T.C.
Radyo ve Televizyon Üst Kurulu

TÜRKÇEDE DERİN ÖĞRENME İLE KONUŞMA TANIMA UYGULAMALARI

UZMANLIK TEZİ

Recep Sinan ARSLAN

ANKARA
HAZİRAN/2018

T.C.
Radyo ve Televizyon Üst Kurulu

TÜRKÇEDE DERİN ÖĞRENME İLE KONUŞMA TANIMA UYGULAMALARI

UZMANLIK TEZİ

Recep Sinan ARSLAN

DANIŞMAN
Üst Kurul Uzmanı
Fatih ÖZTÜRK

ANKARA
HAZİRAN/2018

Radyo ve Televizyon Üst Kurulu Başkanlığına,

Bu çalışma, Tez Değerlendirme Komisyonu tarafından oy birliği/oy çokluğu ile Uzmanlık Tezi olarak kabul edilmiştir.

	Adı Soyadı	İmza
Başkan :		
Üye :		
Üye :		
Üye :		
Üye :		
Üye :		

ONAY

…./…./2018

Üst Kurul Başkanı

TEZ TESLİM TUTANAĞI VE DOĞRULUK BEYANI

Radyo ve Televizyon Üst Kurulunda görev yapan Uzman Yardımcısı olarak, Radyo ve Televizyon Üst Kurulu Uzman Yardımcılığı Giriş ve Yeterlik Sınavları ile Uzmanlığa Atanma, Yerleştirme, Görev, Yetki ve Çalışma Usul ve Esasları Hakkında Yönetmelik'in 19 uncu maddesinin 4 üncü fıkrasına istinaden çıkartılan Radyo ve Televizyon Üst Kurulu Tez Hazırlama Yönergesi'ne uygun olarak hazırlamış olduğum uzmanlık tezi ilişikte sunulmuştur.

Bu uzmanlık tezindeki bütün bilgileri akademik kurallara ve etik davranış ilkelerine uygun olarak toplayıp sunduğuma; ayrıca, bu kural ve ilkelerin gereği olarak, çalışmada bana ait olmayan tüm veri, düşünce ve sonuçları andığımı ve kaynağını gösterdiğimi beyan ederim.

Bilgilerinizi ve gereğini arz ederim.

…./…./2018

Recep Sinan ARSLAN
Üst Kurul Uzman Yardımcısı

Uzmanlık Tezinin Adı:
Türkçede Derin Öğrenme ile Konuşma Tanıma Uygulamaları

EK:
TEZ (6 adet)

Tezi Teslim Alan
Daire Başkanı
…./…./ 2018

TEŞEKKÜR

Bu çalışmanın gerçekleştirilmesi sürecindeki değerli katkılarından dolayı Üst Kurul Başkanı Sayın Prof. Dr. İlhan YERLİKAYA ve Üst Kurul Üyelerine, tez hazırlama sürecinde maddi ve manevi desteğini esirgemeyen Sevgili Eşim Gamze Arslan'a teşekkür ederim.

Recep Sinan ARSLAN

Ankara 2018

İÇİNDEKİLER

TEZ TESLİM TUTANAĞI VE DOĞRULUK BEYANI ..... **İİ**
TEŞEKKÜR ..... **İİİ**
İÇİNDEKİLER ..... **İV**
KISALTMALAR ..... **Vİİ**
ŞEKİLLER ..... **Vİİİ**
TABLOLAR ..... **İX**
GİRİŞ **1**
TEZİN KONUSU 3
TEZİN KAPSAM 3
TEZİN AMACI 3
TEZİN ÖNEMİ 4
TEZİN YÖNTEM 4
TEZİN SINIRLILIKLARI 4
VERİ TOPLAMA VE VERİ ANALİZ TEKNİKLERİ 5
BİRİNCİ BÖLÜM **7**
KONUŞMA TANIMA TEORİSİ VE SINIFLANDIRMASI **7**
1.1 KONUŞMA TANIMA TEORİSİ 7
1.2 KONUŞMA TANIMA SİSTEMLERİ SINIFLANDIRMASI 11
1.2.1 KONUŞMACI TANIMA 12
1.2.2 KONUŞMACI DİLİ TANIMA 13
1.2.3 KONUŞMA TANIMA SİSTEMLERİ SINIFLANDIRMASI 19
1.2.3.1 Yalıtık Sözcük Tanıma *20*
1.2.3.2 Bağlı Sözcük Tanıma *20*
1.2.3.3 Sürekli Konuşma Tanıma(SKT) *20*
1.2.3.4 Spontan Konuşma Tanıma *21*
1.2.4 KONUŞMACI MODU 21
1.2.4.1 Konuşmacı Bağımlı *22*
1.2.4.2 Konuşmacı Bağımsız *22*
1.2.5 SÖZLÜK BOYUTU 22
1.3 AKUSTİK ÖN UÇ 23
1.3.1 ÖZELLİK ÇIKARIMI 23
1.3.2 ÖZELLİK ÇIKARIM ALGORİTMALARI 26
1.3.2.1 Mel Frekans Kepstral Katsayıları(MFKK) *26*
1.3.2.2 Gerçek Kepstral Katsayıları *28*
1.3.2.3 Doğrusal Öngörülü Kodlama *29*
1.3.2.4 Türetilmiş Öngörülü Kodlama(TÖK) *29*
1.3.2.5 Algısal Doğrusal Öngörülü Kepstral Katsayıları(ADÖKK) *30*
1.3.2.6 Birincil Bileşen Analizi *31*
1.3.2.7 Doğrusal Ayrık Analiz *31*
1.3.2.8 Bağımsız Bileşen Analizi *31*
1.4 DİL MODELİ TASARIMI 32
1.4.1 KELİME TABANLI DİL MODELİ 32
1.4.2 KÖK-EKLER TABANLI DİL MODELİ 33
1.4.3 MORF TABANLI DİL MODELİ 33
1.4.4 HİBRİT DİL MODELİ 34

İKİNCİ BÖLÜM 35

KONUŞMA TANIMA SİSTEMLERİ YAKLAŞIMLARI VE TASARIM ALTYAPISI 35

2.1 KONUŞMA TANIMA YAKLAŞIMLARI 35
2.1.1 AKUSTİK FONETİK YAKLAŞIM 36
2.1.2 ÖRÜNTÜ TANIMA YAKLAŞIMI 37
2.1.2.1 Şablon Temelli Yaklaşım 38
2.1.2.2 Dinamik Zaman Bükmesi 40
2.1.2.3 Vektör Kuantasyonu(VK) 41
2.1.2.4 İstatistiksel Yaklaşım 41
2.1.3 YAPAY SİNİR AĞI(YSA)YAKLAŞIMI 42
2.1.4 KARAR DESTEK VEKTÖRLERİ(KDV) 44
2.1.5 YAPAY ZEKA YAKLAŞIMI 44
2.1.6 BİLGİYE DAYALI YAKLAŞIM 45
2.2 KONUŞMA VERİTABANLARI 46
2.2.1 Resource Management Complete Set 2.0 51
2.2.2 TIMIT ve Türevleri 51
2.2.3 TI46 52
2.2.4 SWITCHBOARD 53
2.2.5 ATIS 54
2.3 PERFORMANS ANALİZİ 54
2.4 SES TANIMA İŞLEMLERİNDE KULLANILAN ARAÇLAR 55
2.5 SES TANIMA UYGULAMA ÖRNEKLERİ 58
2.6 OTOMATİK KONUŞMA TANIMA SİSTEM TASARIMI ZORLUKLARI 60

ÜÇÜNCÜ BÖLÜM 64

DERİN SİNİR AĞI TEMELLİ SES TANIMA SİSTEMLERİ 64

3.1 DERİN SİNİR AĞI TÜRLERİ 65
3.1.1 Konvolüsyon Sinir Ağları 66
3.1.2 Tekrarlayan Sinir Ağları(TSA) 68
3.1.3 Kısıtlı Boltzman Makinesi(KBM) 69
3.1.4 Derin İnanç Ağları 71
3.1.5 Derin Konveks Ağlar(DKA) 72
3.2 ÖĞRENME SÜRECİ 74
3.2.1 Denetimli Öğrenme 74
3.2.2 Denetimsiz Öğrenme 75
3.2.3 Takviyeli Öğrenme 75
3.3 AKTİVASYON FONKSİYONLARI 76
3.3.1 ReLU 76
3.3.2 Sigmoid 77
3.3.3 Tanh 78
3.3.4 Adım(Step) 79
3.3.5 Gauss 79
3.4 HATA FONKSİYONLARI 80
3.4.1 Ortalama Hata Karesi 80
3.4.2 Ortalama Mutlak Hata(Mean Absolute Error) 81
3.4.3 Karekök Ortalama(Root Mean Square) 82
3.4.4 Dereceli Azalma(Gradient Descent) 82
3.5 DERİN ÖĞRENME SİSTEMLERİ SINIFLANDIRMASI 83
3.5.1 Üretken Öğrenme Yaklaşımı(SMM-GMM) 83
3.5.2 Ayrık Öğrenme Yaklaşımı(HMM-ANN) 84
3.5.3 Derin Öğrenme Yaklaşımı(HMM-DNN) 85
3.6 DERİN ÖĞRENME KÜTÜPHANELERİ 86

3.6.1 TensorFlow ... *87*
3.6.2 Theano ... *88*
3.6.3 Torch ... *89*
3.6.4 Caffe ... *90*
3.6.5 Keras ... *90*
3.6.6 Anakonda ... *91*
3.7 DERİN ÖĞRENME TEMELLİ SÜREKLİ KONUŞMA TANIMA ÜRÜNLERİ KARŞILAŞTIRMASI ... 92
3.7.1 Bing Konuşma Tanıma Uygulaması ... *92*
3.7.2 Switchboard Konuşma Tanıma Uygulaması ... *93*
3.7.3 Google Konuşma Tanıma Uygulaması ... *95*
3.7.4 Youtube Konuşma Tanıma Uygulaması ... *96*
3.7.5 İngilizce Yayın Haber Konuşma Tanıma Uygulaması ... *96*

DÖRDÜNCÜ BÖLÜM ... 98

SES TANIMA SİSTEMLERİNE İLİŞKİN TARİHSEL GELİŞİM VE TÜRKÇE KONUŞMA TANIMA LİTERATÜR ÖZETİ ... 98

4.1 1920-1959 ... 98
4.2 1960-1970 ... 99
4.3 1970-1980 ... 100
4.4 1980-1990 ... 101
4.5 1990-2000 ... 104
4.6 2000-2009 ... 108
4.7 TÜRKÇE KONUŞMA TANIMA SİSTEMLERİ LİTERATÜR ÖZETİ ... 113

BEŞİNCİ BÖLÜM ... 118

SONUÇ VE ÖNERİLER ... 118

KAYNAKÇA ... 127

ÖZET ... 134

ABSTRACT ... 135

ÖZGEÇMİŞ ... 136

EK-1 ... 137

KISALTMALAR

OKT	: Otomatik Konuşma Tanıma
ODT	: Otomatik Dil Tanımlama
SKT	: Sürekli Konuşma Tanıma
DÖK	: Doğrusal Öngörülü Kodlama
RTÜK	: Radyo ve Televizyon Üst Kurulu
s.	: Sayfa
MFKK	: Mel Frekans Kepstral Katsayıları
HFD	: Hızlı Fourier Dönüşümü
Hz	: Hertz
AFD	: Ayrık Fourier Dönüşümü
AKD	: Ayrık Kosinüs Dönüşümü
TÖK	: Türetilmiş Öngörülü Kodlama
ADÖKK	: Algısal Doğrusal Öngörülü Kepstral Katsayıları
LD	: Levinson-Durbin
VK	: Vektör Kuantasyonu
SMM	: Saklı Markov Model
YSA	: Yapay Sinir Ağı
KDV	: Karar Destek Vektörleri
LDC	: Linguistic Data Consortium
HTK	: Hidden Markov Model Toolkit
SRILM	: SRI Dil Modelleme Aracı
KSA	: Konvolüsyon Sinir Ağı
TSA	: Tekrarlayan Sinir Ağı
DKA	: Derin Konveks Ağ
DSA	: Derin Sinir Ağları

ŞEKİLLER

Sayfa

Şekil 1 Konuşma Tanıma Sistemi Genel Yapısı 8
Şekil 2 Konuşma İşleme Sistemleri Sınıflandırma Şeması 11
Şekil 3 Konuşmacı Tanıma Sınıflandırma Şeması 13
Şekil 4 Bir gerçek zamanlı konuşma tanıma için dil tanımlama sistemi örneğidir. Çıktı olarak en olabilir 3 dili belirlemektedir. Konuşma tanıma sistemi belirlenmiş 3 dil için modelleri sisteme yükler ve ses işlemesi sonrasında hangi dil olduğuna karar verilir 15
Şekil 5 Çok dilli olarak hizmet veren viren asistanlık hizmetinin veya acil yardım sisteminin ön uç yapısıdır. Gelen çağrının hangi dilde olduğunu tespit ederek ilgili çağrı operatörüne bildirir 16
Şekil 6 Dil Tanımlamanın iki safhası gösterilmektedir. Eğitim sırasında konuşma dalga formları analiz edilir. Dil bağımlı modeller üretilir. Tanıma aşamasında, yeni bir ses işlenir ve eğitim sırasında üretilen modeller ile karşılaştırılır[13]. 18
Şekil 7 MFKK Özellik Çıkarımı Adımları 27
Şekil 8 Kelime Tabanlı Dil Modeli Yapısı 32
Şekil 9 Örüntü Tanıma Yaklaşımı Akış Şeması 38
Şekil 10 Şablon Temelli Yaklaşım 39
Şekil 11 Konvolüsyonel Sinir Ağı Şeması 67
Şekil 12 Kısıtlı Boltzman Makinesi Modeli 70
Şekil 13 Derin Konveks Ağ Yapısı 73
Şekil 14 ReLU Grafiği 76
Şekil 15 Sigmoid Grafiği 77
Şekil 16 tanh Grafiği 78
Şekil 17 Adım Fonsiyonu Grafiği 79
Şekil 18 Gauss Grafiği 80
Şekil 19 Üç Kütüphanenin Test Sonuçları 89

TABLOLAR

Sayfa

Tablo 1 Özellik Çıkarımı için Kullanılabilecek Metodlar ve Bazı Özellikleri 24
Tablo 2 Konuşma Tanıma Yaklaşımları Genel Tablosu..... 35
Tablo 3 İngilizce Dili ile Hazırlanmış Konuşma Veri Tabanları..... 49
Tablo 4 Türkçe Dili için Hazırlanmış Akustik ve Dil Modeli Veritabanları 50
Tablo 5 Ayrık Eğitim yöntemi ile eğitilmiş iki güçlü GMM-HMM tabanlı sistem ile 5 farklı Derin inanç ağı-derin sinir ağı akustik modelinin karşılaştırması. Konuşmacı bağımsız 309 saatlik eğitim verisi ve tek geçişli kod çözme tekniği ile 2000 saatlik konuşmacı 94
Tablo 6 Geniş Sözcük Dağarcığına sahip görevler için DSA-SMM ile GMM-SMM modellerinin hata oranlarının karşılaştırılması..... 97
Tablo 7 Türkçe Dili ile İlgili Çalışmaların Tür ve Yıllara göre Dağılımı..... 114
Tablo 8 Test Ortamı Topolojisi..... 121
Tablo 9 50 Adet Deney Ortamı Performans Değerlendirmesi..... 122
Tablo 10 Gizli Katman Sayısının Performansa Etkisi 124
Tablo 11 Epoch sayısının Performansa Etkisi 125

GİRİŞ

Konuşma, akciğerlerden gelen havanın boğaz, ağız ve burun yoluyla biçimlenerek insanların duygu ve düşüncelerini ifade etmesi ve bunun insanların duyabileceği şekilde seslendirilmesidir. İnsanların sahip olduğu konuşma yapısı bilim dünyasında ciddi olarak araştırılmakta olup, bu alanda hızlı bir yol kat edilmiştir[1].

Ses işleme sistemleri içerisinde bir tür olan konuşma tanıma alanı, gelişen teknolojik gelişmeler ile üzerinde ciddi araştırmaların yapıldığı bir alan haline gelmiştir. Bu sistemlerde amaç, insan sesinin bir mikrofon vasıtası ile alınarak bilgisayarlara iletilmesi ve bilgisayarların bu sinyali algılayarak tanımasının sağlanmasıdır. Bu sayede, insanların klavye benzeri bilgisayar girdi aletlerini kullanarak bilgisayarlara hükmedebilmesi yerinde daha hızlı ve kolay bir şekilde ses ile bu komutların verilmesi mümkün olabilmektedir. Çeşitli teknoloji firmalarının ses ile bilgisayarların yönetilmesi sürecini devrim olarak nitelendirmesi sebebiyle, 1950'li yıllardan itibaren gelişen ve üzerinde çalışma yapılan bir alan olmuştur.

Günümüze kadar konuşma tanıma sistemleri üzerine ciddi araştırmalar yapılmış olsa da, henüz yeterli düzeyde gelişme kaydedilememiştir. Uzun süredir yapılan çalışmalar insanları umutlandırmakla birlikte, bu teknolojik devriminin gerçekleştirilmesi üzerine çalışmak oldukça ilgi çekicidir.

Konuşma tanıma sistemlerinin ticari olarak ciddi getirilere sahip olabilecek potansiyele sahip olması da bu alanda teknoloji şirketlerinin çalışmasında önemli bir noktadır. Bu sebeple, günümüzde işitme engellilerin bilgisayar ile iletişimde

[1] Yakar, Özlem ve Aşlıyan, Rıfat (2016), "*Saklı Markov Modeli Kullanarak Türkçe Konuşma Tanıma",* Adnan Menderes Üniversitesi, Yüksek Lisans Tezi.

bulunmaları, ses sinyallarinin metne dönüştürülmesi, havaalanı operasyonları, robotlar, mekanik araçların uzaktan yönetilmesi, zeki telefon sistemleri gibi birçok alanda ürünler geliştirilmekte ve insanların kullanımına sunulmaktadır. Tanıma işleminin bilgisayar tarafından yapılabilmesi için ciddi tezlerin ortaya konulması ve bunların matematiksel, istatistiksel veya psikolojik yapılarak çözümlenmesi gereklidir. Bu süreçler, üniversite veya araştırma geliştirme faaliyeti yürüten özel şirketler bünyesinde oluşturulan ekipler ile yönetilmektedir.

İnsanların biyolojik yapılarının oldukça değişken olması sebebiyle çıkardıkları sesin özellikleri birbirinden oldukça farklıdır. Farklı tonda veya vurguda sesler insanlara ait bir takım özelliklerin ortaya çıkarılmasında kullanılabilmektedir. Bu farklılık insanlar arasından ciddi bir ayırt edici unsur olmaktadır ve ses tanıma sistemlerinde kullanılabilir bir veri olarak değerlendirilmektedir.

Ses işleme üzerine yapılan çalışmaların bir türü olan otomatik konuşma tanıma(OKT) sistemleri, konuşma sinyallerinin işlenerek kelimeye dönüştürülmesi sürecinin bilgisayarlar tarafından yapılması işlemidir. Konuşma tanıma teknolojileri sayesinde bilgisayarlar sözlü komutları takip edebilmekte, dilini anlayabilir duruma gelebilmektedir. Konuşma tanıma alanında yapılan geliştirmelerin ana amacı makinelere sesli girdiler verilebilir ortamlar yaratabilmektedir. Konuşmanın iletişimde birincil yol olması sebebiyle, konuşma kabiliyetine sahip veya konuşmanın tanınması, anlaşılması işlemlerinin bilgisayar tarafından yapılabilir hale getirilmesi üzerine son 60 yıldır ciddi çalışmalar yapılmaktadır[2]. Konuşma işaretlerinin istatiksel modellenmesi temelinde otomatik konuşma tanıma sistemleri çok geniş bir alanda kullanılır hale gelmiştir.

[2] Sadaoki, Furui (2005), "50 years of Progress in Speech and Speaker Recognition Research", *ECTI Transactions on Computer and Information Technology*, s. 1-2.

TEZİN KONUSU

Bu tez çalışmasında, Ses sinyallerinin işlenerek otomatik olarak metne dönüşüm işleminin gerçekleştirilmesi için bir yazılım tasarımı yapılması konusu ele alınmıştır. Yazılım Türkçe dili için çalışabilir özellikte olacaktır ve sonuçta belirli özellikte seslerin metne dönüşümü yazılım tarafından gerçekleştirilecektir.

TEZİN KAPSAM

Türkçe dili için belirli problemlere çözüm üretebilir bir ses tanıma sisteminin tasarlanması ve yazılımının gerçekleştirilmesi tezin kapsamı dahilindedir. Tez süreci boyunca yapılacak çalışmalar ile her türlü sesi metne dönüştürecek bir mimarinin geliştirilmesinin mümkün olmaması sebebiyle, problem belli özelliklere sahip olmaları durumlarında ele alınmış olup, bu özelliklere sahip olanlar ses tanıma sürecine dahil edilmiştir. 10 adet sınıflandırılmış cümle ve 100 adet sınıflandırımış cümle ile çalışmalar gerçekleştirilmiştir. Kısıtlı bir veriseti ile çalışabilir sistem geliştirilmiştir. Böylece sınıflandırma performansında belirli bir düzeyin üzerinde kalınmıştır.

TEZİN AMACI

Bu tez çalışmasında amaç, ses tanıma uygulaması geliştirmek için ihtiyaç duyulacak bilgilerin edinilmesi ve bu bilgiler ışığında bir modellemenin yapılmasıdır. Bu modelleme Türkçe dili için çalışabilir bir sistem olacaktır ve var olan RTÜK sistemlerinde entegre edilebilir yapıda geliştirilecektir. Böylece, sesin metne dönüştürülmesi ile ilgili yapılan çalışmaların daha hızlı ve insan gücü olmaksızın gerçekleştirilmesi mümkün olacaktır. SKAAS sistemlerine ile ilişkilendirilebilir bir yapı olması halinde kurumumuzda kullanılabilir bir yazılım modülü gerçekleştirilmiş olacaktır.

TEZİN ÖNEMİ

Kurumumuz bünyesinde gerek tv yayınları gerekse radyo yayınlarının sesten metne dönüştürülmesi süreçleri bilgisayarlar tarafından gerçekleştirilmemektedir. Bu tez çalışmasının önemi, bu işlemlerin bilgisayarlar tarafından hızlı ve doğru bir şekilde gerçekleştirilmesi ve böylece kurumumuz personellerinin bu işlemler yerine sonuçların değerlendirilmesi süreçlerinde daha fazla çalışmalarını sağlayabilmektir. Tam ve her durum için çalışabilir bir sistem ortaya çıkması halinde tüm deşifre işlemleri artık bilgisayar sistemleri tarafından gerçekleştirilebilecektir.

TEZİN YÖNTEM

Bu çalışmanın gerçekleştirilmesi sürecinde, öncelikle bu konuda yapılmış olan daha önceki çalışmaların incelenmiştir. Yapılması planlanan modele benzer makalelerin gözden geçirilmesi ile konu ile ilgili bilgi edinilmesi sağlanmıştır. Sonrasında bu bilgiler ışığında ihtiyaç duyulan yazılım teknolojileri, verisetleri ve örnek yapay zeka modelleri incelenerek en iyi sistemler tespit edilmiştir. Örnek bir model geliştirilerek sunucu üzerinde test edilmiş ve sonuçlar değerlendirilerek test ve gözlem döngüsü devam ettirilmiştir. Tez süresinde ortaya çıkan değerler incelenerek sonuçları bu çalışma içerisinde ortaya konulmuştur. Bundan sonra, ortaya çıkan model geliştirilerek daha iyi tanıma performansı sunabilecek modellemeler yapılabilecektir.

TEZİN SINIRLILIKLARI

Bu çalışmanın geliştirme sürecinde ses tanıma sistemi, derin öğrenme mimarisinde tasarlanmıştır. Derin öğrenme mimarisinin doğası gereği belirli kısıtları bulunmakta olup, yazılım geliştirme süreçleri de bu durumdan etkilenmiştir. Veri setinin sınırlı ve erişiminin zor olması diğer bir problemdir. Türkçe dili için

hazırlanmış ses setleri ücretsiz olarak sunulmamakta olup, yüksek ücretler ile dağıtımı yapılmaktadır. Veri setinin elde edilmesi sonrasında, veri seti modellenen sinir ağında eğitilmesi sürecinde yüksek performansa sahip sunuculara ihtiyaç duyulmaktadır. Aksi halde öğrenme süreci oldukça uzun sürmekte olup, sonuçların elde edilmesi için 1-2 haftalık sürelerde beklemek gerekmektedir. Son olarak, verisetinin elde edilmesi, modelin eğitilmesi sonucunda çıktıların sınıflandırma performanslarının düşük kalması sebebiyle modellerde güncelleme yapılarak bu işlemlerin tekrarlanması oldukça zordur. Bahse konu bu sınırlılıklar hem Türkçe dili ile çalışılması hem de Derin öğrenme yapılarıının genel kısıtlarını içermektedir.

VERİ TOPLAMA VE VERİ ANALİZ TEKNİKLERİ

Tez çalışmasının ilk aşamasında konu ile ilgili bilgi edinilmesi, benzer çalışmaların incelenebilmesi için makale araştırması sürecinde, IEEE XPlore, ScienceDirect gibi indekslerden yararlanılmıştır. Bu indeksler konuya göre filtrelenerek detaylı olarak taranmış ve benzer makaleler bulunarak incelenmiştir. Sonrasında, yazılım geliştirme aşamalarının nasıl olacağına dair bilgi edinmek üzere Google TensorFlow kütüphane sayfası detaylı olarak incelenmiş olup, örnek kodlar Github üzerinden indirelerek çalıştırılmıştır. Bu işlemler sonrasında yeni bir model geliştirilmesi için elde edilen veriler incelenmiş, örnek çalışmalar ışığında yeni bir yazılım geliştirilmiştir. Bu yazılım TUBİTAK TRUBA sunucularında test edilmiş ve sonuçlar bu tez çalışmasında gösterilmiştir.

Bu tez çalışmasının birinci bölümünde konuşma tanıma teorisi ve sınıflandırmasından detaylı olarak bahsedilmiştir. İkinci bölümde konuşma tanıma sistemlerine çözüm üretmek için kullanılabilecek yaklaşımlar ve tasarım yapısı anlatılmıştır. Böylece konuşma tanıma sistemlerinin genel çerçevesinin çizilmesi amaçlanmıştır. Çalışmanın üçüncü bölümünde bir yapay sinir ağı türü olan derin sinir ağları açıklanmış olup, dördüncü bölümde konu ile ilgili literatür çalışmaları yıl bazlı

olarak özetlenmiştir. Son olarak beşinci bölümde, yapılması planlanan modelin detayları, deneyler ve sonuçlar açıklanarak, RTÜK sistemlerine uygulanabilirliği noktasında yapılması gereken çalışmalar anlatılmıştır.

BİRİNCİ BÖLÜM

KONUŞMA TANIMA TEORİSİ VE SINIFLANDIRMASI

1.1 KONUŞMA TANIMA TEORİSİ

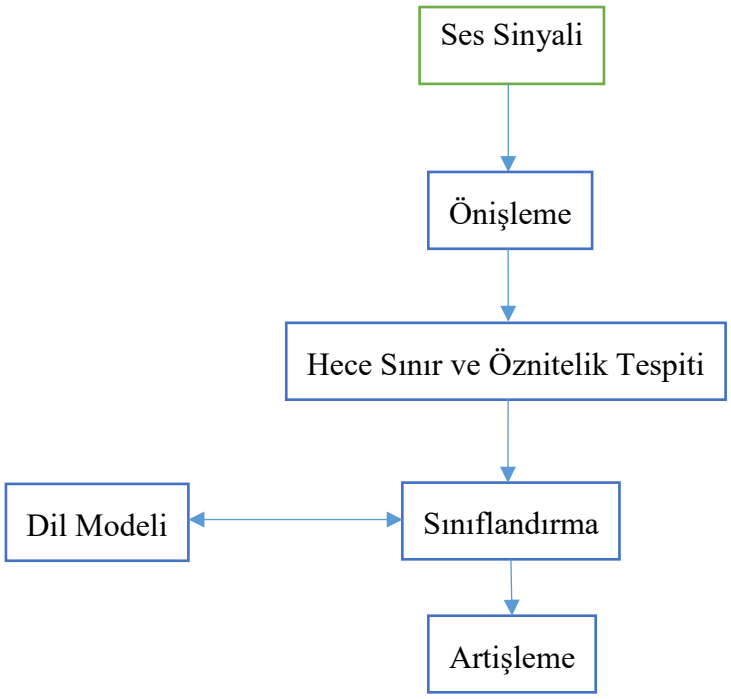
Otomatik konuşma tanıma ses sinyallerinin kelime dizilerine çevrilmeleri ve bu işlemin bazı algoritmik yaklaşımlar kullanan bilgisayar programları kullanılması halidir.

Konuşma işleme araştırmalarında insanların iletişim kabiliyetlerini taklit eden mekanik modeller yaratılması düşünülmektedir. Ses insanların birbirleriyle iletişimlerinde ve karşılıklı bilgi alışverişlerinde kullandıkları en doğal yoldur ve bu durum sinyal işleme sistemlerinin önemli araştırma alanlarından biridir. Konuşma tanıma teknolojileri, bilgisayarların insanların sesli komutlarını takip etme veya konuşmalarını anlama imkanı tanır. Ses tanıma araştırmalarının ana amacı bilgisayarlara girdi olarak ses sinyallerini verebilmektir. İnsanların konuşma yeteneklerinin taklit edilmesine yönelik, insan makine etkileşimi ile basit görevleri otomatikleştirmek veya otomatik konuşma tanıma sistemleri ile seslerin metne dönüştürülmesine kadar birçok alanda araştırma yapılmaktadır[3].

[3] Tran, Dat Tat (2000), *Fuzzy Approaches to Speech and Speaker Recognition,* Canberra: Canberra Üniversitesi, Doktora Tezi.

Sesin istatistiksel modellemesindeki gelişmeler ile birlikte, otomatik konuşma tanıma sistemleri çağrı merkezleri ses işlemesi, insan-makine arayüzü gerektiren görevler, seyahat bilgileri, hisse senedi işlemleri, fiyat teklifleri, hava durumu raporları, veri girişi, sesli dikte, bilgilere erişim, gibi birçok farklı alanda kullanılmaya başlanmıştır. OKT sistemlerine ilişkin olarak son 60 yılda ciddi ilerleme kaydedilmiş olmakla birlikte hala çözülmesi gereken birçok problem ve geliştirilmesi gereken çok konu bulunmaktadır[4].

Konuşma tanıma sistemlerinin modellenmesinde aşağıdaki gibi bir akış izlenmektedir. Sinyal olarak verilen girdinin ön işlem, modelleme, dil modeli oluşturma ve tanıma bölümlerini içeren bir şema olarak gösterilmiştir.



Şekil 1 Konuşma Tanıma Sistemi Genel Yapısı

Ses sinyalinin işlenerek anlamlandırılabilmesi için Şekil-1'deki bir akış içerisinde çeşitli işlemlere tabi tutulması gerekmektedir. Giriş sinyali sisteme

[4] Anusuya, M.A.ve Katti, S.K. (2009), "Speech Recogniton by Machine: A Review", *International Journal of Computer Science and Information Security*, s. 181-205.

mikrofon vasıtası ile alındıktan sonra sesin örneklemlerinin çıkarılması, sayısallaştırılması, gerekli hallerde çeşitli filtrelerden geçirilmesi, etiketlenmesi ve modellenebilir hale getirilmesi işlemi önişleme modülünde gerçekleştirilmektedir. Burada amaç, sesin işlenerek daha yalın, konuşma değişkenliklerinden ve gürültüden arındırılmış bir halini üretebilmektir. Bir sonraki adımda elde edilen örneklenmiş ses sinyalinin parametrelerinn yakalanır ve sesin belirli zaman aralıklarında sabit özelliklerinin çıkarılması için gerekli hesaplamalar yapılır. Burada elde edilen veriler tanıma için kritik öneme sahip değerlerdir ve önemli ipuçları sağlarlar. Bir sonrası adımda parametlerden tahmini yapılan kelimenin dil modellemesi ile elde edilen yapı ile karşılaştırılması işlemi yapılır. Böylece daha önce sözcük kütüğü yardımıyla dil modelleri oluşturulmuş metin verisi içerisinde akustik modele en uygun seçim yapılarak, tanındığı düşünülen kelime ortaya çıkmış olur. Bu işlemlerin son aşaması da artişleme safhasıdır. Bu adımda hangi dil ile çalışma yapılıyor ise o dile özgü gramer bilgileri veya çeşitli temel dil kuralları ile ortaya çıkan kelimenin veya cümlenin değerlendirilme işlemi yapılır. Dile ait yapılar kullanarak daha iyi bir tanıma yapılması amaçlanır. Ses tanıma teknolojilerinin günümüzde kullanılan en temel akışı bu şekildedir. Farklılıklar akustik ve dil modeli yaratılması için kullanılan yaklaşımlardan kaynaklanmaktadır[4].

Otomatik konuşma tanıma teorisi, verilen bir akustik X dizisi için, W kelime dizilerini bulmayı amaçlamaktadır.

Konuşma cümleleri W = (w1, w2, ..) şeklinde belirtilen kelimelerin dizisi olarak gösterilir. W_t ayrık bir t zamanında söylenmiş belirli bir kelimedir. Kelimelerin dizisi söylenen sesli ifade ile bağlantılıdır ve bu sesli ifade X olarak gösterilen akustik seslerin dizisidir[5].

Geniş sözcük dağarcıklı sürekli konuşma tanıma için standart yaklaşıma göre (1)'de verilmiş olan formül kullanılır.

[5] Yalçın, Nursal (2008), "Konuşma tanıma teorisi ve teknikleri", *Kastamonu eğitim dergisi*, s. 249-266.

$$P(W/A) = \arg\max P(W/A) \quad (1)$$

Buna göre, olası bir P(W|A) olasılıklı A sözcük dizisinin W bir akustik gözlem dizisi ürettiği varsayılmaktadır. Sonrasında, akustik gözlem sırasına bağlı olarak söylenmiş olan kelime dizisinin çözülmesi ve maximum olasılıklı dizinin tanınmış sözcük olarak ortaya konulması yapılmaktadır.

Bayes kuralına göre 1 numaralı formül aşağıdaki şekilde düzenlenmektedir.

$$P(W|A) = P(A/W)P(W) / P(A) \quad (2)$$

P(A), W'den bağımsız olarak alınırsa, Maximum olasılık dağılımı tahminine göre (2) numaralı formül çözümlenecek olursa, sonuçta

$$W = \arg\max P(A/W)P(W) \quad (3)$$

formülü elde edilir.

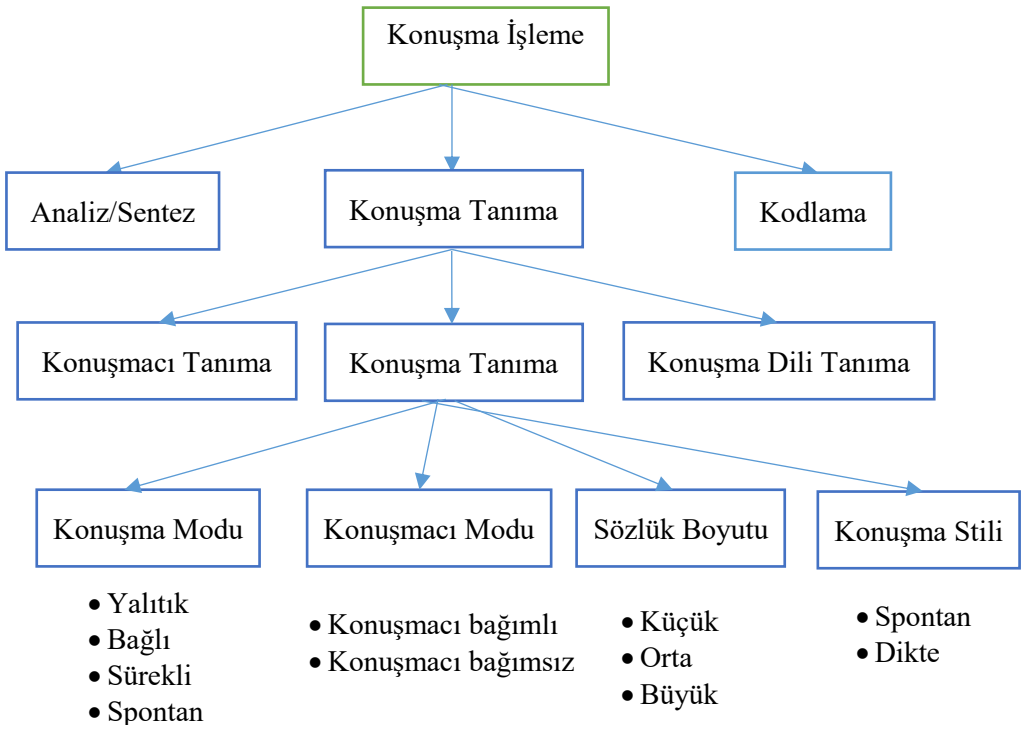
Ses analizinin teorik altyapısını açıklamada kullanılan en yaygın matematik modelidir. Bu formül incelendiğinde,

P(A/W) olarak tanımlanan denklemin ilk modeli, akustik model olarak isimlendirilir. Sözcük dizesine göre düzenlenmiş akustik model dizisinin olasılığını ifade eder. Büyük ölçekli konuşma tanıma sistemleri için, alt sözcük konuşma birimleri ile istatistiksel modeller oluşturmak, bu alt birimlerin bir araya getirilmesi ile kelimelerin oluşturulması ve kelimeleri bir araya getirerek istenilen sözcük dizilerinin oluşturulması akustik model olasılıklarının değerlendirilmesi ile yapılmaktadır. P(W) olarak tanımlanan ikincil terim dil modeli olarak isimlendirilir. Sözlü ifadeler dizisiyle

ilgili olasılığı tanımlar. Bu tür dil modelleri dilin sözdizimsel ve semantik kısıtlamalarını içerebilir[6].

1.2 KONUŞMA TANIMA SİSTEMLERİ SINIFLANDIRMASI

Konuşma tanıma teorisinin temeli oluşturan konuşma işleme uygulamalarının sınıflandırma şeması Şekil-2'de verilmiştir.



Şekil 2 Konuşma İşleme Sistemleri Sınıflandırma Şeması

Ses işleme araştırmaları içerinde bulunan konuşma işleme sistemlerinde konuşma tanıma bölümü ciddi bir yer kaplamaktadır. Konuşma tanıma bölümü, bu alandaki en zor alanlardan birisidir. Bunun yanında ses üretimi ve analizi ile ses kodlaması üzerine yapılan çalışmalarda bulunmaktadır.

[6] Anusuya, M.A. ve Katti, S.K., 2009, s. 181-205.

Konuşma sinyalinin işlenmesi ile ilgili yapılacak çalışmalarda bu sınıflandırmaya uygun bir sınıf seçilerek araştırma geliştirme çalışmalarının yapılması beklenmektedir. Çünkü bu sınıflandırma içerisindeki her bir türün kendine özel problemleri ve çözüm yöntemleri bulunmaktadır. Böylece her bir araştırmada bir alan için iyi sonuçlar elde edilmesi mümkün olacaktır.

