyasının zaman-genlik dalga formu gösterimi

Şekil 5.6'den yararlanılarak bir ses sinyali içerisindeki sessizliklerin bulunabilir olduğu anlaşılmaktadır. Dalga formu üzerinde yer alan örnekleme frekansı göz önüne alındığında belirli bir değerin altı sessizlik olarak nitelendirilebilir. Bu nedenle önerdiğimiz yöntemde ses sinyalinin örnekleme frekansına göre bir hesaplama yapılmıştır (Eş. 5.8).

$$Eşik\ Değeri = \frac{Örneklem\ Frekansı}{1000} \tag{5.8}$$

Eş. 5.8'de görüldüğü gibi sessizlik ifadesinin bulunması amacıyla örnekleme frekansının belirli bir oranı alınmaktadır. Böylelikle elde edilen frekans bilgisinin aşağısında kalan değerler sessizlik olarak algılanmıştır. Denklem 5.8'de belirtilen oran değiştirilerek farklı amaçlar için kullanılabilir. Gerçekleştirdiğimiz deneylerde bu oranın en uygun sonucu verdiği görülmüştür.

Belirtilen yöntemde ele alınması gereken temel sorun durağan ve aralıklı konuşmalarda sessizlik bilgisinin elde edilmesidir. Bu sorunu çözebilmek için sessizlik bilgisi ile karşılaşıldığı ilk andan itibaren arka planda bir zaman sayacı tutulmaktadır. Bu sayaç sessizlik bilgisi boyunca değiştirilmektedir. Sayaç belirli bir eşik değerine ulaştığı anda sessizlik bilgisinin var olduğu sonucuna varılmaktadır. Böylelikle durağan konuşan, hece veya harf kayıplarının yaşandığı konuşmaları içeren konuşmalarda sessizlik bilgisi başarılı bir şekilde elde edilmiş olur.

Belirtilen yöntemdeki ana amaç başlangıçtan itibaren sessizlik bilgisi ile karşılaşıldığı ana kadar ki konuşmanın ana konuşma içerisinden ayrılmasıdır. Ana konuşma dosyasından ayrılan her bir parçaya benzersiz bir kimlik verilerek geçici bellekte tutulması sağlanmıştır. Kimlik bilgisi aynı zamanda hangi ses parçasının hangi sırada olduğu bilgisini de tutmaktadır. Böylelikle parçalar arasında bir bütünlük sağlanmıştır.

Çalışmada sessizlik bilgisi ile ilgili iki yaklaşım karşılaştırılmıştır. Bu yaklaşımlardan ilki geliştirilen DSA tabanlı OKT sistemidir. Diğeri ise çalışma kapsamında sessizlik ile ilgili yöntem olan sessizlik bilgisinin ortadan kaldırıldığı ve konuşmanın parçalara ayrıldığı yaklaşımdır. Bu yaklaşımda bilinmesi gereken diğer bir konu OKT sistemine daha küçük konuşma parçalarının verildiğidir. Deneysel işlemler KHO temel alınarak değerlendirildiğinde Çizelge 5.6'daki sonuçlar elde edilmiştir. Deney işlemlerinde DSA tabanlı modelde en iyi sonucu veren Doğrulanmış HS veri kümesi kullanılmıştır.

Çizelge 5.6. Sessizliğin kaldırılması uygulamasının KHO sonuçları

Uygulama yöntemi	KHO (%)	Örneklem frekansı
DSA Tabanlı Türkçe OKT	18,7	16 kHz
Ön işlemli Türkçe OKT	17,4	16 kHz

Çizelge 5.6'da belirtilen uygulama yönteminde iki farklı yöntem test edilmiştir. DSA tabanlı Türkçe OKT yönteminde konuşma bilgisi OKT sistemine direk verilmiş herhangi bir ön işlem yapılmamıştır. Ön işlemli OKT yönteminde ise sessizlik bilgisi ortadan kaldırılmış ve

konuşma parçalara ayrılmıştır. Gerçekleştirilen deneylerde çalışmada sunulan ön işlemli OKT'nin kelime hata oranı açısından daha iyi sonuç verdiği görülmüştür. Ön işlemli OKT tanıma başarımını kötü etkileyecek akustik bilgileri dikkate almamıştır.

Ön işlem bileşeni üzerinde yapılan deneyler AM için farklı yaklaşımların sunulması gerektiğini ortaya koymaktadır. Sadece sesbirimlerin sınıflandırılmasında değil aynı zamanda gereksiz akustik sinyallerin AM'ye verilmemesi gerekir. Bu nedenle ön işlem adımlarında ses sinyalinin iyileştirilmesi, gürültülerin temizlenmesi vb. işlemlerin yapılması gerekmektedir.

AM ile ilgili deneyler genellikle akustik ortam bilgilerinden elde edilen sesbirimler üzerinedir. Ancak sadece sesbirim çıktılarının elde edilmesi tek başına yeterli değildir. DM bu noktada deneylerimize yön vermekte ve OKT'nin çıktı başarımını arttıracak yöntemler sunmaktadır. Çizelge 5.4'te istatistiksel tabanlı farklı n-gram modellerinin OKT sistemine etkisi verilmiştir. Ancak Türkçe'yi sadece istatistiksel bir model ile modellemek yeterli değildir. Bu nedenle ilk olarak OKT çıkışındaki KHO oranını tespit etmek için temel DM, 3-gram olarak geliştirilmiştir. Ardından temel DM'in en iyi sonuç veren (n-best) listeleri tekrar revize edilerek KHO hesaplanmıştır. Bu işlem için öncelikle en temel algoritma olan Kneser-Ney smoothing algoritması kullanılmıştır [138]. Kneser-Ney smoothing ile OKT sistemindeki dil modelinin n-best listeleri güncellenmiştir. Bu işlem bi-gram üzerinden gerçekleştirilmiştir. DM için kelime sınırlaması yapılmamıştır. Veri kümelerinde belirtilen kelime sayısı kadar kelime kullanılmıştır. Çizelge 5.7'de temel DM ve n-best güncellemesi sonrasında elde edilen KHO sonuçları verilmiştir.

Çizelge 5.7. Türkçe OKT için İstatistiksel DM yapısına ait KHO (%) sonuçları

Veri Kümesi Adı	3-Gram DM	Bi-Gram	Önerilen Yöntem
ODTÜ Veri Kümesi	64,5	63,5	65,0
Boğaziçi Veri Kümesi	22,6	21,3	21,0
HS Veri Kümesi	18,7	16,5	16,2

Çizelge 5.7'de gösterilen 3-gram DM, Türkçe OKT sisteminde kullanılan n-gram tabanlı DM'yi temsil etmektedir. Bi-gram, Kneser-Ney smoothing işlemi sonucu elde edilen değerdir. Önerilen yöntem ise kısaca, 1-skip-2-gram DM işlemini temsil etmektedir. Sonuçlar incelendiğinde ODTÜ veri kümesi ile geliştirilen Türkçe OKT sisteminde önerilen yöntemin uygulanmasının KHO'yu arttırdığı görülmüştür. Bunun nedeni veri kümesi içerisinde iki veya üç kelimeden oluşan cümlelerin çok fazla olmasıdır. Önerilen yöntemin amacı geçmiş bilgilerden daha fazla yararlanmaktır. Diğer veri kümeleri incelendiğinde ise HS veri kümesinin (doğrulanmış) hem boyut hem de içerdiği cümlelerin uzun olması açısından daha iyi sonuçlar verdiği görülmüştür.

Önerilen yöntemin farklı DM geliştirme yöntemleri üzerindeki etkisi de araştırılmıştır. Bu nedenle temel olarak 3-gram ve İBSA tabanlı bir DM geliştirilmiştir. Önerilen yöntem ise 1-skip-3-gram ve 2-skip-4-gram olacak şekilde geliştirilmiştir. İstatiksel dil modelinde olduğu gibi kelime sınırlaması yapılmamıştır. İBSA tabanlı modelin eğitim zamanını azaltabilmek için sınıf-tabanlı çıkış katmanı kullanılmıştır. İBSA modeli iki saklı katmana sahip olacak şekilde tasarlanmıştır. İBSA DM yapısı temel olarak geliştirilen 3-gram DM ile enterpolasyon işlemine alınmıştır. Çizelge 5.8'de önerilen yöntemin İBSA ile hazırlanan DM yapısına etkisi verilmiştir.

Çizelge 5.8. Türkçe OKT için İBSA DM yapısına ait KHO (%) sonuçları

Veri Kümesi Adı	3-Gram DM	İBSA	Önerilen Yöntem	Önerilen Yöntem*
ODTÜ Veri Kümesi	64,5	63,0	62,4	63,9
Boğaziçi Veri Kümesi	22,6	20,8	20,0	19,7
HS Veri Kümesi	18,7	16,0	15,6	14,4

Çizelge 5.8'de Önerilen Yöntem* ile 2-skip-4-gram ifade edilmiştir. Başarım oranları incelendiğinde en iyi KHO sonucunun Önerilen Yöntem* ile elde edildiği görülmektedir. Çizelge 5.8'de gösterildiği gibi ODTÜ veri kümesi, İBSA modellemesinde önerilen yönteme göre daha iyi sonuç vermektedir. En iyi KHO sonucu ise HS veri kümesi ile elde edilmiştir. Bu sonuç temel olarak iki durum ile açıklanabilir. Birincisi temel DM'nin önerilen yöntem

ile geliştirilmiş DM ile yakın eşleşmiş değerleri barındırmasıdır. Diğeri ise eğitim işleminin HS veri kümesi verileri ile daha başarılı olmasıdır.

Çizelge 5.9. Türkçe OKT için TSA-UKSB DM yapısına ait KHO (%) sonuçları

Veri Kümesi Adı	3-Gram DM	TSA-UKSB	Önerilen Yöntem	Önerilen Yöntem*
ODTÜ Veri Kümesi	64,5	62,8	62,2	64,6
Boğaziçi Veri Kümesi	22,6	20,3	20,1	19,2
HS Veri Kümesi	18,7	15,8	14,9	14,3

DM için önerilen yöntem, literatürde sıklıkla kullanılan n-gram ve İBSA ile modellenmiştir. Ancak literatürde UKSB ile birlikte kullanılan TSA yapısının DM için başarılı sonuçlar verdiği gösterilmiştir [92,139,140]. Bu nedenle önerilen yöntem son olarak TSA-UKSB yapısı üzerinde test edilmiştir. İBSA ile benzer şekilde kelime boyutunda sınırlama yapılmamıştır. Eğitim işleminin hızlı gerçekleşmesi için sınıf tabanlı bir yaklaşım kullanılmıştır. TSA-UKSB yapısının kodlayıcı tarafında 7 katmanlı bir TSA-UKSB bulunup UKSB'lerin ünite sayısı 512 olarak belirlenmiştir. Çözücü tarafında ise 7 katmanlı ve 512 boyutlu TSA-UKSB ağı bulunmaktadır. Önerilen yöntem TSA-UKSB yapısına uygulanmış ve elde edilen sonuçlar Çizelge 5.9'da verilmiştir.

Çizelge 5.9'da görüldüğü gibi önerilen yöntem, TSA-UKSB yapısında da başarılı sonuçlar vermiştir. Veri kümesi boyutu arttığında önerilen yaklaşımın daha net sonuçlar verdiği görülmüştür. Ancak KHO oranındaki başarım, hızlı işlem yapma yeteneği ile aynı oranda başarılı değildir. Önerilen yöntem çok fazla hesaplama gerektirmektedir. İşlem sürelerinin azalması en uygun şekle sokma teknikleri veya daha hızlı bilgisayar altyapısı ile mümkün olabilir. Özellikle n-best listelerinin tekrar revize edilmesi için fazladan süre harcanmaktadır. Bu nedenle deşifre aşamasında en uygun şekle sokma teknikleri geliştirilmelidir.

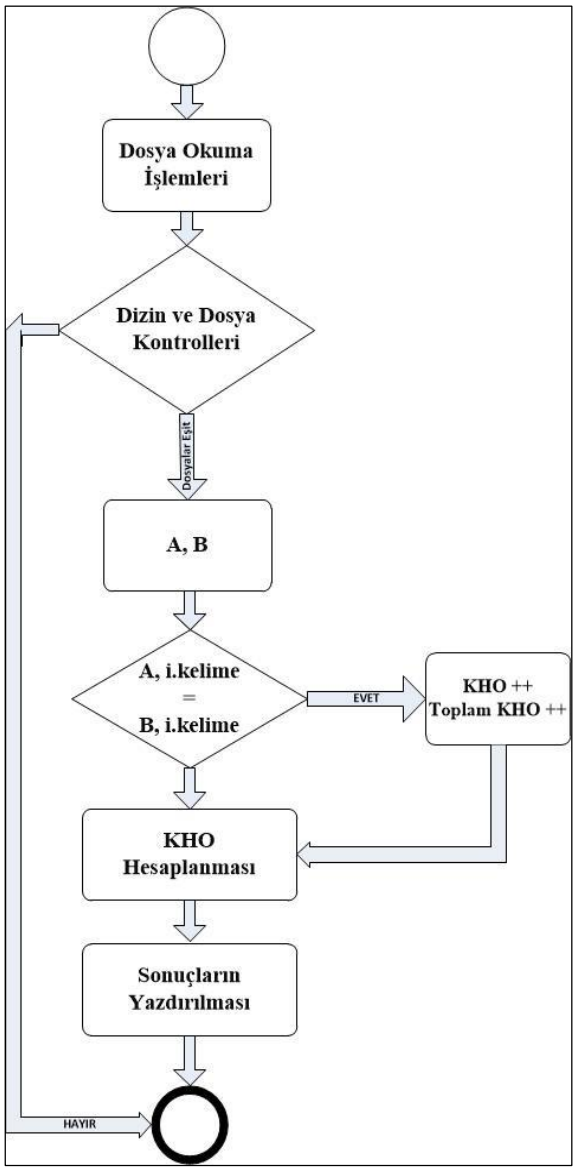
Gerçekleştirdiğimiz deneylerde AM'nin ve DM'nin Türkçe OKT sistemine etkisi net olarak ortaya konulmuştur. Ayrıca her iki modelde de iyileştirmeler yapılmıştır. OKT'nin diğer

bileşeni olan okunuş sözlüğünün etkisi ise AM ile doğrudan ilişkilidir. Dolayısıyla AM'de okunuşu gerçekleştirilen bütün kelimeler mutlaka sözlüğe eklenmelidir. Aksi durumda sözlük doğru kestirimler yapamayabilir. Bu nedenle okunuş sözlüğüne, kelimelerin okunuşu, aynı kelimelerin farklı okunuşları, yabancı kelimelerin okunuşu ve kısaltmaların okunuşu eklenmiştir.

Geniş kelime dağarcığına sahip Türkçe OKT sistemi için bütün bileşenler hazırlanmıştır. OKT için gerekli bütün bileşenler tek tek ele alınarak geliştirilmiştir. Geniş kelime dağarcığı üzerine deneyler gerçekleştirebilmek için veri kümesi hazırlama bölümünde belirtildiği gibi bir test veri kümesi oluşturulmuştur. Test veri kümesinin detayları Çizelge 4.3'te verilmiştir. literatürdeki birçok çalışmanın alana özgü (hukuk, medikal, spor vb.) yapıldığı görülmektedir. Alana özgü çalışmalarda konuşulacak kelime dağarcığı belirli bir kümeye aittir. Ancak kendiliğinden konuşmaları belirli bir kümeye ait bilgiler ile gerçekleştirmek mümkün değildir. Bu nedenle tez çalışması kapsamında geniş kelime dağarcığına sahip bir OKT sistemi geliştirilmiştir. OKT için gerekli olan AM ve DM geniş kelime dağarcığına sahip olacak ve uzun bağımlılıkları modelleyecek şekilde geliştirilmiştir. AM üzerinde yaptığımız eğitim ve DM üzerinde gerçekleştirdiğimiz iyileştirmeler başarılı sonuçlar vermiştir.

Bu çalışmada, geniş kelime dağarcığı deneylerini gerçekleştirebilmek için bir deney prosedürü hazırlanmıştır. Bu prosedür için KHO hesabı java dili kullanılarak kodlanmış ve bir deney prosedürü uygulaması geliştirilmiştir. Geliştirilen uygulama referans metin ile OKT çıkışındaki metni karşılaştırmakta ve bir KHO oranı hesaplamaktadır. Karşılaştırma işlemi için ayrıca Google Cloud Speech-to-Text API kullanılmıştır. Google servisinin kullanılmasındaki ana amaç bu çalışmada geliştirilen geniş kelime dağarcığına sahip OKT sisteminin başarımını karşılaştırmalı olarak elde edebilmektir. Test prosedürü için geliştirilen uygulama Şekil 5.7'de verilen dizin yapısını kullanmaktadır. Şekil 5.7'deki dizin yapısını kullanan uygulama KHO hesabını Şekil 5.8'de verilen akış şemasına göre gerçekleştirmektedir. Çalışma içerisinde şekil 5.8'de gösterilen akış şemasına göre gerekli hesaplamalar yapılmış ve yazdırılmıştır. Öncelikle dizin ve dosya kontrolleri gerçekleştirilmiştir. Her bir kelimenin hata durumu kontrol edilmiş ve sonunda KHO sonucu yazdırılmıştır.

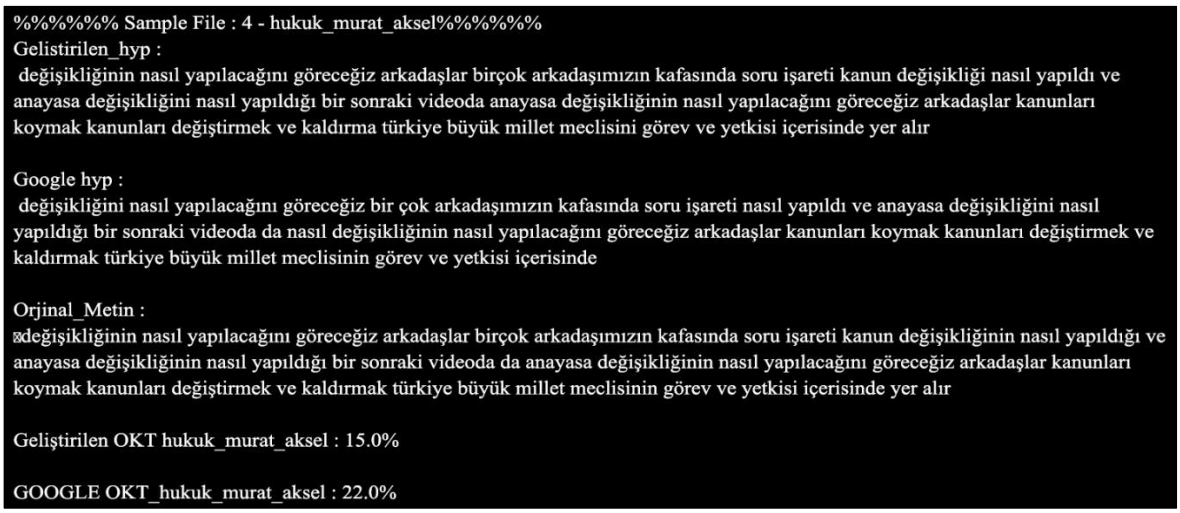
Şekil 5.7. Test prosedürü dizin yapısı



Şekil 5.8. Deney prosedürü akış şeması

Şekil 5.7'de gösterilen dizine göre "Transcript" içerisinde yer alan metin dosyaları bire bir gerçek kullanıcılar tarafından metne aktarılmış konuşma ifadelerini içermektedir. "Wave" ise metin dosyalarında belirtilen konuşma kayıtlarını ifade etmektedir. "Our_OKT" içerisinde "Wave" klasöründe yer alan konuşma kayıtlarının tez çalışması kapsamında

geliştirilen OKT sonuçları bulunmaktadır. “Wave” klasöründeki dosyalar Google’a gönderilmiş ve sonuçlar “Google_Results_Orig” klasörüne yerleştirilmiştir. Google sonuçları üzerinde sonradan işlem uygulanarak sonuçlar içerisindeki noktalama işaretleri kaldırılmış, büyük harfler küçük harfe dönüştürülmüş ayrıca rakamların metin karşılıkları yazılarak “Google_Results” içerisindeki dosyalar elde edilmiştir. Böylelikle bütün metin dosyalarının yazılışı birbirine benzetilerek daha net KHO hesabı gerçekleştirilmiştir. Örnek KHO hesabı çıktısı ise Şekil 5.9’da verilmiştir. Şekil 5.9’da hukuk alanında konuşan bir konuşmacının metne aktarılmış hali gösterilmiştir.