Bu bölümde, yukarıda verilmiş olan sınıflandırmaya uygun olarak her bir tür için açıklamalar verilmiştir.

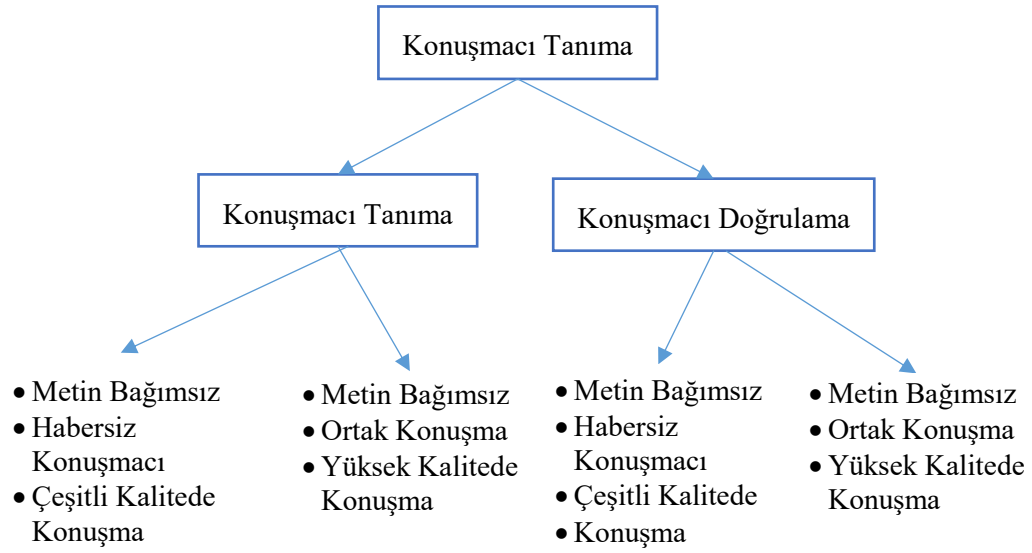
1.2.1 Konuşmacı Tanıma

Otomatik konuşmacı tanıma, kişinin kimliğini sesiyle doğrulamak için bilgisayarların kullanılmasıdır. Bu alanla ilişkili olarak ses doğrulama, konuşmacı kimlik doğrulama, ses kimlik doğrulaması veya konuşmacı doğrulama gibi farklı tabirler kullanılmaktadır. Otomatik konuşmacı tanımlamada, tanıma öncesinde herhangi bir kimlik iddiası bulunulmaz ve sistem kişinin kim olduğunu, kişinin hangi gruba üye olduğunu veya kişinin daha önce tanınan biri olup olmadığına karar verir.

Konuşmacı doğrulaması, bir sesin kime ait olduğunu karar vermek olarak tanımlanır. Bu bir sesin belirli bir kişi olup olmadığına karar veren veya bir grup insan arasında yer aldığına karar vermeye göre farklı ve zor bir problem olarak karşıya çıkmaktadır. Konuşmacı doğrulamasında, bir kimlik iddiası yapılır. Metne bağımlı tanımada, sabit girdi sistem tarafından bilinir ve sabitlenebilir. Kişi mikrofona sabit metni okur ve sinyal kullanıcıların kimlik iddiasını kabul veya reddetmek veya karar vermeden önce yetersiz veri olduğuna karar verilir. Daha önce sisteme kaydedilmiş olan ses ile tanınacak ses arasında bir karşılaştırma ile kimlik doğrulaması yapılmaya çalışılır. Sesin mikrofon ile alınması sırasında ortam gürültüsü, sesin farklı versiyonları, mikrofonun yansıtıcı özellikleri de dikkate alınması gerekmektedir. Bu

farklı durumları temsil eden ortamlar yaratılır ve öncesinde sisteme aktarılır ve ses modelleri oluşturulur. Bu şekilde model sayısındaki artış beraberinde test süresini de artırmaktadır. Bu tip insan hatalarının değerlendirildiği sistemler yüksek doğrulukla performans sunmaktadırlar[7].

Konuşmacı tanıma sistemlerinin sınıflandırması aşağıdaki şemada (Şekil-3) gösterilmiştir.



Şekil 3 Konuşmacı Tanıma Sınıflandırma Şeması

1.2.2 Konuşmacı Dili Tanıma

Otomatik dil tanımlama(ODT), belirli bir konuşmacının konuştuğu dilin otomatik olarak tanımlanması süreci olarak tanımlanır[8]. ODT sistemleri günlük çok dil kullanılan çeviri sistemlerinde veya acil arama yönlendirmesi yapan yerel operatörlerini müdahale süresinin kritik olabileceği uygulamalarda kullanılabilir.

[7] Campbell, Joseph P. (1997), "Speaker recognition: A tutorial", *Proceeedings of the IEEE*, s. 1437-1462.

[8] Y.K., Muthusamy, E., Barnard ve R.A., Cole (1994), "Reviewing automatic language identification", *IEEE Signal Processing Maganize*, s. 33-41.

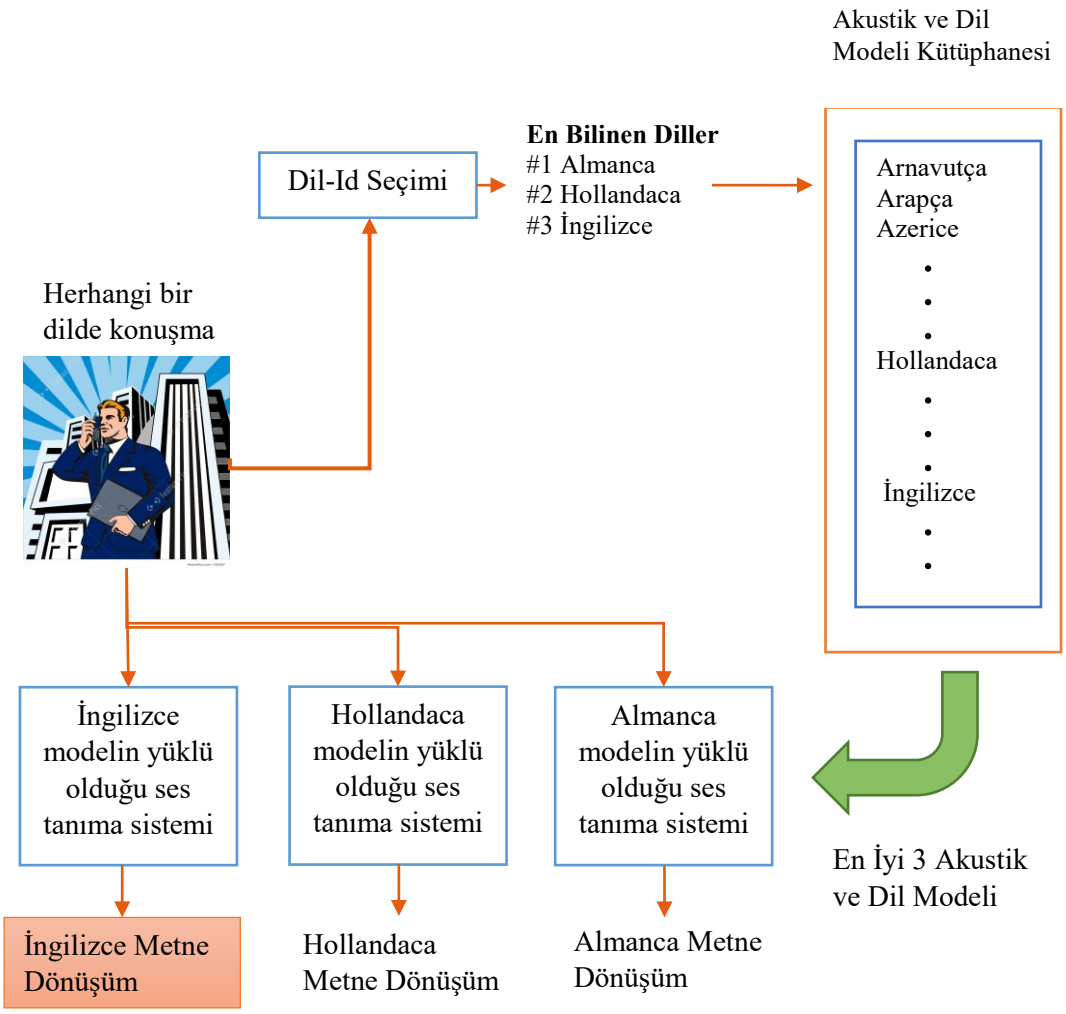
Günümüzdeki birçok konuşmacı dili tanımlama sistemleri üst düzey yaklaşımlarla karar verme yoluna gitse de, akustik modellemeye dayalı sistemler hala iyi sonuç vermektedirler ve güvenilen sistemler olarak kalmaktadırlar[9].

Ses doğrulamasındaki gelişmeler, vektör çıkarıcıların kullanıldığı ve çeşitli sınıflandırıcılarla karar verme mekanizmasına sahip modeller ortaya çıkarmıştır. Akustik düzeyde, sığ mimariler veya katmanlı sinir ağları kullanılmakla birlikte günümüzde araştırmacılar daha çok sinir ağı tasarımları ile modelleme yapma yoluna gitmektedir ve eğer yeterli eğitim verisi ile çalışılıyor ise başarılı sonuçlar ürettiğini kanıtlamışlardır [10].

Dil kimliği belirleme uygulamaları iki ana kategoriye ayrılır. Makine sistemleri için önişleme ve insan dinleyen sistemler için ön işlemedir.

[9] Torres, Pedro, Terry, Gleason, Alan, McCree, Douglas, Reynolds, Elliot, Singer ve Douglas, Sturim Fred Richardson (2010), "The MITLL NEST LRE 2009 Language Recognition System", *International Conference on Acoustics, Speech and Signal Processing: IEEE*, Dallas.

[10] Javier Gonzales, Domingues, Ignacio Lopez, Moreno, Javier Franco, Pedroso ve Daniel Ramos (2010), "Multilevel and Session Variability Compensated Language Recognition: TVS-UAM Systems at NIST LRE 2009", *IEEE Journal of Selected Topics in Signal Processing, s.* 1084-1093.

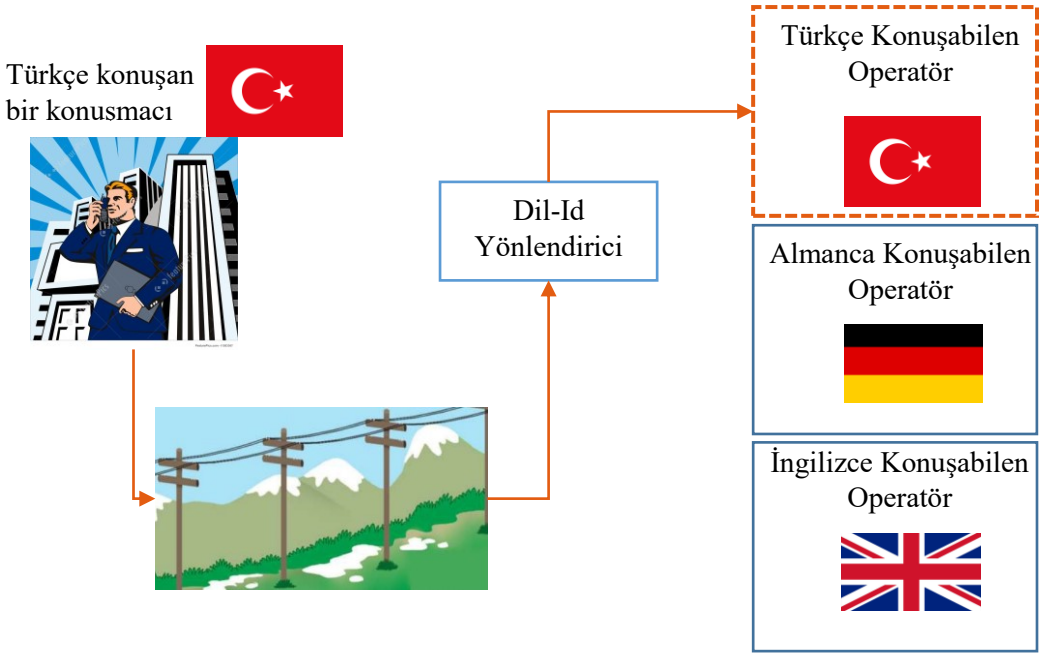


Şekil 4 Bir gerçek zamanlı konuşma tanıma için dil tanımlama sistemi örneğidir. Çıktı olarak en olabilir 3 dili belirlemektedir. Konuşma tanıma sistemi belirlenmiş 3 dil için modelleri sisteme yükler ve ses işlemesi sonrasında hangi dil olduğuna karar verilir[11]

Şekil-4'te örnek bir otel giriş veya uluslararası bir havaalanında çok dilli ses kontrol sistemine sahip bilgi alım sistemi örneği gösterilmektedir. Sisteme giriş olarak ses haricinde bir veri girişi yok ise, sistem hangi dil ile konuşulduğuna karar verebilmelidir ve verilen komutları anlayabilmelidir. Bu sistemler aynı anda birkaç dili birlikte tanıyabilen ve farklı diller için hazırlanmış tanıma sistemlerinin paralel olarak çalıştığı bir mimaride tasarlanmalıdır. Kullanılacak olan farklı dil sayısının artışı beraberinde daha fazla donanım ve gerçek zamanlı çalışabilecek sistemlerin

[11] Marc A., Zissman ve Kay M.,Berkling (2001), "Automatic Language Identification", Speech Communication, s. 115-124.

gereksinimi olacaktır. Alternatif olarak, konuşma tanıma işlemi öncesinde bir dil kimliği belirleme sisteme çalıştırılabilir. Bu sayede sesin öncelikle dil tespiti yapılır ve ses tanıma cihazına sadece o dilin modelleri yüklenerek tanıma işlemi gerçekleştirilir. Donanımsal ve işlemsel olarak avantajlı olmaktadır. Bu işlemler bittikten sonra, son bir dil kimliği belirleme işlemi yapılır.



Şekil 5 Çok dilli olarak hizmet veren viren asistanlık hizmetinin veya acil yardım sisteminin ön uç yapısıdır. Gelen çağrının hangi dilde olduğunu tespit ederek ilgili çağrı operatörüne bildirir[12]

Şekil-5'te insanları dinleyen sistemler için bir önişleme prosedürü gösterilmiştir. Bu tip sistemler hali hazırda polis şubelerinde, acil servis hizmetlerinde kullanılmaktadır. Bir dil hattı tercümanı gelen çağrıyı ilgili operatöre iletmektedir. Bir çağrı geldiğinde Türkçe konuşmuyor ise, önce ilgili tercümana yönlendirilir. Bu tür sistemler kurulurken, çalışacak kişilerin doğru seçilmesi mecburi hallerde değişikliğe gidilmesi gerekmektedir. Gelen çağrının dilinin kısa sürede tespit edilmesi ihtiyacı acil çağrılar için hayati bir konudur[12].

[12] A.g.m. s. 115-124.

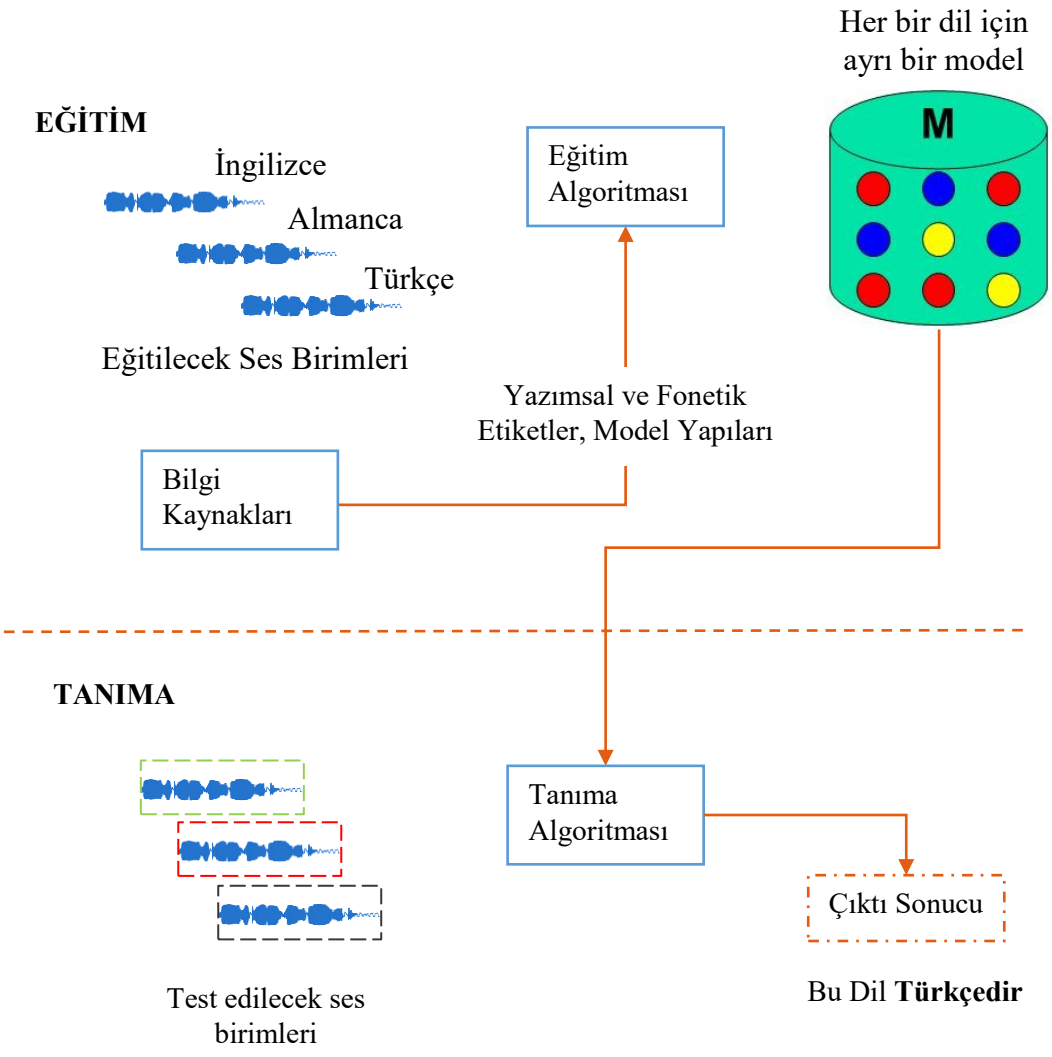
İnsanlar ve makineler bir dili diğerinden ayırmak için çeşitli ipuçlarından yararlanırlar. Dillerin farklı durumları dil bilim literatürlerinde tanımlanır ve araştırmacılar bu tanımlara atıfta bulunarak çalışmalarını yürütürler. Bir dili diğerinden ayırt etmek için temelde dil bilim çalışmalarına temel oluşturan aşağıdaki parametreler kullanılmaktadır:

Fonoloji: Bir fonem bir dildeki fonolojik birimin altta yatan zihinsel temsilidir ve akustik-fonetik birimlerin gerçekleştiği durumdur. Konuşmacının, konuşmayı düşündüğü anda üretilen gerçek sestir. Dilbilimciler tarafından ortaya konulan araştırmalarda, telefon ve ses birim setleri birçok dil için ortak olmakla birlikte birçok durum için farklı ses birimleri ile de karşılaşılır. Özellikle kullanım sıklıkları dilden dile farklılık gösterir. Fonoloji, bir dil için izin verilen fonem dizilerini düzenleyen kuralları ifade eder.

Morfoloji: Kelime kökleri ve sözlükler dilden dile farklılık gösterir. Her dilin kendi sözcük dağarcığı ve kelimeleri oluşturma biçimi vardır. Bu biçimlerin değerlendirilmesi yoluyla dillerin birbirinden ayrılması mümkün olabilmektedir.

Sözdizimi: Cümle kalıpları dilden dile farklılık gösterir. Aynı yazılımlı iki kelime farklı dillerde farklı şekillerde kullanılabilirler. Bu farklı dizilim dikkate alınarak dil belirlemesi yapılabilmektedir.

Aruz(ölçü): Süre özellikleri, sesler arası adım aralıkları ve stres kalıpları dilden dile farklılık gösterir.



Şekil 6 Dil Tanımlamanın iki safhası gösterilmektedir. Eğitim sırasında konuşma dalga formları analiz edilir. Dil bağımlı modeller üretilir. Tanıma aşamasında, yeni bir ses işlenir ve eğitim sırasında üretilen modeller ile karşılaştırılır[13].

Şekil-6'da dil tanımanın iki adımı gösterilmektedir. Eğitim aşamasında sistem çeşitli dillerden konuşma örneklerini alır. Her bir eğitim konuşması özellik vektörlerine dönüştürülür. Genel olarak 20 ms lik pencerelerde örnekleme yapılarak özellik vektörleri oluşturulur. 10 ms pencere kaydırılarak adım adım özellik vektörleri hesaplanır ve konuşma sinyali hakkında ceptral ve spektral veriler elde edilir. Eğitim

algoritması bu tür vektörlerin bir dizisini analiz eder ve her dil için bir model oluşturur. Bu modeller o dili temsil edecek şekilde sistemde saklanır[13].

Şekil 6'da gösterilen tanıma safhasında, yeni bir girdi için özellikler vektörleri çıkarılır ve daha önce sistemdeki vektörler ile karşılaştırılır. Sistemdeki modeller arasında olma olasılıkları hesaplanır ve maksimum olasılık modeli bulunur. Böylece yeni konuşmasının diline karar verilir.

Bu konuda kritik husus dillerin modellenmesidir. Dil ile ilgili çıkarılan özellikler yöntemden yönteme farklılık göstermektedir ve araştırmanın gelişmişliği bu farklılıkla ortaya çıkmaktadır. Modelleme ile daha iyi sonuçlar almak için fonem tabanlı özellik çıkarımından yararlanılır. Konuşulan sesleri temsil eden sembol dizisi üretilmesi veya konuşma metninin doğrudan teleffuzla eşlenilerek eğitilmesi yöntemleri izlenmektedir. Bu işlemler pahalı ve zaman alıcı süreçlerdir ve dil konusunda ciddi dil bilim bilgisine ihtiyaç duyulmaktadır. Bu şekilde dil bilim bilgisi ile konuşmacı dili tanıma yazılımlar birleştirilerek tüm bu işlemleri otomatik olarak gerçekleştirebilecek yazılımlar gerçeklenebilmektedir.

1.2.3 Konuşma Tanıma Sistemleri Sınıflandırması

Konuşma tanıma sistemleri girdi olarak aldıkları birimlere göre farklı sınıflara ayrılmaktadırlar. Farklı araştırma alanlarında daha özel sınıflandırmalar yapılmakla birlikte temelde 4 gruba ayrılırlar[14].

[13] A.g.m. s. 115-124.
[14] Anusuya, M.A. ve Katti, S.K., 2009, s. 181-205.

1.2.3.1 Yalıtık Sözcük Tanıma

Yalıtık sözcük tanıma, ses sinyali içerisindeki kelime söyleyişleri arasında belli bir miktarda sessiz bölümün olmasını zorunlu kılan bir türdür. Bu sistemlerde bir zaman periyodunda tek bir kelime veya sesin sisteme girdi olarak verilmesi beklenir. Ses tanıma yapısında, dinle ve dinleme durumları vardır ve konuşmacı her bir ses birimi sonrasında duraksamalı ve sonrasında ikincil ses birimini seslendirmelidir. Bu yapıdaki mimariler genelde komut temelli ses tanıma sistemlerinde kullanılırlar. Bir makine veya sisteme ileri, geri gibi temel komutları veya aç kapa gibi görevlerin verildiği durumlarda kullanılırlar. Bu görevlerde bir komut sonrasında ikinci komut verilmeden önce doğal bir bekleme süresi vardır ve yalıtık sözcük tanıma yapılarına uygun bir problem türüdür.

1.2.3.2 Bağlı Sözcük Tanıma

Yalıtık sözcük tanıma sistemlerindeki sesler arasındaki belirgin duraksamaların biraz daha az olabildiği sistemlerdir. Kısa duraksamalar ile konuşmaya devam edilebilmektedir. Ancak yine de sesler arasında sistemlerin anlayabileceği miktarda sessiz bölüme ihtiyaç duyulmaktadır.

1.2.3.3 Sürekli Konuşma Tanıma(SKT)

Kullanıcıların hemen hemen sürekli konuşarak, bilgisayar ile iletişimde bulunabildiği türdeki sistemlere sürekli konuşma tanıma sistemleri denilmektedir. Bir tür bilgisayar diktesi olarak adlandırılabilir. Bu tip sistemlerde birçok ciddi problem ile karşılaşılmaktadır. Bunlardan en önemlisi ise, sürekli ses sinyalinin girdi olarak alınması ve bu sinyal içerisindeki kelime sınırlarının belirlenmesi işlemidir. Ses sinyali içerisindeki sınırların belirlenmesi çoğu halde mümkün olmakla birlikte, ciddi işlem

ve hesaplamaya ihtiyaç duyulmaktadır. Günümüz konuşma tanıma sistemlerinde bu tipteki problemlere çözüm üretmek amaçlanmaktadır.

Sürekli konuşma tanıma sistemleri, ayrık sözcük tanıma sistemlerine göre zordur. Çünkü sözcük sınırları belirgin değildir. Tanınacak söz birimleri, önüne ve arkasına gelen ses birimlerinden etkilenirler. Hızlı konuşma halinde bazı sözcüklerin kaybedilmesi söz konusudur.

1.2.3.4 Spontan Konuşma Tanıma

İnsanların doğal konuşma halleriyle ve herhangi bir hazırlık olmaksızın seslendirmelerinin bilgisayarlar tarafından işlendiği sistemlerdir. Bu tarz yapılarda sürekli konuşma tanıma sistemlerinin özelliklerinin yanında, konuşmacıların doğal hallerinde konuşurken kullandıkları "ımm, aaa, uuu" gibi sesleri veya harf olarak tanımlanamayacak gürültü vb. durumlarında algılanabilmesi mümkün olmalıdır. İnsanların doğal konuşma hallerine ait videolarda bir dikte işlemi yapılması düşünülecek ise, bu tarz bir yapının kurgulanması gereklidir.

1.2.4 Konuşmacı Modu

Konuşma tanıma sistemleri, seslendirmeyi yapan kişinin değişkenlik durumuna göre konuşmacı bağımlı ve bağımsız olarak iki farklı türe ayrılmaktadır[15].

[15] Yalçın, Nursal, 2008, s. 249-266.

1.2.4.1 Konuşmacı Bağımlı

Konuşmacı bağımlı sistemlerde tek kişinin referans şablonlarının oluşturulması ile probleme çözüm üretilmeye çalışılır. Yeni bir kişininde tanınabilmesi için referans şablonlarının güncellenmesi gerekir. Farklı kişilerin konuşması halinde tanıma performanslarında ciddi düşüşlerle karşılaşılır.

1.2.4.2 Konuşmacı Bağımsız

Sistem herhangi bir kişi tarafından seslendirilen sesli ifadeyi tanıyabilir. Doğal sistem tasarımlarında kişiden bağımsız sistemler kurulması gerekmektedir. Kurgulanacak olan konuşma tanıma sistemlerinde doğal olarak kullanımının daha yaygın olması için konuşmacı bağımsız olması beklenir. Ancak bu sistemleri geliştirmek daha zor bir problemdir. Bu sebeple sistem tasarımlarında öncelikle kişi bağımlı sistemler tercih edilmektedir.

1.2.5 Sözlük Boyutu

Bir konuşma tanıma sistemince tanınabilen sözcüklerin kümesine sistem sözlüğü denilmektedir. Kelime altı birimlerle çalışıldığı durumlarda ise ses kümesi olarak adlandırılmaktadır. Sözlük boyutunun büyümesi tanıma başarısını olumsuz yönde doğrudan etkiler. Sözcük dağarcığının artırılması daha fazla işlemci gücü ve daha fazla kelime hata oranı demektir.

Sözlükler boyutlarına göre farklı şekilde sınıflandırılabilmektedir[16]. Genel olarak:

[16] A.g.m. s. 249-266.

- 100’den az sözcük içeren sözlükler küçük dağarcıklı,
- 100-1000 arası sözcük içerenler orta dağarcıklı,
- 1000’den fazla sözcük içeren geniş(büyük) dağarcıklı olarak gruplandırılabilir.

1.3 AKUSTİK ÖN UÇ

Bu bölümde akustik ön uç tasarımında kullanılan algoritmalardan bahsedilmektedir.

1.3.1 Özellik Çıkarımı

Konuşma tanıma sistemi tasarımlarında kullanılan özellik çıkarımı işlemi 3 aşamada gerçekleştirilir. Öznitelik çıkarmaktaki ana amaç giriş sinyalini temsil edebilecek daha kompakt bir özellik vektör dizisini elde edebilmektir[17].

Özellik çıkarımı temel olarak 3 adımda gerçekleştirilir:

- Birinci adım konuşma analizi veya akustik ön uç olarak isimlendirilir. Sinyalin belirli bir tür spektrumunun zamansal analizini gerçekleştirilir ve kısa konuşma aralıkları için güç spektrumunun penceresini tanımlayan ham özellikleri üretir.
- İkinci aşamada, statik ve dinamik özelliklerden oluşan genişletilmiş bir özellik vektörü oluşturulur.
- Son aşamada, üretilen genişletilmiş özellik vektörlerinden daha kompakt ve sağlam vektörler üretilir ve daha sonra tanıyıcılarda kullanılır.

Her bir uygulama için belirlenmiş bir özellik ve özellik grubu yoktur ancak elde edilen özellik vektörlerinin en az aşağıdaki özellikleri sağlaması beklenir:

[17] Karpagavalli, S ve Chandra, E. (2016), "A Review on Automatic Speech Recogniton Architecture and Approaches", *International Journal of Signal Processing, Image Processing and Pattern Recognition,* s. 393-404.

- Benzer sesli konuşma seslerinin ayırt edilebilmesine imkân tanıyacak farklılıkları göstermelidir.
- Aşırı miktarda eğitim verisine ihtiyaç duymadan akustik modellerin otomatik oluşturulmasına imkân vermelidir.
- Konuşmacı veya konuşma ortamında büyük ölçüde değişiklik olmasa dahi bu değişimi istatistiksel özellikler olarak gösterebilmelidir.

Gelen her bir girdinin istatistiksel olarak bazı bilgileri bulundurması, her bir ses işleme segmentinin bilgilerinin azaltılması mekanizmalarında önemlidir. Ses sinyalini nispeten az sayıda parametreye ve özelliklere dönüştürür. Bu özellikler sayesinde, her bir segmentin benzer segment özelliklerini karşılaştırarak gruplandırabilecek karakteristik tanımlamalar yapılabilmelidir.

Akustik özellik çıkarımı için temelde aşağıdaki tabloda gösterilmiş methodlar kullanılmaktadır.[18]

Tablo 1 Özellik Çıkarımı için Kullanılabilecek Metodlar ve Bazı Özellikleri

Method	**Özellikler**	**Açıklamalar**
Birincil Bileşen Analizi	Doğrusal olmayan özellik çıkarım metodu, lineer haritalama, hızlı, özvektör temelli	Geleneksel, Karhuneu-Loeve metodu olarakta bilinir. Gaussian verileri için uygun bir yöntemdir.
Doğrusal Diskriminant Analizi	Doğrusal olmayan özellik çıkarım metodu, gözetimli lineer haritalama, hızlı, özvektör temelli	Sınıflandırma için Birincil Bileşen Analizinden daha iyi bir yöntemdir.
Bağımsız Bileşenler Analizi	Doğrusal olmayan özellik çıkarım metodu, lineer haritalama, tekrarlı Gaussian olmayan veri gerekliliği	Gaussian verileri dışındaki verilerin karşılaştırmasında kullanılabilir.

[18] J, Li, Deng, Y. ve Gong, R.H. Umbach (2014), "An Overview of Noise-Robust Automatic Speech Recognititional", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, s. 745-777.

Doğrusal Tahmin Kodlaması	Statik özellik çıkarım metodu	
Kepstral Analizi	Statik özellik çıkarım metodu, güç spektrumu	Spektrum pencerelerinin temsili için kullanılır.
Mel-Frekans Analizi	Statik özellik çıkarım metodu, spektral analiz	Spektral analiz, öznel bir frekans ölçeği boyunca sabit bir çözünürlük ile yapılır.
Filtre Bankası Analizi	Filtreler gerekli frekanslara göre ayarlanır.	
Mel-Frekans Kepstrum Analizi	Güç spektrumları Fourier analizi ile hesaplanır.	
Çekirdek Tabanlı Özellik Çıkarım Metodu	Doğrusal olmayan dönüşüm	Boyut azaltımı sayesinde daha iyi bir sınıflandırma yapılır ve gürültü ve tekrarlı özelliklerin silinmesinde ve sınıflandırma hata oranlarının düşürülmesinde kullanılır.
Dalgacık(Wavelet)	Fourier dönüşümüne göre daha iyi zaman çözünürlüğü	Fourier dönüşümün sabit bant genişliği yerine yüksek frekanslarda fourier dönüşümünden daha iyi zaman çözünürlüğüne izin veren orantılı frekans ile yer değiştirir.
Dinamik Özellik Çıkarımı DÖK, MFKK	Hızlanma ve delta katsayıları, normal DÖK ve MFKK katsayılarının 2. veya 3. Dereceden türevlerini içerir.	
Spektral Çıkarım	Gürbüz özellik çıkarım metodu	
Kepstral Ortalama Çıkarım	Gürbüz özellik çıkarım metodu	
RASTA Filtreleme	Gürültü içeren konuşmalar için uygudur.	
Entegre Fonem Altuzay Yöntemi	Birincil Bileşen Analizi, Lineer Kodlama Analizi ve Bağımsız Bileşenler analizinin karışımıdır.	Diğer tüm metodlara göre Yüksek Doğruluk.

1.3.2 Özellik Çıkarım Algoritmaları

Konuşmacı tanıma, konuşma tanıma, biyometrik sistemler gibi birçok uygulamada özellikler çıkarımı için yaygın kullanılan birçok algoritma bulunmaktadır[19].

1.3.2.1 Mel Frekans Kepstral Katsayıları(MFKK)

MFKK en popüler özellik çıkarma algoritmalarından biridir. Konuşmacı tanıma ve konuşma tanıma sistemlerinde kullanılır. MFKK tekniği insanların işitsel sistemlerini taklit eden bir yapıdadır. Ses sinyalleri için seslerin frekans içeriğinin algılanışına göre, doğrusal bir yapıda değildir[20].

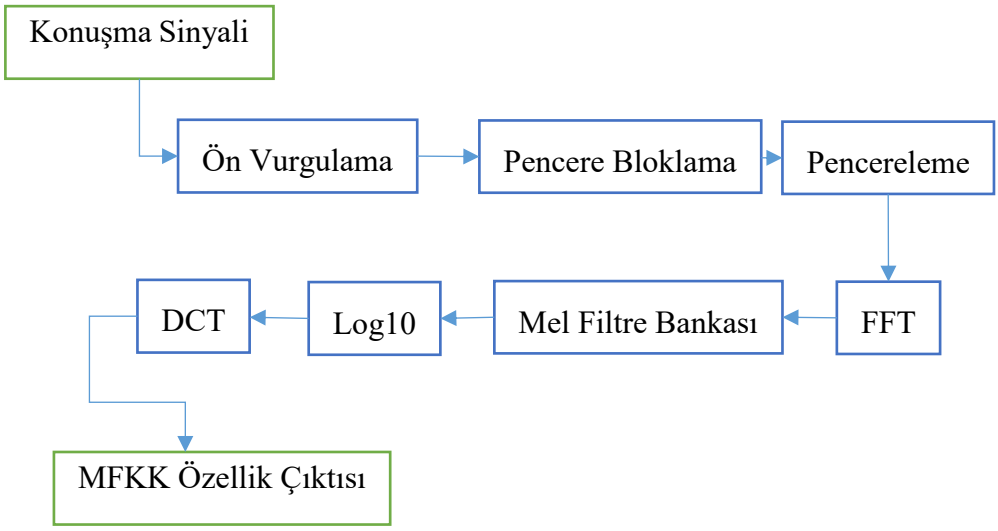
İnsan algılaması 1000 Hz üzerindeki frekanslara daha az duyarlıdır. Bu yüzden gerçek frekansın her bir tonu için Mel ölçeği ile öznel bir adım ölçümü yapılır. 1 Khz üzerinde doğrusal ölçeği takip etmeyen insan algılama davranışları sebebiyle 1000 Hz üzerini ölçekleyen bir Mel skalası kullanılır.

Bir Mel ölçeği, 1000 Hz'e kadar doğrusallık, üzerinde logaritmik değer gösterir. İnsan kulağı bir sinyalin statik ve dinamik özelliğine duyarlıdır ve MFKK çoğunlukla statik özelliklere odaklanmaktadır. Her bir karedeki bilgiyi dinamik olarak takip etmek için karaler arası MFKK farkları alınarak işlem yapılır.

MFKK öznitelik çıkarım işlem adımları aşağıdaki şekilde (Şekil-7) gösterilmiştir[21].

---

[19] R., Rashmi (2014), "Review of Algorithms and Applications in Speech Recognition System", *International Journal of Computer Science and Information Technologies*, s. 5258-5262.
[20] Tiwari, Vibha (2010), "MFCC and its applications in speeaker recognition", *International Journal on Emerging Technologies*, s. 19-22.
[21] Karpagavalli, S ve Chandra, E., 2016, s. 393-404.



Şekil 7 MFKK Özellik Çıkarımı Adımları

Ön Vurgulama: Bu aşama girdi işaretinin yüksek frekanslarda enerjisini yükseltmek için kullanılır. Böylece Saklı markov model eğitim ve tanıma aşamasında her bir bölgenin tanımlanabilir olması sağlamış olur.

Pencereleme: Giriş sinyali ayrık zaman dilimlerine ayrılır. X milisaniye genişlikte, Y milisaniye uzunlukta birbirini takip eden pencereler kullanılır. Hamming pencereleme dikdörtgen şeklindeki keskin değişimlerle ilişkili kenar etkilerini minimuma indirmek için yaygın olarak kullanılır.

Ayrık Fourier Dönüşümü(AFD): Pencerelenmiş ses sinyaline uygulanır, sinyalin büyüklüğü ve faz gösterimi ile sonuçlanır.

Mel Filtre Bankası: Ayrık Fourier dönüşümünden elde edilen spektrum her frekansta bilgi içerirken, insan işitme seviyesi 1000 Hz üzerindeki frekanslarda daha az duyarlıdır. Bu konseptin otomatik konuşma tanıma sistemlerinin performansına doğrudan etkisi vardır. Bu nedenle spektrum bir logaritmik Mel ölçeği kullanılarak kırpılır.

Her bir mel frekansı denklem (4) kullanılarak yapılır.

$$mel(f) = 1127ln\left(1 + \frac{f}{700}\right) \quad (4)$$

AFD spektrumunda bu etkiyi oluşturmak için, üçgen filtreler olarak bilinen filtre kümesi, 1000 Hz'in altında eşit aralıklarla dağılmış ve logaritmik olarak 1000 Hz in üzerinde olan filtreler kullanılarak oluşturulur. AFD sinyalinin her bir mel filtresi tarafından filtrelenmesi Mel spektrumu olarak bilinir.