Şekil 5.9. Karşılaştırmalı KHO sonuçları

Şekil 5.9’da hukuk alanında konuşan bir konuşmacının metne aktarılmış konuşmaları yer almaktadır. Görüldüğü gibi bu çalışmada geliştirilen OKT ve Google OKT farklı sonuçlar vermektedir. Bu deney prosedürü Çizelge 4.3’te detayları verilen veri kümesine uygulanmış ve başarılı sonuçlar alınmıştır. Çizelge 5.10’da Google ile yapılan karşılaştırmalı geniş kelime dağarcığı deneylerine ait istatistikler verilmiştir.

Çizelge 5.10’da HS OKT Sistemi, bu çalışma kapsamında geliştirilen geniş kelime dağarcığına sahip Türkçe OKT’yi temsil etmektedir. Sonuçlar her bir konuşma kaydı için ele alınmıştır. Konuşma kayıtları çıktısında en düşük KHO, HS OKT sistemi ile elde edilmiştir. Ayrıca ortalama KHO hesabında da yine HS OKT sistemi başarılıdır. En yüksek KHO ise Google OKT sisteminde oluşmuştur. Deneyler incelendiğinde Google’ın öncelikle

ses kalitesini ve bazı durumları değerlendirdiği ardından OKT sistemine verdiği görülmüştür. Bu nedenle KHO yüksek çıkmıştır.

Çizelge 5.10. Geniş kelime dağarcığına sahip Türkçe OKT için karşılaştırmalı sonuçlar

Türkçe OKT Sisteminin Adı	En Düşük KHO (%)	En Yüksek KHO (%)	Ortalama KHO (%)
HS OKT Sistemi	3,0	93,0	18,7
Google OKT Sistemi	4,0	100	23,5

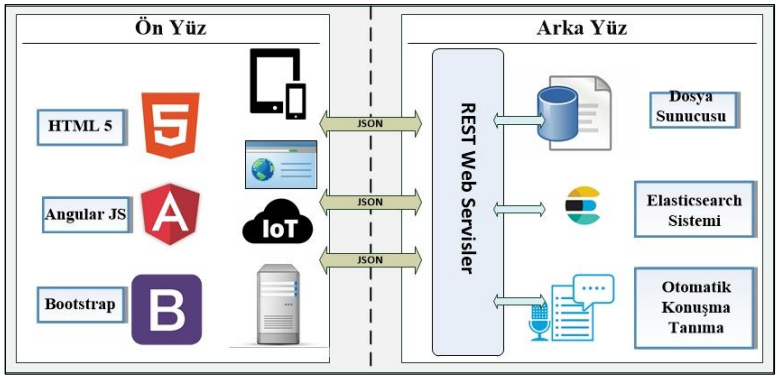
5.8. Web Servis Tabanlı Türkçe OKT Sisteminin Geliştirilmesi

OKT sistemlerinin başarımının artması ve arayüz teknolojilerinin gelişmesi bu alanda yapılan çalışmalara yeni bir boyut kazandırmıştır. Kolay erişilebilir olması ekonomik kazanımların ortaya çıkmasını sağlamış ve bu alana yapılan yatırımları arttırmıştır. Amazon firması "Amazon Triscribe", Microsoft firması "Azure Bing Speech", Google firması "Cloud Speech-to-Text", IBM ise "Watson Speech-to-Text" projesini başlatmıştır [62]. Bu kapsamda bulut platformu üzerinden sunulan UPA'lar sayesinde farklı uygulamalar geliştirilmiştir. Özellikle Google'un sunduğu "Ok Google", Microsoft'un sunduğu "Cortana", IBM'in sunduğu "Watson" ve Apple'ın sunduğu "Siri" kişisel asistan uygulamaları en iyi örneklerdendir. Ancak kullanıcılar bulut tabanlı OKT sistemlerinin gizliliğinden ve güvenliğinden tam anlamı ile emin değildirler. Bu nedenle çevrimdışı OKT sistemlerin yetersiz donanım kaynağına rağmen başarımlarını arttıracak yaklaşımlar gerçekleştirilmelidir [63].

Bu çalışmada, geniş kelime dağarcığına sahip Türkçe OKT'nin kolay kullanımı ve yerel sunucularda rahatlıkla çalışabilmesi için bir Web Servis Tabanlı Türkçe OKT Platformu (WSTTOKTP, Web Service Based Turkish OKT Platform) geliştirilmiştir. Uzaktan sesli komut sistemleri, multimedya varlıklarının metne aktarılıp raporlanması vb. tüm yönetim gereksinimleri geliştirilen WSTTOKTP sayesinde gerçekleştirilebilmektedir. Teknolojik ilerlemeler dikkate alınarak daha esnek ve erişimi daha kolay hale getiren iyileştirmeler

yapılmıştır. Teknoloji alanındaki gelişmeler detaylandırılarak WSTTOKTP üzerine uygulanmıştır.

WSTTOKTP'ın amacı mümkün olduğu kadar düşük kaynaklar ile gerekli işlemleri gerçekleştirmektir. Diğer bir amaç ise mümkün olduğu kadar çok kullanıcıya erişebilmektir. Bu amaçla kolay kullanılabilir, ölçeklenebilir, güvenli, performansı yüksek ve katmanlı bir mimariye sahip WSTTOKTP geliştirilmiştir. WSTTOKTP tasarımı yapılırken literatürden farklı olarak masaüstü veya konsol uygulaması yerine TDT web servis yaklaşımına sahip bir OKT platformu geliştirilmiştir. Geliştirilen platform birden fazla bileşeni bir arada içermektedir. WSTTOKTP'ın ön yüz ve arka yüz olarak ikiye ayrılmış temel mimarisi Şekil 5.10'da verilmiştir.



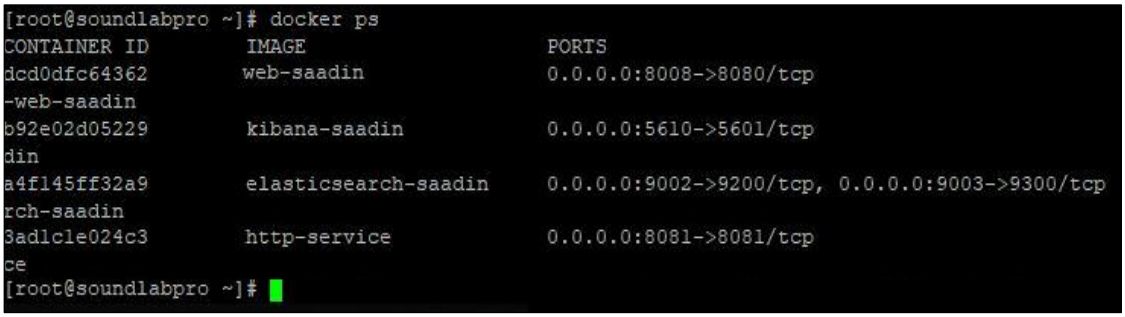
Şekil 5.10. WSTTOKTP ön yüz ve arka yüz olarak ikiye ayrılmış temel mimarisi

Şekil 5.10'da görüldüğü gibi web servisler kullanıcıdan gelen istekleri değerlendirmekte ve gerçekleştirilmesini sağlamaktadır. Web servisler Java programlarının geliştirilmesi ve çalıştırılması konusunda verimlilik artışı sağlamayı hedefleyen yeni özellikler, iyileştirme ve hata düzeltmeleri içeren Java 8 versiyonu kullanılarak geliştirilmiştir. Web servisler Apache Yazılım Vakfı tarafından geliştirilmiş açık kaynaklı bir Java Sunucu uygulaması olan Apache Tomcat sunucusu üzerinde çalıştırılmıştır.

Kullanıcı arayüzü, TypeScript tabanlı bir açık kaynaklı web uygulama çerçevesi olan Angular 7 kullanılarak hazırlanmıştır. Arayüz ayrıca Hiper Metin İşaretleme Dili (HMİD, Hypertext Markup Language) 5 ve açık kaynak kodlu, web sayfaları veya uygulamaları

geliştirmek için kullanılabilecek araçları içeren Bootstrap ile desteklenmiştir. Arayüz, Angular uygulamalarını başlatmak, geliştirmek, iskelet oluşturmak ve bakımını yapmak için kullandığınız bir komut satırı arabirim aracı olan Angular komut satırı arayüzü üzerinde çalıştırılmıştır. Geliştirilen platformda OKT işlemi sonucunda elde edilen metin verileri Lucene kütüphanesine dayanan, Hiper Metin Transfer Protokolü (HMTP, Hypertext Transfer Protocol) web arayüzü ve şema içermeyen JNB belgelerini kullanan dağıtılmış, çok kullanıcılı bir tam metin arama motoru olan Elasticsearch sistemine gönderilmiştir.

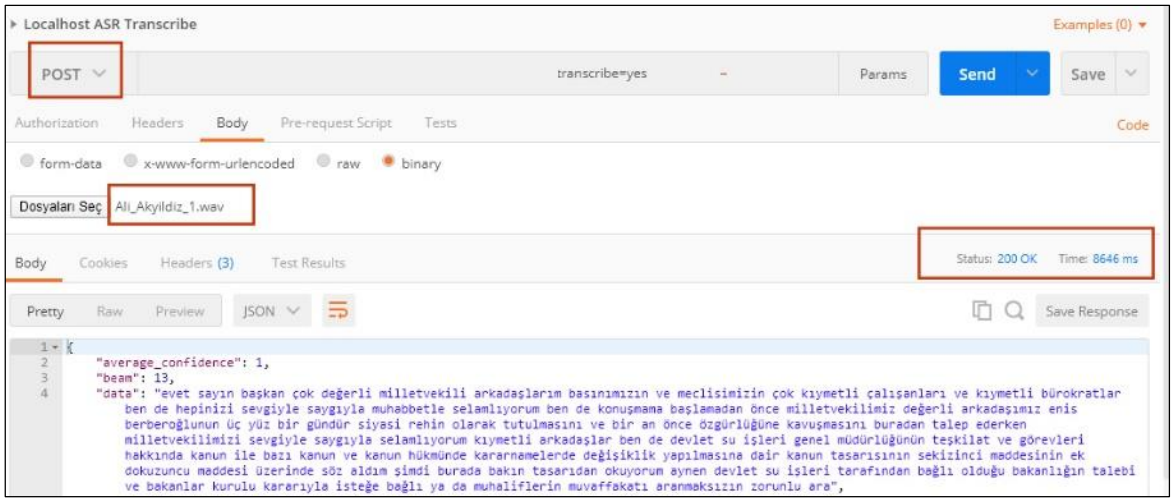
Şekil 5.10'da belirtilen teknolojilerin her birini bir sanal makinada çalıştırmak mümkündür. Bu kullanım şekli bir nevi kolaylık sağlayacaktır ancak birden fazla sanal makinayı aynı anda çalıştırmak yetersiz kaynak problemini ortaya çıkmaktadır. Bu nedenle sanal makine tercihi yapılmamıştır. Bu kullanımın yerine Docker sistemi kullanılmıştır. Sanal makineler kullanıcı operasyon çağrılarını ana sistem operasyonu çağrılarına çevirebilmek için oldukça zaman ve enerji harcamaktadır. Docker'da ise bu kayıp yoktur. Şekil 5.11'de OKT platformunu oluşturan her bileşenin farklı bir Docker paketi olarak hazırlandığı gösterilmiştir.



Şekil 5.11. OKT platform Docker yapısı

Şekil 5.11'de belirtilen Docker yapılarından olan http-service ile temel olarak Kaldi ile geliştirilen OKT sisteminin deşifre işlemi gerçekleştirilmiştir. Bu Docker yapısının tek amacı kendisine gelen konuşma kayıtlarını metne aktarmaktır. Şekil 5.11'de görülen web-saadin isimli Docker yapısı ile ön işlemler ve kullanıcı ile etkileşimi sağlayacak arayüzler geliştirilmiştir. Elasticseearch-saadin ile belirtilen Docker yapısında ise platform üzerinden elde edilen bütün veriler saklanmıştır. Konuşma kayıtlarına karşılık gelen metinler, kullanıcı bilgileri ve giriş-çıkış kayıtlarının tamamı elastic üzerinde saklanmıştır. Diğer yandan kibana

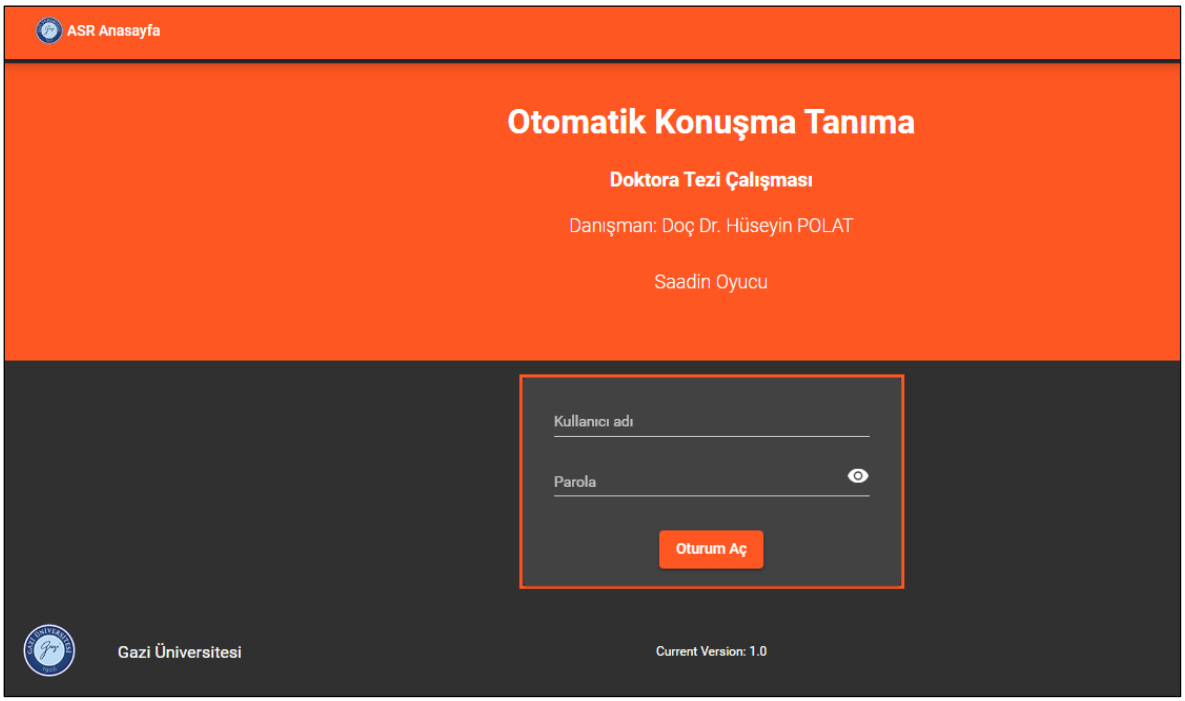
ile elastic sistemindeki verilerin görselleştirilmesi ve yönetilmesi sağlanmıştır. Sistemin tümünü test edebilmek için farklı yaklaşımlar kullanılmıştır. Geliştirilen OKT platformun web servis test işlemlerinde öncelikle Postman kullanılmıştır. Postman kullanımı arayüzü Şekil 5.12'de verilmiştir.



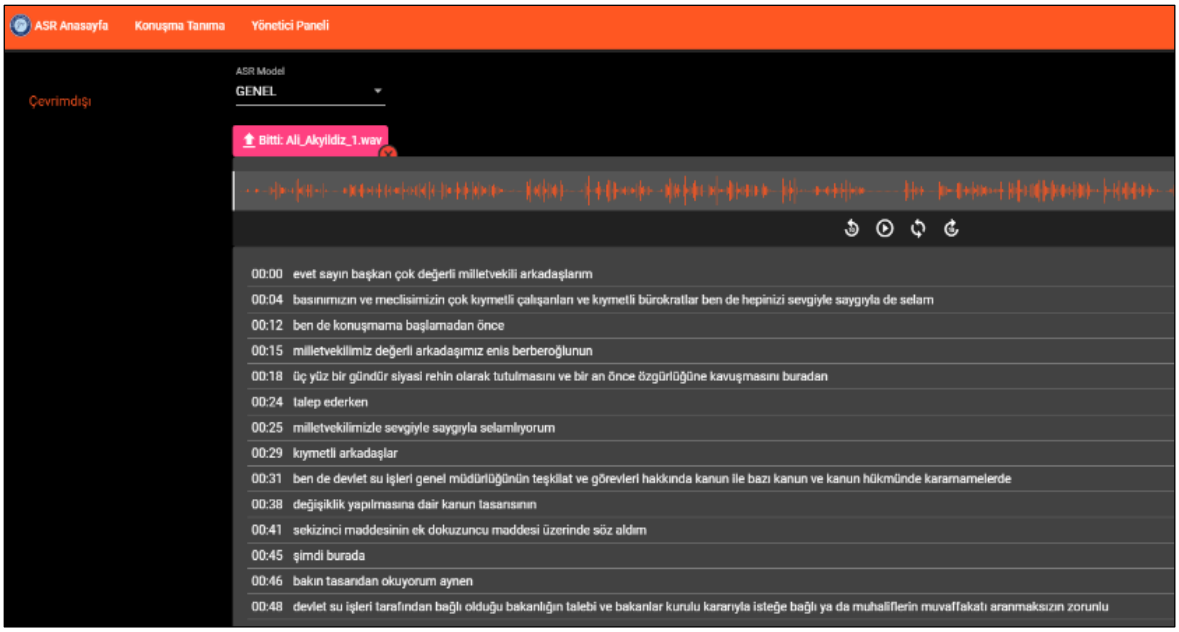
Şekil 5.12. Web servis test işlemleri

Şekil 5.12'de verildiği gibi POST metodu ile gönderilen WAV dosyasına karşılık bir metin değeri geri dönmektedir. Postman sistemin cevap süresini vermektedir. Bu nedenle 100 adet 59 saniye uzunluğunda konuşma kaydı hazırlanmış ve OKT platform web servisine gönderilmiştir. Böylelikle platformun cevap verme süresi hakkında bilgi elde edilmiştir. 59 saniyelik bir konuşma kaydı için platformun ortalama 9,6 saniyede cevap verdiği görülmüştür.

Web servis yapısının test işlemlerinden sonra ikinci aşamada arayüz test işlemleri gerçekleştirilmiştir. Duyarlı tasarıma sahip bir yaklaşım ile geliştirilen WSTTOKTP arayüzünü test etmek için birçok araç bulunmaktadır. Bu test araçlarından en yaygın kullanılanı Google Chrome web tarayıcısıdır. Chrome geliştiricilere cihaz modu ve mobil simülasyon araçlarını ücretsiz olarak sunmaktadır. Bu çalışmada geliştirilen WSTTOKTP arayüzleri hem Chrome üzerinde hem de fiziksel diğer ortamlarda (akıllı telefon, tablet, dizüstü bilgisayar) test edilmiş ve başarılı sonuçlar alınmıştır. Geliştirilen WSTTOKTP arayüzü farklı akıllı telefonlarda ve web tarayıcısında Şekil 5.13,14 ve 15'teki gibi bir görünüme sahiptir.