Log: Mel Filtre Bankasında kullanımı ile Mel Spektrumu co-efficient değerleri üretilir.

Ayrık Kosinüs Dönüşümü(AKD): MFKK değerlerini elde etmedeki son aşamadır. Mel spektrum katsayıları üzerinde ayrık kosinüs dönüşüm yapılmalıdır. AKD çıktıları Mel-cepstral katsayıları 13. sıradadır.

Delta MFKK Özellikleri: Çerçeveden çerçeveye değişen konuşma özeliklerini yakalamak için, MFKK katsayılarının birinci ve ikinci türevi hesaplanıp kullanılır.

1.3.2.2 Gerçek Kepstral Katsayıları

Gerçek kepstral sayıları hesaplamak için, sinyal zaman domeninden frekans domenine alınır ve bu dönüşüm için Hızlı Fourier Dönüşümü(HFD) kullanılır[22].

[22] Genevieve I, Sapijaszko ve Wasfy B., Mikhael (2012), "An Overview of Recent Windows Based Feautre Extraction Algorithms for Speech Recognition", *Circuits and Systems (MWSCAS), 2012 IEEE 55th International Midwest Symposium on : IEEE*, Boise.

Sonuçların alınması ve ters fourier dönüşümü ile sinyalin gerçek keptrumu elde edilir. Matematiksel formülü aşağıdaki gibidir(Denklem (5)).

Gerçek Kepstrum = IFFT (log (FFT (s (n)))) (5)

1.3.2.3 Doğrusal Öngörülü Kodlama

Bu yöntemde konuşma sinyali, formantları tahmin edilerek analiz edilir. Aynı zamanda, formantların konuşma sinyaline etkileri de ortadan kaldırılır ve kalan işaretin yoğunluğu ve frekansı tahmin edilir. Formantları kaldırma işlemine test filtreleme denir[23].

Doğrusal öngörülü kodlama(DÖK) yönteminde, konuşma sinyalinin her örneği önceki örneklerin doğrusal birleşimi olarak iletilir. Formantlar, fark denkleminin katsayıları ile tanımlanır.

1.3.2.4 Türetilmiş Öngörülü Kodlama(TÖK)

Türetilmiş öngörülü kodlama konuşma sinyalinde özelliği ayıklamak için yaygın olarak kullanılan bir metoddur. DÖK parametreleri, ses çerçevesinin enerji ve frekans spektrumunu etkili bir şekilde tanımlayabilir.

Akustik sinyal spektrumu, modelleme ve desen tanımanın temelinde, frekans spektrumunun hızlı değişimini dikkate alabilen kısa sürede karakter tanımlaması yapan logaritmaların belirlenmesidir ve bu işlem için orjinal spektrumda kepstral sayıları türetilir.

[23] Tiwari, Vibha, 2010, s. 19-22.

En yaygın kullanılan spektral ölçümlerden birisi DÖK'ten türetilen TÖK katsayıları ve bu katsayıların regresyon değerleridir. TÖK insanların ses yolundaki biyolojik değişikliklere göre farklılık gösterir ve DÖK parametrelerinden iterasyon yoluyla hesaplanır.

1.3.2.5 Algısal Doğrusal Öngörülü Kepstral Katsayıları(ADÖKK)

Algısal doğrusal öngörülü kepstral katsayıları, konuşma analizi penceresinin büyüklük spektrumuna dayanmaktadır. MFKK ve DÖK gibi zamansal bir yöntemdir. Katsayıları hesaplamak için aşağıdaki adımlar uygulanır:

- Konuşma penceresinin güç spektrumu hesaplanır.
- 8 Khz örnekleme frekansı için, 23 kritik bantta sonuçların gruplandırılması gerçeklenir.
- Ses yüksekliği dengeleme ve küp kök sıkıştırma işlemi güç yasasını taklit etmek üzere yapılır.
- Ters Hızlı Fourier Dönüşümü yapılır.
- Levinson-Durbin(LD) algoritması ile analiz yapılır.
- LD katsayılarını kepstral katsayılara dönüştürülür.
- Bark frekansı ile Hz frekansı arasındaki ilişki bu şekilde tanımlanır.

$$f(\text{bark}) = 6 * \text{arcsin}\, h(f(\text{Hz})/600) \qquad (2)$$

1.3.2.6 Birincil Bileşen Analizi

Doğrusal olmayan bir özellik çıkarma yöntemidir. Doğrusal haritalamaya göre hızlı, öz vektör tabanlı bir yöntemdir. Gauss verileri için iyi sonuçlar üretebilmektedir[24].

1.3.2.7 Doğrusal Ayrık Analiz

Doğrusal olmayan bir özellik çıkarımı yöntemidir. Gözetimli doğrusal haritalamadan hızlı ve özvektör tabanlı bir metoddur. Sınıflandırma problemlerinde Birincil Bileşen Analizine göre daha iyi bir yöntemdir.

1.3.2.8 Bağımsız Bileşen Analizi

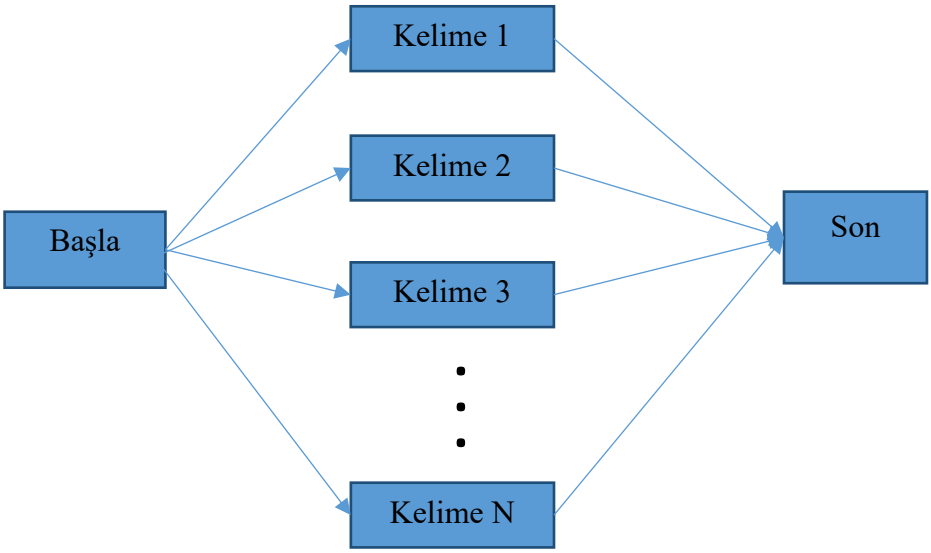
Doğrusal olmayan bir özellik çıkarımı yöntemidir. Lineer haritalama içeren ve iteratif Gauss olmayan bir metoddur.

[24] Santosh K., Gaikwad, Bharti W., Gawali ve Pravin, Yannawar (2010), "A Review on Speech Recognition Technique", *International Journal of Computer Applications*.

1.4 DİL MODELİ TASARIMI

1.4.1 Kelime Tabanlı Dil Modeli

Kelime tabanlı dil modellerinde, konuşma tanıma birimi olarak kelime kullanılmaktadır. Şekil-8'de genel yapısı gösterilmektedir[25].



Şekil 8 Kelime Tabanlı Dil Modeli Yapısı

Bu yöntemde morfolojik veya istatistiksel yöntemlerin aksine kelimeler doğrudan kullanılarak dil modeli oluşturulur ve olasılıkları hesaplanır. N-gram dil modelleri kelimeler kullanılarak hesaplanır.

Kelime tabanlı dil modeli üretme de temelde aşağıdaki adımlar uygulanır.

1- Aşağıdaki kriterleri sağlayan kelime çiftleri elde edilir.

[25] Susman, D., Köprü, S. ve Yazıcı, A. (2012), "Kısıtlı Ses külliyatı ile türkçe geniş dağarcıklı sürekli konuşma tanıma", *Signal Processing and Communications Applications Conference(SIU), s.1-4*

a. Türkçe dil bilgisi kurallarına göre bitişik yazılması gereken kelime çiftleri
b. Sıkça birlikte kullanılan kelime çiftleri hazırlanır.

2- 1. Adımda elde edilmiş olan tüm kelime çiftlerini birleştir.
3- N-gram modelleri kelime tabanlı olarak üretilir.
4- Metin ve ses külliyatındaki her bir kelimenin kök kısmını elde edilir.
5- N-gram dil modelini kök tabanlı olarak elde edilir.
6- 3. Ve 5. Adımdaki dil modelleri karşılaştır ve sonuçları alınır.

1.4.2 Kök-Ekler Tabanlı Dil Modeli

Kök-ekler tabanlı dil modellerinde, tanıma birimleri kök ve kelimenin kökten sonraki kısmını içerir. Dil modeli olasılıkları kök ve ekler üzerinden oluşturulur.[26]

Kök-ekler tabanlı dil modeli oluşturmak için temelde aşağıdaki adımlar uygulanır.

1- Ses ve metin külliyatında geçen her kelime morfolojik ayrıştırıcı ile kök ve eklerine ayrılır.
2- Her bir kelime için kök kısmı en uzun olan ayrışmayı seçilir.
3- Sıkça birlikte kullanılan kelime çiftlerini birleştir.
4- Tek harften oluşan kök ve ekleri birleştir.
5- İşlenmiş kök ve ekleri kullanarak N-gram modellerini oluştur.

1.4.3 Morf Tabanlı Dil Modeli

Kelimelerin otomatik olarak morflarına ayrılması işlemi, yeterince büyük bir derlem kullanıldığında söz konusu derlemin, üzerinde çalışılan dilin yapısıyla ilgili

[26] A.g.m. s.1-4.

istatistiki verileri içereceği fikrini temel alan ve gözetimsiz şekilde çalışan algoritmalar kullanılarak gerçekleştirlir. Elde edilen bu morfem yapıları ses tanıma temel birimini oluşturur ve tanıma işlemi bu birimler üzerinden gerçekleştirilir.

1.4.4 Hibrit Dil Modeli

Kelime tabanlı ve kök-ekler tabanlı dil modellerinin birlikte kullanılmasına dayanır. Hibrit modellerin üretilmesinde aşağıdaki temel olarak aşağıdaki adımlar uygulanır.

1- Metin külliyatından en sık geçen kelime grubu belirlenir ve kök-ekler ayrıştırması yapılır.
2- Geriye kalan kelimeleri kök-ekler tabanlı ayrıştırma kullanılan yaklaşımlar kullanılarak ayrıştırılır.
3- Kelimeler kökleri ve ekleri baz alınarak n gram modeller oluşturulur.

İKİNCİ BÖLÜM

KONUŞMA TANIMA SİSTEMLERİ YAKLAŞIMLARI VE TASARIM ALTYAPISI

2.1 KONUŞMA TANIMA YAKLAŞIMLARI

Konuşma Tanıma sistemlerinin tasarımında temel olarak 5 farklı tür yaklaşım bulunmakta olup, genel tablosu aşağıda gösterildiği gibidir.

Tablo 2 Konuşma Tanıma Yaklaşımları Genel Tablosu

Yaklaşım	**Temsiliyet(Gösterim)**	**Tanıma Fonksiyonu**	**Kriter**
*Akustik Fonetik Yaklaşım	Fonetik/Segment ve Etiketleme	Olasılıksal sözlü erişim	Log olasılık oranı
*Desen Tanıma Yaklaşımı			
**Şablon Temelli	Konuşma örnekleri, pikseller, eğriler	Korelasyon, uzaklık ölçümü	Sınıflandırma hatası
**Dinamik Zaman Bükmesi	Spektral vektör dizileri	Dinamik bükme, optimum algoritma	Benzersizlik ölçümü
**Vektör Kuantasyonu	Spektral vektör seti	Kümeleme algoritması	Öklid farkı

**Olasılıksal	Özellikler	Diskrimant fonksiyonu	Sınıflandırma hatası
*Yapay Sinir Ağı	Ses özellikleri/perseptron/kurallar/birimler/prosedürler	Ağ fonksiyonları	Ortalama karesel hata
*Karar Destek Vektörleri	Çekirdek temelli özellikler	Maksimal marj hiperplane, radyal temel fonksiyon sınıflandırıcıs	Genelleştirme hatasının minimize edilmesi
*Yapay Zeka Yaklaşımı	Bilgi tabanlı yaklaşım		Kelime hata olasılığı

2.1.1 Akustik Fonetik Yaklaşım

Konuşma tanımaya ilişkin olarak ilk çalışmalar, konuşma seslerini bulma ve bu seslere uygun etiketler sağlama üzerine kurulmuştur. Konuşma dilinde sınırlı, ayırıcı fonetik birimlerin var olduğunu ve bu ünitelerin geniş anlamda bir dizi akustik özellik ile karakterize edildiğini varsayan teori akustik fonetik yaklaşımın temelini oluşturur. Fonetik birimlerin akustik özellikleri hem konuşulan ses hem de komşu seslerle oldukça değişken olup, bu yaklaşımda değişkenliği yöneten kuralların basit olduğu varsayılır ve makine tarafından bu birimler kolay bir şekilde öğrenilir.

Akustik fonetik yaklaşımın ilk adımı, farklı fonetik birimlerin geniş akustik özelliklerini tanımlayan bir dizi kümeye dönüştüren özellik algılama ile birlikte konuşmanın spektral analizinin yapılmasıdır. Bir sonraki adımda, konuşma sinyalinin standart ses bölgelerine ayrılması ve her bir bölüm için bir veya birden fazla etiket eklenmesidir. Her bir fonetik çerçeve bölütleme ve etiketlenmeye tabi tutulur. Son adımda, bölümlendirme ve etiketleme için üretilen fonetik etiket dizilerinden geçerli kelime veya kelime dizisini belirlemeye çalışılır.

Sözcük dizisinin belirlenmesi sonrasında, doğrulama fazında, dile özgü sözdizimi ve semantik kısıtlamalar göz önünde bulundurularak bulunan sözcükler

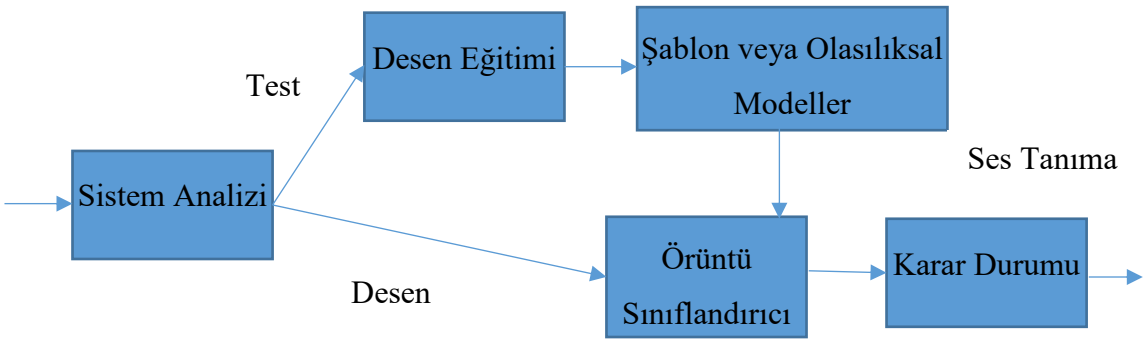
içinde eleme işlemi yapılır ve en doğru kelime dizisi belirlenir. Bu yaklaşım birçok farklı ticari uygulamada yaygınlıkla tercih edilmiştir[27].

2.1.2 Örüntü Tanıma Yaklaşımı

Bu yaklaşım ilk defa 1975 yılında (Itakura, 1975)[28] çalışmasında karşılaşılmaktadır. Sonrasında 1989 yılında (Rabiner, 1989)[29] ve 1993 yılında (Rabiner ve Juang, 1993)[30] çalışmalarında karşımıza çıkmaktadır. Önerilen bu yaklaşımda desenlerin eğitimi ve modellerin karşılaştırılması yapılır. Bu modelde iyi bir matematiksel altyapı, ses temsilini doğru yapan desenlerin üretimi ve doğru etiketleme önemlidir. Bu sayede güvenli bir desen karşılaştırması ile doğru sonuçlar alınması mümkün olmaktadır. Desenlerin oluşturulması için kullanılan veriler, her bir konuşma penceresinin temsili elde edilerek oluşturulur. Bu temsil şablon temelli veye istatiksel modelleme ile yapılabilmektedir. Pencereleme işlemi ses, kelime veya cümle için yapılabilmektedir. Desenlerin karşılaştırılması aşamasında, konuşma sesi haricindeki gürültü vb. diğer sesler daha önce oluşturulmuş ve sistemin eğitiminde kullanılmış desenler ile karşılaştırılır. Böylece doğru metne ulaşılmaya çalışılır.

Bu yaklaşım son yıllarda konuşma tanıma problemlerinin çözümünde yaygın kullanılan bir yöntem haline gelmiştir. Bu yöntemin akış diyagramı aşağıdaki şekilde verildiği gibidir (Şekil-9). Bu yaklaşımı temel alan 4 farklı yöntem bulunmaktadır. Bu bölümde bu yöntemler açıklanmaktadır.

[27] Tran, Dat Tat, 2000, Doktora Tezi.
[28] Itakura, F. (1975), "Minimum Prediction Residula Applied to Speech Recognition", *IEEE Trans Acoustics, Speech and Signal Processing*, s. 67-72.
[29] Rabiner, L.R. (1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *IEEE Trans Acoustics, Speech and Signal Processing*, s. 257-286.
[30] Rabiner, L.R. ve Juang, B.H. (1993), *Fundamentals of Speech Recognition,* Printice-Hall.



Şekil 9 Örüntü Tanıma Yaklaşımı Akış Şeması

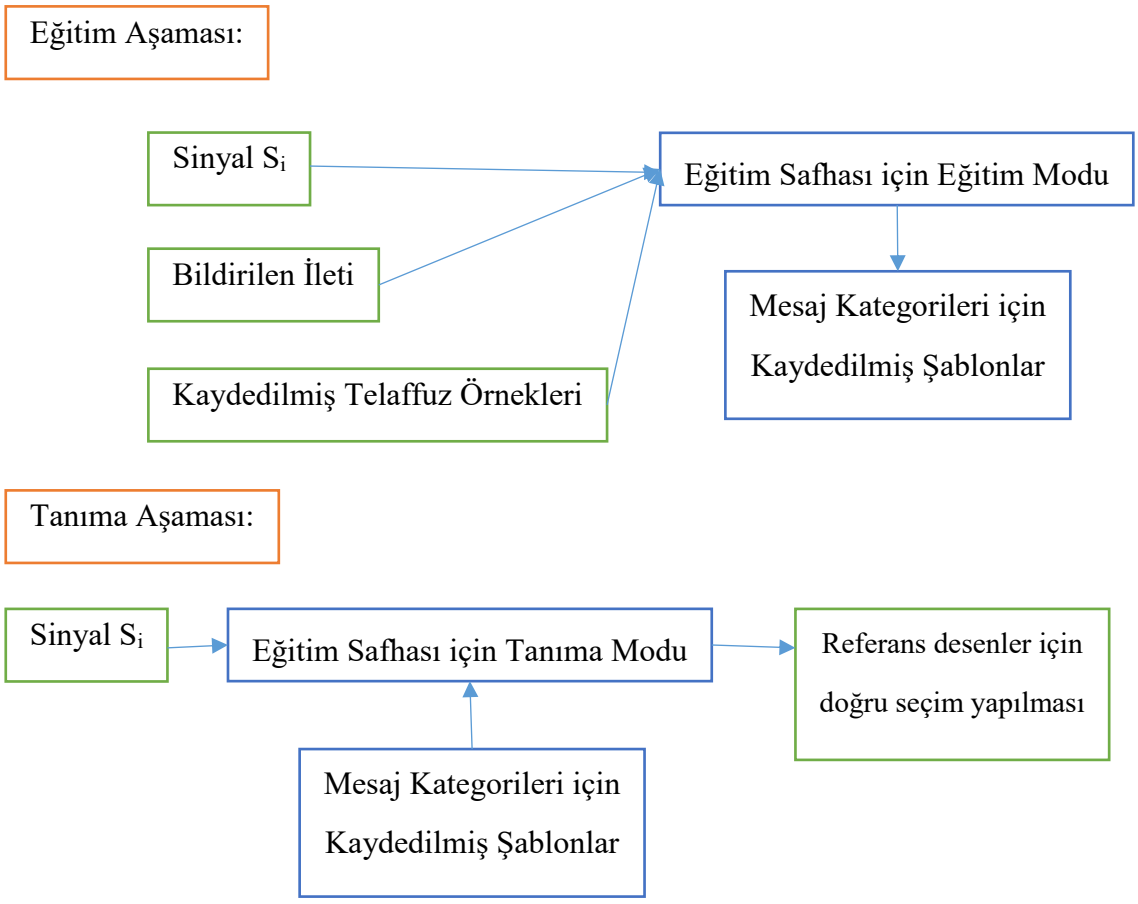
2.1.2.1 Şablon Temelli Yaklaşım

Şablon tabanlı yaklaşım son 60 yılda oldukça gelişen ve üzerinde çalışmalar yapılan bir yöntem olmuştur. Basit bir mantığı vardır. Daha önceden hazırlanmış konuşma deseni prototipleri ile referans kalıplar olarak oluşturulur ve sisteme kayıt edilir. Sonrasında, tanınması istenen konuşma kalıbı referans kalıplar ile karşılaştırılır ve en iyi eşleşme bulunur. Böylece konuşma tanıma işlemi gerçekleştirilmiş olur.

Bu tipteki konuşma tanıma sistemleri için, kullanılabilecek tüm sözcükler için öncesinde kalıplarının oluşturulmuş olması gerekmektedir. Her bir kelimenin referans şablonuna ihtiyaç duyulması ve bir dilde milyonlar seviyesinde kelime haznelerinin bulunması oldukça ciddi bir problem olarak karşıya çıkmaktadır. Tüm kelimelerin şablonlarının çıkarılması zahmetli, pahalı ve pratik olmayan bir yoldur. Bu dezavantajlı olduğu noktaların yanında diğer yöntemlerdeki gibi sözcüklerin hecelere ayrılmasında karşılaşılan problemlerden veya fonem gibi akustik değişkenlerden daha çok etkilenen tanıma birimlerinden yararlanarak tanıma işlemi gerçekleşetirmek yerine sözcüklerden yararlanılması oldukça iyi bir seçimdir.

Şablon temelli bu yaklaşımın kullanıdığı iki farklı fikir vardır. Birincisinde, her bir kelime için konuşma pencerelerinin tipik bir dizisini üretmek ve kalıpları karşılaştırma içinde spektral mesafe ölçümleri sonuçlarına güvenmektedir. Diğer bir fikir ise konuşmacı bağımlı bir model oluşturup, konuşmacının konuşma oranlarındaki farklılıları dikkate alarak, her bir sözcüğün aynı kişi tarafından tekrarlanması ve hizalama yoluyla dinamik programlama yöntemlerini kullanmaktır.

Şablon temelli yaklaşımın şekil-10'da akış şeması gösterilmektedir[31].



Şekil 10 Şablon Temelli Yaklaşım

[31] Moore, Roger K. (1994), "Twenty Things we still don't know about Speech", P*rogress and Prospects of Speech Reserach an Technology*.

Bu şema incelendiğinde, eğitim ve tanıma aşamasının iki ayrı süreç olduğu görülmektedir. Eğitim aşamasında tüm örnek şablonlar oluşturulmakta ve bu şablonlar tanıma aşamasında kullanılmaktadır. Basit ancak bazı problemler için oldukça etkili yöntemlerden birisidir.

2.1.2.2 Dinamik Zaman Bükmesi

Dinamik zaman bükmesi, zaman ve hız açısından değişebilen iki dizi arasındaki benzerliği ölçmek için kullanılan bir algoritmadır. Bir video da kişinin yavaş yürümesi ve aynı kişinin farklı video da hızlı yürümesi arasındaki yürüyüş kalıplarındaki benzerlikler değişmez. Dinamik zaman bükmesi video, ses veya grafikler gibi doğrusal gösterime dönüştürülebilen tüm veriler için uygulanabilir.

Otomatik konuşma tanıma sistemlerinde ses hız değişimlerinin kontrol edilmesi gereklidir ve bu algoritma için uygulanacak alanlardan biridir. İki zaman serisi arasında optimum eşlenmenin yakalanması için kullanılır. Zaman domenindeki doğrusal olmayan değişimlerden bağımsız olarak eşleşmenin sağlanması için iyi bir yöntemdir.

Bu algoritma genelde saklı markov modelleri bağlamında kullanılır. Zaman domenindeki eşleşmenin monotonluğu dizilerin eşleştirilmesinde önemli bir problemdir. Süreklilik dinamik zaman bükmesi yönetim için daha az önemlidir ve yeterli uzunlukta seriler olmak kaydıyla, dizi içerisindeki eksik bilgiler ile eşleşme yakalamak mümkündür. Bu işlemlerin optimizasyonu dinamik programlama kullanılarak yapılır.

2.1.2.3 Vektör Kuantasyonu(VK)

Otomatik konuşma tanıma sistemlerinde sıklıkla kullanılan bir yöntemdir. Konuşma kodlayıcılar açısından veri azaltımına imkan tanıması sebebiyle oldukça kullanışlıdır. OKT sistemleri için VK'nın faydası, referans modeller ve kodlama arama için daha maliyetli yaklaşımlara göre kullanılabilir bir yöntem olmasıdır. Oldukça verimli ve hızlı bir yöntemdir. Modelin eğitimi sırasında, her bir sözcüğün, bir kaç kez tekrarlanması ile oluşturulan VK kod listeleri oluşturulur. Test konuşması tüm bu kod listeleri içinde değerlendirilir ve OKT sistemi en düşük mesafe miktarını veren kelimeyi seçer[32].

Basit VK sistemlerinde, kod çizelgelerinde açık zaman bilgisi bulunmaz. Yani tanıma birimlerinin her bir kelime için zamansal sırası ve süresi göz ardı edilir. Çünkü kod çizelge girdileri eğitim kelimelerinin herhangi bir bölümünden oluşturulabilir. Bunun yanında, sesli harfler gibi daha sık rastlanan kelime bölümleri için ortalama mesafe hesabı yapıldığı için bu sık kullanımı dikkate alıcı bazı algoritmik yaklaşımlar kullanılmaktadır.

VK yönteminin kullandığı ve sık geçen kelime bölümleri ile az rastlanan bölümleri dikkate alan bu yaklaşım diğer OKT sistemlerindeki arama yaklaşımlarına göre daha avantajlı sonuçlar doğurabilmektedir.

2.1.2.4 İstatistiksel Yaklaşım

Belirsizliklerin olduğu veya yeterince veriseti olmaması sebebiyle eksik veri ile çalışılması gereken durumlarda olasılıksal modeller kullanılması daha uygun olmaktadır. Konuşma tanıma için belirsizlikler birçok farklı sebepten

[32] Moore, Roger K. (1994), "Twenty Things we still don't know about Speech", P*rogress and Prospects of Speech Reserach an Technology*.

kaynaklanmaktadır. Örneğin, birbirine karıştırılabilir sesler içermesi, konuşmacı değişiklikleri, içeriğe bağlı etkiler ve benzer ses özelliklerine sahip kelimelerdir. Böyle durumlar için olasılıksal modeller uygun bir yaklaşımdır.

Günümüzdeki en popüler olasılıksal yaklaşım Saklı Markov Modellemesidir(SMM). Bu model, sonlu durum markov model ve çıktı dağılımları kümesiyle karakterize edilir. Markov zincir modellemelerinde geçiş parametresi olarak zamansal değişkenler, çıkış dağılım modelinde spektral değişkenler önemli bir noktadır. Bu iki tipteki değişkenler ses tanıma uygulamalarının temelini oluşturur.

Şablon temelli yaklaşıma kıyasla, saklı markov modellemesi daha genel ve iyi bir matematiksel temel taşır. Şablon tabanlı model sadece kimlik kovaryans matrisleri ve eğim ile ilişkili topoloji içeren sürekli yoğunluklu bir Saklı markov türüdür. Şablonlar daha az örnek ile eğitilmesine rağmen, SMM'lerin olasılık formülasyonundan yoksundurlar ve genellikler SMM'leri daha düşük performans göstermektedir.

Bilgiye dayalı yaklaşımlarla karşılaştırıldığında, SMM'ler bilgi kaynaklarının mimariye entegrasyonunda kolaylık sağlar. Bunun olumsuz yanı, SMM'lerin tanıma sürecinde fazla bilgi sağlamamasıdır. Sonuç olarak, SMM sisteminin performansını artırmak için yapılan hataları analiz etmek genellikle zordur. Bunun yanında, elde edilen bilgilerin sisteme dâhil edilmesi SMM tabanlı sistemleri önemli derecede geliştirmektedir.

2.1.3 Yapay Sinir Ağı(YSA)Yaklaşımı

Yapay sinir ağı yaklaşımı çıkarılan akustik öznitelik kümesine dayalı olarak konuşmayı görüntüleme, analiz etme ve karakterini belirleme süreçlerinin bilgisayarlar tarafından yapılmasını sağlamaya yönelik bir yöntemdir. Bu sınıftaki

yöntemler fonetik, söz dizimsel, semantik ve hatta bölütleme ve etiketleme için pragmatik bilgiyi kullanırlar ve fonetik olaylar arasındaki ilişkileri öğrenmeyi amaçlarlar. Bu yaklaşımın odak noktası bilginin temsil edilmesi ve bilgi kaynaklarının entegre bir şekilde kullanılmasıdır.

Bu yöntem henüz ticari sistemlerde yaygın olarak kullanılmamaktadır. Sinir ağları şeklinde bağlantılı bir modelleme yeni gelişmekte olan bir modeldir ve araştırılmaktadır. Bu modellerde konuşma birimleri, kuralları veya procedürleri doğrudan kodlanmaz. Bunun yerine birçok basit matematiksel hesaplama birimi arasında dağıtılır. Böylece belirsizlikler tek bir ünite de olasılık veya olasılık yoğunluk fonksiyonu olarak değil birçok üniteye dağıtık şekilde etkinlik modeliyle modellenmiştir.

Dağıtık yapıdaki her bir işlem biriminin yapısı basittir ve herhangi bir işlev yapacak şekilde programlanmaz. Bu tür küçük işlem birimlerinden oluşan ağlar içerisinde gerçekleştirilen hesaplama tarzı, sinir sistemindeki hesaplama şekilleri ile benzerlik gösterir. Bu türdeki bağlantılı modellere sinir ağı denilmektir. Dağıtık işlem veya paralel işlem terimleri sinir ağı modellerinde karşımıza çıkmaktadır. Olasılıksal modellerin aksine, sinir ağı modelleri iyi eğitim ve öğrenme stratejilerine sahiptir. Ağın zamanla yeni veriler ışığında optimize edilmesine imkân tanır. Bununla birlikte, olasılıksal modeller gibi varsayımlar yapılması gerekmez. Çok katmanlı sinir ağları, oldukça karmaşık ve doğrusal olmayan sınıflandırıcılar veya haritalama fonksiyonları üretmek için kullanılırlar. İşlem birimlerinin basit ve verimli olmaları donanımsal olarak işlem görmelerini kolaylaştırır. Diğer taraftan, yüksek miktarda eğitim verisinin ağa dahil edilmesi yüksek miktarda işlem gücü ve pahayı beraberinde getirir.

Sinir ağı şeklindeki bağlantılı modeller her ne kadar ses tanıma alanı ile ilgili cazip fırsatlar sunuyor olsa da günümüzde daha pratik ve sonuç alınabilir diğer modeller hala popülerdir. Sinir ağlarının çözülmeyi bekleyen veri ve işlem gücüne ilişkin sorunlarının çözülmesi yapay ağlarının gelişimini hızlandıracak noktadır.

2.1.4 Karar Destek Vektörleri(KDV)

Karar destek makineleri ayrık yaklaşım kullanan örüntü tanıma temelli bir yöntemdir. Doğrusal ve doğrusal olmayan hiper düzlemler ile veri sınıflandırması yapar. Bununla birlikte KDV yanlızca sabit uzunlukta veri vektörlerini sınıflandırabildiğinden, değişken uzunluktaki veri sınıflandırmalarında kullanılamaz. Değişken uzunlukta veriler sisteme verilmeden önce sabit uzunlukta vektörlere dönüştürülmelidirler.

KDV maksimum marj ayarlama fonksiyonlarına sahip bir doğrusal sınıflandırıcıdır Bu fonksiyon sınıflandırıcının genelleştirilmesine yardımcı olur. Sınıflandırıcı birçok özelliği yok sayma eğilimindedir. Konveksiyonel istatistik ve sinir ağı metodları model karmaşıklığını az sayıda özellik kullanarak kontrol eder. KDV ise veri boyutlarını kontrol ederek model karmaşıklığını kontrol eder. Bu yöntem boyutsallıktan bağımsızdır ve çok yük boyutlu veri kümelerini kullanabilir ve eğitim sırasında özellik seçiminin yapılmasını mümkün kılar. Tüm doğrusal olmayan özellikleri kaydırma yaparak bilinen boyutlarda doğrusal modeller oluşturur ve kullanır.

2.1.5 Yapay Zeka Yaklaşımı

Yapay zeka yaklaşımı akustik model yaklaşım ile örüntü tabanlı yaklaşımın bir karışımıdır. Akustik fonetik yaklaşım ve örüntü tabanlı yaklaşımın temel konseptini ve fikrini kullanmaktadır. Bilgi tabanlı yaklaşım dile bağlı, fonetik ve spektrogramla ilgili verileri kullanır.

Bazı konuşma tanıma araştırmacıları, konuşma seslerini sınıflandırma kurallarını geliştirmek için akustik fonetik bilgiyi kullanan sistemler geliştirdiler. Şablon temelli yaklaşımlar çeşitli konuşma tanıma sistemlerinin tasarımında çok etkili

olmasına rağmen, konuşma dikte uygulamalarında yetersiz kalması sebebiyle hata analizi ve bilgiye dayalı sistemlerin geliştirilmesini gerekli kıldılar. Yeni modellerin geliştirilmesi için dil ve fonetik yapıya ilişkin farklı anlayışlar sağladı.

Uzmanların konuşma bilgisinin bir tanıma sistemine doğrudan ve açık biçiminde dahil ediliği yöntemdir. Bu bilgi genellikle spektrogramların dikkatle incelenmesinden elde edilir ve kurallar veya prosedürler kullanılarak dahil edilir. Büyük miktarda uzman verilerin elde edilmesinin güç olması sebebiyle bu yaklaşımın sınırlı bir başarısı elde edilebilmiştir.

Konuşmaya ilişkin uzman verisi, fonetik, fonotaktik, sözlü erişim, sözdizimi, anlambilim ve pragmatik gibi bir çok verinin birleştirilmesinden meydana gelir. Alternatif olarakta, bağımsız ve asenkron bilgi kaynaklarını optimum düzeyde birleştirme bir problem olarak ortada durmaktadır. Hem olasılıksal hem şablon tabanlı modellerde bilgi tabanlı yaklaşımlar kullanılmıştır. Bu şekilde bilgi tabanlı ve algoritmik yaklaşımların birlikte kullanılması problemlerin çözülmesinde iyi bir yöntemdir. Uygun bir girdi gösteriminin seçilmesinde, konuşma birimlerinin tanımlanmasında veya tanımlama algoritmasının tasarımında önemli bir rol oynar.

2.1.6 Bilgiye Dayalı Yaklaşım

Konuşma farklılıkları konusunda uzman kişilerin bilgilerinin sisteme tanıtılmasını gerektirir. Bu yaklaşım modelleme varyasyonları avantajına sahiptir ancak uzmanlık bilgisinin elde edilmesi ve başarıyla kullanılabilmesi zor bir süreçtir.

Bu sebeple pratik bir yaklaşım değildir. Uzmanlık bilgisinin otomatik öğrenme ile elde edilebilmesi üzerine çalışmalar yapılmıştır[33,34].

2.2 KONUŞMA VERİTABANLARI

Konuşma ve dil tanımlama, doğrulama, konuşma sentezi, kodlama ve analiz gibi birçok alanda ses işleme süreçleri yürütülmektedir. Bu çalışmaların yapılabilmesi içinde büyük miktarda kayıtlı veri tabanlarına ihtiyaç duyulmaktadır. Bu veritabanları kullanılarak doğrudan geliştirme yapılabilmekte ve yeniden kayıt yerinde bu veritabanlarının kullanımı yaygın olarak tercih edilmektedir.

Konuşma içerikleri ve konuşmacılar arasındaki değişkenlik bir konuşma veri tabanı için oldukça önemli parametrelerdir. Konuşmacı çeşitliliği konuşma tanıma performanslarını doğrudan etkileyen bir husustur ve bu çeşitlilik konuşma hızından, duygusal ve zihinsel değişkenlerden veya çevresel etkiler sebebiyle oluşmaktadır. Konuşma veri tabanları temelde tekli veya çoklu oturum olarak sınıflandırılabilirler. Tekli oturum olarak hazırlanan veri tabanlarında kayıt gürültü ve diğer kişilerden yalıtılmış bir ortamda kaydedilmektedir. Çünkü akustik ortamın mümkün olduğunca kayıt edilmesi gereken sese etki etmemesi istenir. Diğer taraftan çoklu oturum veri tabanlarında suni gürültü bulunan ses kabinlerinde veya ofis/ev ortamı gibi doğal gürültüye sahip ortamlarda kaydedilirler. Ayrıca, konuşma veri tabanları amaçlarına göre veya anadili olmayan kişilerden seçilecek şekilde oluşturulabilir.

[33] Keh Yih, Su ve Chin Hui, Lee (1994), "Speech Recognition using weighted HMM and subspace projection approaches", *IEEE Transactions on Speech and Audio Processing*, s. 69-79.

[34] Santosh K., Gaikwad, Bharti W., Gawali ve Pravin, Yannawar (2010), "A Review on Speech Recognition Technique", *International Journal of Computer Applications*.

Konuşma veritabanlarının oluşturulması esnasında bazı kriterler veritabanın kapsamını ve kullanılma amacını etkilemektedir. Bu kriterler özetle aşağıdaki gibidir:

- Veriseti kayıt aşamasında farklı söyleyiş tarzlarının değerlendirilmesi gerekmektedir. Bu farklı söyleyiş tarzlarının farklı seslendiriciler tarafından yapılması ile tek bir seslendiricinin farklı söyleyiş tarzlarını kendisinin yapılması durumunda üretilen veritabanlarında ciddi parametre değişiklikleri olmaktadır.

- Verisetlerinin üretilmesi aşamasında tek bir seslendiricinin kullanılması ile farklı seslendiricilerin kullanılması arasında duygusal, zihinsel ve çevresel etkiler kaynaklı farklılıklar ortaya çıkmaktadır. Farklı konuşmacıların ürettiği seslerde, ses yolu eklemleri, dudaklar ve burun deliğine bağlı durumlar bu değişkenliği oluşturmaktadır.

- Çok oturumlu ses kayıt sistemlerinde oluşturulan verisetleri konuşmacı bağımlı ses tanıma sistemleri için oldukça uygun olmaktadır.

- Akustik ortama göre veritabanları ya ses kabinleri gibi çevresel gürültüden arındırılmış ortamlarda ya da ofis/ev gibi doğal gürültülü ortamlarda kaydedilirler. Böylece farklı ortamlarda tanıma performans ölçümleri yapılabilir, sonuçlarda daha geniş açıdan değerlendirilebilmektedir. Aksi halde, gürültüsüz ortamda oluşturulan verisetleri için geliştirilmiş modeller, ev/ofis ortamında alınan test örneklerinde ciddi performans kayıpları yaşayabilmektedirler.

- Konuşma tanıma veritabanları bazı hallerde konuşma tanıma uygulamaları geliştirmek için tasarlanmış olmasına rağmen, bazı durumlarda da konuşma tanıma veya konuşulan dil tanıma gibi farklı problemler içinde hazırlanmış olabilmektedir. Buna en iyi örnek SIVA, Polycost ve YOHO verisetleridir. Bu sebeple, verisetinin

oluşturulmadan önce hangi amaçla kullanılacağının belirlenmesi oldukça önemlidir çünkü her bir problem için verisetinden beklentiler farklı olabilmektedir.

- Birçok veritabanından kayıtlar o dili anadili olarak konuşan kişiler seçilerek yapılır. Ancak bazı durumlarda o dili sonradan öğrenmiş kişilerin konuşmalarıda verisetlerine eklenerek, daha anlamlı setler oluşturulur.

Bu şekilde verisetlerinin oluşturulmasında kullanılan seslendiriciler, kayıt ortamı, kayıt amacı, dile ve seslendiriciye özgü fiziksel durumlar etki edici unsur olarak düşünülmekte ve veriseti özetlerinde paylaşılmaktadırlar.

Bugüne kadar hazırlanmış ve araştırmacıların kullanımına sunulmuş olan veri tabanlarının çoğu İngilizce dilindeki konuşmalardan oluşmaktadır ve bu veritabanları belirli platformalar üzerinden dağıtılmaktadır. Bunlar içerisinde en yaygın olanları LDC[35], Arctic[36], Harvard[37] ve CSTR[38] sistemleridir.

LDC veri tabanları, duygusal konuşma, ses ve metin içeren verilerin araştırmacıların dağıtımına sunulduğu setlerden oluşmaktadır. Konuşma örnekleri, öfkeli, mutlu ve tiksinti gibi farklı güçlü duygulara sahiptir ve sınır değerlerinin kısa bir konuşma içinde bile çarpıcı biçimde değişebileceği durumları içerir[39]. Bu durum üretilen sistemlerin objektif sonuçlar üretebilmesi için oldukça önemlidir. Belli

[35] *Emotional prosody speech and transcripts database from Linguistic Data Consortium(LDC),* http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2002S28 (01.01.2018).
[36] *CMU Arctic Database,* http://www.festvox.org/cmu arctic/, (01.01.2018).
[37] *Harvard Sentences Database in Open Speech Repository,* http://www. voiptroubleshooter.com/open speech/american.html, (01.01.2018).
[38] P.C., Bagshaw, S.M., Hiller ve M. A., Jack (1993), *Enhanced Pitch Tracking and The Processing of F0 Contours for Computer Aided Intonation Teaching,* Teknik Rapor: Center for Speech Technology Research, Edinburgh.
[39] H., Ba, N., Yang, I., Demirkol ve W., Heinzelman (2012), "Bana: A hybrid approach for noise resilient pitch detection", *Statistical Signal Processing Workshop (SSP)*, s. 369-372.

özellikteki kayıtlar ile çalışan sistemler bu tarz verisetlerinde oldukça düşük doğrulukta sınıflandırma yapabilmektedir.

LDC üzerinden dağıtımı yapılan İngilizce ve Türkçe veri tabanları aşağıdaki tablolarda verilmiştir (Tablo 3-4).

Tablo 3 İngilizce Dili ile Hazırlanmış Konuşma Veri Tabanları

İsmi	**Konuşmacı Sayısı**	**İçerik Sayısı**	**Konuşma Tipi**	**Kayıt Ortamı**	**Örnekleme Frekansı(Khz)**	**Konuşma Türü**
TI Digits	326	2500 rakam	Okuma	Sessiz oda	20	Kelime
TIMIT	630	6300 cümle	Okuma	Sessiz oda	16	Fonem
NTIMIT	630	6300 cümle	Okuma	Telefon	8	Fonem
RM1	144	15024 cümle	Okuma	Sessiz oda	20	Cümle
Switch Board	69	35 diyalog	Spontan konuşma	Telefon	8	Kelime
TI-46	16	19136 yalıtık kelime	Okuma	Sessiz oda	16	Cümle
ATC	100	30000 diyalog	Spontan	Radio frekansı	8	Cümle
ATIS2	351	12000 ses	spontan	Ofis ortamı	16	Cümle

Tablo 4 Türkçe Dili için Hazırlanmış Akustik ve Dil Modeli Veritabanları

İsmi	Konuşmacı Sayısı	İçerik Sayısı	Konuşma Tipi	Kayıt Ortamı	Örnekleme Frekansı(Khz)	Veritabanı Türü
Middle East Technical University Turkish Microphone Speech v 1.0(2006)	120(60 erkek, 60 kadın)	2462 cümle	Okuma	Sessiz oda	16	Akustik
Turkish Broadcast News Speech and Transcripts(2012)	-	130 saatlik konuşma verisi	Haber Kaydı	Stüdyo	16	Akustik
NIST Language Recognition Evaluation Test Set(2014)	-	215 saatlik konuşma verisi	Haber ve telefon görüşme kaydı	Telefon	8	Akustik
CoNLL Shared Task - Ten Languages(2015) METU-Sabanci Turkish Treebank	3444	4087 cümle	Okuma	Sessiz ortam	16	Akustik
IARPA Babel Turkish Language Pack IARPA-babel105b-v0.5(2016)	-	213 saatlik diyalog verisi	Spontan konuşma	Günlük ortam	8	Akustik
Multi-Language Conversational Telephone Speech(2017)	87 diyalog	18.6 saatlik diyalog verisi	Spontan konuşma	Telefon	8	Akustik

TS Wikipedia(2015)	-	1.6M sayfa metin	-	-	-	Dil Verisi
ECI Multilingual Text(1994)	-	20000 kelime	-	-	-	Dil Verisi

2.2.1 Resource Management Complete Set 2.0

DARPA tarafından hazırlanmış olan bu veriseti, sürekli konuşma tanıma sistemlerini geliştirmek ve değerlendirmek için hazırlanmış konuşmalardan oluşur. Konuşmacıya bağımlı eğitim verisi ve konuşmacıdan bağımsız eğitim, test ve değerlendirme verilerini içerir. Konuşmalar deniz kaynaklarının yönetimi görevinden elde edilmiştir. 160'dan fazla konuşmacı ve 25000 civarı kelimeden oluşmaktadır. 16 bit çözünürlükte, 16 khz konuşma dalga formu ile kaydedilmiştir. ISO-9660 veri formatına uygundur.

2.2.2 TIMIT ve Türevleri

Farklı cinsiyet ve lehçelerdeki Amerikalı konuşmacılardan alınan sesli ifadeler derlemesidir. Verilen bir öğenin tasvir edilmesi istenilmiştir. Akustik-fonetik bilgi içerir ve otomatik konuşma tanıma sistemlerinde kullanılmak üzere hazırlanmıştır. Texas Instrument ile MIT tarafından hazırlandığı için TIMIT olarak isimlendirilmiştir. Akustik-fonetik çalışmalar ve otomatik konuşma tanıma sistemi geliştirilmesi için okuma ile oluşturulmuş bir verisetidir. Ses tanımanın yanında, konuşmacı tanıma uygulamalarda da kullanılabilen birkaç büyük verisetinden biridir. Bir ses kabininde geniş bant kulaklık ile tek seansta kaydedilmiştir. 630 konuşmacı 8 ana lehçe ile Amerikan İngilizcesi aksanında kullanarak kayıt alınmıştır. 16 bit çözünürlük 16 khz

konuşma dalga formu ile kayıt yapılmıştır ve her bir kelimenin sesli ve metinsel karşılıklarını içermektedir.

TIMIT veritabanının CTIMIT, FFMTIMIT, HTIMIT, NTIMIT, VIDTIMIT gibi türevleri bulunmaktadır. Bu veritabanlarında sesler kablolu veya kablosuz telefon gibi farklı kayıt ortamlarından toplanmıştır. TIMIT gibi türevleri de tek oturumlu sistemlerdir, her bir kullanıcı bir kere kayıt edilir. Bu sebeple konuşmacı tanıma problemlerinde kullanılabilir bir veriseti değildir.

Bu türevler içinde en yaygın kullanılan VIDTIMIT video içeriğine sahip 43 konudan oluşan bir verisetidir. 3 oturumludur. Otomatik konuşma tanıma ile birlikte görüntü işlemeyi de kullanan yöntemler için faydalı bir settir. 2 sistemin entegre çalışabildiği bu yapılarda daha yüksek doğrulukta tanıma yapılabilmesi mümkün olmaktadır.

2.2.3 TI46

Texas Instruments tarafından oluşturulmuştur. 16 konuşmacı(8 kadın, 8 erkek) tarafından TI alfabesini ve TI-20 alfabesi kullanılarak oluşturulan metinlerin sesletiminden oluşmaktadır. TI-20, 10 adet komut(gir, sil, git, yardım, hayır, robot, dur, başlat, evet) içerir. TI alfabesi ise a-z arasında 26 harften oluşan bir alfabedir. Her bir sözcük öğesi için her bir konuşmacı 10 kere tekrar eder ve bu veri eğitim seti için kullanılır. Test seti için her öğe 8 tekrarlı olarak kaydedilir.

2.2.4 SWITCHBOARD

Çok konuşmacılı telefon konuşmalarını içeren geniş bir veritabanıdır. Konuşma tanıma ve dil araştırmalarında kullanılmak üzere hazırlanmış olmasına rağmen çok çeşitli konuşmacı içermesi, konuşma verileri, telefon setleri ve kayıt koşulları sebebiyle konuşmacı tanıma uygulamalarında da kullanılabilir bir veriseti haline gelmiştir.

Ortalama olarak 6 dk'lık konuşmalardan 2430 adet telefon görüşmesini içermektedir. Her iki cinsiyetten 500 konuşmacının seslendirdiği 240 saatlik kayıt ve bunun 3 milyon kelimeden oluşan yazı halini içermektedir. Amerikan İngilizce aksanında seslendirme yapılmıştır.

Veriler 8 khz 8 bit olarak kodlanmıştır ve 2 kanallı olarak kayıt edilmiştir. Konuşmacı kimliğinin belirlenmesi problemleri araştırmalarında kullanılabilir özelliklere sahiptir. Konuşmacı karakteristik özelliklerini içeren sesler içeren bir verisetidir. Her bir yazılı metinde zaman hizalamasına ihtiyaç duyulur, her bir kelimenin bağlangıç zamanı ve süresini saniye olarak hesaplaması yapılır. 48 kişiden 20 kereden daha fazla sayıda ortalama olarak 1 saatlik konuşma verisi elde edilir ve bunlar eğitim ve modelleme için kullanılabilmektedir. 100 den fazla kişi 10 kere civarında veri üretimi sağlanmıştır.

Veriseti hazırlanması sürecinde her bir telefon görüşmesi için konuşan kişinin demografik özellikleri, tarih/saati, telefon görüşmesine ilişkin bilgi ilişkisel veritabanlarına kaydedilir. Bu şekilde veriseti dağıtımı yapılır. Katılan konuşmacıların, ses lehçe ve konuşma tarzının yanında yaş, cinsiyet, eğitim, mevcut konut ve ikamet yerleri gibi bilgiler de alınmıştır.

2.2.5 ATIS

ATIS veritabanı otomatik konuşma tanıma sistemlerinin kelime hata performanslarının değerlendirilmesinde yaygın olarak tercih edilmektedir. Spontan konuşma tanıma uygulamaları için kullanılır.

2.3 PERFORMANS ANALİZİ

Konuşma tanıma sistemlerinde çalışmalar sonrasında elde edilen sonuçların değerlendirilmesinde birçok farklı yöntem bulunmakla birlikte çalışmaların genelince aşağıda gösterilmiş olan yaklaşım kullanılmaktadır. Yapılan testlerin sonuçları cümle, sözcük ve hassasiyet bazında değerlendirilmekte ve sonuçta sistemin hata değeri tespit edilmektedir.

$$Cümle\ Hata\ Oranı(SRR) = \frac{H}{N} X100$$

$$Sözcük\ Tanıma\ Oranı(WRR) = \frac{N - S - D}{N} X100$$

$$Hassasiyet(ACC) = \frac{N - S - D - I}{N} X100$$

N: Toplam tanınması beklenen cümle/kelime sayısı

H: Doğru tanınan cümle sayısı

S: Hatalı tanınan kelime sayısı

D: Sözcüğü yanlışlıkla silme sayısı

I: Araya fazladan koyulan kelimelerin sayısı

Formülasyon incelenecek olursa, cümle hata oranı en basit yöntemdir. Verilen cümlelerin doğruluk yüzdesi sistemin performansı olarak değerlendirilmektedir.

Sözcük tanıma oranı ise, araştırmalarda kullanılan yaygın yaklaşımdır. Basit bir değerlendirme yapmak yerine, toplam tanınacak kelime sayısından, hatalı tanınan kelime sayısı ve yanlışlıkla silinen kelime sayısının çıkarılması ile elde edilen değerin oransal karşılığı tanıma performansı değeri olarak belirlenmektedir. Son olarak hassasiyet değerinde ise, sözcük tanıma oranı hesabına fazladan koyulan kelime sayısı da katılarak sistemin değerlendirilmesi yapılmaktadır.

2.4 SES TANIMA İŞLEMLERİNDE KULLANILAN ARAÇLAR

Otomatik konuşma tanıma sistemleri oluşturmak için kullanılabilecek birçok farklı araç bulunmaktadır. Bu araçların tamamı açık kaynak kodlu olarak dağıtılmaktadır ve tüm araştırmacılar istedikleri modeli istedikleri geliştirme ortamında yapabilmektedirler. Bu araçlardan yaygın olarak kullanılanlar yöntemleri ile birlikte aşağıda verilmiştir.[40, 41]

- PRAAT[42], ücretsiz olarak dağıtımı yapılan ve tüm işletim sistemlerinde çalışabilen bir altyapıdır ve seslerin kaydedilmesi ve analizinde kullanılmaktadır. Mono ve Stereo kayıtları işleyebilmektedir.
- CSL[43], bilgisayarlı ses laboratuvarı tarafından sunulan ve ses ve sinyal işleme konularında oldukça geniş bir işlem gücü sunan hem donanımsal hem yazılımsal altyapıya sahip bir araçtır. Veri depolama için sağlam bir sisteme ve konuşma analizi için çok yönlü bir yazılım desteğine sahiptir.
- HTK ilk olarak Cambridge üniversitesinde geliştirilmiştir. 1993 yılında Entropic araştırma laboratuvarında tarafından HTK projesi satın alınmıştır ve geliştirilmeye devam edilmiştir. Sonrasında, projesinin son aşamasında Microsoft HTK aracının telif haklarını almıştır ve açık kaynak kodlu olarak

[40] Shipra J., Arora ve Hisar Rishi Pal, Singh (2012), "Automatic Speech Recognition: A Review", *International Journal of Computer Applications*.
[41] Karpagavalli, S ve Chandra, E., 2016, s. 393-404.
[42] *PRAAT, h*ttp://www.fon.hum.uva.nl/praat/ (01.01.2018).
[43] *CSL*, http://www.csl.uni-bremen.de/cms/en/research/automatic-speech-recognition, (01.01.2018).

araştırmacıların hizmetine sunulmuştur. Gizli Markov Model araç seti(HTK) markov modellerini oluşturmak ve değiştirmek için kullanılan taşınabilir bir araçtır. Aracın geliştirilmesi ve ilk kullanım zamanlarında sadece konuşma tanıma araştırmalarında kullanılan bir araç iken zamanla konuşma sentezi, karakter tanıma ve DNA dizilimi gibi problemlerin çözümünde de kullanılmaya başlanılmıştır. C dili ile hazırlanmış kütüphane modülleri ve araçlarını içerir. Konuşma analizi, HMM eğitimi, test ve sonuçların analizi için ayrı ayrı çözümler sunar. Hem sürekli hem karışık Gauss dağılımlarını kullanarak HMM modellerini oluşturmaya imkan tanır.

- SPHINX[44],Java dili ile geliştirilmiştir. Ses tanıma araştırmaları için esnek bir altyapı sunmaktadır. Ses tanıma araştırmalarında yaygın olarak tercih edilen, kullanıcı dostu arayüz sunmaktadır. Yazılım bilgisi gerektirmeden konuşma tanıma modülleri geliştirilebilmektedir.
- SCARF[45], bölütlenmiş koşullu rastgele durumlar yardımıyla ses tanıma yapmak üzere tasarlanmış bir yazılımdır.
- MICROPHONES, ses veritabanlarının kayıtlarında kullanılmaktadır. Sony ve I-ball gibi markalarında gürültüsüz ses kayıt yapmaya imkan tanıyan geliştirmeleri bulunmaktadır.
- RWTH(Aachen otomatik konuşma tanıma sistemi)[46], geliştirilmiş ATROS(iATROS)[47], İdiap Juicer[48] ve SHoUT[49] ses tanıma aracı bazı az bilenen araçlar da araştırmacıların hizmetine sunulmuştur[50].
- SRI dil modelleme araç seti(SRILM)[51], konuşma tanıma, istatistiksel etiketleme, segmentasyon ve makine çevirisinde kullanılmak üzere istatiksel dil modelleri oluşturmak için kullanılır.

[44] *SPHINX,* https://cmusphinx.github.io/ (01.01.2018).
[45] Geoffrey, Zweig ve Patrick, Nguyen (2009), *SCARF: A Segmental CRF Speech Recognition System*, Teknik Rapor-Microsoft.
[46] *Aachen University Speech Recognition System,* https://www-i6.informatik.rwth-aachen.de/rwth-asr/ , (01.01.2018).
[47] *IATROS,* https://github.com/PRHLT/iatros-decoder, (01.01.2018).
[48] *IDIAP Juicer,* https://github.com/idiap/juicer, (01.01.2018).
[49] *SHoUT,* http://shout-toolkit.sourceforge.net/ (01.01.2018).
[50] C, Gaida, P., Lange, R., Petrick, P., Proba, A., Malatawy ve A.D., Suendermann-Oeft (2014), *Comparing open-source speech recognitions toolkits,* Teknik Rapor, Stuttgart.
[51] *SRILM,* https://www.sri.com/engage/products-solutions/sri-language-modeling-toolkit (01.01.2018).

SRI araç seti,

- Dil modellerini uygulayan, veri yapılarını ve çeşitli yardımcı işlevleri destekleyen bir dizi C++ sınıf kütüphanesi,
- Dil modellerini eğitmek ve bunları metin üzerinde veri, etiketleme veya segmentlere ayırmada test etmek gibi standart görevleri yerine getirmek için kitaplıkların üzerine kurulmuş bir program,
- Küçük görevleri kolaylaştıran çeşitli komut dizisi topluluğu bileşenleri içermektedir.
- UNIX ve Windows platformlarda çalışır. Çeşitli istatistiksel modelleme uygulamalarında kullanılır.

SRI araç seti, aşağıdaki poblemleri çözmek için kullanılabilmektedir:

- Konuşma Tanıma: SRI Otomatik konuşma tanıma motoru (DynaSpeak), konuşma metni hizalaması aracı SailAligni, Makine çevirisi araçları RWTH-ASR, Julius[52], Sphinx ve LIUM[53] gibi tanıma sistemleri ile entegrasyon yapabilmektedir.
- Makine Çevirisi: Dil modelleme desteği, Cunei makine çeviri platformu sunulmaktadır.
- Etiketleme ve bölümlendirme fonksiyonlarına sahiptir.
- Belge işleme: El yazısı tanıma özelliği bulunmaktadır.
- Visual Studio, Python, OCaml gibi dil ve IDE'ler ile birlikte çalışabilmektedir.

- KALDI, sonlu durum transdüserlerine dayalı bir konuşma tanıma sistemidir. Tam tanıma modelleri oluşturmak için uygundur. Yeterli dokümanlara sahiptir ve birçok işlem için hazır komut dosyaları sunmaktadır. Bu sebeple, hızlı bir şekilde geliştirme yapabilmek mümkün olmaktadır[54]. C++ dili ile yazılmıştır,

[52] *JULIUS,* https://github.com/julius-speech/julius (01.01.2018).
[53] *TED-LIUM, http://www-lium.univ-lemans.fr/en/content/ted-lium-corpus (01.01.2018).*
[54] D., Povey ve A., Ghoshal (2011),"The Kaldi Speech Recognition Toolkit", *Proceedings of ASRU, s.* 1-4.

Apache Lisansı kapsamındaki bir araçtır. HTK ile benzer özelliklere sahiptir. Daha modern ve esnek kod geliştirme, kolay modifiye edilebilirlik ve genişletilebilirlik sağlanması amaçlanmıştır[55].

KALDI altyapısının en önemli özellikleri şunlardır:

- Sonlu durum dönüştürücüleri(finite state transducer) ile kod düzeyinde entegrasyon,
- Kapsamlı lineer cebir desteği,
- Genişletilebilir tasarım,
- Açık Kaynak Lisanslı,
- LDC altyapısında sunulan veri tabanları ile tam uyumlu geliştirme tarifleri bulunmaktadır. Tam uyum sayesinde, araştırmacılar serbest lisans ile birlikte, kendi kodlarını geliştirebilirler ve istedikleri örnek senaryolarını uygulayabilir durumdadırlar.

2.5 SES TANIMA UYGULAMA ÖRNEKLERİ

Otomatik konuşma tanıma sistemleri ile ilgili yapılan çalışmaların artması, bu çalışmaların sonucunda iyi sonuçlar alınması sebebiyle günümüzde endüstriyel ürün olarak da yaygınlaşmaya başlamaktadır. Ticari alanlarda ses tanıma sistemlerinin kullanıldığı alanlar ve bu alanlarda girdi ve çıktı parametreleri olarak nelerin belirlendiği aşağıda verilmektedir.

- Telefon ile iletişim sektörü: Kullanıcıların, firmaların çağrı merkezlerini aramaları halinde, isteklere herhangi bir operatör olmadan bilgisayar tarafından cevap verilmesi, problemlere çözüm üretilmesi uygulamalarıdır. Bu uygulamalarda girdi parametresi olarak kullanıcıların spontan konuşma sinyali

[55] *KALDI*, http://kaldi-asr.org/doc/about.html (01.01.2018).

olurken, çıktı bir komut dizisi veya sesin metne dönüştürülmüş hali olabilmektedir.

- Eğitim sektörü: Yabancı dil öğrenmek isteyen öğrencilerin o dile ait telaffuzların doğru olup olmadığının kontrolünün bilgisayarlar tarafından yapılması işleminde kullanılabilmektedir. Bu uygulamalarda girdi olarak ses sinyali verilirken, çıktı telaffuzların kontrolü olacaktır. Ayrıca, fiziksel olarak handikaplı kişilerin bilgisayar ile iletişimlerinin sağlanabilmesi için klayve benzeri girdi birimlerinin kullanılmaksızın bilgisayar ile iletişim sağlanabilmesi mümkün olabilmektedir. Bunun yanında bu alandaki en popüler konu olan sözlü anlatıların bilgisayarlar tarafından yazılı hale getirilmesi ve böylece insan hatalarından bağımsız doğru ve hızlı veri üretiminin sağlanması mümkün olabilmektedir.
- İnternet üzerinden eğitim sektörü: Bilgisayarlar ve video oyunları, kumar, hassas cerrahi işlemleri yapan uygulamalarda ses tanıma sistemleri kullanılabilmektedir.
- Ev hizmetleri: Fırın, buzdolabı, çamaşır ve bulaşık makinelerinde ses ile komut verme örnekleri bulunmaktadır.
- Askeri sektörler: Yüksek performanslı savaş uçaklarının, helikopterlerin, tankların yönetimi, hava trafik kontrolörlerinin eğitimi gibi süreçlerde kullanılabilmektedir.
- Yapay Zekâ Sektörü: Robotik araştırmalarda ses sinyallerinin işlenmesi projeleri ile entegrasyon kurulması çalışmaları yapılmaktadır.
- Sağlık Sektörü: Sağlık hizmetlerinin yürütümü, sağlık verilerinin dikte işlemlerinde kullanılmaktadır.
- Dikte İşlemleri: Sürekli konuşma dizilerinin metne dönüştürülmesi işlemlerinde kullanılır.
- Tercüme İşleri: Bir dilden başka bir dile çeviri faaliyetlerini yürüten gelişmiş uygulamalarda kullanılmıştır.

Yukarıda verilmiş olan tüm bu alanların yanında, büyük veri işleme ve bilgi işlem gücünün doğrusal olarak büyümesi ile birlikte otomatik konuşma tanıma sistemlerine ilişkin daha zorlu uygulamaların gerçekleştirildiği aşamaya geçilmiştir.

İnsanların yaygın olarak kullandığı mobil telefonlarda sesli arama, mobil cihazlarla ses ile iletişim kurma(Ör. Siri, Iphone, WinPhone Bing sesli arama, Android Google Now), ev eğlence sistemlerinin ses ile kontrolü(Kinect on Xbox) veya konuşma tanıma merkezli bilgi işlem uygulamaları gibi ileri seviye çalışmalara rastlamak mümkündür. [56]. Ayrıca bu alanda dikte sistemlerini, sesli kullanıcı arabirimlerini, sesli arama, arama yönlendirme, ev aletleri kontrolü, basit veri girişi, eller ve gözler ile çalışan uygulamalar, özürlü insanlar için öğrenme sistemleri gibi örnek uygulamalara da rastlanılmaktadır[57].

2.6 OTOMATİK KONUŞMA TANIMA SİSTEM TASARIMI ZORLUKLARI

Bölüm 2.5'te bahsedilmiş olan konuşma tanıma sistemlerinin en basit örneklerini dahi sıfırdan geliştirmek oldukça zaman alıcı çalışmaları ve karmaşık matematiksel problemlerin çözülmesini gerektirmektedir. Bunun yanında son bölümde bahsedilmiş olan ileri seviye geliştirme örnekleri için ise uzun süreli ekip çalışmalarına ve büyük veri seviyesinde verisetlerine ihtiyaç duyulmaktadır.

Bu alanda yeni bir çalışma yapılmak istenilmesi durumunda aşağıda verilmiş olan başlıklardaki gibi zorluklarla karşılaşılması olasıdır[58].

Konuşmanın insanlar tarafından anlaşılması durumu

İnsanlar bir sesi dinlerlerken konuşma hakkındaki daha önceki bilgileri de kullanırlar. Sözcükler keyfi diziler halinde kullanılmazlar, bir dil bilgisi yapısı vardır ve insanlar henüz gelmemiş kelimleri tahmin edebilirler. Ancak otomatik konuşma

[56] Karpagavalli, S ve Chandra, E., 2016, s. 393-404.
[57] J, Li, L., Deng, R.H., Umback ve Yifan, Gong (2015), *Robust Automatic Speech Recognition: A Bridge to Practical Applications*, *Academic Press*.
[58] Shipra J., Arora ve Hisar Rishi Pal, Singh (2012), "Automatic Speech Recognition: A Review", *International Journal of Computer Applications*.

tanıma sistemlerinde sadece ses sinyalleri vardır ve eldeki bu bilgiler kullanılarak kelime tanıması yapılması beklenir. Bunun temel problemlerde dahi çözüm üretmemesi sebebiyle, dil bilgisi yapıları kullanılarak modeller oluşturulabilir ve kelime tahminini geliştirmek için bazı istatistiksel modellemeler kullanılabilir. Ancak konuşmacıların ve dünyada ki bilgi birikiminin tamamının nasıl modellenebileceğine dair ciddi sorunlar vardır. Burada hedef tüm dünya bilgisini modellemek yerine bir insana en yakın nasıl bir model üretilebileceğinin araştırılmasıdır. En yakın model ile en iyi tanıma performansı yakalamak mümkün olabilmektedir.

Konuşulan dil ile yazılı dilin birbirine birebir eşit olmaması durumu

Konuşma tanıma sistemleri üzerine yapılan ilk çalışmalarda, konuşulan dil ile yazılı dil arasında fark olmadığı düşünülmekteydi. Ancak, insanlar konuşurken, bir yazılı metin oluşturma durumlarına göre daha az dikkat ederler ve konuşma performanslarında ciddi hatalar yapabilirler. Konuşma yazılı dilin aksine iki yönlü bir iletişimdir ve konuşmada kayıplar olabilmektedir. Normal bir konuşma tekrarlar, dil sürçmesi, konuşmanın ortasında konu değişiklikleri gibi durumlar olabilmektedir. Son yıllarda konuşulan dil ile yazılı dil arasındaki farkın önemli olduğu ve bu durumun ciddi olarak incelenmesi ve çözümlenmesi gereken bir problem olduğuna dikkat edilmesi gerektiği belirtilmiştir.

Gürültü Durumu

Konuşma sesleri, arka planda başka bir konuşmacı sesi, bir odada TV sesi ile birlikte veya bir dış ses ile birlikte karışık olmaktadır. Konuşma sinyalinin dışındaki tüm bu sesler gürültü olarak adlandırılmaktadır. Bu sebeple de konuşma tanıma sistemleri tasarımında bu gürültüyü filtreleyecek tasarım yapılmalıdır. Bilinen bu dış seslerin yanında konuşma sinyalinin yayıldığı alanda bulunan cisimlerden yankılanması ve bir kaç milisaniye sonra mikrofona ses olarak girdi olduğu durumlar olabilmektedir. Echo etkisi olarak adlandırılan bu durumda da otomatik konuşma

tanıma sistemlerinde önemlidir. Sesin sinyal halinde ki bu etkilenme durumları ciddi olarak değerlendirilmeli ve çalışmalarda dikkate alınmalıdır.

Beden Dili

İnsanlar birbirleriyle iletişim kurarlarken el kol hareketlerini, hareketli gözlerini veya duruş şekillerini de kullanırlar. Böylece söylenilen sözlerin anlaşılması oldukça kolaylaştırırlar. Klasik tipteki otomatik konuşma sistemleri bu tipteki bir veriyi işleyememektedi ve sadece ses verisi ile tanıma yapması beklenmektedir. Ancak insanların iletişim modellerine benzer yapılar oluşturabilmek için insan beden dilinin görüntü işleme teknikleri ile veya sensör bazlı sistemler yardımıyla alınarak işlenmesine yönelik çalışmalar yapılmaktadır. Böylece ses verisi ile görüntü verisi ortak kullanılarak ses tanıma performansları yükseltilebilmektedir.

Kanal Değişikliği

Farklı mikrofon tipleri, zaman içinden değişken gürültü ve konuşmacı değişenliğinden bilgisayarda ayrık temsiline kadar akustik ses içeriğini etkileyen herşey kanal değişkenliği olarak adlandırılır. Bu etki elde edilen verilerde değişkenliğe sebep olmaktadır.

Konuşmacı Değişkenliği

İnsanların sesleri birbirlerinden farklı olduğu gibi, tek bir kişinin de farklı zamanlarda sesleri birbirinden farklı olmaktadır. Konuşmacıların aşağıda verilmiş olan temel özelliklerindeki değişkenlikler tanıma sistemi modellemesinde dikkate alınması gereken bir unsurdur. Aşağıda konuşmacı değişkenliğine ilişkin farklılıları doğurabilecek temel özellikler verilmiştir.

Konuşma Stili

Tüm konuşmacılar farklı şekilde konuşurlar, kelimeleri farklı şekilde telaffuz ederler ve telaffuz şekli farklı durumlarda da birbirinden farklılık gösterirler. Evde veya kamusal alanlarda bile insanlar farklı şekilde konuşmaktadırlar. Ayrıca üzgün olma, korku, süpriz veya öfke gibi duygusal durum farklılıklarında da birbirinden farklı şekilde konuşmaktadırlar.

Gerçekleme

Aynı kelimenin tekrar tekrar telaffuz edilmesi halinde ortaya çıkacak konuşma sinyali asla aynı olmayacaktır. Üretilen akustik dalga da farklılıklar çıkacaktır. Konuşmanın gerçekleşmesi zaman içinde değişkenlik gösterebilmektedir.

Söyleyiş Tarzı(Ağız)

Ağız, bir dil içerisindeki farklı gruplarda birbirinden farklı unsurlar barındıran bir söyleyiş yapısıdır. Bölgesel lehçe, konuşmacının yaşamış olduğu bölgeye göre kelime dağarcığındaki, telaffuzundaki ve gramer özelliklerindeki farklılığı gösterir. Sosyal lehçe, konuşmacının dâhil olduğu sosyal çevre sebebiyle kelime dağarcığı, telaffuz ve grammer özelliklerini içerir. Çoğu problem için, lehçe farklılıklarını gözetmek mecburiyeti bulunmaktadır. Hatta aynı dilin iki lehçesini birbirinden farklı dillermiş gibi el alan araştırmalara da rastlanılmaktadır. Aynı dil içindeki farklı lehçeler arasından ciddi değişkenlikler görülmesi sebebiy, geliştirilen uygulamalarda en yaygın lehçe, İstanbul lehçesi gibi bir seçim yapılarak çalışmalar yapılması daha iyi sonuçlar alınmasını sağlamaktadır.

ÜÇÜNCÜ BÖLÜM

DERİN SİNİR AĞI TEMELLİ SES TANIMA SİSTEMLERİ

Derin öğrenme, hiyerarşik mimarilerdeki birçok katmandan oluşan bilgi işlem aşamalarının yapı sınıflandırması ve özellik veya temsil öğrenimi için kullandığı bir makine öğrenme tekniğidir[59].

Derin öğrenme, makine öğreniminin bir alanıdır ve yapay sinir ağlarının gelişiminde önemli bir adımdır. Yapay zeka, grafik modelleme, optimizasyon, model tanıma veya sinyal işleme alanlarında kullanılmaktadır. Çok katmanlı makine öğrenme teknikleri kullanılarak verilerden gözetimli veya gözetimsiz olarak anlam çıkarma işlemidir. Birden fazla gizli katmana sahip yapay sinir ağıdır. Sinir ağlarının temel yapısı basit bir kavramsal yapı olarak kabul edilmesine rağmen, derin mimariler modellerin kapasitesindeki artık ve eğitimdeki zorluklarla beraber ciddi ilerlemeler kaydedilmesini sağlamıştır[60].

Derin sinir ağlarının kullanımı ile klasik yapay sinir ağı yaklaşımlarına kıyasla daha iyi başarılar yakalanmıştır. Yapay sinir ağlarının popüler olduğu 90'lı yıllara göre donanım olanaklarının gelişmesi ile birlikte o zaman tek katmanlı sinir ağını eğitmek ciddi zaman alırken şimdi çok katmanlı sinir ağları daha kısa sürede eğitilebilmektedir.

---

[59] Ji, Wan, Dayong, Wang, Steven, Hoi, Pengcheng, Wu, Jianke, Zhu, Jintao, Li ve Yongdong, Zhang (2014), "Deep learning for content-based image retrieval: A comprehensive study", *Proceedings of the 22nd ACM International Conference on Multimedia:* ACM, s. 157-166, Orlando.

[60] Anwer, Ali ve Mahmood, Ogur (2017), *Derin Öğrenme Yöntemleri ile Göğüs Kanseri Teşhisi,* THK Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Ankara.

Böylece daha yaygın hale gelmiştir. Günümüzde GPU mimarisinin derin öğrenme süreçlerine dahil edilmesi, paralel matris işlemlerinde olduça hızlı sonuçlar alınabilmesi sayesinde daha kısa eğitim süreleri yakalanmıştır[61].

İnternet ortamında dağıtımı yapılan verinin büyümesi, bu verilerin eğitim aşamasında kullanılabilir olması ve GPU mimarisi ile oldukça iyi modeller ortaya çıkmıştır. Facebook ortamında günde ortalama 350 milyon görüntü paylaşılırken, Youtube üzerinde her dakika da 100 saatlik video verisi yüklenmektedir. Derin öğrenmede verilerden deneyim kazanıldığı için bilgisayarın problemleri daha iyi bir şekilde çözebilmesi için gerçek dünyayı daha iyi yansıtacak verilerin kullanılması oldukça önemlidir. Bazı problemlere yönelik olarak internet üzerinde dağıtımı yapılan veriler artık gerçek dünyayı birebir yansıtacak duruma gelmiştir. Böylece iyi bir eğitim ile faydalı uygulamalar üretilebilmektedir. (Ör. Sesi metne dönüştüren uygulamalar)

Derin öğrenmenin problem çözme yeteneği YSA' nın doğrusal olmayan problemleri çözebilme yeteneğinden gelmektedir. Modelin derinliğine paralel olarak artan katman sayısı karmaşık problemlerin çözümüne zemin hazırlamaktadır. Ayrıca derin öğrenme modelleri az sayıda değişiklik ile farklı alanlardaki problemlerin çözümünde esneklik sağlamaktadır. Konuşma tanıma uygulamalarında özelliklerin çıkarılması, seslerin sınıflandırılması gibi işlemlerde derin öğrenme yapıları kullanılmaktadır[62].

3.1 DERİN SİNİR AĞI TÜRLERİ

Derin sinir ağları günümüzdeki birçok farklı problemin çözümünde araştırmacılar tarafından kullanılmaktadır. Ancak yapı olarak benzese de çalışma sistemlerinde ki farklılıklardan kaynaklı olarak farklı türlerde sinir ağı modelleri

[61] YİĞİT, Altuğ (2017), "*İş Süreçlerinde İnsan Görüsünün Derin öğrenme ile destekleme",* Trakya Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi.
[62] A.g.e., Yüksek Lisans Tezi

bulunmaktadır. Her bir konunun çözümünde daha iyi sonuçlar üreten türde bir sinir ağı yapısı olabilmektedir. Bu modellerden en sık kullanılanları aşağıda verilmiştir[63]:

- Konvolüsyonel Sinir Ağları
- Tekrarlı Sinir Ağları
- Kısıtlı Boltzman Makinesi
- Derin İnanç Ağları
- Derin Konveks Ağlar
- Derin Sinir Ağları

Her bir ağ türü için detaylı açıklamaları bu bölümde anlatılmıştır. Böylece ortaya konulan problem için en uygun derin sinir ağının seçilmesi veya birkaç modeli birlikte kullanan hibrit modellerin yaratılması mümkün olabilecektir.