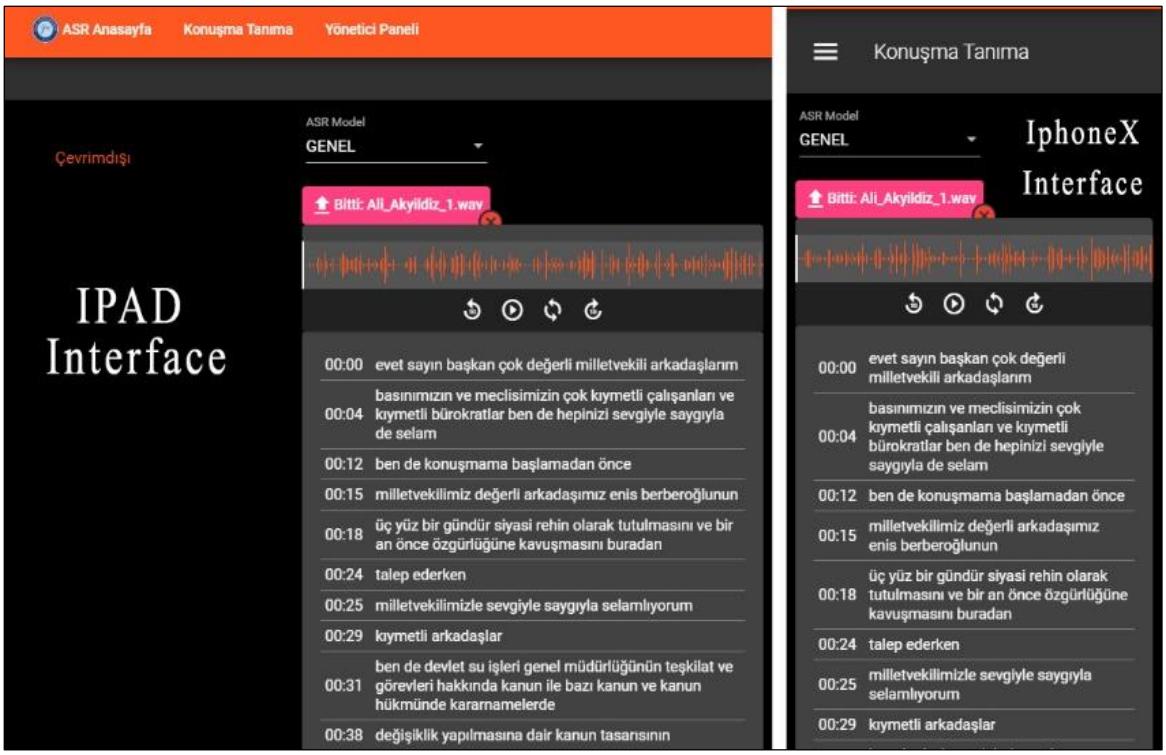
Şekil 5.13. WSTTOKTP kullanıcı girişi arayüzü



Şekil 5.14. WSTTOKTP web arayüz görünümü

Bu çalışmada web servis yapısı oluşturulmuştur. Web servis yapısındaki metotlara ve verilere erişim servis katmanı üzerinden sağlanmıştır Web servis ile etkileşimi sağlayabilmek için bir arayüz geliştirilmiştir. Farklı teknolojileri bir arada kullanılarak geliştirilen platform iki farklı şekilde kullanılabilmektedir. Bunlardan ilki standart bir web

tarayıcısı aracılığıyla ve ikincisi ise bir API aracılığıyla kullanılabilmektedir. İlk yaklaşımda teknik uzmanlığa ihtiyaç duymadan konuşma kayıtları Türkçe OKT sistemine aktarılabilmektedir. İkinci yaklaşımda ise arayüz kullanmaya gerek kalmadan uygulamaların ve diğer servislerin platforma erişimi sağlanmıştır. Böylelikle sayıları gittikçe artan mobil cihazlarda, nesnelerin interneti ekosisteminde ve aynı zamanda diğer erişim cihazlarında sorunsuz çalışabilen bir platform geliştirilmiştir. Geliştirilen sistemin bütün bileşenleri 32 çekirdekli bir sunucuda çalışmaktadır. Bu çalışma sonucunda OKT sisteminin web servis tabanlı bir arayüz ile kullanıcılara, farklı uygulamalara veya cihazlara sunulabileceği gösterilmiştir.



Şekil 5.15. WSTTOKTP mobil arayüz görünümü

6. SONUÇ VE ÖNERİLER

Bu çalışmada, mevcut OKT yöntemleri ve yaklaşımları incelenerek OKT alanındaki gelişmeler detaylı olarak sunulmuştur. Araştırmacıların bu alanda yaptıkları çalışmalarda kullandıkları yaklaşımlar, yöntemler, veri kümeleri, başarım ölçütleri ve bu alanda karşılaştıkları zorluklar ele alınmıştır. Araştırmacıların hangi dil üzerine çalıştığı ve çalışılan dil üzerindeki OKT zorlukları belirtilmiştir. Araştırmalar sonucu elde edilen bilgiler doğrultusunda; akustik ortamlara karşı dayanıklılık, bilinmeyen kelimelerin tespiti, Türkçe OKT'nin geniş dağarcık düzeyindeki başarısı, yetersiz kaynak durumu ve OKT üzerine uygulanabilecek hesaplamalı mimariler üzerine değerlendirmelere yer verilmiştir.

OKT sistemleri için Türkçe yetersiz kaynak diller arasında yer almaktadır. Bu nedenle tez çalışmasında öncelikle yetersiz kaynak durumunu telafi etmeye yönelik çalışmalar gerçekleştirilmiştir. OKT için klasik olarak elle yazıya dökülen veri kümesi hazırlama tekniklerine alternatif bir yaklaşım sunulmuştur. Sunulan yaklaşımda üç farklı yöntem kullanılmıştır. İlk yöntemde, önceden kaydedilmiş filmlerin alt yazı belgeleri kullanılmıştır. İkinci yöntemde, veriler mobil bir uygulama aracılığıyla gerçek kullanıcılar tarafından elde edilmiştir. Üçüncü yöntemde ise transfer öğrenme yaklaşımı kullanılmıştır. Böylece, Türkçe OKT sistemleri için en geniş veri kümesi hazırlama ve doğrulama yaklaşımı sunulmuştur. Bu yaklaşım ile Türkçe için mevcut en geniş veri kümesi HS veri kümesi (350 saat) elde edilmiştir.

Elde edilen veri kümesi konuşma ifadesi, ortam gürültüsü ve dış mekân konuşması gibi farklı konuşma örneklerini içermektedir. Bu nedenle elde edilen veri kümesi ile geliştirilen OKT sistemi akustik ortamlara karşı dayanıklılık göstermektedir. Elde edilen veri kümesi kullanılarak geliştirilen akustik modelde çevresel etkiler daha iyi temsil edilmiştir. Ayrıca veri kümesi geniş kelime dağarcığına sahiptir. Dolayısıyla Türkçe OKT sistemlerindeki dağarcık dışı kelimelerin sayısı azaltılmıştır. Ancak Türkçedeki uzun bağımlılıkların modellenebilmesi için sadece veri kümesinin geliştirilmesi tek başına yeterli değildir. Bu nedenle konuşma dosyalarının OKT sistemine verilmeden önce işlenebilmesi için bir ön işlem modülü geliştirilmiştir. Ön işlem modülünün temel amacı AM başarımını arttıracak görevleri yerine getirmektir.

Ön işlem modülü, OKT sistemlerinin başarımını negatif yönde etkileyecek akustik bilgilerin ortadan kaldırılması için geliştirilmiştir. Geliştirilen modül ile konuşma içerisindeki sessizlik alanları kaldırılmıştır. Ayrıca akustik bilgide uzun bağımlılıkları engellemek adına konuşma bilgisi parçalara ayrılmıştır. Parçalanan her konuşma bilgisi ayrı olarak sırası ile OKT sistemine verilmiştir. OKT sisteminin çıkışında konuşma parçalarına karşılık gelen metin çıktıları birleştirilerek sunulmuştur. Gerçekleştirilen deneylerde sessizliğin kaldırılması işleminin OKT sistemlerinin başarımını doğrudan etkilediği görülmüştür. Türkçe OKT sisteminin KHO başarımında %1,3'lük bir kazanım elde edilmiştir. Ancak bazı durumlarda konuşmanın parçalara ayrılması işleminin başarımı kötü etkilendiği gözlemlenmiştir. Bunun nedeni sesbirim bilgisinin tam olarak çıkarılamayacağı kadar küçük konuşma parçalarının OKT sistemine verilmesidir. Ayrıca bazı durağan konuşmacıların heceler arasında sessiz kalmaları sadece hecelerin parçalara ayrılmasına neden olmuştur. Bu nedenle ön işlem modülünde konuşmacıların yapısal özellikleri yani durağan, kekeme veya çok hızlı konuşan konuşmacıların tespit edilmesi gerekmektedir.

Akustik bilgideki uzun bağımlılıklar için önerilen ön işlem modülünden sonra OKT için gerekli olan AM ve DM farklı bir yaklaşım ile ele alınmıştır. Öncelikle AM için önerilen modellerde DSA tabanlı yaklaşımların GKM-SMM tabanlı yaklaşımlara göre daha iyi sonuç verdiği görülmüştür. DSA tabanlı yaklaşımlar ile geliştirilen modellerde KHO başarımında %6'lık bir kazanım sağlanmıştır. DM'nin modellenmesi ise Türkçe'nin sondan eklemesi yapısı ve uzun bağımlılıklar içermesi nedeniyle zorlu bir görevdir. Bu nedenle uzun bağımlılıkları modelleyebilecek, okunuş sözlüğü çıkışındaki kelimelerin dizilimi için daha iyi sonuçlar verebilecek bir DM iyileştirmesi gerçekleştirilmiştir.

Çalışmada, cümle düzeyinde bir DM iyileştirme yöntemi önerilmiştir. Önerilen yöntem, mevcut olarak dil modellemesinde kullanılan istatistiksel ve YSA tabanlı olarak geliştirilen DM'lere uygulanmıştır. Deneysel sonuçlar önerilen yaklaşımın KHO'nında %4,4'lük bir iyileşme sağladığını göstermiştir. Ayrıca Kenser-Ney smoting gibi iyileştirme algoritmalarının kullanıldığı OKT sistemlerine göre daha iyi sonuçlar verdiği görülmüştür. Sonuç olarak temelde bir düzene sahip olan DM'lerin performansının en iyi listelerinin yeniden düzenlenmesi ile iyileştirilebileceği gösterilmiştir. Gerçekleştirilen deneylerde önerilen yöntemin, istatistiksel ve UKSB-TSA tabanlı DM'lere uygulandığında tutarlı bir performans iyileştirmesi sağladığı görülmüştür.

Çalışma kapsamında gerçekleştirilen deneylerde okunuş sözlüğüne veri kümesindeki bütün kelimeler eklenmiştir. Ayrıca Türkçede sıklıkla kullanılan yabancı kelimeler ve birden farklı okunuşa sahip kelimeler de okunuş sözlüğüne dâhil edilmiştir. Okunuş sözlüğünün hazırlanması, AM ve DM'nin geliştirilmesi için mevcutta bulunan farklı veri kümeleri ve yeni hazırlanan veri kümesi kullanılmıştır. Veri kümesi boyutunun artması, OKT sonucunu doğrudan etkilemiş ve önerilen yöntemler geniş veri kümelerinde daha iyi sonuçlar vermiştir. Özellikle geniş veri kümeleri ile geliştirilen AM ve DM'nin deneysel sonuçları incelendiğinde DSA tabanlı yaklaşımların klasik GKM tabanlı yaklaşımlardan daha iyi sonuç verdiği görülmüştür. DSA tabanlı geliştirilen modellerde, model boyutundaki ve derinliğindeki bir artışın hata oranını düşürdüğü gözlemlenmiştir. Ancak model derinliğinin belirli bir sınırı bulunduğu ve model boyutunun artmasının deşifre işleminde çok zaman aldığı görülmüştür. Birden fazla gizli katmana sahip olmak önemlidir. Ancak DSA'ların büyük miktarda gizli ünite ile oluşturdukları farkların, hesaplanan tüm ölçüm standartlarının aksine biraz küçük olduğu gözlemlenmiştir.

Deneysel çalışmalar sadece mevcut test veri kümesi üzerinden değerlendirilmemiştir. Geniş kelime dağarcığı deneylerinde kullanılmak üzere siyaset, bilim, teknoloji, magazin, spor, sanat, sağlık ve hukuk alanında konuşmalar içeren bir test veri kümesi hazırlanmıştır. Her alandan konuşmalar içeren geniş kelime dağarcığına sahip bu test veri kümesi üzerinde konuşma tanıma deneyleri gerçekleştirilmiştir. Geniş kelime dağarcığına sahip deneylerde %18,7'lik KHO elde edilmiştir.

Geniş kelime dağarcığına sahip deney sonuçlarını karşılaştırmalı olarak sunabilmek için Google servislerinden yararlanılmıştır. Google Türkçe OKT programlama arayüzü, karşılaştırmalı sonuçlar elde edebilmek için kullanılmıştır. Geniş kelime dağarcığına sahip olarak hazırlanan test veri kümesinde Google'ın %23,5'lik KHO verdiği görülmüştür. Bu durum çalışma kapsamında geliştirilen Türkçe OKT sisteminin geniş kelime dağarcığına sahip konuşmalarda Google OKT servislerinden daha iyi sonuçlar verdiğini göstermiştir.

Sonuç olarak bu tez çalışmanın dört ana çıktısı mevcuttur. Bunlardan ilki, Türkçe'deki mevcut en geniş konuşma ve bu konuşmalara ait metinleri içeren veri kümesinin hazırlanmasıdır. İkinci çıktı ise akustik modele girdi olarak verilecek konuşma dosyalarında

başarımı negatif yönde etkileyecek akustik bilgilerin ortadan kaldırılmasıdır. Ayrıca akustik bilgide uzun bağımlılıkları engellemek için konuşma dosyasının parçalara ayrılmasıdır. Üçüncü çıktı ise mevcut en geniş okunuş sözlüğünün hazırlanması ve cümle düzeyinde bir DM iyileştirme yönteminin geliştirilmedir. Son çıktı ise kullanıcıların rahatlıkla erişebileceği web servis tabanlı Türkçe OKT platformunun geliştirilmesidir. Web servis yapısındaki metotlara ve verilere erişim servis katmanı üzerinden sağlanmıştır Web servis ile etkileşimi sağlayabilmek için bir arayüz geliştirilmiştir. Farklı teknolojileri bir arada kullanılarak geliştirilen platform iki farklı şekilde kullanılabilmektedir. Bunlardan ilki standart bir web tarayıcısı aracılığıyla ve ikincisi ise bir programlama arayüzü aracılığıyla kullanılabilmektedir. İlk yaklaşımda teknik uzmanlığa ihtiyaç duymadan konuşma kayıtları Türkçe OKT sistemine aktarılabilmektedir. İkinci yaklaşımda ise arayüz kullanmaya gerek kalmadan uygulamaların ve diğer servislerin platforma erişimi sağlanmıştır. Böylelikle sayıları gittikçe artan mobil cihazlarda, nesnelerin interneti ekosisteminde ve aynı zamanda diğer erişim cihazlarında sorunsuz çalışabilen bir platform geliştirilmiştir.

Gelecek çalışmalarda araştırmacılar, bu tez çalışmasında elde edilen veri kümesi üzerinde daha fazla deneyin yapılması ve veri kümesinin daha da genişletilmesi üzerine çalışmalar gerçekleştirebilirler. Gürültülü ortamlarda Türkçe konuşma tanıma ve konuşma tanıma alanında kendi kendine öğrenme ile ilgili yaklaşımlar sunabilirler. Ön işlem modülünde hangi dilin konuşulduğu tespit edilip konuşma dosyalarının o dile ait OKT sistemine girdi olarak verilmesi çalışılabilir. Ayrıca bir bütün olarak sunulan AM, DM ve okunuş sözlüğünün parçalar halinde bağımsız olarak geliştirilmesi sağlanabilir. Böylelikle Sadece AM, DM veya okunuş sözlüğüne rahatlıkla ekleme veya çıkarma işlemi yapılabilir.

Gelecek çalışmalarda en büyük payın hem görsel (örneğin dudak hareketleri) hem de işitsel (akustik bilgi - sesbirimler) verilerden yararlanılarak geliştirilen OKT sistemlerinin olacağı düşünülmektedir. Ancak mevcut durumda Türkçe OKT alanı için hem görsel hem de işitsel verilerin eşleştirildiği bir veri kümesi bulunmamaktadır. Veri kümesinin hazırlanması, görsel ve işitsel verilerden yararlanan OKT sistemlerinin geliştirilmesi Türkçe için konuşma tanıma alanında önemli gelişmeler sağlayacaktır.

KAYNAKLAR

1. Gönenç, E. Ö. (2012). İletişimin tarihsel süreci. *İletişim Fakültesi Dergisi*, 1(28), 87-101.

2. Çerçi, A. (2014). Telaffuz vurgu ve tonlama konularının dinleme destekli öğretimi. *Journal of Turkish Studies*, 9(3), 467-467.

3. Prakoso, H., Ferdiana, R. and Hartanto, R. (2017). *Indonesian automatic speech recognition system using CMUSphinx toolkit and limited dataset*. International Symposium on Electronics and Smart Devices, 283-286.

4. Çakır, M. Y. (2017). *Gerçek zamanlı yüksek kalitede ses tanıma*. Yüksek Lisans Tezi, Sabahattin Zaim Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 10-55.

5. Xiao, Q. (2007). Biometrics-technology, application, challenge, and computational intelligence solutions. *IEEE Computational Intelligence Magazine*, 2(2), 5-10.

6. Miao, Y. (2014). Kaldi+PDNN: Building DNN-based ASR systems with kaldi and PDNN. *arXiv CoRR:1401.6*, 1-4.

7. Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357-366.

8. Narang, S. and Divya Gupta, M. (2015). International journal of computer science and mobile computing speech feature extraction techniques: a review. *International Journal of Computer Science and Mobile Computing*, 4(3), 107-114.

9. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A. and Vanhoucke, V. (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.

10. Aubert, X. L. (2002). An overview of decoding techniques for large vocabulary continuous speech recognition. *Computer Speech and Language*, 16(1), 89-114.

11. Stolcke, A. (2000). *Entropy-based pruning of backoff language models*. DARPA Broadcast News Transcription and Understanding Workshop, 270-274.

12. Holmes, J. N., Holmes, W. J. and Garner, P. N. (1997). *Using formant frequencies in speech recognition*. European Conference on Speech Communication and Technology, 2083-2086.

13. Olson, H. F. and Belar, H. (1956). Phonetic typewriter. *Journal of the Acoustical Society of America*, 28(6), 1072-1081.

14. Fry, D. B. (1959). Theoretical aspects of mechanical speech recognition. *Journal of the British Institution of Radio Engineers*, 19(4), 211-218.

15. Forgie, J. W. and Forgie, C. D. (1959). Results obtained from a vowel recognition computer program. *Journal of the Acoustical Society of America*, 31(11), 1480-1489.

16. J., S. (1961). Recognition of Japanese vowels - preliminary to the recognition of speech. *Journal of the Radio Research Laboratory*, 37(8), 193-212.

17. Sakoe, H. (1979). Two-level dp-matching-a dynamic programming-based pattern matching algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(6), 588-595.

18. Martin, T. B., Nelson, A. L. and Zadell, H. J. (1964). *Speech Recognition By Feature-Abstraction Techniques.* United States of America: Defense Technical Information Center, 1-10.

19. Vintsyuk, T. K. (1968). Dpeech discrimination by dynamic programming. *Kibernetika*, 4(1), 59-89.

20. Chiba, H. S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1), 63-72.

21. Velichko, V. M. and Zagoruyko, N. G. (1970). Automatic recognition of 200 words. *International Journal of Man-Machine Studies*, 2(3), 223-234.

22. Jelinek, E. F. (1985). The development of an experimental discrete dictation recognizer. *Informatik-Anwendungen-Trends und Perspektiven*, 73(11), 1616-1624.

23. Reddy, D. R. (1966). Approach to computer speech recognition by direct analysis of the speech wave. *The Journal of the Acoustical Society of America*, 40(5), 1273-1273.

24. Rabiner, L. R., Levinson, S. E., Rosenberg, A. E. and Wilpon, J. G. (1979). Speaker-ındependent recognition of ısolated words using clustering techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(4), 336-349.

25. Furui, S., Kikuchi, T., Shinnaka, Y. and Hori, C. (2004). Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing*, 12(4), 401-408.

26. Klatt, D. H. (1977). Review of the ARPA speech understanding project, *Journal of the Acoustical Society of America*. 62(6), 1345-1366.

27. Bridle, J. S., Brown, M. D. and Chamberlain, R. M. (1983). Continuous connected word recognition using whole word templates. *Radio and Electronic Engineer*, 53(4), 167-175.

28. Myers, C. S. and Rabiner, L. R. (1981). A level building dynamic time warping algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2), 284-297.

29. Liu, D., Fei, S., Hou, Z., Zhang, H. and Sun, C. (2007). *Advances in Neural Networks.* Germany: Springer, 493-600.

30. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE,* 37(3), 257-286.