3.1.1 Konvolüsyon Sinir Ağları

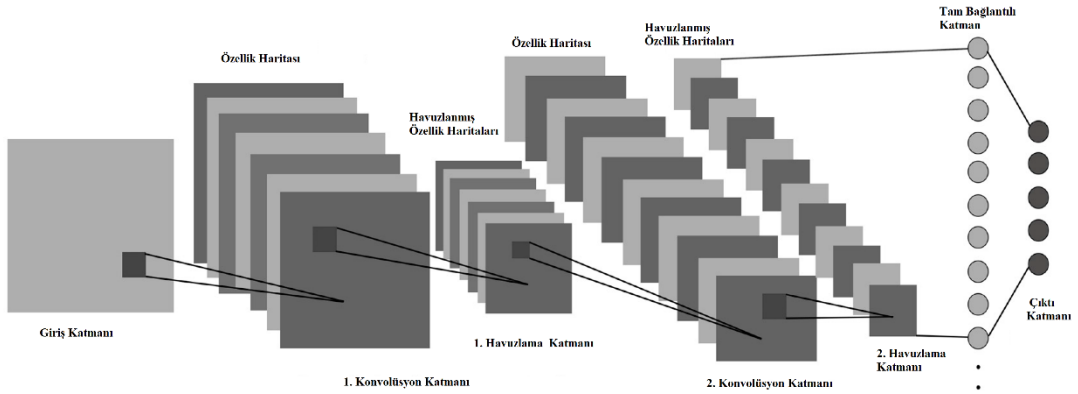
Konvolüsyon sinir ağı, her bir modülün konvolüsyon katmanı ve örnekleme katmanından oluştuğu ayrık yapıdaki sinir ağı türüdür. Her bir konvolüsyon katmanı üst üst yapılandırılır. Bu türdeki ağlar, dönüşümlü konvolüsyon katman ve küçültme(pooling) katmanından oluşan özel bir ağ yapısıdır[64].

Konvolüsyon katmanı birçok ağırlığı paylaşır ve çıkış katmanı pooling(örnekleme) tarafından işlenir ve veri miktarı küçültülerek bir sonraki konvolüsyon katmanına verilir. KSA için girdi katmanında bir dizi özellik haritası

[63] Renu, Karule ve M.A., Potey (2015), "Deep Architecture for Speech Processing", *International Journal of Innovative Research in Computer and Communication Engineering*.
[64] Ossama, Abdel-Hamid (2014), "Convolutional Neural Networks for Speech Recognition", *Audio, Speech and Language Processing*, s. 1533-1545.

verilmelidir. Seslere ilişkin çıkartılan özellik vektörlerinin organizasyonu sinir ağının çalışması için önemlidir.

Konvolüsyon ağlarda girdi katmanı sinyalin spektrogramı olarak düşünülebilir. Konvolüsyon ve örnekleme katmanları haritalar üzerinde aktivasyon özelliklerini uygular. Girdi, konvolüsyon ve örnekleme katmanları sonrasında çıktı katmanı tam bağlı sinir ağı olarak tasarlanabilir.



Şekil 11 Konvolüsyonel Sinir Ağı Şeması

Şekil-11'de standart bir konvolüsyon sinir ağı gösterilmektedir.3 katmanlı konvolüsyon, örnekleme ve tam bağlantılı son bölümden oluşur. Konvolüsyon katmanında her bir nöron önceki katmandan girdi alır, nöronlar aynı ağırlıktaki özellik haritalarını paylaşırlar. Yerel girdiler ağırlık matrisi ile çarpılır ve katmanlarda kullanılan filtreler ile çarpılarak bir sonraki katman iletilir.

Ağırlık matrisi tüm girdiler üzerinde uygulanır ve konvolüsyon katmanı farklı lokal desenlerden lokal filtrelere göre üretilen özellik haritasını içerir. Konuşma tanıma için girdi vektörü 2 boyutludur ve frekans ile domain ekseni bulunur. Konvolüsyon katmanından sonra bir örnekleme katmanı bulunur ve konvolüsyon katmanından local bir bölüme örnekleme yapar.

Örnekleme işlemi maximum ve ortalama alma fonksiyonu şeklinde olabilmektedir. Maximum alma işlemi her bir örnekleme bölgesi için o bölgedeki en yüksek değer alma işlemidir. Örnekleme işlemi sadece hesaplama karmaşıklığı azaltmak için değildir. Gürbüz bir sistemin yaratılmasını da sağlar.

Düşük ve yüksek frekans bölgelerinde konuşma sinyallerinin davranışları değişiklik gösterir. Yapılan çalışmalar konvolüsyon tabakasındaki gizli ünitelerin sayısının artmasının otomatik konuşma tanıma sistemlerinde kelime hata oranını azalttığı gösterilmiştir.

3.1.2 Tekrarlayan Sinir Ağları(TSA)

Tekrarlayan sinir ağları derin mimarinin bir türüdür. Ses ve metin şeklinde sıralı verilerin modellenmesinde oldukça güçlüdür. Tekrarlayan sinir ağları derin mimaride verinin uzunluğu kadar derin olabilmektedir. Konuşma tanıma araştırmalarında TSA'larının SMM ile birlikte kullanılması yerine uçtan uca sadece TSA ile tasarlanması yoluna gidilmektedir[65].

Bu modelde, sinir ağının derinliği zaman bazlı olarak ifade edilmektedir. Bu sebeple sonsuz derinlik olarak tanımlanabilmektedir. TSA'nda her bir katman bir zaman dilimini ifade eder ve her bir zaman dilimine ait katman bir sonraki zaman dilimindeki katman ile bağlıdır. Her bir adımın geçiş fonksiyonu sabittir.

Standart bir TSA'nın matematiksel denklemi aşağıdaki gibidir.

$$h_i = \sigma\left(W_{hh\,h_{i-1}} + W_{hx}x_i + b_h\right)$$

[65] Renu, Karule ve M.A., Potey (2015), "Deep Architecture for Speech Processing", *International Journal of Innovative Research in Computer and Communication Engineering*.

$$y_i = W_{yh} h_i$$

hi = i'inci adımdaki gizli katman

xi = i inci zaman adımındaki girdi katmanı

yi = i'inci zaman adımındaki çıktı katmanı

W = ağırlık matrisi, b bias

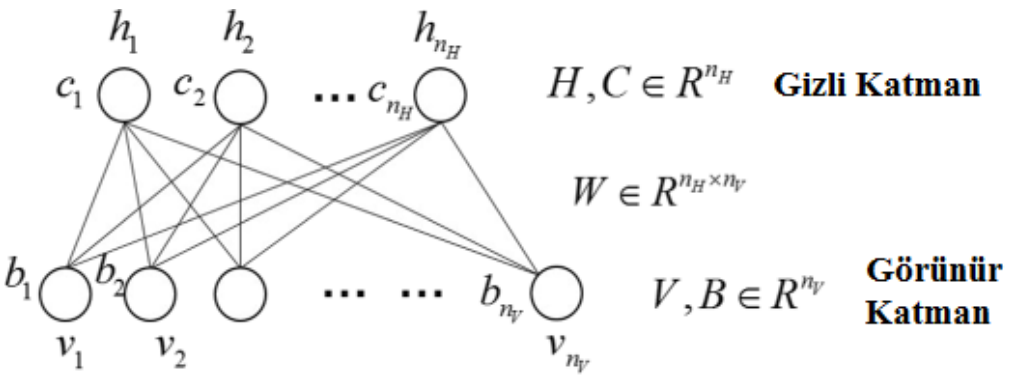
Teorik olarak tekrarlayan sinir ağları güçlüdürler ancak eğitim aşamaları zordur ve bu zorluğu aşmak için eğitim kapasiteleri güçlendirilmelidir. Bu işlemin yapılabilmesi için çeşitli optimizasyon teknikleri uygulanır ve ağ mimarisini değiştirerek daha kolay bir eğitim yapılması sağlanılmaya çalışılır.

3.1.3 Kısıtlı Boltzman Makinesi(KBM)

Derin Boltzman makineleri üreten bir modeldir ve üretken veya ayrımcı yaklaşımları uygulamak için kullanılırlar. Araştırmalarda, üretken yaklaşımların konuşmacı tanımada, ayrımcı yaklaşımların konuşma tanıma da iyi olduğu gösterilmiştir.

Kısıtlı Boltzman makineleri Bolztman makinelerinin bir türüdür. Bir gizli olmayan girdi katmanı, bir gizli katman ve bias biriminden oluşan olasılıksal bir sinir ağıdır. Her bir girdi katmanındaki birimler, gizli katmandaki tüm birimler ile bağlıdır ve bias katmanındaki tüm birimlerde çıktı katmanına bağlıdır. Kısıtlı olma durumu bu aşamada devreye girmektedir. Ağ içinde 2 tane gizli katman veya girdi katmanı bulunmamaktadır. Böylece öğrenme süreci daha kolay bir hale getirilmeye çalışılmaktadır.

Kısıtlı sinir ağları, derin inanç ağlarının yapıtaşıdır ve araştırmacılar için iyi bir modelleme türüdür. Olasılıksal dağılımı temsil eden üretken bir modeldir. Şekil-13'te yapısı gösterilmiştir.



Şekil 12 Kısıtlı Boltzman Makinesi Modeli

Bir kısıtlı Bolztman makinesi, gözlemlenebilir verileri temsil etmek için m adet görünür birimi içeren V= (V1, V2, ... Vm) ve n gizli birimi temsil eden H = (H1, h2, .. Hn) arasındaki bağımlılığı temsil eder.

İkili KBM, rastgele değişkenler (V,H); (v,h) ε $\{0,1\}^{m+n}$ değerlerini alınır ve aağıdaki eşitliğe göre ortak olasılık dağılımını verir.

$$p(v,h) = \frac{1}{z}\, e^{-E(v,h)}$$

Enerji Fonksiyonu $E(v,h) = \sum_{i=i}^{n}.\sum_{j=1}^{m}.\, w_{i,j}v_j - \sum_{j=1}^{m} b_j v_j - \sum_{i=1}^{n} c_i h_i$, i ε {1,2, .. , n} ve j ε {1,2, .. , m}, wi,j Vj ile hi değeleri arasındaki kenarla ilişkili gerçek değerli ağırlık ve bj ile ci gerçek değerli bias değeri (j. görünür ve i. gizli değişken ile ilgilidir)

3.1.4 Derin İnanç Ağları

Derin inanç ağları kısıtlı Boltzman makinelerinin birden fazlası bir araya getirilerek oluşturulur ve ciddi miktarda veri kullanılarak eğitilir. Derin inanç ağları eğitim verisinin derin hiyerarşik bir temsilini öğrenebilen modellerdir. Bu ağlar gözetimsiz öğrenme ile eğitimin mümkün olduğu yapıdadırlar.

Her bir katman için kısıtlı boltzman makineleri kullanılır. Eğitim sonrasında sınıflandırmayı gerçekleştirme esnasında gözetimli şekilde eğitime devam edilebilir.

Derin inanç ağlarının gözetimsiz olarak eğitilmesi süreci aşağıdaki adımları içerir.[66]

- Giriş katmanında gizli olmayan bir katman olarak tasarlanır ve ham girdiyi modelleyen bir kısıtlı boltzman makinesi olarak eğitilir.
- Birinci katman, ikinci katmanda veya veri olarak kullanılacak girdinin hazırlanması için kullanılır.
- İkinci katmanda veriler, örneklem alınarak veya gizli katmanların ortalama aktivasyonları hesaplanarak elde edilir. İkinci katmanın kısıtlı boltzman makineleri olarak eğitilmesinde, ikinci katmanda üretilen veriler kullanılır ve eğitim süreci tamamlanır.
- Yukarıdaki süreçler tüm katmanlarda tekrarlı olarak yapılır. Ya örnek alınarak ya da ortalama değerler kullanılarak veriler elde edilir.

Bu işlemler sırasında tüm parametler her adımda gözetimli eğitim kriterlerine göre ayarlanır.

[66] Deng, Li (2012), *Three Classes of Deep Learning Architectures and Their Applications*, Araştırma Raporu-Microsoft Research, Washington.

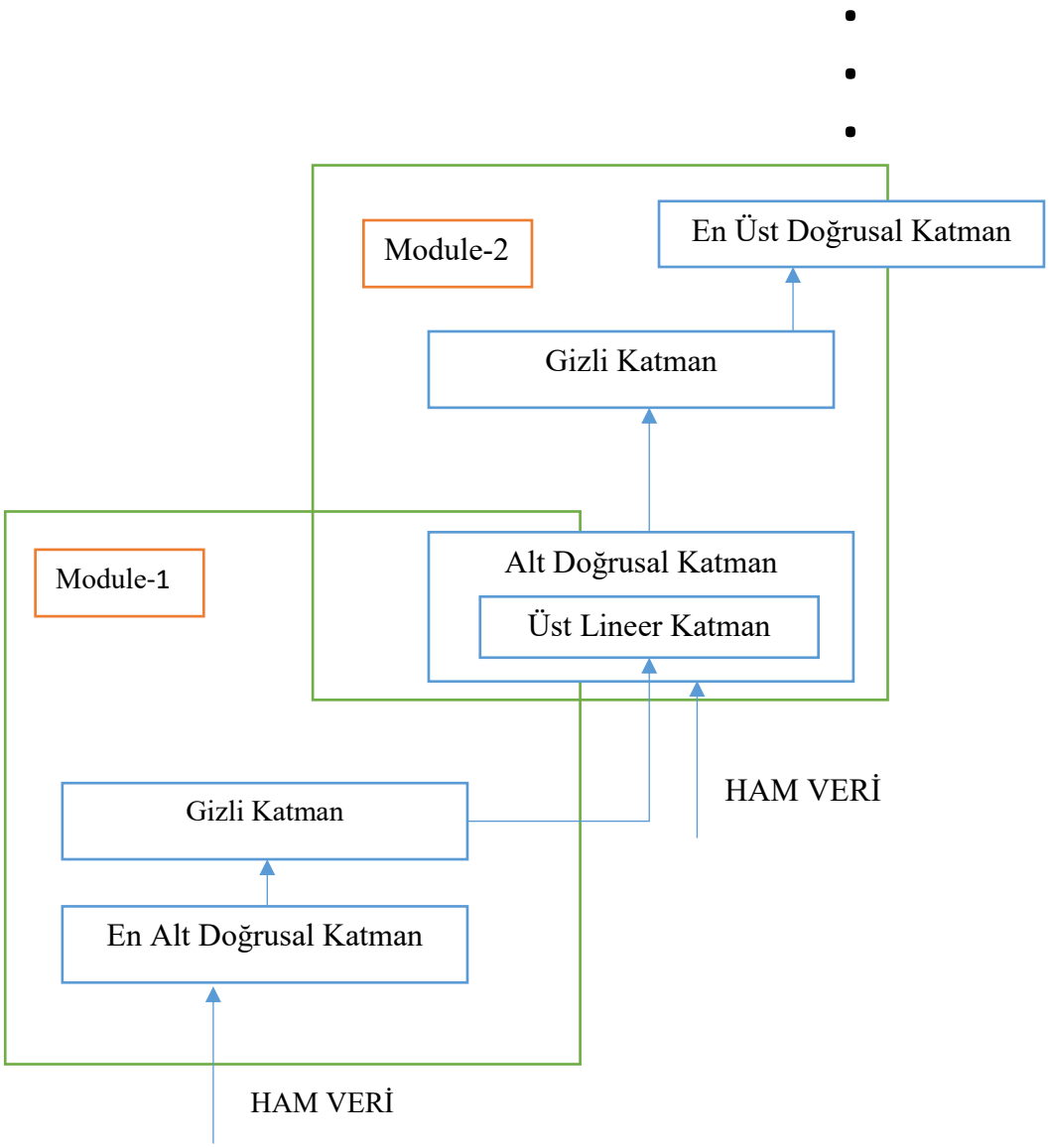
3.1.5 Derin Konveks Ağlar(DKA)

Derin inanç ağları konuşma tanıma da dahil olmak üzere bir çok sınıflandırma probleminde güçlü bir yaklaşım olarak önerilmiştir. Ancak bu ağların eğitiminde hesaplamalar oldukça zordur. Bu ağları eğitmek için kullanılan geleneksel tekniklerden en yaygın olanı kademeli azaltma algoritmasıdır ve bu işlemin paralel olarak yapılması mümkün değildir.

2012 yılında bu sorunu çözmek üzere farklı bir derin mimari model önerilmiştir. Bu mimari de Derin Konveks ağlar olarak adlandırılmıştır[67]. Bu mimarinin arkasındaki fikir, fonksiyonlar veya sınıflandırıcıların öncelikle basit modüller halinde tasarlanması sonrasında bunların bir araya getirilmesi ile daha karmaşık fonksiyonlar veya sınıflandırıcıların oluşturulmasıdır. Böylece kademeli bir geçiş ile istenilen düzeyde tanıma performansı yakalanabilmektedir.

Konveks denilmesinden kasıt çıkış ağırlıklarının dışbükey optimizasyonuna tabi tutulmasıdır. Aşağıdaki şekilde bu modelin bir örneği gösterilmektedir[68].

[67] A.g.m., Araştırma Raporu
[68] Li Deng ve Dong, Yu (2011), *Deep Convex Net: A Scalable Architecture for Speech Pattern Classification*, INTERSPEECH Microsoft Research.



Şekil 13 Derin Konveks Ağ Yapısı

DKA, her modülün iç katman içeriklerine göre birbirinin devamı şeklinde tasarlandığı yapıdan oluşur.

- İlk doğrusal katman, giriş özelliklerinin boyutuna eşit olacak şekilde doğrusal girdi birimlerini içerir. Ses tanıma problemi için girdi verileri bir konuşma dalga formu veya güç spektrumu olabilmektedir.
- Gizli katman, doğrusal olmayan birimlerden oluşur.
- İkinci doğrusal katman sınıflandırmanın yapıldığı katmandır.

3.2 ÖĞRENME SÜRECİ

Makine öğrenmesi sistemlerinden birisi olan derin öğrenme sistemlerinde verilerden kendi kendine öğrenme mümkündür. Sistemlerin eğitimleri aşamasında kullanılacak verileri oldukça önemlidir ve problemi ne kadar iyi yansıtıyor ise o kadar iyi sonuçlar alınmasını sağlayacaktır. Eğitim aşamasında ağ, bu veriler ışığında eğitilir ve ağırlık değerleri istenen sonuçlara göre güncellenir. Böylece en uygun matematiksel fonksiyon elde edilmiş olur. Bir problem için oluşturulmuş model, tekrar eğitilerek farklı bir problem çözümünde kullanılabilir.

Öğrenme yapıları elde bulunan veri miktarı, ağın eğitilebileceği sistem performansı, problemin türüne göre farklı şekillerde olabilmektedir. Her bir tür için ağın eğitimi yapılmakla birlikte ağın eğitim modeli ve süresi farklılık göstermektedir. Öğrenme süreçleri denetimli, denetimsiz ve takviyeli öğrenme olmak üzere üçe ayrılmaktadır[69].

3.2.1 Denetimli Öğrenme

Denetimli öğrenmede veriler makineye etiketli olarak gönderilmektedir. Bir verinin girdi olarak verilmesi halinde çıktı olarak ne çıkması gerektiği de sisteme verilir. Sınıflandırma ve regresyon problemlerinde kullanılmaktadır. Öğrenme modelinin amacı hata oranını düşürmektir. Hata oranının düşmesi, ağın ağırlık değerlerinin optimum noktaya getirilmesi ile mümkün olmaktadır. Ağın eğitiminin başlangıcında, minimum hata verebilecek şekilde bir tahmin yapılır. Sonrasında herbir adımda istenen değer ile çıktı değeri arasındaki fark hata değeri belirlenir. Çıkış katmanındaki tüm düğümlerin toplam hata miktarı maliyet fonksiyonları ile

[69] YİĞİT, Altuğ, 2017, Yüksek Lisans Tezi.

hesaplanmaktadır. Sonrasında ağırlık değerleri bu hata miktarına göre güncellenerek ikinci eğitim aşamasına geçilmektedir[65].

3.2.2 Denetimsiz Öğrenme

Denetimsiz öğrenmede, modele verilen veriler etiketsizdir. Denetimli öğrenmedeki gibi her bir girdi için beklenen çıktı değeri verilmez. Sistemin kendi kendine verileri anlamlandırması beklenmektedir. Ağın ağırlık güncellemesi başlangıçta belirlenen kurallara göre yapılmaktadır.

Derin öğrenme sistemlerinde amaç giriş ve çıkışlar arasındaki doğrusal olmayan haritalamanın çıkarılmasının yanında girdi vektörlerinin eğilimlerinin belirlenmesi ve öğrenilmesidir. Denetimsiz öğrenmede otomatik kodlayıcılar kullanılmaktadır. Bu sistem girdi verilerinin ifade ettiği anlamları belirlemektedir[70]. Böylece etiketsiz verilerin anlamlandırılması ve öğrenilmesi mümkün olmaktadır.

3.2.3 Takviyeli Öğrenme

Takviyeli öğrenme sürecinde, denetimsiz öğrenme olduğu gibi verilerin etkiletleri bilinmez. Sadece elde edilen çıktının doğru veya yanlış olarak değerlendirilmesi sisteme girdi olarak verilir. Her bir adımda doğru ve yanlış durumlarına göre ağırlıklar güncellenmektedir. Genellikle oyun geliştirme, örüntü tanıma gibi problemlerin çözümünde kullanılmaktadır.

[70] YİĞİT, Altuğ, 2017, Yüksek Lisans Tezi.

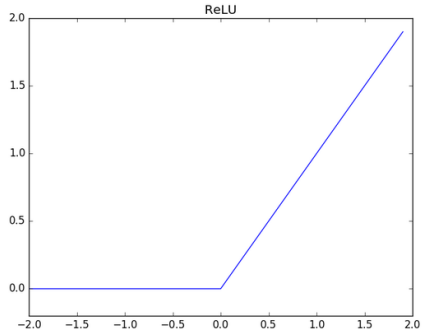
3.3 AKTİVASYON FONKSİYONLARI

Aktivasyon fonksiyonları yapay sinir ağlarında doğrusal olmayan fonksiyonlardır. Nöronların çıkışlarında doğrusal olmayan verilerin üretilmesini ve daha karmaşık fonksiyonları temsil edebilecek modellerin üretilebilmesini sağlar. Aktivasyon fonksiyonu olarak doğrusal bir fonksiyon seçilecek olursa, oluşturulan model sadece doğrusal problemleri çözebilecek şekilde olacaktır[71].

Nöronların ateşlenmesi, aktivasyon fonksiyonlarına göre seçilir ve eşik değerin altında kalındığı durumlarda ateşleme olmaz. Böylece bu nöronun çıkışı genel sistem çıkışına etki etmemiş olur. Araştırmalarda kullanılan birçok farklı aktivasyon fonksiyonu bulunmaktadır. Aktivasyon fonksiyonlarının seçimi probleme göre değişmektedir ve uzmanlık bilgisine veya istatistiki bilgiye ihtiyaç duyulmaktadır. Bu sebeple birçok farklı aktivasyon fonksiyonu incelenerek doğru bir seçim yapılmalıdır. Bu bölümde bazı aktivasyon fonksiyonlarının detayları verilmiştir.

3.3.1 ReLU

Fonksiyona girdi olarak verilen negatif değerlerin sıfır ve sıfırdan büyük değeri kendi değerine atayarak eşik değer oluşturan bir fonksiyondur. En önemli avantajlı



Şekil 14 ReLU Grafiği

[71] YİĞİT, Altuğ, 2017, Yüksek Lisans Tezi.

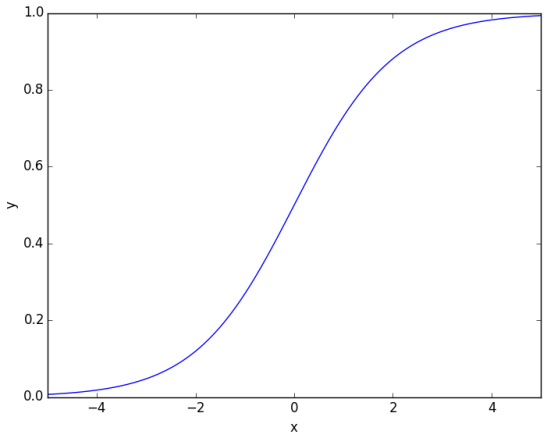
noktası hızlı bir şekilde aktivasyon işlemi yapabilmesidir. Aşağıda fonksiyonun denklemi ve grafiği gösterilmiştir[72].

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

3.3.2 Sigmoid

Sigmoid, verileri silmeden verilerin aşırı değerlenmesini azaltmanın bir yolunu temsil eder. Bağımsız değişkenleri 0 ile 1 arasında basit olasılıklara dönüştüren bir fonksiyondur ve çıkışların büyük kısmı 0 veya 1'e çok yakın olarak üretilmektedir. Aşağıda fonksiyonun denklemi ve grafiği gösterilmiştir[73].

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



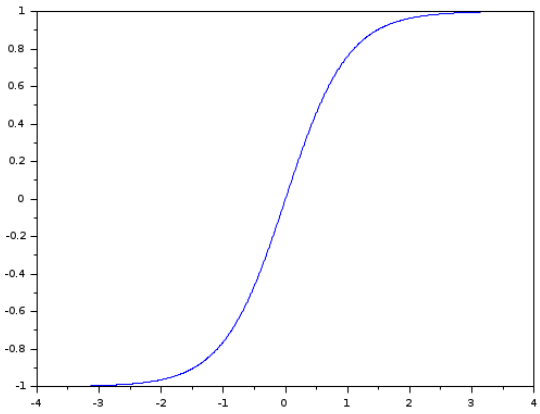
Şekil 15 Sigmoid Grafiği

[72] Anwer, Ali ve Mahmood, Ogur, 2017, Yüksek Lisans Tezi.
[73] YİĞİT, Altuğ, 2017, Yüksek Lisans Tezi.

3.3.3 Tanh

Hiperbolik bir fonksiyondur. Tanjant bir dik üçgenin ters ve bitişik kenarları arasındaki bir oranı temsil ettiği gibi, tanh hiperbolik sinüsün hiperbolik kosinüse oranını temsil eder. Normalize edilmiş aralığı -1 ile 1 arasındadır. Avantajlı olduğu husus negatif sayılarla daha kolay işlem yapabilmesidir. Aşağıda fonksiyonun denklemi ve grafiği gösterilmiştir[74].

$$f(x) = \tanh(x) = \frac{2}{1 + e^{-2x}} - 1$$

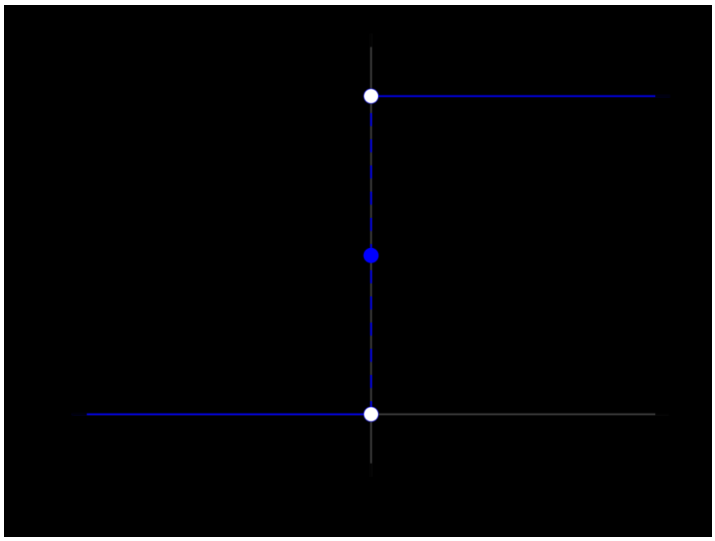


Şekil 16 Tanh Grafiği

[74] YİĞİT, Altuğ, 2017, Yüksek Lisans Tezi.

3.3.4 Adım(Step)

Gelen girdinin belirlenen bir eşit değerinin altında veya üstünde olması durumunda, nöron çıktısı 1 veya 0 değerini almaktadır[75]. Aşağıda grafiği gösterilmiştir.



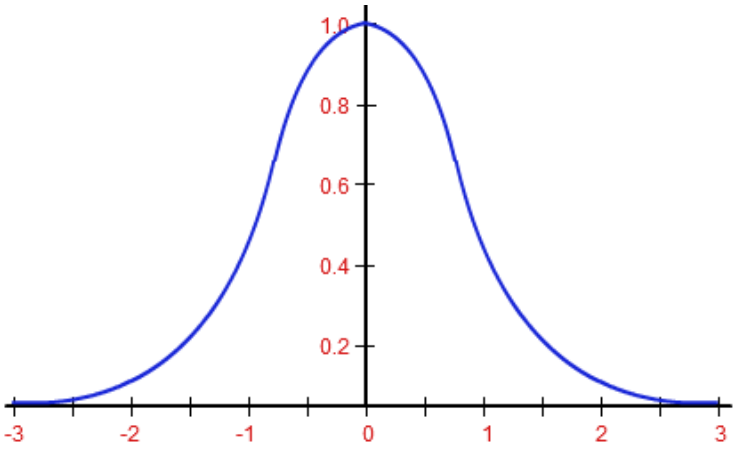
Şekil 17 Adım Fonsiyonu Grafiği

3.3.5 Gauss

Sürekli çan eğrisi şeklindedir. Net girdinin seçilen bir ortalama değerine ne kadar yakın olduğuna bağlı olarak nöron çıkışı(yüksek/alçak) sınıf üyeliği (1/0) olarak yorumlanır. Aşağıda fonksiyonun denklemi ve grafiği gösterilmiştir[71].

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}}\, e^{\frac{-(x-u)^2}{2\sigma^2}}$$

[75] Anwer, Ali ve Mahmood, Ogur, 2017, Yüksek Lisans Tezi.



Şekil 18 Gauss Grafiği

3.4 HATA FONKSİYONLARI

Eğitim sürecinde ağırlıkların güncellemesinde kullanılacak hata miktarını tespit etmek için kullanılan fonksiyonlardır. Amaç hata miktarını minimize etmektir. Araştırmalarda kullanılmakta olan birçok hata fonksiyonu bulunamaktadır. Bunun seçimi probleme göre olmaktadır[76]. Bu bölümde bu fonksiyonlardan bazıları detaylı olarak anlatılmıştır.

3.4.1 Ortalama Hata Karesi

Regresyon problemlerinde kullanılmaktadır. Matematiksel ifadesi aşağıdaki gibidir.

[76] YİĞİT, Altuğ, 2017, Yüksek Lisans Tezi.

$$OHK(w) = \frac{1}{N}\sum_{\text{İ}=1}^{N}(yi - \hat{y}i)^2$$

$$OHK(w) = \frac{1}{N}\sum_{\text{İ}=1}^{N}(yi - f(xi, w))^2$$

$$OHK(w) = \frac{1}{N}\sum_{\text{İ}=1}^{N}(ei)^2$$

Eğitim verileri içerisinden her bir i örneği için x_i girdisi ile w ağırlık değerleri birlikte alınır, ağın çıktı değeri ŷi değeri elde edilir. yi ile ŷi değeri arasındaki fark hesaplanarak e değeri bulunur. Bu değerin karelerinin toplamı alınır ve toplam sayıya bölünerek ortalama değeri bulunur. Performans ölçütü olarak kullanılmaktadır.

3.4.2 Ortalama Mutlak Hata(Mean Absolute Error)

Veri sayısı kadar hata değerinin mutlak değeri alınarak toplanır ve ortalaması hesaplanır. Aşağıda matematiksel modeli gösterilmektedir.

$$OMH = \frac{1}{N}\sum_{\text{İ}=1}^{N}|yi - \hat{y}i|$$

$$OMH = \frac{1}{N}\sum_{\text{İ}=1}^{N}|ei|$$

Bu formülasyon ile ŷi değerinin yi değerine mümkün olduğunca yakınsaması amaçlanır. Böylece hata değeri daha azaltılmış olacaktır.

3.4.3 Karekök Ortalama(Root Mean Square)

Veri sayısı kadar hatanın karesi alınır, toplanır ve sonucun tekrar karekökü alınır. Matematiksel formülü aşağıdaki gibidir[73].

$$KO = \sqrt{\frac{1}{N}\sum_{\text{İ}=1}^{N} |yi - \hat{y}i|^2}$$

$$KO = \sqrt{\frac{1}{N}\sum_{\text{İ}=1}^{N} |ei|^2}$$

Karekök ortalaması veya ortalama mutlak hata birbirinin yerine kullanılan metodlardır. Karekök ortalama farklı olarak tüm ağın hata dağılımının tam olarak yansıtılmasına imkân tanımaktadır. Büyük miktarda veri olduğu durumlarda bu yöntemin kullanılması daha uygundur.

3.4.4 Dereceli Azalma(Gradient Descent)

Eğim azalımı yapay sinir ağları ile derin öğrenme yöntemlerinde ağın öğrenmesini iyileştirme yöntemidir. Hata fonksiyonunun minimum olması için en iyi ağırlık ve bias değerlerinin bulunması amaçlanmaktadır. Bir ağ eğitiminde temel modeller için aşağıdaki adımlar uygulanmaktadır [77].

1- Başlangıçta rasgele bir ağırlık belirlenir.

[77] YİĞİT, Altuğ, 2017, Yüksek Lisans Tezi.

2- Eğitim aşamasında hata fonksiyonu sonucunu minimum yapacak şekilde ağırlık güncellenir.
3- Bir sonraki eğitim aşamasına geçilir. Ve bu işlemler hata çıktısı için belirlenen değere ulaşana kadar devam eder. Burada kritik husus bu hata değerinin hiçbir zaman sıfır olması beklenmemektedir. Bu durumda modelin veriyi ezberlediği sonucuna varılmaktadır.

Her bir epokta hem ağırlıklar hem de bias değeri güncellenir. Ağırlık güncellemesi, eğitim katsayısının hata fonksiyonunun ağırlığına göre türevi öğrenme oranı ile çağrılarak bulunan değere göre yapılır. Bulunan bu değer bir önceki değer ile karşılaştırılarak farkına göre bir sonraki adıma geçilir.

Eğitim oranı, öğrenme adımlarında ne kadar büyüklük ile güncelleme yapılacağını etkileyen bir değerdir. Küçük değerli bir öğrenme oranı belirlendiğinde iyi bir öğrenme yakalamak için çok fazla sayıda işlem ve zaman gerektirecektir. Büyük değerli bir oran belirlendiğinde ise istenen optimum öğrenme seviyesine erişilemeyecektir. Bu nedenle bu oranın en iyi şekilde seçilmesi beklenmektedir.

3.5 DERİN ÖĞRENME SİSTEMLERİ SINIFLANDIRMASI

3.5.1 Üretken Öğrenme Yaklaşımı(SMM-GMM)

Otomatik konuşma tanıma sistemlerinde, en yaygın öğrenme yaklaşımı Gauss Karışım Modeline dayalı gizli markov modellerine dayanmaktadır. Geleneksel konuşma tanıma sistemleri konuşma sinyallerinin ardışık yapısını temsil etmek için Gauss karışık model tabanlı gizli Markov modellerini kullanır. Bu modellemenin kullanılmasının sebebi konuşma sinyalinin durağan bir sinyal şeklinde veya kısa süreli

durağan sinyal şeklinde görüntülenebilmesidir. Böylece, ses sinyali kısa bir zaman ölçeğinde durağan bir süreç olarak ele alınabilmektedir.[78]

Her bir SMM durumu, ses dalgasının spektral bir temsili modellemek için Gauss karışımını kullanır.

GMM-SMM standart olarak $\theta = (A, B, \pi)$ şeklinde formüle edilir.

π = Olasılık Vektörü

A = (a_{ij}) bir durumun geçiş olasılık matrisi

B = {(b1, b2, .. , bn)} bj = her bir j durumunun gauss karşık modelini temsil eder.

Durumlar tipik olarak sesteki her bir fonemin alt segmenti ile ilişkilendirilir.

3.5.2 Ayrık Öğrenme Yaklaşımı(HMM-ANN)

Ayrık öğrenme, ayrık bir modelin üretken bir modele uygulanmasını ya da ayrık model ile eğitimi içerir. Tanıma sürecinin son katmanında doğrusal olmayan "softmax" fonksiyonunun çok katmanlı perceptron ile sinir ağlarında kullanılmasına denilmektedir. 1990'lı yıllarda popüler yaklaşımlardan biridir.

Çok katmanlı perceptronların çıktıları koşullu olasılık ile yorumlanabildiğinden, saklı markov model girdi olarak verilir ve iyi bir ayırt edici melez model yaratılmış olur. Esas olan, çok katmanlı perceptronların öğrenme zorluğundan dolayı üretilen araştırma dizisini saklı markov modele özelik olarak

[78] Karpagavalli, S ve Chandra, E., 2016, s. 393-404.

vermek yoluna gidilmiştir[79]. Geri yayılımlı türev temelli hata azaltma yöntemlerini kullanan sinir ağları, 1980'li yıllarda akustik modelleme için yaygın olarak kullanılmıştır. Sinir ağları saklı markov model gibi akustik özelliklerin istatistiksel özellikleri için herhangi bir varsayım yapmaz.

Sinir ağları, konuşma özelliklerini değerlendirerek tahmin yaptığı için etkili bir ayırt edicilik eğitimine imkan tanır. Ancak yalıtık sözcük tanıma gibi kısa tanıma birimlerinin sınıflandırılmalarında etkin olmalarına rağmen, sürekli konuşma tanıma birimlerinin tanınması problemlerinde yeterli başarı göstermez. Bunun genel sebebide zamansal bağımlılıkların modellenmesinde beceri eksikliği olması durumudur.

3.5.3 Derin Öğrenme Yaklaşımı(HMM-DNN)

Derin öğrenme, gözetimsiz özellik öğrenimi veya makine öğrenmesinin yerine kullanan bir kavram olmuştur. Konuşma tanıma için ana teknolojidir ve Gauss karışımlarının konuşma tanıma ve özellik kodlaması için kullanım yaygınlığını azaltmıştır.

Derin öğrenme ilk olarak üretken öğrenmenin bir türü olarak düşünülebilir. Çünkü verilerin üst düzey özellikleri ve istatistiksel dağılımlarının karakterize edilmesi için kullanılan derin mimarilerden oluşur. Ancak Bayes kuralının uygulandığı derin öğrenme mimarilerinin ayrık öğrenme yaklaşımına kaydığı bir yapıda oluşabilir.

Derin öğrenme yapısının birinci tipinde, derin otomatik kodlayıcılar, derin Bolztman makineleri, toplam ürün ağları, derin inanç ağlarının özgün biçimi ve alt katmanlarda çarpanlarına ayrılmış daha üst düzey Boltzman makinesi uzantısı olan çeşitli biçimleri bulunmaktadır.

[79] N., Morgan, Q., Zhu, A., Stolcke, K., Sonmez, S., Sivadas, T., Shinozaki, M., Ostendorf, P., Jain, H., Hermansky, D., Ellis, G., Doddington, B., Chen, O., Cretin, H., Bourlard ve M., Athineos (2005), "Pushing the envelope Aside Speech Recognition", *IEEE Signal Processing Maganize*, s. 81-88.

İkinci tipteki derin öğrenme yapıları ayrık öğrenme yaklaşımına yakındır. Desen sınıflandırması için ayırt edici bir güç sağlamayı ve görünür verilerden sınıf etiketlerinin belirlenmesi yoluna gider. Derin yapılandırılmış CRF, tandem MLP mimarisi, derin konveks ağları, tensor ağlar ve algılama tabanlı otomatik konuşma tanıma yapıları örnek olarak verilebilir.

Üçüncü tip olarak da iki yaklaşımı birlikte kullanan hibrit yapılarak oluşturulmasıdır [80,81].

3.6 DERİN ÖĞRENME KÜTÜPHANELERİ

Derin öğrenme sistemlerinin kurulabilmesi, geliştirilmesi ve çalışabilir uygulamaların geliştirilebilmesi için bir takım konularda uzmanlık bilgisi ve beraberinde donanımsal ve yazılımsal altyapıya ihtiyaç duyulmaktadır. Eğitim aşamasında sistem kaynaklarının yeterliliği performansa doğrudan etki eden bir unsurdur. Örnek olarak, işlemcilerinin paralel olarak kullanılması, kümeleme algoritmaları vb. durumlar eğitim ve tanıma sürecininin hızlandırılmasında önemlidir. Bu şekilde karmaşık modellerin oluşturulması belirli bir bilgi birikimi gerektirmektedir[82].

Derin öğrenme kütüphaneleri, bu bilgi birikimi gereknimini azaltmak, kısa ve temel eğitimler ile karmaşık modeller yaratmak ve uygulamaya geçirmek için oldukça

[80] Dong, Yu ve Li, Deng (2015), *Automatic Speech Recognition- A deep learning approach,* Springer-Verlag, London.
[81] Li, Deng ve Xing, Li (2013), "Machine Learning Paradigms for Speech Recognition: An Overview", *IEEE Transactions on Audio, Speech and Language Processing*, s. 1060-1089.

[82] YİĞİT, Altuğ, 2017, Yüksek Lisans Tezi.

faydalıdır. Piyasa da kullanılabilir birçok derin öğrenme kütüphanesi bulunmaktadır. Bu çalışmada en yaygın olanlarından bahsedilecektir.

3.6.1 TensorFlow

Tensorflow, Apache 2.0 lisansı ile açık kaynak kodlu olarak dağıtımı yapılan bir derin öğrenme kütüphanesidir. Google tarafından, derin öğrenme araştırmalarında kullanılmak üzere geliştirilmiştir. Hesaplama sistemi veri akış mimarisi kullanılarak tasarlanmştır. Çizelgelerin düğümleri matematiksel işlemleri temsil ederken, bağlantılar çok boyutlu dizileri temsil eden tensörlerden oluşmaktadır[78]. Java, Python, C++ gibi diller için arayüzleri mevcuttur. Python ile programlamaya en uygun kütüphanelerden birisidir.

Tensorflow kütüphanesi, gömülü sistemlerde, mobil cihazlarda ve işlemci tabanlı kişisel bilgisayarlarda çalışabilir şekilde tasarlanmıştır. Kütüphanenin sağladığı ortam sayesinde, herhangi bir problemi çözmek üzere tasarlanmış olan modeli ufak değişikliklerle farklı bir problem içinde kullanılabilmektedir. Konuşma tanıma başta olmak üzere, görüntü işleme, doğal dil işleme, robotik gibi araştırma alanlarında kullanılmaktadır. Hesaplamaların paralel olarak CPU veya GPU altyapısında yapmaya imkan tanımaktadır. CUDA platform desteği bulunmaktadır[83].

2017 yılındaki bir çalışmada[84] , AlexNet modeli üzerinden farklı derin öğrenme kütüphaneleri test edilmiştir. Verilen her bir yığın için hesaplama süreleri karşılaştırılmıştır. Tensorflow kütüphanesinin bu karşılaştırmada en kısa sürede sınıflandırma yapabildiği gösterilmiştir.

---

[83] Team, TensorFlow (2016), *Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems,* White Paper- Google Research.

[84] Shaohuai, Shi, Qiang, Wang, Pengfei, Xu ve Xiaowen, Chu (2017), "Benchmarking State-of-the-Art Deep Learning Software Tools", *Cloud Computing and Big Data (CCBD), 2016 7th International Conference on: IEEE*, Macau.

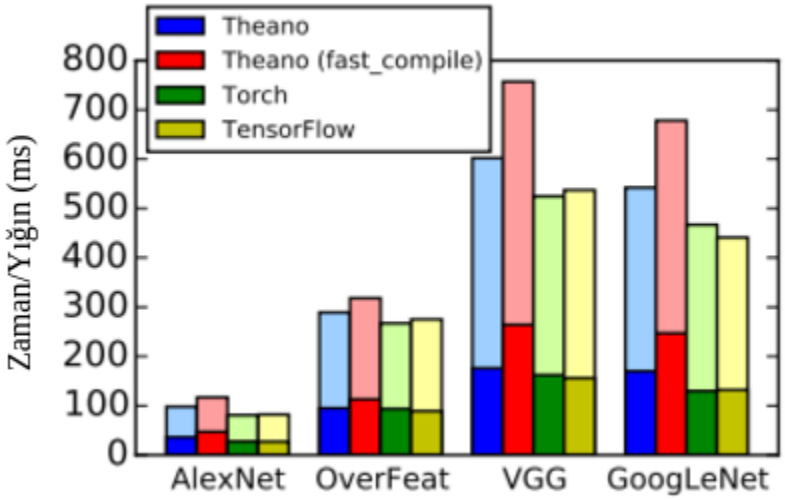
3.6.2 Theano

Montreal Üniversitesi tarafından geliştirilen, BSD lisansı ile açık kaynak kodlu dağıtımı yapılan Python temelli bir derin öğrenme kütüphanesidir. Numpy numerik kütüphanesi ile doğal makine dilinin hızlı işlem yapabilme kapasitesinin birleştirilmesi ile etkin hesaplamalar yapmaya imkan tanıyan bir kütüphanedir.

Numpy kütüphanesi, dizilerin tanımlanması, indekslenmesi, yeniden boyutlandırılması gibi fonksiyonlara sahiptir. C dili ile yazılmıştır. Hem CPU hemde GPU mimarisi ile çalışabilmektedir. Hesaplama öncesinde işlem birimleri C++ veya GPU'da çalışmak üzere CUDA yapısına dönüştürülür ve Python modülleri halinde derleme işlemi yapılır[85]. Bu işlem sayesinde matematiksel ifadeler optimize edilir, gereksiz hesaplamalardan kaçınılır.

Theano üzerine inşa edilen derin öğrenme mimarisi, yüksek seviyeli diller ile geliştirme yapmaya izin veren Pylearn, Keras gibi kütüphaneler ile birlikte gelmektedir. AlexNet testlerinde TensorFlow kütüphanesine göre biraz daha uzun hesaplama süresi ile çalıştığı anlaşılmıştır. Şekil-19'da test sonuçları gösterilmektedir.

[85] James, Bergstra, Olivier, Breuleux, Frédéric ,Bastien, Pascal, Lamblin, Razvan, Pascanuve Guillaume, Desjardins (2010), "Theano: A CPU and GPU Math Compiler in Python", *proc. of the 9th python in science conference.*



Şekil 19 Üç Kütüphanenin Test Sonuçları [86]

3.6.3 Torch

Basit bir geliştirme dilini sahip, açık kaynak kodlu olarak dağıtımı yapılan hesaplama kütüphanesidir. Derin öğrenme problemlerinin çözümünde sıklıkla kullanılmaktadır. Hem CPU hem de GPU mimarisinde işlem yapmaya imkan tanımaktadır. Diğer kütüphanelerden farklı olarak Lua dili ile geliştirme yapılmaktadır. Bu sayede oldukça esnek bir yapıda makine öğrenmesi yapılabilmektedir. C dili içerisinde kolayca entegre edilecek kullanılabilir. Bunun yanında bu kütüphaneye özgü Lua, en hızlı yorumlanabilen dillerden birisidir. Tablolar hem sayı değerleri ile hemde metin değerleri ile indekslenebilmektedir. Oyun geliştirme alanında yaygın olarak kullanılmakta olan bu dil, derin öğrenme alanında Python diline göre daha az tercih edilmektdir. LuaJIT kullanılarak Lua dilindeki kodlar kolaylıkla C dili ile birlikte kullanılabilir hale gelmektedir[87].

[86] YİĞİT, Altuğ, 2017, Yüksek Lisans Tezi.
[87] Ronan, Collobert, Kavukcuoglu, Koray ve Clement, Farabet (2011), "Torch7: A Matlab-like Environment for Machine", *BigLearn NIPS Workshop.*

AlexNet modeli ile yapılan testler sonucunda hesaplama sürelerine bakıldığında bazı testlerde en iyi sonuçları Torch vermiştir. Caffe kütüphanesi ile benzer sonuçlara sahiptir.

3.6.4 Caffe

BSD lisansı altında açık kaynak kodlu olarak dağıtımı yapılan bir C++ kütüphanesidir. Berkeley Görüntü ve Öğrenme merkezinde geliştirilmiştir. Arayüzleri kullanılarak hem Python dili ile hemde Matlab dili ile geliştirme yapabilmek mümkündür. CPU ve GPU altyapısında paralel işlem yapmaya olanak tanımaktadır. Endüstriyel ve bilimsel çalışmalarda yaygın olarak tercih edilmektedir[88].

Konvolüsyonel sinir ağlarının eğitiminde kullanılmaktadır. Görüntü sınıflandırma, özellik çıkarma gibi işlemlerin yapılabildiği bu kütüphane ile ilgili detaylı dokümanlara ulaşabilmek mümkündür. Uygulama geliştirme aşamasında oldukça pratiktir ve hızlı bir şekilde işlem tamamlanarak kullanılabilir hale getirilebilir. Çok fazla kod yazmadan model tasarımları mümkündür.

3.6.5 Keras

François Chollet tarafından geliştirilen yüksek seviyeli sinir ağı uygulamaların programlarında kullanılmaya uygun bir derin öğrenme kütüphanesidir. Python dili ile geliştirilmiştir. Tensorflow, CNTK ve Theano kütüphanelerinin üzerinde çalışabilmektedir. Konvolüsyonel sinir ağları ve tekrarlı sinir ağlarının birlikte tanımlanması mümkündür. CPU ve GPU üzerinde çalışabilmektedir. Birbirine

[88] YİĞİT, Altuğ, 2017, Yüksek Lisans Tezi.

bağlanan, birbirinden bağımsız veya isteğe bağlı yapılandırılabilir modül dizilerinden oluşmaktadır[89].

Sinir ağlarının katmanları, maliyet fonksiyonları, optimizasyon teknikleri, aktifvasyon fonksiyonlar birbirinden bağımsız olarak tasarlanabilir. Bu sayede genişletilebilir bir yapıdadır. Bu kütüphane ile ilgili detaylı dokümanlara ulaşabilmektedir. Bu sayede kolay bir şekilde kütüphanenin anlaşılması ve geliştirme yapılması mümkündür.

3.6.6 Anakonda

Anakonda, paket yöneticisi, çevre yöneticisi, Python ile birlikte birçok açık kaynak kodlu dağıtımı içerisinde barındıran bir platformdur. Ücretsiz olup, topluluk desteği sunmaktadır. Günümüzdeki yaygın olarak kullanılmakta olup, açık kaynak kodlu uygulama geliştirme projelerinde tercih edilmektedir[90].

Bu platform kullanılarak Python dilinin farklı sürümleri ile geliştirme yapılabilir.

[89] YİĞİT, Altuğ, 2017, Yüksek Lisans Tezi.
[90] Anwer, Ali ve Mahmood, Ogur, 2017, Yüksek Lisans Tezi.

3.7 DERİN ÖĞRENME TEMELLİ SÜREKLİ KONUŞMA TANIMA ÜRÜNLERİ KARŞILAŞTIRMASI

Derin İnanç Ağları ile Derin Sinir ağlarının TIMIT görevlerine 2009 yılında başlaması, geniş sözcük dağarcığı sahip büyük uygulamaların yaygınlaşmasını sağladı. Yazılım ile ilgili büyük firmaların bu alanda çalışmaları hızlandı[91].

Bu bölümde bu alandaki en büyük 5 konuşma tanıma uygulamasının detayları anlatılacaktır. 3 farklı konuşma grubu 5 farklı benchmark testi ile bu uygulamalar değerlendirilmiştir.

Uygulama geliştirme sürecinde, TIMIT veritabanındaki monofoneların işlenmeden önce triphone ile değiştirilmeleri önemlidir. İçerik bağımlı durumların incelenmesinde, her bir bölütün sahip olduğu bilginin fazla olması, karar ağacı kümelemesinde mantıksal sınıflandırmayı mümkün kılması sebebiyle triphonelar tercih edilmektedir. İçerik bağımlı saklı markov modelleri kullanılarak, 2 gizli katman ile herhangi bir ön eğitim olmaksızın GMM-HMM sistemleri tasarlandığında gayet iyi sonuçlar alınabilmektedir. Gizli katman sayısının artması ve ön eğitim işlemlerinin yapılması sistemin performansını doğrudan etkilemektedir[92].

3.7.1 Bing Konuşma Tanıma Uygulaması

Geniş dağarcıklı ses tanıma için, Derin sinir ağları kullanılarak oluşturulan akustik modellerin kullanıldığı ilk başarılı çalışma Bing mobil sesli arama

[91] Geoffrey, Hinton, Li, Deng, Dong, Yu, George E., Dahl, Ebdel-rahman, Mohamed, Navdeep, Jaitly, Andrew, Senior, Vincent, Vanhoucke, Patrick, Nguyen, Tara N., Sainath ve Brian, Kingsbury (2012), "Deep Neural Networks for Acoustic Modelling in Speech Recognition: The Shared Views of Four Research Groups", *IEEE Signal Processing Maganize*, s. 92-97.

[92] Frand, Seide, Gang, Li ve Dong, Yu (2011), "Conversational Speech Transcription Using context-dependent Deep neural networks", *In Proceedings of the Annual Conference of International Speech Communication Association(INTTERSPEECH)*, s. 437-440.

uygulamasından alınan veriler ile yapılmıştır. Gürültü, müzik, yan konuşmalar, vurgular, özensiz telaffuzlar, tereddütler, tekrarlamalar, kesintiler veya cep telefonu farklılıklarının neden olduğu akustik değişkenliğin yüksek olduğu 24 saatlik bir eğitim verisi kullanılmıştır.

Sonuçlara göre[93] en iyi akustik model ile içerik bağımlı durumlar için test setinde %69.6 cümle doğruluğuna ulaşıldığı(DNN-HMM), aynı model için GMM-HMM kullanıldığı durumlarda %63.8'lik bir doğru tanıma yakalandığı gösterilmiştir. Model tasarımında, her katmanda ön eğitimden geçirilmiş 2040 node kullanılmıştır. Monofone yerine triphoneların kullanılmasının ciddi olarak performans artışı sağladığı gösterilmiştir. Seslerin girdi olarak verilmeden önce hizalanmalarının sistemin hata oranında düşmeye fayda sağladığı anlaşılmıştır. Hizalama işlemi hem GMM-HMM model için hemde DNN-HMM model için fayda sağlamıştır.

Bu çalışmasının devamında eğitim verisi 48 saate çıkarılmıştır ve sinir ağında bazı düzenlemeler yapılmıştır. Beklendiği gibi ön eğitimin sinir ağlarında faydalı olduğu gösterilmiştir.

3.7.2 Switchboard Konuşma Tanıma Uygulaması

Bing ses tanıma uygulamasında kullanılan DSA-SMM modeli herhangi bir değişiklik olmaksızın Switchboard uygulamasında uygulanmıştır. Bing uygulamasında 48 saatlik eğitim verisi ile bağlı trifon setleri ile eğitim sağlanmışken bu çalışmada 300 saatin üzerinde eğitim verisi ve binlerce trifon seti kullanılmıştır.

[93] G., Dahl, D., Yu, L., Deng ve A., Acero (2012), "Context-Dependent Pre-Training Deep Neural Networks for Large-Vocabulary Speech Recognition", *IEEE Transactions on Audio, Speech and Language Processing*, s. 30-42.

Ayrıca Switchboard sisteminin konuşmadan metne dönüşüm sağlayan bir açık uygulaması bulunmaktadır[94].

Sistem temelde 309 saatlik eğitim verisi ile eğitilmiştir. Pencereli ortalama-varyans normalizasyonu kullanılmış olan 13 boyutlu PLP özellikleri kullanılmış ve bu değerlerinin 3. dereceden türevleri ile birleştirilmiştir. Sonrasında doğrusal diskriminant analizi ile 39 boyutlu bir vektör haline dönüştürülmüştür. Konuşmacı bağımsız trifonlar soldan sağa 3 durum topoloji ile kullanılmıştır ve 9304 durum arasında paylaştırıldı. Sistem başlangıçta 40 Gauss karışımı içermektedir. Her bir katmanda 2048 birim içeren 7 gizli katmanlı model oluşturulmuştur.

Sonuçta, DNN-HMM ile eğitilmiş modellerde olasılıksal modellere göre (SMM-GMM) %18.5'lik bir performans artışı yakalanmıştır. Derin sinir ağlarını önceden eğitmek daha iyi sonuçlar alınmasına imkan tanırken, kritik bir husus değildir. En fazla %1'lik bir kazanç sağlamaktadır. Ancak daha az kaynağın bulunduğu diller için etiketli verilerin az olduğu durumlarda ön eğitimin daha faydalı olacağı değerlendirilmiştir[95].

Tablo 5 Ayrık Eğitim yöntemi ile eğitilmiş iki güçlü GMM-HMM tabanlı sistem ile 5 farklı Derin inanç ağı-derin sinir ağı akustik modelinin karşılaştırması. Konuşmacı bağımsız 309 saatlik eğitim verisi ve tek geçişli kod çözme tekniği ile 2000 saatlik konuşmacı

Modelleme Tekniği	**Girdi Sayısı**	**Kelime Hata Oranı Hub500-SWB**	**Kelime Hata Oranı R035-FSH**
GMM, Her bir SMM durumu için 40 Gausian, Ayrık Eğitim, 390 saatlik konuşmacı bağımsız veri	29.4	23.6	27.4
Yapay Sinir Ağı, 1 gizli katman, Her katmanda 4634 çekirdek	43.6	26.0	29.4
+ 2x5 komşu pencereler	45.1	22.4	25.7

[94] Geoffrey, Hinton, Li, Deng, Dong, Yu, George E., Dahl, Ebdel-rahman, Mohamed, Navdeep, Jaitly, Andrew, Senior, Vincent, Vanhoucke, Patrick, Nguyen, Tara N., Sainath ve Brian, Kingsbury, 2012, s. 92-97.

[95] Dong , Yu, Li, Deng ve G, Dahl (2010), "Roles of Pre-Training and Fine-Tuning in Context-Dependent DBN-HMMs for Real-World Speech Recognition", *Deep Learning and Unsupervised Feature Learning*.

Derin İnanç ağları- Derin Sinir Ağları, 7 gizli katman, Her bir katmanda 2048 çekirdek	45.1	17.1	19.6
+ hizalama ile durum güncellemesi	45.1	16.4	18.6
+ Seyrekleştirme	15.2 nz	16.1	18.5
GMM, Her bir HMM durum için 72 Gausian, Ayrık Eğitim, 2000 Saatlik konuşmacı eğitimli ve veri	102.4	17.1	18.6

3.7.3 Google Konuşma Tanıma Uygulaması

Google ses tanıma uygulaması, mobil cihazlarda sesli arama sorguları, kısa mesajlar, epostalar ve bazı kullanıcı işlemlerini yapar. Arama sorguları ve dikte işlemleri için tasarlanmış dil modeli kullanan geniş sözcük dağarcıklı bir tanıma görevidir. Google bu işlem için, çok geniş bir kelime sözlüğü ile içerik bağımsız çapraz trifonlar içeren saklı markov modellerden oluşan konuşmacı bağımsız GMM-SMM modeller kullanmaktadır. Model temelde 7969 duruma sahiptir ve LDA tarafından dönüştürülmüş akustik girdi kullanmaktadır. 5780 saatlik hizalanmış eğitim verisi, derin sinir ağları kullanılarak oluşturulmuş akustik model 7969 HMM durumu kullanılarak elde edildi. Her bir katmanda 2560 node, son katmanda softmax katmanı içerir ve 7969 durum içerir.

Canlı ses verileri için anonimleşmiş sözcüklerden oluşan test setinde DSA tabanlı sistemler GMM tabanlı sistemlere kıyasla %12.3 lük bir kelime hata oranı iyileşmesi sağlandığı gözlemlenmiştir[96].

[96] Geoffrey, Hinton, Li, Deng, Dong, Yu, George E., Dahl, Ebdel-rahman, Mohamed, Navdeep, Jaitly, Andrew, Senior, Vincent, Vanhoucke, Patrick, Nguyen, Tara N., Sainath ve Brian, Kingsbury, 2012, s. 92-97.

3.7.4 Youtube Konuşma Tanıma Uygulaması

Bu uygulamada amaç Youtube verilerinin kopyalanmasıdır. Yukarıda anlatılmış olan mobil ses temelli uygulamaların aksine bu uygulamada yeterince güçlü bir dil modeli bulunmadığı için iyi bir akustik model gerektirmektedir. Bunun için Google'ın 1400 saatlik hizalanmış eğitim verisi kullanılmıştır. LDA tarafından dönüştürülmüş 9 çerçeveli MFKK kullanıldı. 17552 trifon ile karar ağacı kümelemesi yapılmıştır. Kod çözme işlemi sırasında, maksimum olabilirlik doğrusal regresyonu ve maksimum olabilirlik doğrusal regresyon dönüşümleri kullanılmıştır. Derin öğrenme modelinde ilk gizli katmanda 2000 sonrasında her gizli katmanda 1000 düğüm kullanılmıştır[38]. Sonuçta, Derin ağ modellerine göre hata oranında %4.6'lık bir iyileşme sağlanmıştır.

3.7.5 İngilizce Yayın Haber Konuşma Tanıma Uygulaması

Derin sinir ağları İngilizce haber yayıncılığına uygulanmıştır. GMM-SMM sistemler oluşturmak için öncelikle konuşmacıdan bağımsız özellikler oluşturulmuştur. Sonrasında konuşmacıya göre ayarlanmış şekilde eğitim ayrık olarak özelliklere göre yapılmıştır. Konuşmacı bağımsız özellikler doğrusal ayrık analiz kullanılarak elde edilmektedir.

Bu çalışmada hizalanmış 50 saatlik eğitim verisi kullanılarak derin sinir ağı ile oluşturulmuş akustik model eğitilmiştir. Geniş sözcük dağarcılıklı konuşma tanıma sistemleri için konuşmacı bağımlı ve konuşmacı bağımsız özellikler kullanılarak eğitim yapılmıştır. Kurulan mimaride 6 gizli katman, her katman 1024 düğüm bulunmakta ve son katmanda ise softmax fonksiyonu kullanılmaktadır.

Sonuçta, 2200 Gauss durumu için DNN-SMM sistemi için %17.5'lik bir hata oranı yakalanmıştır.

Tablo-6'da bu bölümde bahsedilmiş uygulamaların akustik modelleme sonuçları gösterilmektedir.

Tablo 6 Geniş Sözcük Dağarcığına sahip görevler için DSA-SMM ile GMM-SMM modellerinin hata oranlarının karşılaştırılması

Görev	**Eğitim Verisi Miktarı(Saat)**	**DSA-SMM**	**GMM-SMM, Aynı Veri**	**GMM-SMM, Fazladan Veri İçerir.**
Switchboard(test set 1) Switchboard(test set 2)	309 309	18.5 16.1	27.4 23.6	18.6(2000 saat) 17.1(2000 saat)
English Broadcast News	50	17.5	18.8	
Bing Voice Search(Cümle hata oranı)	24	30.4	36.2	
Google Voice Input	5870	12.3		16.0(5870 saat)
Youtube	1400	47.6	52.3	

DSA-SMM mimarinin aynı miktarda veriyle eğitilen GMM-HMM sistemlerden daha iyi performans gösterdiği ve bazı problemler için bu farkın fazla olduğu gözlemlenmiştir. Hatta bazı problemlerde DNN-HMM, daha fazla veri ile eğitilmiş GMM-HMM'den daha iyi performans göstermektedir.

DÖRDÜNCÜ BÖLÜM

SES TANIMA SİSTEMLERİNE İLİŞKİN TARİHSEL GELİŞİM VE TÜRKÇE KONUŞMA TANIMA LİTERATÜR ÖZETİ

Otomatik konuşma tanıma sistemlerin tarihsel süreç içinde gelişimi 1900'lü yıllarında başlarından gelmektedir. Aşağıdaki tarihsel gelişimi içerisinde dönemsel olarak gelişmeler verilmiştir.

4.1 1920-1959

Konuşma tanıma 1920 yıllarında başından beri çalışılmakta olan konulardan biridir. Bu alandaki ilk tanıma Radio Rex isimli ürün ile yapılmıştır[97].

Ses tanıma ile ilgili konseptlerinin belirlenmesi 1936 yıllarında Bell Lab. yapılmıştır. 1939 yılında ses sentezi ile ilgili NewYork'da ilk sunum gerçekleştirilmiştir.

Otomatik konuşma tanımaya ilişkin ilk sistemlere 1950 yılındaki çalışmada rastlanmaktadır. Bu çalışmalarda akustik fonetik yaklaşım fikri uygulanmaya

[97] Stefan, Windmann ve Reinhold Haeb, Umbach (2009), "Approaches to Iterative Speech Feature Enhancement and Recognition", *IEEE Transactions on Audio, Speech and Language Processing*, s. 974-984.

çalışılmıştır[98]. 1950'li yıllarda yapılan konuşma tanıma sistemlerinde, bir analog filtre bankası ve lojik devrelerin çıkış sinyallerinden her harfin spektral rezonans değeri kullanılmıştır.

1952 yılında Bell laboratuvarları tarafından tekil konuşmacılı yalıtık rakam tanıma sistemi ilk defa gerçekleştirilmiştir[99].

1952 yılları ile 1959 yılı arasında[100, 101, 102] numaralı çalışmalar yapılmıştır. Bu çalışmalarda filtre bankası ve seslerin sistem rezonanslarının zamanla değişiminden hangi sesli harfin seslendirildiğinin tahmin edilmesi yaklaşımı kullanılmıştır.

4.2 1960-1970

1960'lı yıllarda konuşma tanımaya ilişkin birkaç farklı fikir ortaya atılmıştır. Bilgisayarın yavaş olduğu bu nedenle de daha hızlı tanıma yapabilir sistemlerinin kurulması gerektiği anlaşılmıştır[103]. Bunun sonucunda, birkaç Japon laboratuvarında konuşma tanıma çalışmaları için özel amaçlı donanımsal altyapılar kurulmuştur. İlk defa Radyo Araştırma laboratuarında sesli tanıma cihazı kullanılmıştır[104]. Aynı yıl içerisinde Kyoto üniversitesinde ses tanıma cihazı geliştirmesi çalışmaları yapılmıştır.

---

[98] Sadaoki, Furui, 2005, s. 1-2.
[99] K.H., Davis, R., Biddulph ve S., Balashek(1952), "Automatic Recognition of Spoken Digits", *The Journal of The Acoustical Society of America*.
[100] H.F., Olson ve H., Belar (1956), "Phonetic Typewriter", *Journal of Acoustics Soc. am.* , s. 1072-1081.
[101] Fry, D.B. (1959), "Theoritical Aspects of Mechanical Speech Recognition", *Journal of the British Institution of Radio Engineers*, s. 211-218.
[102] James W., Forgie ve Carma D., Forgie (1959), "Results Obtained from a Vowel Recognition Computer Program", *The Journal of the Acoustical Society of America*, s. 1480-1489.
[103] J., Suzuki ve K., Nakata (1961), "Recognition of Japanese Vowels Preliminary to the Recognition of Speech", *Journal of Radio Research Lab.*, s. 193-212.
[104] T., Sakai ve S., Doshita (1962), "The Phonetic TypeWriter, Information Processing", *Proceedings of IFIP Congress*.

Konuşma tanıma konusunda o yıllarda karşılan zor problemlerden biri, konuşma olaylarında zaman ölçeğinde tekdüzelik bulunmamasıdır. Bu konuda ciddi çalışmalar yapılmıştır. Bu alanda ilk çalışmada, tanıma skorlarının değişkenliğini önemli ölçüde azaltan konuşmanın başı ve bitiş zamanını güvenilir bir şekilde algılama yeteneğine sahip normalleştirme yöntemi geliştirilmiştir[105] ve bu yöntemi kullanan ticari ürünler üretilip satılmaya başlanılmıştır.

Aynı yıllarda dinamik zaman bükmesi olarak tanımalanan ve dinamik programlamayı kullanan yöntem geliştirilmiştir[106].

1960'lı yılların sonlarında, günümüzdeki sistemlerin temellerinin atıldığı ve sürekli konuşma tanıma sistemlerinin dinamik fonem incelemesi ile tasarlanması projesi yapılmıştır ve oldukça yüksek seviyede konuşma tanıma başarıları yakalanmıştır[107]. Bu çalışmada kullanılan yöntemler günümüzdeki sürekli konuşma tanıma özelliğine sahip ticari uygulamalarda kullanılmaktadır.

4.3 1970-1980

Yalıtık veya ayrık konuşma tanımaya ilişkin ilk çalışmalar Rusya ve Japonyada yapılmıştır[108, 109]. Japonyada ki araştırmalar dinamik programlama yöntemlerinin başarıyla uygulanabileceğini göstermiştir ve araştırmada düşük bit hızında konuşma kodlamasında başarıyla kullanılan doğrusal öngörülü kodlamadaki spektral

[105] T.B., Martin, A.L., Nelson ve H.J., Zadell (1964), *Speech Recognition by Feature Abstraction Techniques*, Teknik Rapor -Ohio: Air Force Avionics Lab., Ohio.
[106] Vintsyuk, T.K. (1968), *Speech dicrimination by dynamic programming*, Sprink Link- Cybernetics, s. 52-57.
[107] Reddy, D.R. (1966), *An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave*, Teknik Rapor, Stanford University.
[108] V.M., Velichko ve N.G., Zagoruyko (1970), "Automatic Recogniton of 200 Words.", *International Journal of Man-Machine Studies*, s. 223-234.
[109] H., Sakoe ve S., Chiba (1978), "Dynamic Programming algorithm optimization for spoken word recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, s. 43-49.

parametrelere dayanan mesafe ölçümü yöntemiyle konuşma tanıma yapılabileceği gösterilmiştir.

Aynı yıllarda IBM tarafından geniş sözcük haznesine sahip konuşma tanıma sistemleri üzerinde basit veri tabanı sorguları, lazer patent metinlerinin yazımı ve basit notların okunmasına ilişkin çalışmalar yapılmıştır[110].

1970'li yıllarda, AT & T Bell laboratuvarında, konuşma tanıma çalışmaları yapılmıştır. Sözcüklerin tüm farklı varyasyonlarını içeren çalışmalar yapılmıştır ve ihtiyaç duyulabilecek maksimum desen sayılarının belirlenmesi sağlanmıştır. Bu alandaki çalışmalar DARPA tarafından finanse edilmiştir[111]. Ayrıca, CMU tarafından tanıtılan sistemde 1011 kelimelik bir sözcük dağarcığını iyi bir şekilde tanımıştır. Bu sistemde, sözdizimsel bilgiler kullanılmış ve hesaplamayı azaltarak en yakın kelimeyi bulmak için sonlu durum ağı kullanılmıştır. DARPA finansmanı ile geliştirilen diğer sistemlerde paralel asenkron süreçleri kullanan bir konsept kullanılmıştır[112].

4.4 1980-1990

1970'li yıllarda daha çok yalıtık sözcük tanıma üzerine çalışılırken, 1980'li yıllarda sürekli konuşma tanıma üzerine çalışmalar yapılmıştır. Amaç akıcı bir konuşma dizesinin daha güçlü bir sistem ile tanınmasını sağlamaktadır.

[110] Jelinek, F. (1985), "The Development of an experimental discrete dictation recognizer", *Proceedings of IEEE*, s. 1616-1624.
[111] Klatt, Dennis H (1977), "Review of the ARPA Speech Understanding Project", *The Journal of the Acoustical Society of America*, s. 1324-1366.
[112] Lowerre, Bruce T. (1990), "The Harpy Speech Recognition System", *Readings in Speech Recognition*, s. 576-586.

Harf ve rakamların spektrogramlarını kullanarak sürekli konuşma tanıma üzerine ilk çalışma yapılmıştır[113].

İki Seviyeli dinamik programlama yaklaşımı ile akıcı bir düzene sahip kelimelerin tek tek tanınmasına yönelik çalışmalar yapılmıştır[114].

Tanınacak ses deseni ile verisetindeki seslerin eşleştirilmesine yönelik farklı yaklaşımların kendilerine göre avantajları bulunmaktadır. Bu yıllarda daha çok şablon temelli yaklaşım tercih edilirken, istatiksel modellemeye geçiş olmuştur. Çoğunlukla saklı markov model kullanan uygulamalar üzerine çalışılmıştır[115].

Saklı markov model 1980'li yıllarında başlarında iyi anlaşılmış olsa da ortalarında kadar bir geliştirme yapılması mümkün olmamıştır. 1980'li yılların ortalarından itibaren araştırma yayınları ortaya çıkmaya başlamış ve SMM kendisine geniş bir uygulama sahası bulmaya başlamıştır.

Günümüzdeki en pratik konuşma tanıma sistemleri, 1980'li yıllarda geliştirilen istatiksel temelli bu çalışmaların sonuçlarına dayanmaktadır.

Saklı Markov Model

Saklı Markov model, 1980'li yıllarda geliştirilen önemli teknolojilerden birisidir[116]. Gözlemlenmesi normal şartlarda mümkün olamayan olasılıksal bir süreçtir ancak bir dizi gözlem üreten başka bir stokastik süreç aracılığıyla gözlemlenebilmektedir. İlk olarak IBM gibi bir kaç laboratuvarda tanınmış ve

[113] Moore, Roger K. (1994), "Twenty Things we still don't know about Speech", P*rogress and Prospects of Speech Reserach an Technology*.
[114] J.S., Bridle, M.D., Brown ve R.M., Chamberlain (1983), "Continuous Connected word recognition using whoel word templates", *Radio and Electronic Engineer*, s. 167-175.
[115] Rabiner, L.R., 1989, s. 257-286.
[116] Ferguson, J. (1980), *Hidden Markov Models for Speech*, Princeton, New Jersey.

anlaşılarak uygulanmış olmakla birlikte, 1980'li yılların ortalarında ses tanıma ile ilgili çalışan birçok araştırma merkezinde yaygın olarak kullanılan bir yöntem olmuştur.

1970'li yıllaında başlarında Princeton üniversitesinden gizli markov modellemeyi konuşma tanıma da kullanmak üzere matematiksel modelleme icat edilmiştir. Saklı markov model desen eşleştirme stratejisi bu alanda çalışan birçok ticari uygulamada da tercih edilmiş ve o yıllarda ABD savunma sanayii tarafından bu alanda çalışan firmalar ciddi destekler almıştır. Bu alandaki çalışmalar ilk yıllarda yavaş iken sonrasında yaygın bir kullanım alanı yakalamıştır.

Yapay Sinir Ağları

1980'li yılların sonlarında konuşma yanıma yeni bir teknoloji olan sinir ağların kullanılması fikri gelişmiştir. Sinir ağları 1950'li yıllarda ortaya çıkmış olsa da ses tanıma alanındaki pratik uygulamalarda kullanılması 1980'li yılların sonlarında olmuştur. Ses tanıma teknolojisindeki zor ve kısıtlı durumlar anlaşılıp çözüldükçe sinir ağlarının kullanılması yaygınlaşmaya başlamıştır[117].

Darpa Programı

1980'li yılları sonlarında DARPA topluluğu, geniş sözcük dağarcıklı sürekli konuşma tanıma sistemlerinin geliştirilmesi ana amaç olarak ele almıştır. 1000 sözcük ile sürekli konuşma tanıma mimarisinde çalışacak bir sistemin tasarımı için çalışmıştır. Bu araştırmaların sonucunda CMUSPHINX[118] olarak bilinen sistemler, AT&T

[117] Lippmann, R.P. (1987), "An Introduction to Computing with Neural Nets", *IEEE ASSP Magazine*, s. 4-22.
[118] C.H., Lee, L. R., Rabiner, R., Pieraccini ve J.G., Wilpon (1990), "Acoustic Modeling for Large Vocabulary Speech Recognition", *Computer Speech and Language*, s. 127-165.

laboratuvarları, BYBLOS[119] sistemi, MIT[120] laboratuvarları tarafından katkı verilerek oluşturulmuştur. SPHINX sistemi HMM'in istatistiksel metodunu Harpy sistemine entegre etmiştir. Böylece içerik bağımlı fonem tabanlı modelllerin geliştirilmesi ve eğitilmesinin mümkün olması sağlanmıştır.

4.5 1990-2000

1990'lı yıllarda desen tanıma ile ilgili çalışmalar yapılmıştır. Geleneksel Bayes algoritmasını takip eden ve veri dağılımının hesaplanmasına ihtiyaç duyan model tasarımı için, örneklem tanıma hatasını minimize eden optimizasyon teknikleri üzerinde çalışılmaya başlanmıştır[121]. Bayes teorisinde konuşma sinyalinin dağılım işlevlerinin doğru seçilememesi veya tanımlanamaması sebebiyle bu şekilde farklı teknikler üzerinde araştırmalar yapılmıştır. Bir tanıma sisteminde, dağılım fonksiyonun Bayes teorisine göre verisetine en iyi şekilde uydurulması yerine en düşük hata oranının elde edilmesinin gerekli olduğu anlaşılmıştır. Örnek olarak, maksimum benzerlik bilgisi kriteri incelendiğinde, akustik bilgi ile bir eğitim seti üzerinde ortalaması alınan doğru sözcük simgesi arasında karşılıklı bilgi en üst düzeye getirilmeye çalışılmaktadır.

Konuşma tanıma sistemi tasarımlarında bir kritik hususta, eğitim için kullanılacak verisetinin seçimidir. Temel konuşma birimlerinin(sözcük veya sözcük altı birimler) karakterize edilmesinde gözetimli öğrenme parametlerinin(şablonlar,

[119] Y., Chow, M., Dunham, O., Kimball, M., Krasner, G., Kubala, J., Makhoul, P., Price, S. ve Roucos, R., Schwartz (1987), "System, BYBLOSS: The BBN Continuous Speech Recognition", *Acoustics, Speech and Signal Processing*.

[120] Victor, Zue, James, Glass, Michael, Philips ve Stephanie, Seneff (1989), "The MIT SUMMIT Speech Recognition system: a Progress Report", *Workshop on Speech and Natural Language:HLT", s.* 179-189, Stroudsburg.

[121] Bing, Hwang Juang ve Sadaoki, Furui (2000), "Automatic Recognition and understanding of spoken language a first step towards natural human machine communication", *Proceedings of the IEEE, s.* 1142-1165.

istatistiksel modeller), etiketli konuşma örneklerinin veya cümlelerinin kullanılması mümkündür.

Darpa

Darpa programı 1990'lı yıllarda devam etmiş ve doğal dil işleme alanında çalışmalar yapmıştır. Hava ulaşım bilgilerinin yönetimi servislerine önem verilmiştir. Sonrasında, program konuşmayı anlamaya yönelik farklı alanların kullanımı, haber yayınlarının veya diyalog halindeki konuşmaların yazı haline getirilmeleri alanlarında çalışmalara yönelik olmuştur. "Switchboard" olarak isimlendiren santral görevi en zorlu problemlerden birisi üzerine çalışılmıştır. Bu problemde, diyolog halinde ve spontan olarak seslendirilen ve eksik kelimeler, tereddütler ve onarımlar gibi sesi etkileyen durumlarıda içeren konuşmaların çözümlenmesi yapılmaya çalışılmıştır. Bu uygulamalar, bilgi çıkarımı teknolojileri ile entegre edilerek, otomatik sesli belge indeksleme ve bilgi çıkarımı uygulamaları geliştirilmiştir.