31. Gales, M. J. F. (1998). Maximum likelihood linear transformations for HMM-based speech recognition, *Computer Speech & Language*, 12(2), 75-98.

32. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review,* 65(6), 386-408.

33. Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning represenations by back propogation errors. *Letters to Nature*, 5(3), 533-536.

34. Juang, B. H. and Furui, S. (2000). Automatic recognition and understanding of spoken language - A first step toward natural human-machine communication. *Proceedings of the IEEE*, 88(8), 1142-1165.

35. Furui, S. (2005). Recent progress in corpus-based spontaneous speech recognition. *IEICE Transactions on Information and Systems*, 88(3), 366-375.

36. Koo, M. W., Lee, C. H. and Juang, B. H. (2001). Speech recognition and utterance verification based on a generalized confidence score. *IEEE Transactions on Speech and Audio Processing*, 9(8), 821-832.

37. Kawahara, T., Lee, C. H. and Juang, B. H. (1996). *Key-Phrase Detection and Verification for Flexible Speech Understanding*. International Conference on Spoken Language Processing, 861-864.

38. Lin, H. and Ou, Z. (2006). *Switching auxiliary chains for speech recognition based on dynamic bayesian networks*. International Conference on Pattern Recognition, 258-261.

39. Robinson, T. and Fallside, F. (1991). A recurrent error propagation network speech recognition system. *Computer Speech and Language*, 5(3), 259-274.

40. Shahin, M., Ahmed, B., Mckechnie, J., Ballard, K. and Gutierrez-osuna, R. (2014). *A comparison of gmm-hmm and dnn-hmm based pronunciation verification techniques for use in the assessment of childhood apraxia of speech*. Annual Conference of the International Speech Communication Association, 1583-1587.

41. Mohamed, A., Dahl, G. E. and Hinton, G. (2012). Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1), 14-22.

42. Chen, F. and Jokinen, K. (2010). *Speech Technology: Theory And Applications*. United States of America: Springer, 1-331.

43. Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. United States of America: MIT Press, 95-250.

44. Kimanuka, U. A. and Büyük, O. (2018). Turkish speech recognition based on deep neural networks. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 22(1), 319-329.

45. Fohr, D., Mella, O. and Illina, I. (2017). *New paradigm in speech recognition: deep neural networks*. IEEE International Conference on Information Systems and Economic Intelligence, 1-7.

46. Kat, L. W. and Fung, P. (1999). *Fast accent ıdentification and accented speech recognition*. IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings, 221-224.

47. Thimmaraja, Y. G. and Jayanna, H. S. (2017). *Creating language and acoustic models using kaldi to build an automatic speech recognition system for kannada language*. IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, 161-165.

48. Chan, H. Y. and Woodland, P. (2004). *Improving broadcast news transcription by lightly supervised discriminative training*. IEEE International Conference on Acoustics, Speech, and Signal Processing, 3-6.

49. Sainath, T., Mohamed, A. R., Kingsbury, B. and Ramabhadran, B. (2013). *Deep convolutional neural networks for lvcsr*. IEEE International Conference on Acoustics, Speech and Signal Processing, 8614-8618.

50. Chan, W. and Lane, I. (2015). *Deep convolutional neural networks for acoustic modeling in low resource languages*. IEEE International Conference on Acoustics, Speech and Signal Processing, 2056-2060.

51. Arisoy, E. and Arslan, L. M. (2005). Turkish *dictation system for broadcast news applications*. European Signal Processing Conference, 1351-1354.

52. Raina, R. and Madhavan, A. (2009). *Large-scale deep unsupervised learning using graphics processors.* International Conference On Machine Learning, 873-880.

53. Rosdi, F. (2017). *Assessing automatic speech recognition in measuring speech intelligibility : a study of Malay speakers with speech impairments.* International Conference on Electrical Engineering and Informatics, 1-6.

54. Tong, R., Wang, L. and Ma, B. (2017). *Transfer learning for children's speech recognition.* International Conference on Asian Language Processing, 36-39.

55. Lin, S., Tsunakawa, T., Nishida, M. and Nishimura, M. (2017). *DNN-based feature transformation for speech recognition using throat microphone*. Asia-Pacific Signal and Information Processing Association, 1-4.

56. Baniardalan, F. and 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.

57. Gali, J., Šumarac, D., Jovi, S. T. and Markovi, B. (2017). *Recognition of bimodal produced speech based on support vector machines*. Telecommunication Forum, 1-4.

58. Shahnawazuddin, S., Deepak, K. T., Pradhan, G. and Sinha, R. (2017). *Enhancing noise and pitch robustness of children's ASR.* IEEE International Conference on Acoustics, Speech and Signal Processing, 5225-5229.

59. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T. and Rubinstein M. (2018). Looking to listen at the cocktail party: a speaker-ındependent audio-visual model for speech separation. *ACM Transaction Graphic*, 37(4), 112-122.

60. Nasereddin, H. H. O., Graph, D. A., Dynamic, D. B. N. and Networks, B. (2017). *Classification techniques for automatic speech recognition (ASR) algorithms used with real time speech translation.* Computing Conference, 200-207.

61. Li, Q., Zheng, J., Tsai, A. and Zhou, Q. (2002). Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *IEEE Transactions on Speech and Audio Processing*, 10(3), 146-157.

62. Moon, D., Chung, S., Kwon, S., Seo, J. and Shina, J. (2019). Comparison and utilization of point cloud generated from photogrammetry and laser scanning: 3D world model for smart heavy equipment planning. *Automation in Construction,* 98(2), 322-331

63. Bumbalek, Z., Zelenka, J. and Kencl, L. (2012). *Cloud-Based Assistive Speech-Transcription Services*. Germany: Springer, 113-116.

64. Yu, B., Li, H. and Fang, C. (2012). Speech emotion recognition based on optimized support vector machine. *Journal of Software,* 7(12), 2726-2733.

65. Sivanagaraja, T., Ho, M. K., Khong, A. W. H. and Wang, Y. (2017). *End-to-End speech emotion recognition using multi-scale convolution networks*. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 1-4.

66. Lefter, I. and Jonker, C. M. (2017). *Aggression recognition using overlapping speech.* Seventh International Conference on Affective Computing and Intelligent Interaction, 299-304.

67. Engineering, E., Mara, U. T. and Pauh, P. (2017). *Automatic gender recognition using linear prediction coefficients and artificial neural network on speech signal*. IEEE International Conference on Control System, Computing and Engineering, 24-26.

68. Liu, W. K. ans Kung, P. ((1999).). F*ast accent identification and accented speech recognition*. IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings, 221-224.

69. Sagha, H. and Deng, J. (2017). *The effect of personality trait , age , and gender on the performance of automatic speech valence recognition*. Seventh International Conference on Affective Computing and Intelligent Interaction, 86-91.

70. Zheng, Y. (2005). *Accent detection and speech recognition for Shanghai-accented Mandarin.* Annual Conference of the International Speech Communication Association, 7-10.

71. Schultz, T. (2002). *Globalphone: a multilingual speech and text database developed at karlsruhe university.* Annual Conference of the International Speech Communication Association, 345-348.

72. Salor, Ö., Pellom, B. L., Ciloglu, T. and Demirekler, M. (2007). Turkish speech corpora and recognition tools developed by porting SONIC: Towards multilingual speech recognition. *Computer Speech and Language*, 21(4), 580-593.

73. Amodei, D., Anubhai, R., Battenberg, E., Case, C. and Casper. J. (2015). *Deep speech 2: end-to-end speech recognition in English and Mandarin*. International Conference on Learning Representations, 1-28.

74. Li, L. (2013). *Hybrid deep neural network - hidden markov model (dnn-hmm) based speech emotion recognition.* Humaine Association Conference on Affective Computing and Intelligent Interaction, 312-317.

75. Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, 61(1), 85-117.

76. Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56(1), 85-100.

77. Oyucu, S., Sever, H., and Polat, H. (2019). Otomatik konuşma tanımaya genel bakış, yaklaşımlar ve zorluklar: Türkçe konuşma tanımanın gelecekteki yolu. *Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji*, 7(4), 834-854.

78. Işık, G. and Artuner, H. (2020). Turkish dialect recognition using acoustic and phonotactic features in deep learning architectures. *Bilişim Teknolojileri Dergisi*, 13(3), 207-216.

79. Shafran, I., Rose, R. and Park, F. (2003). Robust speech detection and segmentation for real-time asr applications. United States of America: Izhak Shafran & Richard Rose AT&T Labs Research, 432-435.

80. Ochiai, T., Watanabe, S. and Katagiri, S. (2017). Does s*peech enhancement work with end-to-end ASR objectives?: Experimental analysis of multichannel end-to-end ASR.* IEEE International Workshop on Machine Learning for Signal Processing, 1-5.

81. Huang, X. and Deng, L. (2010). A*n Overview of Modern Speech Recognition, Handbook of Natural Language Processing*. United Kingdom: Chapman & Hall, 339-367.

82. Huang, C., Chang, E., Zhou, J. and Lee, K. (2000). A*ccent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition.* Annual Conference of the International Speech Communication Association, 818-821.

83. Winursito, A., Hidayat, R. and Bejo, A. (2018). *Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition.* International Conference on Information and Communications Technology, 379-383.

84. Divyesh, S., Mistry, A. V. and Kulkarni, P. (2013). Overview: speech recognition technology, mel- frequency cepstral coefficients (MFCC), artificial neural network (ANN). *International Journal of Engineerıng Research & Technology*, 2(10), 1994-2001.

85. Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. *International Journal for Advance Research in Engineering and Technology*, 1(6), 1-5.

86. Gupta, H. and Gupta, D. (2016). *LPC and LPCC method of feature extraction in speech recognition system.* International Conference - Cloud System and Big Data Engineering, Confluence, 498-502.

87. Haridas, A.V., Marimuthu, R. and Sivakumar, V.G. (2018). A critical review and analysis on techniques of speech recognition: The road ahead. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, 22(1), 39-57.

88. Hinton, G. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6), 82-97.

89. İnternet: Turkish broadcast news speech and transcripts-linguistic data consortium. Url: https://catalog.ldc.upenn.edu/LDC2012S06, Son Erişim Tarihi: 11.12.2019.