1990'lı yıllarda düzenlenmiş olan DARPA'nn ses ve doğal dil işleme çalıştaylarında konuşma tanıma sistemlerinde gelişmelerin hızlandırılmasına yönelik çalışmalar yapılmıştır. Var olan konuşma tanıma sistemlerinin işlevlerinin ve performanslarının geliştirilmesine yönelik araştırmalara odaklanılmıştır. Sürekli konuşma tanıma ve gürültülü ortamlarda konuşma tanıma problemlerine çözüm üretilmeye çalışılmıştır.

Saklı Markov Model

Saklı markov model tabanı konuşma tanıma için daha gürbüz ve farklı sesleri birbirinden ayırt etme gücünde gelişme sağlayabilmek için ağırlıklandırılmış saklı markov modeli önerilmiştir[122]. Kelime tabanlı modeller için fonetik ve fonenik

[122] L.R., Bahl, P.F., Brown, P.V., de Souza, R.L., Mercer, M.A., Picheny (1993), "A Method for the construction of acoustic markov models for words", *IEEE Transactions on Speech and Audio Processing*, s. 443-452.

modellerin birlike kullanılması yoluna gidilmiştir. Saklı markov model ile öğrenmeli vektör birlikte kullanılarak farklı hibrit modeller önerilmiştir. Öğrenmeli Vektör Kuantasyonu statik kalıpları sınıflandırmak için yüksek derecede ayırt edici referans vektörleri üretmede önemli katkı sağlamıştır. Makine öğrenmesi için parametre tahmininde ileri-geri beslemeli algoritmaların verimsiz olması olması sebebiyle, parametre tahmininde hataları en az indiren bir sinir ağı temelli vektör kuantasyonu önerilmiştir[123].

Vektör Kuantasyonu yaklaşımı ile ses tanıma da kelime hata oranlarında ciddi iyileşmeleri gözlemlenmiştir. Saklı markov modelini temel alan birçok çalışma sunulmuştur. Bunların içerisinde en önemlisi, Maksimum eş bilgi elde edilen MVA yöntemi olmuştur. Bu yöntemde, fonemin konuşma tanıma alanında bir küme olarak ele alınması ve her fonem için bir gauss modeli yaratılmasını önerilkmiştir. Böylece gauss modelleme ile daha etkili bir sınıflandırma yapığı ve kelime hata oranlarında ciddi iyileşmeler olduğu gözlemlenmiştir. Gauss Seçimi, belirli bir girdi vektörü için Gauss bileşen olasılıklarının alt kümesini seçerek yükü azaltmak ve iyi Gauss altkümelerini elde etmek için yeni teknikler önerilmiştir.

Gürbüz Konuşma Tanıma

Arka plan sesleri, eğitim ve test koşulları, konuşmanın kişi bağlı özelliği, mikrofon tipi, iletim kanalları, oda yankılanması gibi sorunlara daha sağlam çözümler üreten konuşma tanıma sistemlerinin geliştirimlesi için çalışmalar yapılmıştır. Gürbüz konuşma tanıma sistemleri geliştirmek için maksimum olabilirlik doğrusal regresyonu, model ayrıştırma, paralel model kompozisyonu ve yapısal maksimum posteriori metodları kullanılmıştır.

[123] Adoram, Erell ve Mitchel, Weintraub (1993), "Filterbank-Energy Estimation Using Mixture and Markov Models for Recognition of Noisy Speech", *IEEE Transactions on speech and Audio Processing*, s. 68-76.

Bu alanda ilk çalışmada[124] çevresel gürültü, kanal bozulmaları gibi sese olumsuz etkisi olan durumları ortadan kaldırmak için maksimum olabilirlik doğrusal regresyonunu kullanan bir yöntem önerilmiştir.

[125] numaralı çalışmada, test ifadeleri ile akustik modeldeki bozulmaları en aza indirerek tanıma performasını artırmak için akustik uyuşmazlıkları mimimuma indiren maksimum olabilirlik yöntemini kullanan bir yaklaşım önerilmiştir.

[126] numaralı çalışmada, cochlear bant geçiren filtreler kullanan ve sinyal frekans bilgisinin sıfır geçiş aralıklarıyla elde edildiği doğrusal olmayan işlemlerden oluşan bir model önerilmiştir.

Gürültülü Konuşma Tanıma

Gürültülü konuşma tanımaya ilişkin bu yıllarda fazla çalışma yapılmamıştır. Filtreli log enerjilerinin minimum ortalama kare hatası olarak adlandırılan yöntem ile tahmin edilmesinde gelişme sağlanmıştır. Spekral tahmin algoritması ile gürültüye dayanıklı ses tanıma sistemi çalışmaları yapılmıştır. Filtre bankası tabanlı sistemler için uyarlanmıştır. Hesaplama farklarını minimuma indirmeye yönelik bir tasarım yapılmıştır[127].

[124] Bling, Hwang Juang ve Rahim, M.G. (1996),"Signal bias removal by maximu likelihood estimation for robust telephone speech recognition", *IEEE Transactions on Speech and Audio Processing*, s.19-25.
[125] Sankar, A. ve Chin-Hui, Lee (1996), "A Maximum likelihood approach to stochastic matching for robust speech recognition", *IEEE Transactions on Speech and Audio Processing*, s. 190-202.
[126] Doh-Suk, Kim, Soo-Young, Lee ve R.M., Kil (1999), "Auditory processing of speech signals for robust speech recognition in real-world noisy environments", *IEEE Transactions on Speech and Audio Processing*, s. 55-69.
[127] Adoram, Erell ve Mitchel, Weintraub, 1993, s. 68-76.

4.6 2000-2009

2000'li yıllarda, değişken Bayes tahmini ve kümeleme teknikleri geliştirilmiştir. Adaptif öğrenme probleminin çözümü ve OKT için aktif öğrenme algoritması önerilmiştir[128].

Spontan ve okuma sesleri için tanıma yapılırken üretilen akustik özellikler arasındaki farklar analiz edilmiştir[129].

Konuşma tanımada sürekli sürekli yoğunluklu saklı markov model geniş marj tahmin problemini çözmek için yeni bir optimizasyon tekniği önerilmiştir[130].

Darpa

Verimli ekonomik yeniden kullanılabilir sesten metne dönüştürme programı otomatik olarak konuşmayı metne dönüştürme için daha doğru ve etkin bir yöntem olarak ortaya koyulmuştur. Cümle sınırlarının belirlenmesi, nefes alma ve verme seslerinin tespitini yapmayı içermektdir. Doğal olarak seslendirilmiş olan ve çok dilli olarak kayıt edilen insan seslerinin tanınmasını hedeflenmiştir. Makinelerin konuşulan sesteki önemli bilgileri algılamasını, özetlemesini veya tercüme etmesini otomatik olarak yapmak ve böylece kişilere sinyalleri dinlemek yerine önlerine sunulan özet metinleri okuyarak söyleneni anlamalarını sağlamak amaçlanmıştır.

Spontan Konuşma Tanıma

Haber programlarıdan belli bir tonda belirli bir metnin okunması gibi kalıplara bağlı konuşmalar %95 doğrulukla tanınabilecek sistemler geliştirilmiştir. Ancak bu

[128] M., Afify, Feng, Liu, Hui, Jiang ve O., Siohan (2005), "A new verification-based fast-match for large vocabulary continuous speech recognition", *IEEE Transactions on Speech and Audio Processing*, s. 546-553.
[129] Simon, King, Joe, Frankel, Karen, Livescu, Eric, McDermott, Korin, Richmond ve Mirjam, Wester (2007), "Speech production knowledge in automatic speech recognition", *The Journal of the Acoustical Society of America*.
[130] Marc A., Zissman ve Kay M.,Berkling, 2001, s. 115-124.

oran son teknolojinin kullanıldığı sistemlerde bile spontan konuşmalar için ciddi olarak düşmektedir. Konuşma tanıma sistemlerinin kullanımının yaygınlaşması spontan konuşma sistemlerinin performans geliştirilmesi ile mümkündür.

Spontan konuşma için bir kaç proje yürütülmüştür. Japonyada "Spontan Konuşma: Kütük ve İşleme Teknolojileri" isimli 5 yıllık ulusal bir proje önerilmiştir[131]. 7 milyon kelimeden oluşan ve 700 saatlik bir konuşmayı içeren dünyanın en büyük spontan konuşma veriseti hazırlanmıştır. Yeni çalışmaların tamamı bu veriseti üzerinde yapılmıştır. Bu teknikler esnek akustik modelleme, cümle sınırların algılama, teleffuz modelleme, akustik ve dil modeli adaptasyonu ve otomatik konuşma özetlemesini içermektedir.

Spontan konuşma sürekli konuşma tanıma performansını:

1) doğrudan spontan konuşma,

2) metin okumasına yakın konuşulması ile elde edilen spontan konuşma,

3) spontan konuşma setleri ile eğitim yapılması etkilemektedir.

Bu çalışmalar sonucunda, spontan konuşma tanıma sistemlerinde konuşma doğruluğunu artırmak için daha fazla spotan eğitim verisi ile modellerin eğitilmesi ve spontan konuşmasının bilenen etkilerinin modellere eklenmesinin gerektiği anlaşılmıştr.

Gürbüz Konuşma Tanıma

Konuşma tanıma sistemlerinin sağlamlığını artırmak için, özellikle spontan konuşma sistemlerinde kelime doğrulama ve güven ölçütleri üzerine çalışılmıştır. Diyalog içeren konuşma tanıma uygulamalarında insana benzer etkileşimlere sahip olmak için her olayı öngörecek bir model oluşturulmalı olduğu tespit edilmiştir.

[131] Furui, Sadaoki (2005), "Recent Progress in Corpus-Based Spontaneous Speech Recognition.", *IEICE Transactions of Informations and Systems*, s. 366-375.

Spontan konuşma için semantik olarak çalışan ve konuşmanın gerekli ve gereksiz bölümlerini birbirinden ayırabilen yaklaşımlar araştırılmıştır[132].

Birleşik tanıma ve doğrulama stratejileri düzgün konuşma sesleri için çalışmaktadır. Geleneksel saklı markov modellerinden daha iyi akustik modeller inşa edebilmek için dinamik Bayes ağlarının kullanımı araştırılmıştır[133].

Sürekli yoğunluklu Saklı markov model tabanlı gürbüz konuşma tanıma için min-maks karar kuralının viterbi aramasıyla birleştirildiği model önerilmiştir.

Telefon konuşmalarında geniş dağarcıklı srürekli konuşma tanıma sistemlerinde hece seviyesinde akustik birim kullanan ilk gürbüz tanıma sistemi geliştirimiştir[134].

Gizli markov modeli için yeni bir regresyon tabanlı Bayes kestirim sınıflandırması geliştirilmiştir[135].

Multimodal Konuşma Tanıma

İnsanlar birbirleriyle iletişimlerinde multimodal iletişim kullanmaktadırlar. Konuşmanın anlaşılabilirliği çalışmaları, hem göresel hem de sesli bilgiye sahip olmanın özellikle mesaj karmaşık olduğunda veya iletişim gürültülü bir ortamda gerçekleştiğinde, başarılı bir bilgi transfer oranı yakalandığı göstermiştir.

---

[132] Tatsuya, Kawahara, Chin-Hui, Lee ve Biing-Hwang, Juang (1998), "Key Phrase Detection and Verification for Flexible Speech Understanding", *IEEE Transactions of Speech and Audio Proc.*, s. 558-568.

[133] C., Mayers ve L., Rabiner (1981), "Connected word recognition using a level building dynamic time warping algorithm", *Acouistics, Speech and Signal Processing*.

[134] M., Karnjanedecha ve S.A., Zahorian (2001), "Signal modeling for high-performance robust isolated word recogniton", *IEEE Transactions on Speech and Audio Processing*, s. 647-654.

[135] Rita, Singh, Bhiksha, Raj ve Richard, M Stern (2002), "Automatic Generation of Subword Units for Speech Recognition Systems", *IEEE Transactions on Speech and Audio Processing*, s. 89-99.

Konuşma tanıma sistemlerinde özelikle dudak bilgisinin kullanılması araştırılmış ve gürültülü ortamlarda ses tanıma başarısının düştüğü durumlarda daha iyi tanıma performası verdiği gösterilmiştir[136].

Gürültülü konuşma için seçim yaklaşımını içeren yeni bir altuzay modellemesi önerilmiştir. Bu modelde, gürültülü konuşma için faktör analizi temsilinin genelleştirilmesi üzerine çalışmalar yapılmıştır. Alt uzay seçimiyle küçük konuşma birimlerine eşdeğer öz değer vektörlerinin test edilmesi ile daha yüksek tanıma doğruluğu sağlanmıştır[137]. Eğitim ve test verilerinde çevre etkisi tanıma performansını ciddi olarak etkilerken, çoklu dil modelleme ile bu probleme çözüm üretilir.

Modelleme Teknikleri

Saklı Markov Model tabanlı sürekli konuşma tanıma da konuşma doğrulama için bir dizi akustik modelleme ve kod çözme tekniği geliştirilmiştir[138].

Konuşmanın kısa zaman özelliklerini özetlemek için yüksek boyutlu özellik vektörlerine dayanan otomatik konuşma tanıma için SMM hakkında çalışılmıştır.[139] Gözlemlenen konuşma işaretlerinin olasılığını en yükseğe çıkarmak veya sınıflandırma hatalarını en aza indirgemek için iki farklı metod önerilmiştir.

Ayrık eğitim algoritması ile karar destek vektörlerini kullanan saklı markov modellerinin gelişmiş sınıflandırılması yapılmıştır.

---

[136] Kimura, Shinta (1999), *Advances in Speech Recognition Technologies,* Teknik Rapor, FUJITSU.
[137] Bellgarda, Jerome R. (2000), "Large vocabulary Speech Recognition with Multispan statistical Language Models", *IEEE Transactions of Speech and Audio Processing*, Sayı: 200, s. 76-84.
[138] Eduardo, Lledia ve Richard C., Rose (2000), "Utterance Verification in Continuous Speech Recognition: Decoding and Training Procedures", *IEEE Transactions on Speech and Audio Processing*, s. 126-139.
[139] L.K., Saul ve M.G., Rahim(2000), "Maximum likelihood and minimum classification error factor analysis for automatic speech recognition", *IEEE Transactions on Speech and Audio Processing*, s. 115-125.

Ayrık ve sürekli çıktıların sınıflandırmalarını iyileştirmek için karar destek vektörlerini kullanan ayırt edici bir algoritma sunulmuştur[140]. Algoritma sistemin tabanında maksimum olasılık seti kullanırken, sonuçların tekrar değerlendirilmesi için karar destek sistemleri kullanılmaktadır.

Saklı markov modellerine bulanık mantık yaklaşımının kullanılması önerilmiştir. Konuşmacı ve konuşma tanıma uygulamalarında bulunık beklenti maksimizasyonu algoritması kullanılmıştır[141]. Bu yaklaşım için Baum Welch algoritması kullanımı önerilmiştir. Ayrık ve sürekli saklı markov parametrelerinin tahmininin nasıl yapılacağı bu algoritma ile açıklanmış ve 100 kelime için konuşma tanıma performansı incelenmiştir. Bulanık mantık tabanlı saklı markov modelininin geleneksel modele göre daha iyi sonuçlar verdiği gözlemlenmiştir.

Gürültülü Konuşma Tanıma

Gürültülü konuşma tanıma için konuşma özelliği artırımında kullanılan yeni bir yaklaşım önerilmiştir[142]. Temiz konuşma dağılımı için anahtarlamalı doğrusal dinamik modelleme araştırılmıştır. Bu yaklaşım ile gürültülü konuşmalar için gürbüz tanıma doğruluğu yakalanmış ve daha da geliştirilebileceği öngörülmüştür.

[140] Andras, Kocsor ve Laszlo, Toth (2004), "Kernel Based Feature Extraction with a Speech Technology Application", *IEEE Transactions on Signal Processing*, s. 2250-2263.
[141] Norden, E. Huang, Zheng, Shen, Steven, R. Long, Manli, C., Wu, Hsing, H. Shih, Quanan, Zheng, Nai-Chyuan, Yen, Chi Chao, Tung ve Henry H., Liu (1998), "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis", *Proceedings of the Royal Society a Mathematical, Physical and Engineering Sciences*, s. 903-995.
[142] Chia Yu, Wan ve Lin Shan, Lee (2008), "Histogram-based Quantization for Robust and/or Discributed Speech Recognition", *IEEE Transactions on Audio, Speech and Language Processing,* s. 859-873.

Veri Güdümlü Yaklaşımı

Bileşik sözcük üretimi için yeni bir yaklaşım önerilmiştir[143]. Bileşik sözcük üretme konuşma tanıma hatalarının azaltılması ve var olan hataların kısa sürede düzeltilmesi için kullanılabilmektedir.

4.7 TÜRKÇE KONUŞMA TANIMA SİSTEMLERİ LİTERATÜR ÖZETİ

Türkçe özgür dizilime imkân veren bir dildir. Kurucu unsurlarından kök ve türetici ekler ile birlikte eklemeli bir dil yapısına sahip olması serbest hareket imkânı vermektedir. Bu yöntemle üretilebilecek sözcük formalarının sayısı milyonlar seviyesinde olabilmektedir[144]. Örneğin, bir Türkçe haber sitesinden 10 milyon seviyesinde kelime kütüğüne ulaşması çok kolay olabilmektedir. Bu şekilde geniş dağarcık ciddi problemleri de beraberinde getirir ve dil modeli oluşturulmasında tahmin edilmesi gereken parametrelerin sayısını önemli ölçüde artırmaktadır. Ayrıca, boyutun büyümesi tanıma sırasında karmaşıklığın da artışına neden olmaktadır. Tablo-7'de 1M ve 10M kelime için 3-gram dil modeli oluşturulması halinde karışıklığa sebep olabilme oranları verilmiştir.

Tablo 7 Türkçe Sözlük Kütüğü 3-gram Dil Modeli Karışıklığı

Eğitim Seti	Eğitim Seti Karışıklığı	Test Seti Karışıklığı(1M)
1 Milyon	66.13	1449.81
10 Milyon	94.08	1084.13

[143] Hui, Jiang, K., Kirose ve Qiang, Hue (2000), "A minimax search algorithm for robust continuous speech recognition", *IEEE Transactions on Speech and Audio Processing*, s. 688-694.
[144] Hankamer, J.(1989), Lexical Representation and Process, Morphological Parsing and the Lexicon, MIT Press.

Bu şekilde yüksek değişkenliğe sahip Türkçe konuşma tanıma sistemlerinde hatalı kelime tespit oranları oldukça yüksek olabilmektedir. Bunun sebebi, Türkçe dil haznesinin oldukça büyük olması ve tanıma modellerinde dil modeli oluşturulması sonra dağarcık dışı kelime sayısının fazla olabilmesi durumudur. Türkçe gibi üretken dillerde, geniş kelime dağarcık durumu metnin etiketlenmesi süreçlerinde de önemli bir husustur. İngilizce için, bir metin içerisinde farklı kelime sayısı çok az(100<) ve etiketlenmesi daha kolay olabilmekte iken; Türkçe için bu şekilde sınırlı etiketleme kaçınılmaz bir şekilde bilgi kaybına neden olmaktadır[145].

Türkçe dilinin eklemeli diller kategorisinde yer alması ve İngilizce gibi dillerden oldukça farklı bir yapıya sahip olması sebebiyle bu diller ile ilgili yapılan çalışmalar ve önerilen modeller Türkçe için uygun olmamaktadır. Bu tez çalışmasına başlanılması sürecinde, konu ile ilgili çalışmalar incelendiğinde, özellikle Türkçe için yapılmış çalışmalarda bu ayrımın vurgulandığı görülmüştür. Bu sebeple ses tanımaya ilişkin olarak çalışmalar Türkçe dili özelinde araştırılmış ve incelenmiştir. Bu inceleme sonrasında, Türkçe dili için Tablo-8'de görüldüğü şekli ile yıllara göre bildiri, makale sayılarındaki dağılım gösterilmiştir.

Tablo 7 Türkçe Dili ile İlgili Çalışmaların Tür ve Yıllara göre Dağılımı

	2000-2005	**2005-2010**	**2010-2016**	**2017**
Bildiri	10	19	17	2
Makale	6	10	6	2
Tez	25(2000 öncesi), 15(YL), 3(DR)	9(YL),4(DR)	26(YL),2(DR)	2(YL),1(DR)

[145] Tür, Dilek Z. Hakkani, Oflazer, Kemal ve Tür, Gökhan (1999), "Design for a Turkish treebank", *In proceedings of the 37th annual meeting of the Association for Computational Linguistics*, Maryland.

Çalışmaların yıllara göre değişimi incelendiğinde, son yıllarda Türkçe diline yönelik olarak ses tanıma uygulamalarında çalışmalar olduğu ve 2017 yılında toplam 4 adet çalışmanın yayınlandığı görülmüştür. Özellikle 2010-2016 yılları arasında ciddi sayıda yüksek lisans çalışması ve bir kaç adet doktora çalışması olduğu tespit edilmiştir. 2017 yılında doktora tez çalışması ses tanıma içeren bir çalışma olmakla birlikte coğrafi bilgi sistemleri konulu bir araştırmadır.

Ses tanıma teorisinin anlatıldığı ikinci bölümde ses tanıma sürecinin adımlarına ilişkin olarak verilmiş olan yöntem farklılıkları incelenen 62 adet [1-62][146] yayına göre değerlendirilmiş ve sonuçlar aşağıda verildiği gibi olmuştur.

- Konuşma moduna göre çalışmalar değerlendirildiğinde, akustik model eğitimi için kullanılan veri setleri 3-5 kelimeden oluşan komut temelli çalışmalardan [30, 32, 43, 48][146] başlamakla birlikte LDC üzerinde dağıtımı yapılan veri setlerinin tamamını kullanan çalışmalara kadar değişik tipte kullanımlar bulunmaktadır.
- Konuşma veri tabanların kullanımına bakıldığında, çalışmaların %90'lık kısmında Metu 1.0 olarak adlandırılan akustik veri setini kullanıldığı anlaşılmaktadır.
- Akustik öznitelik çıkarımı için kullanılan tekniklere bakıldığında çalışmaların hemen hemen tamamında MFKK temelli bir özellik çıkarımı yapıldığı anlaşılmıştır. Bunun yanında bazı çalışmalarda özniteliklerin sayılarını artırmak veya ihtiyaç duyulan bazı özniteliklere MFKK yoluyla erişilememesi durumlarını çözmek için bu yöntemle birlikte LPC ve PCA yöntemlerinin birlikte kullanıldığı görülmüştür. Yapılacak çalışmalarda özniteliklerin belirlenmesi sonrasında ihtiyaç olması halinde birden fazla yöntem birlikte kullanılabilmektedir.
- Tanıma birimine göre çalışmalar değerlendirildiğinde, Türkçenin sondan eklemeli dil olması ve dağarcık boyunun çok büyük olması sebebiyle hiç bir

[146] Ek-1

çalışmada kelime tabanlı bir yaklaşımın kullanılmadığı görülmüştür. Çalışmalarda kelime altı dil modelleriyle çalışılmıştır. Araştırmaların bir kısmı hece tabanlı olmakla birlikte, çalışmalarda daha çok morf tabanlı yöntem kullanıldığı tespit edilmiştir. Morf tabanlı sistemlerin daha kısıtlı bir dağarcıkla çalışmayı(Türkçe 30 morf ile temsil edilebilmektedir)mümkün kılması nedeniyle tercih edildiği düşünülmektedir. Böylece hem karmaşıklıktan, hem işlem gücü gereksiniminden kaçınılmıştır. Ancak bu tercihler, hata oranlarında yüksekliği de beraberinde getirmiştir.

- Konuşma tanıma yaklaşımına göre çalışmalar incelendiğinde, bir kısım çalışma haricinde geri kalan çalışmaların olasılıksal modellerle sınıflandırma yapmayı tercih ettiği görülmüştür. Çalışmaların kullandığı algoritmalar incelendiğinde; karar destek vektörü [9, 11, 19, 21, 38, 39][147], dinamik zaman bükmesi [6][147], yapay sinir ağı[15][147], bayes algoritması[25,39][147], bulanık mantık [44][147] gibi farkı seçimlerin olduğu ve 2017 yılındaki bir çalışmada da derin öğrenme yapısının kullanıldığı tespit edilmiştir.
- Dil modelinin oluşturulması için kullanılan kelime haznelerine göre değerlendirildiğinde, en fazla 184 milyon [20][147] kelimeden oluşan sözlüklerin oluşturulduğu ve bu sözlüklerden derlenen kelime grupları ile çalışmaların yürütüldüğü görülmüştür. Kelimelerin internet üzerinden genelde haber kaynaklarından temin edildiği ve amaca göre filtrelenerek kullanıldığı görülmüştür.
- Saklı markov modellerini yaratmak için HTK aracının, N-gram [53, 55, 57, 58, 62][147] şeklinde dil modellerini oluşturmak için SRILM [29,37][147] altyapısının, Akustik model eğitimi için KALDI yazılımının kullanıldığı görülmüştür.
- Örnekleme frekansı tercihinin çoğunlukla 16khz olduğu, bir çalışmada [10][148] 8khz örnekleme frekansı bulunan veri setlerinin tercih edildiği anlaşılmıştır.

Sonuçta tüm bu çalışmaların amacı konuşma sinyalinin işlenerek tanınmasıdır. Bu çalışmaların başarı oranları günümüzde komut sistemli modeller haricinde %100

[147] EK-1
[148] EK-1

seviyelerinde ulaşmış değildir. Türkçe için yapılmış olan literatürdeki araştırmalar incelendiğinde:

- 3-20 arasında komut tanıma üzerine şablon temelli olarak hazırlanmış sistemlerde %100 başarı oranı yakalandığı [30, 40][148],
- Küçük dağarcıkla, kişi bağımlı komut benzeri mekanizmalarda %95 seviyelerinde doğru tanımlamaların yapılabildiği,
- Dağarcığın büyümesi ve kişi bağımlılığından vazgeçilmesi durumlarında başarı seviyelerinde dramatik düşüşlerin olduğu,
- Genel olarak, başarı oranlarında %40'lı seviyelerden başlanılarak, problemin kapsamına göre %95 seviyelerine kadar yükseltilebildiği anlaşılmıştır.

Ayrıca, bu tez çalışmasında Türkçe dili için kullanılması planlanan derin öğrenme algoritmasının kullanıldığı makale ve bildiriler incelenmiştir. Bu inceleme sonucunda, Türkçe dili için bir adet bildiride [25][148] akustik model eğitiminde derin sinir ağlarının kullanıldığından bahsedilmiştir. Çalışma detaylı olarak incelendiğinde,

- Türkçe sözlü ders anlatım videolarının geniş dağarcıklı sürekli konuşma tanıma sistemi tasarımı ile yazılandırılmasının amaçlandığı,
- Akustik model olarak SMM modeli kullanılmıştır. SMM'de akustik öznitelik vektörlerinin koşullanmış olasılık hesaplanmasında, Gauss karışım modeli yerine çok katmanlı yapay sinir ağları tercih edilmiş olup, bu yöntemin konuşma tanıma başarımlarında ciddi ölçüde artış sağlanmasına imkan tanıdığı,
- "Wall Street Journal" ve "Switchboard" İngilizce veri tabanları üzerinde hazırlanan derin öğrenme yöntemlerini içeren eğitimi modellerinin birebir uygulamasının Türkçe diline uygulandığı,
- 4 katmanlı bir yapay sinir ağının tasarlandığı,
- %14.18 kelime hata oranının yakalandığı, değerlendirilmiştir.

BEŞİNCİ BÖLÜM

SONUÇ VE ÖNERİLER

Konuşma tanıma sisteminin baştan sonra tasarlanması, bu tez çalışmasının önceki bölümlerinde de bahsedildiği gibi oldukça fazla aşamaya sahip bir problemdir. Çünkü herhangi bir ses tanıma sistemi tasarımında birçok aşama bulunmaktadır ve her bir aşama için kullanılabilir birden fazla yöntem vardır. En iyi performansa sahip bir modelin yapılabilmesi için tüm bu yaklaşımların gözden geçirilmesi, deneysel çalışmalar ile incelenmesi ve en iyi yönteme karar verilmesi gerekmektedir. Bu sebeple bu tez çalışmasında ses tanıma problemi belirli şartlarda özelleştirilerek modelleme yapılması ve ortaya çalışabilir bir yazılımın ortaya koyulması amaçlanmıştır.

Temel olarak bir konuşma tasarımına başlamadan önce aşağıdaki sorulara cevap vererek, problemin sınırları belirlenmelidir.

1- Konuşma tanıma sisteminin teorisi, sinyal işleme tekniklerinin temel mantığı ve matematiksel modellemesi iyi bir şekilde anlaşılmalıdır.
2- Bölüm-1, şekil-2'de gösterilmiş olan konuşma işleme sınıflandırma şemasına göre öncelikle problemin:
    a. Konuşmacı Tanıma,
    b. Konuşma Tanıma,
    c. Konuşma Dili Tanıma, türlerinden hangisinin yapılacağına karar verilmelidir.

3- Bu tez çalışmasında konuşma tanıma problemi tercih edilmiştir. Konuşma tanıma türü seçildikten sonra:
    a. Konuşma Modu,
    b. Konuşmacı Modu,
    c. Sözlük Boyutu,
    d. Konuşma Stili, başlıklarına göre problem daha özel bir hale getirilmelidir. Aksi halde her bir durum için ayrı ayrı çalışılması ve birbirlerini etkileyecek unsurlarında değerlendirilmesi gerekmektedir.

Bir ses tanıma uygulamasının spontan konuşmacı bağımsız ve geniş bir sözcük haznesi ile konuşulan sesin dikte edilmesi en tercih edilen problem olarak karşımıza çıkmaktadır.

4- İlk 3 adım sonrasında, artık ses tanıma sisteminin tasarımına geçilmektedir. İlk olarak Akustik ön uç sisteminin yapılmasıdır. Bu adımda ses sinyalinin daha işlenebilir bir hale dönüştürülmesi işlemi yani özellik çıkarımı yapılmalıdır. Özellik çıkarılabilmesi için araştırmalarda kullanılan birçok farklı algoritma bulunmaktadır(Tablo-1). Bu algoritmaların probleme göre avantajlı veya dezavantajlı olduğu noktalar vardır. Sistem tasarımında bu yöntemlerden bir veya birkaçı birlikte seçilerek kullanılmalıdır.
5- Akustik ön uç tasarımı sonrasında, dil modeli hazırlanmalıdır. Dil modeli tasarımında en kritik husus "kelime, hece, morf" gibi birimlerden hangisi ile tanıma işleminin gerçekleştirileceğine karar verilmesidir. Bu seçim tüm ses tanıma sistemini doğrudan etkileyen, problemin en başında karar verilmesi gereken bir hususdur.
6- Bu işlemler sonrasında akustik ön uç ve dil modelinin birleştirilerek konuşma seslerinin tanınması sürecinde ilk karar verilmesi gereken husus ne tür bir yaklaşımın kullanılacağıdır. Elde ki bilgilerin işlenerek kullanılmasında (Tablo-2) de gösterilmiş olan çok farklı algoritmalar vardır. Problemin türüne göre her birinin avantajlı ve dezavantajlı olduğu durumlar vardır. Bunlar değerlendirilerek bir seçim yapılmalıdır.
7- Konuşma tanıma sistemi tasarlandıktan sonra, öncelikle sistemin eğitilmesi sonrasında da test edilebilmesi için hem akustik model hemde dil modeli için verisetlerine ihtiyaç duyulmaktadır. Verisetlerinin hazırlanması ayrı bir işlem

olarak değerlendirilmelidir. Ancak ses tanıma araştırmalarında bu aşamada hali hazırda Türkçe ve diğer diller için dağıtımı yapılan verisetlerinin kullanımı tercih edilmektedir. Tablo-3'de Türkçe dışındaki diller ve Tablo-4'te Türkçe dili için kullanılabilir verisetleri listesi verilmiştir. Bunların dağıtımı açık olmayıp ücretli olarak temin edilebilmektedir.

8- Akustik ön uç, dil modeli tasarımlarını oluşturmak için yazılımsal altyapı sunan birçok farklı araç bulunmaktadır. Bölüm 2.4'te gösterilmiş olan araçlardan birisi seçilerek uygulamanın yazılımsal geliştirmesi yapılmalıdır.
9- Bu tez çalışması için sistem kullanımında derin öğrenme algoritmasının daha başarılı olduğu değerlendirilmiştir. Derin öğrenme algoritmaları 5 farklı kategoriye ayrılmakta ve kendisine göre ses tanıma problemlerine çözüm üretebilmektedir. Ses tanıma sistemlerinde en başarılı yaklaşım olan Tekrarlayan Sinir Ağları son dönem araştırmalarda tercih edilen derin öğrenme türü olmuştur.
10- Öğrenme sürecinde eldeki veriye göre denetimli veya denetimsiz yaklaşımlardan birisi tercih edilebilir. Eğer veride girdi ve etiketler var ise denetimli yaklaşım tercih edilmesi faydalı olacaktır. Eğitim aşaması hızlı bir şekilde geçilebilecektir.
11- Derin sinir ağlarının çalışması için gerekli olan aktivasyon ve hata fonksiyonlarının seçimi veya bunların birlikte kullanıldığı hibrit bir modelin oluşturulması gerekmektedir.
12- Derin sinir ağlarını yazılımsal olarak oluşturmak için birçok derin öğrenme kütüphanesi bulunmaktadır. Bunlardan birisi seçilerek sinir ağı oluşturulmalıdır.

Yukarıda verilmiş olan bu 12 adımlı problem çözüm aşamaları sırasıyla değerlendilerek bir konuşma tanıma sistemi tasarımı yapılabilecektir.

Bu tez çalışmasında Türkçe dili için derin öğrenme altyapısını kullanan bir konuşma tanıma sistemi tasarlanması amaçlanmıştır. Bu sistemin aşağıdaki özelliklere sahip olması ve gösterilmiş olan probleme çözüm üretmesi düşünülmüştür.

1- Sistem konuşma tanıma (dikte) işlemi yapacaktır.
2- Öncelikle konuşma tanıma tipine göre yalıtık sözcük tanıma yapması, sonrasında bunun sürekli konuşma tanıma şeklinde çalışması düşünülmektedir.
3- Konuşmacı moduna göre konuşmacı bağımlı olarak çalışacaktır.
4- Sözlük boyutu ilk aşama için küçük seçilecektir.
5- Akustik ön uç tasarımında özellik çıkarım algoritmalarından en yaygın olan ve konuşma tanıma sistemlerinde öncelikli olarak tercih edilen MFKK yöntemi kullanılacaktır. İhtiyaç duyulması halinde Doğrusal Öngörülü Kodlama(DÖK) ile birlikte uygulanabilir.
6- Dil modeli tasarımı harf tabanlı olarak yapılacaktır.
7- Konuşma tanıma yaklaşımı olarak Yapay sinir ağının bir türü olan Derin Sinir Ağları kullanılacaktır.
8- Türkçe konuşma tanıma yapılacağı için, Türkçe konuşma verisetlerinden METU 1.0 kullanılacaktır.
9- Aktivasyon fonksiyonu olarak Sigmoid veya Gauss, Hata fonksiyonu olarak Eğim Azalımı yöntemi kullanılacaktır.
10- Geliştirme Python dili ile TensorFlow Kütüphanesi kullanılarak yapılacaktır.

Yukarıda bahsedilmiş olan kriterler dahilinde Python dili ile konuşmacı bağımlı bir ses tanıma sistemi tasarlanarak sonuçlar değerlendirilmiştir. Bu değerlendirme için oluşturulan ortamın topolojisi Tablo 8'de gösterildiği gibidir.

Tablo 8 Test Ortamı Topolojisi

Optimizasyon Tekniği	**EK**	**EDV**	**SS**	**Epoch**	**PB**	**ÖS**	**Ortalama**
GradientDescent	0.01	0-10	0.1	5000	1	1	0
ProximalGradientdescent	0.01	0-10	0.1	5000	1	1	0
RMSPropOptimizer	0.01	0-10	0.1	5000	1	1	0
MomentumOptimizer	0.005	0-10	0.01	5000	1	1	0
AdadeltaOptimizer	10	0-10	0.1	5000	1	1	0
AdagradOptimizer	0.1	0-10	0.1	5000	1	1	0
AdamOptimizer	0.01	0-10	0.1	5000	1	1	0
FtrlOptimizer	0.01	0-10	0.1	5000	1	1	0

ProximalAdagradOptimizer	0.01	0-10	0.1	5000	1	1	0

EK: Eğitim Katsayısı, EDV: Eğitim ve Doğrulama Veriseti, SS: Standart Sapma, PB: Paket Büyüklüğü, ÖS: Örnek Sayısı

Geliştirme ortamı hazırlandıktan sonra, derin öğrenme algoritmasının çalıştırılabilmesi için Tablo-8'de verilmiş olan tasarım topolojisine karar verilmesi gerekliliği bulunmaktadır. Bu yaklaşımda kullanılacak olan optimizasyon teknikleri için yapılan testler sonucunda her bir teknik için en iyi parametreler belirlenmiş olup, buna göre Tablo 8'de gösterildiği gibi her bir deney için farklı değerler kullanılarak daha iyi performansa sahip bir tasarım yapılması amaçlanmıştır. Performans değerlendirmesine doğrudan olumlu katkısı sebebiyle Epoch sayısı tümü için aynı tutulmuştur. Bu topolojik yaklaşıma göre kişi bağımlı konuşma tanıma testlerinde elde edilen sonuçlar Tablo 9'da gösterildiği gibidir.