90. Sak, H., Saraçlar, M. and Güngör, T. (2012). Morpholexical and discriminative language models for Turkish automatic speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 20(8), 2341-2351.

91. Akin, A. A., Demir, C. and Dogan, M. U. (2012). *Improving sub-word language modeling for Turkish speech recognition*. Signal Processing and Communications Applications Conference, 1-4.

92. Asefisaray, B. (2018). *Uçtan uca konuşma tanıma modeli: Türkçe'deki deneyler,* Doktora Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 10-85.

93. Anusuya, M.A. and Katti, S.K. (2009). Speech recognition by machine: a review. *International Journal of Computer Science and Information Security*, 6(3), 181-205.

94. Dikici, E. and Saraçlar, M. (2016). Semi-supervised and unsupervised discriminative language model training for automatic speech recognition. *Speech Communication*, 83(1), 54-63.

95. Irie, K., Tüske, Z., Alkhouli, T., Schlüter, R. and Ney, H. (2016). *LSTM, GRU, highway and a bit of attention: An empirical overview for language modeling in speech recognition.* Annual Conference of the International Speech Communication Association, 3519-3523.

96. Dalmia, S., Li, X., Metze, F. and Alan W. B. (2018). *Domain robust feature extraction for rapid low resource ASR development.* PhD Thesis, Language Technologies Institute Carnegie Mellon University, 258-265.

97. Inaguma, H., Cho, J., Baskar, M. K., Kawahara, T. and Watanabe, S. (2019). Transfer *learning of language-independent end-to-end ASR with language model fusion.* IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 6096-6100.

98. Brown, P. F., Della Pietra, V. J., deSouza, P. V, Lai, J. C. and Mercer, R. L. (1990). {Class-based}{n-gram} models of natural language. *Computational Linguistics*, 18(1950), 14-18.

99. Sundermeyer, M., Ney, H. and Schluter, R. (2015). From feedforward to recurrent LSTM neural networks for language modeling. *IEEE Transactions on Audio, Speech and Language Processing*, 23(3), 517-529.

100. Karafi, M. and Cernock, J. H. (2010). *Recurrent neural network based language model.* Annual Conference of the International Speech Communication Association, 1045-1048.

101. Zhao, H., Lu, Z. and Poupart, P. (2015). *Self-adaptive hierarchical sentence model*. International Joint Conference on Artificial Intelligence, 4069-4076.

102. De Mulder, W., Bethard, S. and Moens, M. F. (2015). A survey on the application of recurrent neural networks to statistical language modeling. *Computer Speech and Language*, 30(1), 61-98.

103. Popova, I. A. and Stepanova, N. G. (1977). Estimation of inorganic phosphate in presence of phosphocarbohydrates (Russian). *Voprosy Meditsinskoj Khimii*, 23(1), 135-139.

104. Chien, J. T. and Ku, Y. C. (2014). *Bayesian recurrent neural network language model*. IEEE Workshop on Spoken Language Technology, 206-211.

105. Arisoy, E., Sethy, A., Ramabhadran, B. and Chen, S. (2015). *Bidirectional recurrent neural network language models for automatic speech recognition*. IEEE International Conference on Acoustics, Speech and Signal Processing**,** 5421-5425.

106. Akın, A. A. and Akın, M. D. (2007). Zemberek, an open source nlp framework for Turkic languages. *Structure*, 10(1), 1-5.

107. Bocchieri, E. and Caseiro, D. (2010). *Use of geographical meta-data in ASR language and acoustic models*. IEEE International Conference on Acoustics, Speech and Signal Processing, 5118-5121.

108. Hoffmeister, B., Heigold, G., Rybach, D., Schluter, R. and Ney, H. (2012). WFST enabled solutions to ASR problems: Beyond HMM decoding. *IEEE Transactions on Audio, Speech and Language Processing*, 20(2), 551-564.

109. Kang, S. S. (2008). Regulation of early steps of chondrogenesis in the developing limb. *Animal Cells and Systems*, 12(1), 1-9.

110. Grewal, J. K., Krzywinski, M. and Altman, N. (2019). Markov models - hidden Markov models. *Nature Methods*, 16(9), 795-796.

111. Geary, D. N., McLachlan, G. J. and Basford, K.E. (1989). Mixture models: ınference and applications to clustering. *Journal of the Royal Statistical Society,* 152(1), 126.

112. McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115-133.

113. Öztemel, E. (2012). *Yapay Zeka ve Makine Öğrenmesine Genel Bakış*. Türkiye: Papatya Yayıncılık, 32-50.

114. Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology,* 49(11), 1225-1231.

115. Došilovič, F. K., Brčić, M. and Hlupić, N. (2006). *Explainable artificial intelligence: A survey*. International Conference Information Communication Technology Electronic Microelectron, 0210-0215.

116. Beysolow, T. (2017). *Introduction to Deep Learning Using R Introduction to Deep*. United States of America: Springer, 50-125.

117. Lipton, Z. C., Berkowitz, J. and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv:1506.00019*, 1-38.

118. Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R. and Schmidhuber, J. (2017). LSTM: a search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10), 2222-2232.

119. Wrigley, S. N. and Hain, T. (2011). *Making an automatic speech recognition service freely available on the web*. Annual Conference of the International Speech Communication Association, 3325-3326.

120. Braun, S., Neil, D. and Liu, S. C. (2017). *A curriculum learning method for improved noise robustness in automatic speech recognition*. European Signal Processing Conference, 48-552.

121. Bourlard, H. and Morgan, N. (1994). *Connectionist Speech Recognition: A Hybrid Approach*. United States of America: Springer, 27-58.

122. Mahanty, R. and Mahanti, P. K. (2018). *Unleashing Artificial Intelligence onto Big Data.* Handbook of Research on Computational Intelligence Applications in Bioinformatics, 2099-2114.

123. Aggarwal, R. K. and Dave, M. (2011). Acoustic modeling problem for automatic speech recognition system: Advances and refinements (Part II). *International Journal of Speech Technology*, 14(4), 309-320.

124. Poulter, A. J., Johnston, S. J., and Cox, S.J. (2015). *Using the MEAN stack to implement a RESTful service for an Internet of Things application.* IEEE World Forum on Internet of Things, 280-285.

125. Povey, D. (2011). *The Kaldi speech recognition toolkit*. IEEE Workshop on Automatic Speech Recognition and Understanding, 1-4.

126. Erdoğan, H., Büyük, O. and Oflazer, K. (2005). *Incorporating language constraints in sub-word based speech recognition*. IEEE Automatic Speech Recognition and Understanding Workshop, 281-286.

127. Cartwright, R. (2010). Book reviews. *Perspectives in Public Health*, 130(5), 239-239.

128. Arisoy, E., Dutağaci, H. and Arslan, L. M. (2006). A unified language model for large vocabulary continuous speech recognition of Turkish. *Signal Processing*, 86(10), 2844-2862.

129. Leahy, R. L. (2008). A closer look at ACT. *Behavior Therapist*, 31(8), 148-150.

130. Zhang, S., Jiang, H., Xu, M., Hou, J. and Dai, L. (2015). *The fixed-size ordinally-forgetting encoding method for neural network language models*. Natural Language Processing of the Asian Federation of Natural Language Processing, 495-500.

131. Battenberg, E. (2017). *Exploring neural transducers for end-to-end speech recognition.* IEEE Automatic Speech Recognition and Understanding Workshop, 206-213.

132. Hochreiter, S. and Urgen Schmidhuber, J. (1997). Lstm. *Neural Computation*, 9(8), 1735-1780.

133. Cho, K. 2014). *Learning phrase representations using RNN encoder-decoder for statistical machine translation.* Conference on Empirical Methods in Natural Language Processing, 724-1734.

134. Jozefowicz, R., Zaremba, W. and Sutskever, I. (2015). *An empirical exploration of Recurrent Network architectures.* International Conference on Machine Learning, 2332-2340.

135. Stolcke, A. (2002). *Srilm - an extensible language modeling toolkit.* Annual Conference of the International Speech Communication Association, 901-904.

136. Seide, F. and Agarwal, A. (2016). *CNTK: Microsoft's open-source deep-learning toolkit.* International Conference on Knowledge Discovery and Data Mining, 135-2135.

137. Mohri, M., Pereira, F. and Riley, M. (2008). *Speech recognition with weighted finite-state transducers.* Springer Handbook on Speech Processing and Speech Communication, 559-584.

138. Siivola, V. and Hirsimaki, T. (2007). On growing and pruning kneser-ney smoothed - gram models. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5), 1617-1624.

139. Liu, X. (2017). M*ultimodal learning using 3d audio-visual data for audio-visual speech recognition.* EEE International Conference on Acoustics, Speech and Signal Processing, 40-43.

140. Ali, A., Nakov, P., Bell, P. and Renals, S. (2017). *WERd: using social text spelling variants for evaluating dialectal speech recognition.* IEEE Automatic Speech Recognition and Understanding Workshop, 141-148.

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı	: OYUCU, Saadin
Uyruğu	: T.C.
Doğum tarihi ve yeri	: 04.01.1988, Nizip
Medeni hali	: Evli
Telefon	: 0 (312) 202 85 84
e-mail	: saadinoyucu@gmail.com

Eğitim

Derece	Eğitim Birimi	Mezuniyet Tarihi
Doktora	Gazi Üniversitesi / Bilgisayar Mühendisliği	2020
Yüksek lisans	Gazi Üniversitesi / Bilgisayar Mühendisliği	2015
Lisans	Gazi Üniversitesi / Bilgisayar Mühendisliği	2018
Lisans	Gazi Üniversitesi / Bilgisayar Sistemleri Öğrt.	2012

İş Deneyimi

Yıl	Yer	Görev
2014-Halen	Gazi Üniversitesi	Araştırma Görevlisi
2012-2014	LST Yazılım A.Ş.	Yazılım Mühendisi

Yabancı Dil

İngilizce

Yayınlar

1. Polat, H. And Oyucu, S. (2020). Building a speech and text corpus of Turkish: large corpus collection with initial speech recognition results. *Symmetry*, 12(2), 1-19.
2. Polat, H. And Oyucu, S. (2017). Token-based authentication method for M2M platforms. *Turkish Journal of Electrical Engineering & Computer Sciences*, 2017(25), 2956-2967.

Hobiler

Gezi, yüzme, müzik

GAZİ GELECEKTİR..