Tablo 9 50 Adet Deney Ortamı Performans Değerlendirmesi

Optimizasyon Tekniği	**EK**	**SS**	**Epoch**	**EHO <50 %**	**EHO = 0 %**	**DHO <50 %**	**DHO = 0 %**	**OEHO**	**ODHO**
ProximalGradientDescent	0.01	0.1	20000	0.98075	0.5695	0.97895	0.4061	0.032003	0.042267
GradientDescent	0.01	0.1	10000	0.9648	0.5561	0.9613	0.4085	0.047548	0.056826
GradientDescent	0.01	0.1	20000	0.97715	0.57595	0.97545	0.1214	0.034652	0.087925
GradientDescent	0.01	0.1	5000	0.9458	0.5234	0.9368	0.3446	0.069117	0.091203
ProximalGradientdescent	0.01	0.1	5000	0.9238	0.5186	0.9136	0.3644	0.08813	0.102717
RMSPropOptimizer	0.01	0.1	5000	0.957	0.4692	0.9524	0.1248	0.06259	0.102793
GradientDescent	0.01	0.05	5000	0.925	0.5088	0.9168	0.414	0.088354	0.102867
GradientDescent	0.01	0.1	5000	0.9186	0.5028	0.9128	0.377	0.091645	0.103011
GradientDescent	0.01	0.5	5000	0.9456	0.4898	0.9366	0.1232	0.071248	0.114378
GradientDescent	0.01	0.1	5000	0.9056	0.4912	0.8958	0.317	0.106985	0.125648
GradientDescent	0.01	0.1	5000	0.9176	0.4964	0.9054	0.1874	0.096512	0.127473
GradientDescent	0.01	0.01	5000	0.9102	0.4724	0.9026	0.2934	0.103163	0.128285
GradientDescent	0.01	0.01	5000	0.9098	0.472	0.8988	0.114	0.100365	0.146922
ProximalGradientDescent	0.005	0.1	5000	0.8936	0.4328	0.8808	0.2308	0.122935	0.149305
MomentumOptimizer	0.005	0.01	5000	0.8772	0.4686	0.8676	0.3512	0.135143	0.14953
AdadeltaOptimizer	10	0.1	5000	0.9002	0.3954	0.8914	0.1198	0.122551	0.164146
AdagradOptimizer	0.1	0.1	5000	0.9134	0.3084	0.9016	0.0198	0.109114	0.180349
MomentumOptimizer	0.005	0.01	5000	0.9326	0.4976	0.9202	0.0252	0.080695	0.196193
MomentumOptimizer	0.005	0.003	5000	0.9448	0.5212	0.9354	0.0276	0.067885	0.199259
AdamOptimizer	0.01	0.1	5000	0.8484	0.3958	0.8426	0.0444	0.16545	0.225922
AdamOptimizer	0.01	0.1	5000	0.8402	0.4374	0.8332	0.0406	0.166063	0.236744
GradientDescent	0.015	0.1	5000	0.8192	0.3696	0.7966	0.056	0.202467	0.288937
ProximalGradientDescent	0.01	0.1	10000	0.7345	0.288	0.6525	0.0077	0.268174	0.405148
MomentumOptimizer	0.005	0.5	5000	0.933	0.4726	0.8116	0.0064	0.084915	0.405632
GradientDescent	0.005	0.1	20000	0.8386	0.32865	0.68525	0.0058	1.924264	0.410108

Optimizasyon Tekniği	EK	SS	Epoch	EHO <50 %	EHO = 0 %	DHO <50 %	DHO = 0 %	OEHO	ODHO
GradientDescent	0.001	0.1	5000	0.6022	0.0782	0.5754	0.0382	0.423882	0.450963
RMSPropOptimizer	0.0007	0.1	5000	0.523	0.0166	0.4958	0	0.49724	0.530946
GradientDescent	0.01	0.1	5000	0.5436	0.088	0.3908	0.0012	0.429627	0.562687
MomentumOptimizer	0.01	0.1	5000	0.6648	0.1992	0.0664	0.0002	0.335383	0.679394
MomentumOptimizer	0.005	0.1	5000	0.4758	0.0926	0.1024	0	0.472786	0.6861
AdagradOptimizer	0.01	0.1	5000	0.2214	0	0.1568	0	0.690801	0.716531
GradientDescent	0.05	0.5	5000	0.2222	0	0.015	0	0.614395	0.719488
FtrlOptimizer	0.01	0.1	5000	0.1198	0	0.0766	0	0.739715	0.756674
MomentumOptimizer	0.005	0.01	5000	0.8358	0.356	0	0	0.171205	0.770817
ProximalAdagradOptimizer	0.01	0.1	5000	0.0862	0	0.049	0	0.75492	0.775987
ProximalAdagradOptimizer	0.01	0.1	5000	0.068	0	0.0336	0	0.769071	0.78198
GradientDescent	0.05	0.5	5000	0.119	0	0.0002	0	0.629661	0.788874
Adamoptimizer	0.001	0.5	5000	0.0208	0	0.0102	0	0.781129	0.792159
GradientDescent	0.05	0.1	5000	0.2032	0	0.0002	0	0.638674	0.809957
GradientDescent	0.05	0.5	5000	0.006	0	0.0008	0	0.749095	0.816008
GradientDescent	0.0001	0.1	5000	0	0	0	0	0.943463	0.94346
AdamOptimizer	0.0001	0.1	5000	0	0	0	0	0.952913	0.953928
AdadeltaOptimizer	0.1	0.1	5000	0	0	0	0	0.958979	0.957662
MomentumOptimizer	0.1	0.01	5000	0	0	0	0	1.129285	1.111464
GradientDescent	0.5	0.1	5000	0	0	0	0	1.200547	1.172174
AdadeltaOptimizer	0.01	0.1	5000	0	0	0	0	1.505803	1.494673

EK: Eğitim Katsayısı, SS: Standart Sapma, EHO: Eğitim Hata Oranı, DHO: Doğrulama Hata Oranı, ÖS: Örnek Sayısı, OEHO: Ortalama Eğitim Hata Oranı, ODHA: Ortalama Doğrulama Hata Oranı

Sonuçlar incelendiğinde optimizasyon tekniği, öğrenme katsayısı, standart sapma değeri ve eğitimin Epoch değerinin performans değerlendirmesine doğrudan etki ettiği görülmektedir. Tablo 9'da gösterildiği gibi aynı şartlarda aynı veriseti ile yapılan çalışmalarda hata oranları %90 seviyelerinden %10 seviyesine kadar inebilmektedir. Değerlerin doğru seçilmesi ile performans ölçümlerinde ciddi gelişme kaydedilebilmektedir. Ayrıca tabloda bazı test setlerinde sonuçların sıfır olduğu görülmektedir. Bu durum oluşturulan test setinin öğrenme yapmadığı, eğitim aralığında ciddi sapmalar gösterdiği ve %100 hata oranı dolayısı ile tanıma işlemini yapamadığını göstermektedir.

Kişi bağımlı yapılan çalışmalarda en iyi değerler "Proximal Gradient Descent", "Gradint Descent" ve "RMSProp" optimizasyon teknikleri ile elde edilmiştir. Her 3

yöntem ile Epoch sayısının 5000 olarak belirlendiği testlerde doğru sınıflandırma oranları %90 seviyesinin üstüne çıkmaktadır. %90 seviyesinin üzerindeki tüm modellemede sonuçların artık doğruya yakın olduğu kabul edilmektedir.

Bu sonuçların alınmasından sonra, modelin biraz daha geliştirilerek bu tez çalışmasında önerilen derin öğrenme mimarisini daha iyi anlayabilmek açısından gizli katman sayısının değişiminin sonuçlara etkisi değerlendirilmeye çalışılmıştır. Buna göre Tablo 10'da verilen sonuçlar elde edilmiştir.

Tablo 10 Gizli Katman Sayısının Performansa Etkisi

Optimizasyon Tekniği	Gizli Katman	OEHO	ODHO
Gradientdescent	10	0.866773087	0.875516779
Gradientdescent	100	0.469775197	0.498840039
Gradientdescent	200	0.387289847	0.464624745
Gradientdescent	300	0.788613832	0.831364653

OEHO: Ortalama Eğitim Hata Oranı, ODHO: Ortalama Doğrulama Hata Oranı

Tablo 9'da verilmiş olan sonuçlar değerlendirildiğinde en iyi sonuçların "Gradient Descent" optimizasyon ile elde edildiği görülmektedir. Buna göre bu teknik sabit tutularak gizli katman sayısı değiştirilerek Tablo 10'daki sonuçlar elde edilmiştir. Buna göre gizli katman sayısının sürekli artırılmasının performansa sürekli katkı yaptığı söylenemez. Derinliğin artırılması öğrenme durumunda sapmalara neden olmakta ve en iyi sonuçların üretildiği bölümden kaymalara neden olmaktadır. Bu noktada en kritik hususun öğrenme katsayısı ile ağın derinliğinin doğru belirlenmesi gerektiği sonucu elde edilmiştir.

Bu noktada Epoch sayısının artırılmasının sonuçlara doğrudan etki ettiğinden bahsedilmiştir. Buna ilişkin değerlendirmeleri yapabilmek için oluşturulmuş olan model en iyi optimizasyon teknikleri sabit tutularak Epoch değeri değişkenliğinin sonuçlara etkisi değerlendirilmiştir. Tablo 11'de sonuçların detayları verilmiştir.

Tablo 11 Epoch sayısının Performansa Etkisi

Epoch	EHO<50 %	EHO=0 %	DHO <50%	DHO = 0 %	OEHO	ODHO
100	0	0	0	0	0.942347	0.948248
200	0	0	0	0	0.88488	0.444679
500	0.006	0	0	0	0.871097	0.876559
1000	0.661	0.154	0.635	0.113	0.362153	0.397979
5000	0.9238	0.5186	0.9136	0.3644	0.08813	0.102717
10000	0.96	0.5605	0.9564	0.3417	0.050359	0.066309
20000	0.98075	0.5695	0.97895	0.4061	0.032003	0.042267
50000	0.96926	0.53706	0.87012	0.00634	0.050017	0.321762

EHO: Eğitim Hata Oranı, DHO: Doğrulama Hata Oranı, OEHO: Ortalama Eğitim Hata Oranı, ODHA: Ortalama Doğrulama Hata Oranı

Tablo 11 incelendiğinde en iyi sonuçların Epoch değerinin 20000 olarak belirlendiği durumda elde edildiği görülmektedir. Bu durumda Epoch değerinin çok küçük veya gereksiz yere çok yüksek belirlenmesinin sonuçlara olumlu katkı sağlamadığı görülmektedir. Epoch değerinin az seçilmesi modelin eğitim süresini ve maliyetini azaltırken hata oranlarında ciddi artışa sebep olurken, çok seçilmesi de hem eğitim süresi hemde maliyetini artırmaktadır hemde hata oranlarında ciddi artışlara sebep olabilmektedir. Ayrıca sonuçlarda performans artışlarının sağlanması için dil modellemesine de ihtiyaç duyulabilmektedir. %90 ve üzerindeki doğruluklarla tanıma yapılan durumlar birkaç adet harf hatasının düzeltilebilmesi için dil modelinden yararlanılmaktadır. Bu çalışmasının devamında dil modellemesi yapılarak entegre edilmesi durumunda daha iyi sonuçlar yakalanabilecektir.

Bu tez çalışmasın için yapılmış olan çalışmalar ve deneyler genel olarak değerlendirilecek olursa, bu tezin genelinde de bahsedildiği ve sonuç bölümünde örnek bir çalışma üzerinde yapılan test sonuçlarında gösterildiği gibi, herhangi bir tipteki konuşma tanıma modelinin tasarımında birçok parametre bulunmaktadır ve bu parametrelerin sonuçlara ciddi etkileri olmaktadır. En iyi modelin ne olduğunun

belirlenmesi mümkün olmamakla birlikte problemin detaylı olarak belirlenmesi ve deney gözlem yoluyla modellerin optimize edilmesi ile daha iyi performans sonuçları yakalanabilmektedir. Bu tez çalışmasına ilişkin olarak tasarlanmış olan sistemin çalışmalarının tamamlanması ve tüm durumlarda belirli bir performans değerinin yakalanması halinde kurumumuzda kullanılmakta olan Sayısal Kayıt, Arşiv, Analiz Sistemi(SKAAS) ile entegre edilerek videolardan veya seslerden dikte işlemlerinin otomatik olarak yapılması amaçlanmaktadır. Böylece yapılan çalışmanın Radyo ve Televizyon Üst Kurulu süreçlerine katkı sağlaması mümkün olacaktır.

Bu çalışmada gerçek hayatta karşımıza çıkabilecek ses tanıma problemlerin bazı türleri için çözüm üretecek tasarım yapılmıştır. Ancak tv veya radyo yayınları gibi süreklilik arz eden konuşmaların metne dönüştürülmesi için daha geniş verisetleri ve uzun süreli eğitim algoritmalarına ihtiyaç duyulmaktadır. Bu çalışmasının bir sonraki aşamaları için model tasarımının karmaşıklığının ve derinliğinin artırılması ve deney süresinin uzatılması ile daha iyi bir performansa sahip ses tanıma programı geliştirilebilecektir. Bunun yanında derin öğrenme sistemleri için kulanılabilir daha performanslı sunuculara erişim mümkün olması halinde yüksek Epoch değerleri ile çalışılarak sınıflandırma sonuçlarında doğruluk artırılabilecektir.

KAYNAKÇA

A. Kocsor, L. Toth. «Kernel Based Feature Extraction with a Speech Technology Application.» *IEEE Transactions on Signal Processing*, 2004: 2250-2263.

A. Sankar, Chin-Hui Lee. «A Maximum likelihood approach to stochastic matching for robust speech recognition.» *IEEE Transactions on Speech and Audio Processing*, 1996: 190-202.

Aachen University Speech Recognition System. 15 11 2016. https://www-i6.informatik.rwth-aachen.de/rwth-asr/ (erişildi: 01 01, 2018).

Adoram Erell, Mitchel Weintraub. «Filterbank-Energy Estimation Using Mixture and Markov Models for Recognition of Noisy Speech.» *IEEE Transactions on speech and Audio Processing*, 1993: 68-76.

Andras Kocsor, Laszlo Toth. «Kernel Based Feature Extraction with a Speech Technology Application.» *IEEE Transactions on Signal Processing*, 2004: 2250-2263.

Anusuya, M.A., ve S.K. Katti. «Speech Recogniton by Machine: A Review.» *International Journal of Computer Science and Information Security*, 2009: 181-205.

Anwer, Ali Mahmood Ogur. *Derin Öğrenme Yöntemleri ile Göğüs Kanseri Teşhisi.* Yüksek Lisans Tezi, Ankara: THK Üniversitesi Fen Bilimleri Enst., 2017.

Bellgarda, Jerome R. «Large vocabulary Speech Recognition with Multispan statistical Language Models.» *IEEE Transactions of Speech and Audio Processing*, 200: 76-84.

Bing Hwang Juang, Sadaoki Furui. «Automatic Recognition and understanding of spoken language a first step towards natural human machine communication.» *Proceedings of the IEEE*, 2000: 1142-1165.

Bing-Hwang Juang, S. Furui. «Automatic Recognition and understanding of spoken language - a first step toward natural human-machine communication.» *Proceedings of the IEEE*, 2000: 1142-1165.

Bling Hwang Juang, M.G.Rahim. «Signal bias removal by maximu likelihood estimation for robust telephone speech recognition.» *IEEE Transactions on Speech and Audio Processing*, 1996.

C Gaida, P. Lange, R.Petrick, P.Proba, A. Malatawy, A.D. Suendermann-Oeft. *Comparing open-source speech recognitions toolkits.* Teknik Rapor, Stuttgart: DHBW, 2014.

C. Mayers, L. Rabiner. «Connected word recognition using a level building dynamic time warping algorithm.» *Acouistics, Speech and Signal Processing*, 1981.

C.H. Lee, L. R. Rabiner, R.Pieraccini, J.G. Wilpon. «Acoustic Modeling for Large Vocabulary Speech Recognition.» *Computer Speech and Language*, 1990: 127-165.

Campbell, Joseph P. «Speaker recognition: A tutorial.» *Proceeedings of the IEEE*, 1997: 1437-1462.

Chia Yu Wan, Lin Shan Lee. «Histogram-based Quantization for Robust and/or Discributed Speech Recognition.» *IEEE Transactions on Audio, Speech and Language Processing*, 2008: 859-873.

CMU Arctic Database. 01 01 2018. http://www.festvox.org/cmu arctic/ (erişildi: 01 01, 2018).

CSL. 01 01 2018. http://www.csl.uni-bremen.de/cms/en/research/automatic-speech-recognition (erişildi: 01 01, 2018).

D. Povey, A.Ghoshal. «The Kaldi Speech Recognition Toolkit.» *Proceedings of ASRU.* Hawaii, 2011. 1-4.

D. Susman, S. Köprü, A. Yazıcı. «Kısıtlı Ses külliyatı ile türkçe geniş dağarcıklı sürekli konuşma tanıma.» *Signal Processing and Communications Applications Conference(SIU).* 2012.

Deng, Li. *Three Classes of Deep Learning Architectures and Their Applications.* Araştırma Raporu, Washington: Microsoft Research, 2012.

Dilek Z. Hakkani Tür, Kemal Oflazer, Gökhan Tür. «Design for a Turkish treebank.» *In proceedings of the 37th annual meeting of the Association for Computational Linguistics.* Maryland, 1999.

Doh-Suk Kim, Soo-Young Lee, R.M. Kil. «Auditory processing of speech signals for robust speech recognition in real-world noisy environments.» *IEEE Transactions on Speech and Audio Processing*, 1999: 55-69.

Dong Yu, Li Deng. *Automatic Speech Recognition- A deep learning approach.* London: Springer-Verlag, 2015.

Dong Yu, Li Deng, G Dahl. «Roles of Pre-Training and Fine-Tuning in Context-Dependent DBN-HMMs for Real-World Speech Recognition.» *Deep Learnign and Unsupervised Feature Learning*, 2010.

Eduardo Lledia, Richard C. Rose. «Utterance Verification in Continuous Speech Recognition: Decoding and Training Procedures.» *IEEE Transactions on Speech and Audio Processing*, 2000: 126-139.

Emotional prosody speech and transcripts databse from Linguistic Data Consortium(LDC). 01 01 2018. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2002S28 (erişildi: 01 01, 2018).

Ferguson, J. *Hidden Markov Models for Speech.* New Jersey: Princeton, 1980.

Frand Seide, Gang Li, Dong Yu. «Conversational Speech Transcription Using context-dependent Deep neural networks.» *In Proceedings of the Annual Conference of International Speech Communication Association(INTTERSPEECH)*, 2011: 437-440.

Fry, D.B. «Theoritical Aspects of Mechanical Speech Recognition.» *Journal of the British Institution of Radio Engineers* , 1959: 211-218.

Furui, Sadaoki. «Recent Progress in Corpus-Based Spontaneous Speech Recognition.» *IEICE Transactions of Informations and Systems*, 2005: 366-375.

G. Dahl, D. Yu, L. Deng, A. Acero. «Context-Dependet Pre-Training Deep Neural Networks for Large-Vocabulary Speech Recognition.» *IEEE Transactions on Audio, Speech and Language Processing*, 2012: 30-42.

Genevieve I Sapijaszko, Wasfy B. Mikhael. «An Overview of Recent Windows Based Feautre Extraction Algorithms for Speech Recognition.» *Circuits and Systems (MWSCAS), 2012 IEEE 55th International Midwest Symposium on.* Boise: IEEE, 2012.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Ebdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, Brian Kingsbury. «Deep Neural Networks for Acoustic Modelling in Speech Recognition: The Shared Views of Four Research Groups.» *IEEE Signal Processing Maganize*, 2012: 92-97.

Geoffrey Zweig, Patrick Nguyen. *SCARF: A Segmental CRF Speech Recognition System.* Teknik Raor, Microsoft, 2009.

H. Ba, N. Yang, I. Demirkol, and W. Heinzelman. «Bana: A hybrid approach for noise resilient pitch detection.» *Statistical Signal Processing Workshop (SSP)*, 2012: 369-372.

H. Sakoe, S. Chiba. «Dynamic Programming algorithm optimization for spoken word recognition.» *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1978: 43-49.

H.F. Olson, H. Belar. «Phonetic Typewriter.» *Journal of Acoustics Soc. am.* , 1956: 1072-1081.

Hankamer, J. *Lexical Representation and Process, Morphological Parsing and the Lexicon.* MIT Press, 1989.

Harvard Sentences Database in Open Speech Repository. 01 01 2018. http://www. voiptroubleshooter.com/open speech/american.html (erişildi: 01 01, 2018).

HTK. 01 01 2018. http://spandh.dcs.shef.ac.uk/ed_arena/htk/index.html (erişildi: 01 01, 2018).

Hui Jiang, K. Kirose, Qiang Hue. «A minimax search algorithm for robust continuous speech recognition.» *IEEE Transactions on Speech and Audio Processing*, 2000: 688-694.

IATROS. 27 04 2016. https://github.com/PRHLT/iatros-decoder (erişildi: 01 01, 2018).

IDIAP Juicer. 26 11 2015. https://github.com/idiap/juicer (erişildi: 01 01, 2018).

Itakura, F. «Minimum Prediction Residula Applied to Speech Recognition.» *IEEE Trans Acoustics, Speech and Signal Processing*, 1975: 67-72.

J Li, L. Deng, R.H. Umback, Yifan Gong. «Robust Automatic Speech Recognition: A Bridge to Practical Applications.» *Academic Press*, 2015.

J Li, L. Deng, Y. Gong, R.H. Umbach. «An Overview of Noise-Robust Automatic Speech Recognitional.» *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2014: 745-777.

J. Suzuki, K. Nakata. «Recognition of Japanese Vowels Preliminary to the Recognition of Speech .» *Journal of Radio Research Lab.*, 1961: 193-212.

J.S. Bridle, M.D. Brown, R.M. Chamberlain. «Continuous Connected word recognition using whoel word templates.» *Radio and Electronic Engineer*, 1983: 167-175.

James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins,. «Theano: A CPU and GPU Math Compiler in Python.» *PROC. OF THE 9th PYTHON IN SCIENCE CONFERENCE.* 2010.

James W. Forgie, Carma D. Forgie. «Results Obtained from a Vowel Recognition Computer Program.» *The Journal of the Acoustical Society of America*, 1959: 1480-1489.

Javier Gonzales Domingues, Ignacio Lopez Moreno, Javier Franco Pedroso, Daniel Ramos. «Multilevel and Session Variability Compensated Language Recognition: TVS-UAM Systems at NIST LRE 2009.» *IEEE Journal of Selected Topics in Signal Processing*, 2010: 1084-1093.

Jelinek, F. «The Development of an experimental discrete dictation recognizer.» *Proceedings of IEEE*, 1985: 1616-1624.

Ji Wan, Dayong Wang, Steven Hoi, Pengcheng Wu, Jianke Zhu, Jintao Li, Yongdong Zhang. «Deep learning for content-based image retrieval: A comprehensive study.» *Proceedings of the 22nd ACM International Conference on Multimedia: .* Orlando: ACM, 2014. 157-166.

JULIUS. 02 01 2017. https://github.com/julius-speech/julius (erişildi: 01 01, 2018).

K.H. Davis, R. Biddulph, S. Balashek. «Automatic Recognition of Spoken Digits.» *The Journal of The Acoustical Society of America*, 1952.

KALDI. 01 01 2018. http://kaldi-asr.org/doc/about.html (erişildi: 01 01, 2018).

Karpagavalli S, Chandra E. «A Review on Automatic Speech Recogniton Architecture and Approaches.» *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 2016: 393-404.

Keh Yih Su, Chin Hui Lee. «Speech Recognition using weighted HMM and subspace projection approaches.» *IEEE Transactions on Speech and Audio Processing*, 1994: 69-79.

Kimura, Shinta. *Advances in Speech Recognition Technologies.* Teknik Rapor, FUJITSU, 1999.

Klatt, Dennis H. «Review of the ARPA Speech Understanding Project.» *The Journal of the Acoustical Society of America*, 1977: 1324-1366.

L.K.Saul, M.G. Rahim. «Maximum likelihood and minimum classification error factor analysis for automatic speech recognition.» *IEEE Transactions on Speech and Audio Processing*, 2000: 115-125.

L.R. Bahl, P.F. Brown, P.V. de Souza, R.L. Mercer, M.A. Picheny. «A Method for the construction of acoustic markov models for words.» *IEEE Transactions on Speech and Audio Processing*, 1993: 443-452.

L.R. Labiner, B.H. Juang. *Fundamentals of Speech Recognition.* Printice-Hall, 1993.

Li Deng, Dong Yu. «Deep Convex Net: A Scalable Architecture for Speech Pattern Classification.» *INTERSPEECH, Microsoft Research* (INTERSPEECH, Microsoft Research), 2011.

Li Deng, Xing Li. «Machine Learning Paradigms for Speech Recognition: An Overview.» *IEEE Transactions on Audio, Speech and Language Processing*, 2013: 1060-1089.

Lippmann, R.P. «An Introduction to Computing with Neural Nets.» *IEEE ASSP Magazine*, 1987: 4-22.

LIUM. 17 08 2017. https://github.com/StevenLOL/LIUM (erişildi: 01 01, 2018).

Lowerre, Bruce T. «The Harpy Speech Recognition System.» *Readings in Speech Recognition*, 1990: 576-586.

M. Karnjanedecha, S.A. Zahorian. «Signal modeling for high-performance robust isolated word recogniton.» *IEEE Transactions on Speech and Audio Processing*, 2001: 647-654.

M.Afify, Feng Liu, Hui Jiang, O. Siohan. «A new verification-based fast-match for large vocabulary continuous speech recognition.» *IEEE Transactions on Speech and Audio Processing*, 2005: 546-553.

M.D. Skowronski, John Gregory Harris. «Noise robust automatic speech recognition using a discriminative echo state network.» *IEEE Xplore Circuits and Systems*, 2007.

Marc A. Zissman, Kay M. Berkling. «Automatic Language Identification.» *Speech Communication*, 2001: 115-124.

Moore, Roger K. «Twenty Things we still don't know about Speech.» *Progress and Prospects of Speech Reserach an Technology*, 1994.

N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cretin, H. Bourlard, M. Athineos. «Pushing the envelope Aside Speech Recognition.» *IEEE Signal Processing Maganize*, 2005: 81-88.

Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, Henry H. Liu. «The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis.» *Proceedings of the Royal Society a Mathematical, Physical and Engineering Sciences*, 1998: 903-995.

Ossama, Abdel-Hamid. «Convolutional Neural Networks for Speech Recognition.» *Audio, Speech and Language Processing*, 2014: 1533-1545.

P.C. Bagshaw, S.M. Hiller, M. A. Jack. *Enhanced Pitch Tracking and The Processing of F0 Contours for Computer Aided Intonation Teaching.* Teknik Rapor, Edinburgh: Center for Speech Technology Research, 1993.

PRAAT. 01 01 2018. http://www.fon.hum.uva.nl/praat/ (erişildi: 01 01, 2018).

R., Rashmi C. «Review of Algorithms and Applications in Speech Recognition System.» *International Journal of Computer Science and Information Technologies*, 2014: 5258-5262.

Rabiner, L.R. «A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.» *IEEE Trans Acoustics, Speech and Signal Processing*, 1989: 257-286.

Reddy, D.R. *An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave.* Teknik Rapor, Stanford University, 1966.

Renu Karule, M.A. Potey. «Deep Architecture for Speech Processing.» *International Journal of Innovative Research in Computer and Communication Engineering*, 2015.

Rita Singh, Bhiksha Raj, Richard M Stern. «Automatic Generation of Subword Units for Speech Recognition Systems.» *IEEE Transactions on Speech and Audio Processing*, 2002: 89-99.

Ronan Collobert, Koray Kavukcuoglu, Clement Farabet. «Torch7: A Matlab-like Environment for Machine.» *BigLearn NIPS Workshop.* 2011.

Sadaoki, Furui. «50 years of Progress in Speech and Speaker Recognition Research.» *ECTI Transactions on Computer and Information Technology*, 2005: 1-2.

Santosh K. Gaikwad, Bharti W. Gawali, Pravin Yannawar. «A Review on Speech Recognition Technique.» *International Journal of Computer Applications*, 2010.

Shaohuai Shi, Qiang Wang, Pengfei Xu, Xiaowen Chu. «Benchmarking State-of-the-Art Deep Learning Software Tools.» *Cloud Computing and Big Data (CCBD), 2016 7th International Conference on.* Macau: IEEE, 2017.

Shipra J. Arora, Hisar Rishi Pal Singh. «Automatic Speech Recognition: A Review.» *International Journal of Computer Applications*, 2012.

SHoUT. 01 01 2018. http://shout-toolkit.sourceforge.net/ (erişildi: 01 01, 2018).

Simon King, Joe Frankel, Karen Livescu, Eric McDermott, Korin Richmond, Mirjam Wester. «Speech production knowledge in automatic speech recognition.» *The Journal of the Acoustical Society of America*, 2007.

SPHINX. 01 01 2018. https://cmusphinx.github.io/ (erişildi: 01 01, 2018).

SRILM. 01 01 2018. https://www.sri.com/engage/products-solutions/sri-language-modeling-toolkit (erişildi: 01 01, 2018).

Stefan Windmann, Reinhold Haeb Umbach. «Approaches to Iterative Speech Feature Enhancement and Recognition.» *IEEE Transactions on Audio, Speech and Language Processing*, 2009: 974-984.

T.B.Martin, A.L. Nelson, H.J. Zadell. *Speech Recognition by Feature Abstraction Techniques.* Teknik Rapor, Ohio: Air Force Avionics Lab., 1964.

T.Sakai, S. Doshita. «The Phonetic TypeWriter, Information Processing.» *Proceedings of IFIP Congress*, 1962.

Tatsuya Kawahara, Chin-Hui Lee, Biing-Hwang Juang. «Key Phrase Detection and Verification for Flexible Speech Understanding.» *IEEE Transactions of Speech and Audio Proc.* , 1998: 558-568.

Team, TensorFlow. *Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.* White Paper, Google Research, 2016.

TED-LIUM. 01 05 2012. http://www-lium.univ-lemans.fr/en/content/ted-lium-corpus (erişildi: 01 01, 2018).

Tiwari, Vibha. «MFCC and its applications in speeaker recognition.» *International Journal on Emerging Technologies*, 2010: 19-22.

Torres, Pedro, Terry Gleason, Alan McCree, Douglas Reynolds Elliot Singer, ve Douglas Sturim Fred Richardson. «The MITLL NEST LRE 2009 Language Recognition System.» *International Conference on Acoustics, Speech and Signal Processing.* Dallas: IEEE, 2010.

Tran, Dat Tat. *Fuzzy Approaches to Speech and Speaker Recognition.* Doktora Tezi, Canberra: Canberra Üniversitesi, 2000.

V.M. Velichko, N.G. Zagoruyko. «Automatic Recogniton of 200 Words.» *International Journal of Man-Machine Studies*, 1970: 223-234.

Victor Zue, James Glass, Michael Philips, Stephanie Seneff. «The MIT SUMMIT Speech Recognition system: a progress report.» *Workshop on Speech and Natural Language.* Stroudsburg: HLT, 1989. 179-189.

Vintsyuk, T.K. «Speech dicrimination by dynamic programming.» *Sprink Link- Cybernetics*, 1968: 52-57.

Y. Chow, M. Dunham, O. Kimball, M. Krasner, G. Kubala, J. Makhoul, P. Price, S. Roucos, R. Schwartz. «System, BYBLOSS: The BBN Continuous Speech Recognition.» *Acoustics, Speech and Signal Processing*, 1987.

Y.K. Muthusamy, E. Barnard, R.A. Cole. «Reviewing automatic language identification.» *IEEE Signal Processing Maganize*, 1994: 33-41.

Yakar, Özlem, ve Rıfat Aşlıyan. *Saklı Markov Modeli Kullanarak Türkçe Konuşma Tanıma.* Yüksek Lisans Tezi, Aydın: Adnan Menderes Üniversitesi, 2016.

Yalçın, Nursal. «Konuşma tanıma teorisi ve teknikleri.» *Kastamonu eğitim dergisi*, 2008: 249-266.

Yiğit, Altuğ. *İş Süreçlerinde İnsan Görüsünün Derin öğrenme ile destekleme.* Yüksek Lisans Tezi, Edirne: Trakya Üniversitesi Fen bilimleri enstitüsü, 2017.

ÖZET

Bu çalışmada Radyo ve Televizyon Üst Kurulu SKAAS sistemi ile entegre çalışabilecek bir tür konuşma tanıma uygulaması tasarımı yapılması amaçlanmıştır. Verilerin video kayıtlarından alınması, işlenmesi ve sonuçta metin olarak çıktı üretilebilmesi böylece ses deşifre işlemlerinin otomatik hale getirilmesi mümkün olması hedeflenmiştir. Böyle bir sistemin geliştirilmesi için gerekli tüm bilgiler geniş bir perspektif ile açıklanmış, problem sınırları belirlenerek kullanılacak ağ yapısı, aktivasyon fonksiyonları, hata fonksiyonları, kütüphaneler, verisetleri gibi birçok alanda gerekli tercihler belirlenmiştir. Sonuçta, derin öğrenme kullanan konuşmacı bağımlı, küçük kelime haznesi ile çalışacak, akustik ön uç tasarımında Mel frekans kepstral katsayılarının kullanıldığı, öğrenme aşamasında Metu 1.0 Türkçe Ses verisetinin kullanıldığı, sigmoid ve gauss hata fonksiyonlarının tercih edildiği ve Python ile tensorflow kütüphanesinin kullanılacağı bir ses tanıma(dikte) modeli önerilmiştir. Yapılan çalışmalar sonucunda kişi bağımlı testlerde %96 oranında doğru sınıflandırma değeri yakalanmıştır.

Anahtar Kelimeler: Derin Öğrenme, Konuşma Tanıma, Dikte, Tensorflow, Türkçe

ABSTRACT

The goal of this study is to design a speech recognition software that would work in integration with SKAAS – Digital Recording and Archiving - System of the Radio and Television Supreme Council of Turkey. The data are aimed to be received from the video recordings, processed and finally to generate text output making it possible to automate voice-dictation processes. All information required for the development of such a system were explained in a broad perspective, problem boundaries were identified and preferences which are necessary in many fields such as network structure to be used, activation functions, error functions, libraries and datasets were also identified. Consequently, a speaker-dependent voice recognition (dictation) model was proposed, which uses deep learning and will work on a small vocabulary, of which Mel Frequency Kepstral coefficients will be used in acoustic front-end design, where METU 1.0 Turkish Voice Dataset is used learning stage, Sigmoid and Gauss error functions are preferred and Python and TensorFlow libraries will be used. As a result of the studies carried out, a true classification rate of 96 % was achieved during the person-dependent tests.

Key words: Deep Learning, Speech Recognition, Dictation, Tensorflow, Turkish

ÖZGEÇMİŞ

Recep Sinan ARSLAN, 1987 yılında Yozgat'ta doğdu. 2005 yılında Boğazlıyan Anadolu Lisesinden mezun oldu. 2010 yılında Karadeniz Teknik Üniversitesi Bilgisayar Mühendisliği bölümünden mezun oldu. 2015 yılında Yıldırım Beyazıt Üniversitesi Bilgisayar Mühendisliği bölümünden Yüksek Lisans diplomasını aldı. 2015 yılında Gazi Üniversitesi Bilgisayar Mühendisliği bölümünde başlamış olduğu Doktora programında Tez aşamasında olup çalışmalarına devam etmektedir. 2011 yılında Çalışma ve Sosyal Güvenliği Bakanlığında Müfettiş Yardımcısı olarak görev başlamış olup, 2015 yılından itibaren Radyo ve Televizyon Üst Kurulunda Üst Kurul Uzman Yardımcısı olarak çalışmaktadır.

EK-1

Türkçe dili ile ilgili makale listesi

[1] P. C. Bagshaw, S. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching", 1993

[2] H. Ba, N. Yang, I. Demirkol, and W. Heinzelman, "Bana: A hybrid approach for noise resilient pitch detection", *Statistical Signal Processing Workshop (SSP)*, 369–372, 2012

[3] J. Hankamer, "Lexical Representation and Process, Morphological Parsing and the Lexicon", *MIT Press*, 1989

[4] K. Oflazer, D. Z. Hakkani-Tür and G. Tür, "Design for a Turkish treebank", *In proceedings [18]*1999

[5] H.Palaz, A. Kanak, Y. Bicil, M.U. Doğan, T. İslam, "TREN-Türkçe Konuşma Tanıma Platformu", *Signal Processing and Communications Applications Conference(SIU),* 2005

[6] K.Ö. Şen, C. Köse, S. Aras, "Bir Konuşmacının Konuştuğu Dilin Belirlenmesi", 2007

[7] A. A. Akın, C. Demir, M.U. Doğan, "Kelime altı dil modelleme yöntemlerinin türkçe konuşma tanıma için iyileştirilmesi", *Signal Processing and Communications Applications Conference(SIU),* 2012

[8] F. Kirişçioğlu, "Bilgisayar Destekli Bir Dil Programı", 2005

[9] Ç. Oflazoğlu, S. Yıldırım, "Türkçe konuşmada kızgın duygunun akustik bilgi kullanarak tanınması", *Signal Processing and Communications Applications Conference(SIU)*, 2012

[10] G. Karagöz, C. Demiroğlu, "Sesli yanıt çağrı akışında dilbilgisi tabanlı türkçe konuşma tanıma sistemi tanıtımı", *Signal Processing and Communications Applications Conference(SIU),* 2012

[11] C. Aksoylar, S. O. Mutluergil, H. Erdoğan, "Bir türkçe konuşma tanıma sisteminin anatomisi", *Signal Processing and Communications Applications Conference(SIU)*, 2009

[12] E. Arısoy, M. Saraçlar, "Türkçe haber programları için konuşma tanıma", *Signal Processing and Communications Applications Conference(SIU)*, 2007

[13] Ç. Oflazoğlu, S. Yıldırım, "Türçe duyusal konuşma veritabanı", *Signal Processing and Communications Applications Conference(SIU)*, 2011

[14] İ. Uzun, E. Arısoy, R. Edizkan, M. Saraçlar, "Dağıtık yapıda türkçe sürekli konuşma tanıma sisteminde seyrek paket kayıplarının analizi ve telafisi", *Signal Processing and Communications Applications Conference(SIU),* 2008

[15] O. Peker, "Türkçe kelime ile ses tanıma", *Signal Processing and Communications Applications Conference(SIU)*, 2016

[16] İ. Uzun, R. Edizkan, "Dağıtık yapıdaki türkçe süeki konuşma tanıma sisteminin patlatmalı paket kaybı durumda başarımının iyileştirilmesi", *Signal Processing and Communications Applications Conference(SIU),* 2008

[17] D. Susman, S. Köprü, A. Yazıcı, "Kısıtlı ses külliyatı ile türkçe geniş dağarcıklı sürekli konuşma tanıma", *Signal Processing and Communications Applications Conference(SIU),* 2012

[18] A. O. Bayer, T. Çiloğlu, M. T. Yöndem, "Türkçe konuşma tanıma için farklı dil modellerinin incelenmesi", *Signal Processing and Communications Applications Conference(SIU)*, 2006

[19] O. Büyük, A. Haznedaroğlu, L. M. Arslan, "Dil modeli uyarlanabilir türkçe ses tanıma yazılımı", *Signal Processing and Communications Applications Conference(SIU),* 2007

[20] E. Arısoy, M. Saraçlar, "Türkçe geniş dağarcıklı konuşma tanıma sistemleri için örünün yeniden değerlendirilmesi tabanlı dil modellemesi yaklaşımları", *Signal Processing and Communications Applications Conference(SIU)*, 2006

[21] B. Tombaloğlu, H. Erdem, "MFCC-SVM tabanlı türkçe konuşma tanıma sisteminin geliştirilmesi", Signal Processing and Communications Applications Conference(SIU), 2016

[22] H. Çeliktaş, C. Hanilçi, "Türkçe metne bağlı konuşmacı tanıma üzerine bir çalışma", *Signal Processing and Communications Applications Conference(SIU)*, 2017

[23] S. Keser, R. Edizkan, "Altuzay sınıflama ile sesbirim temelli türkçe yalıtık türkçe kelime tanıma", *Signal Processing and Communications Applications Conference(SIU)*, 2009

[24] O. Yılmaz, M. Saraçlar, "Türkçe haber programlarının yazılandırılması için işitsel bölütleme", Signal Processing and Communications Applications Conference(SIU), 2011

[25] E. Arısoy, "Türkçe sözlü ders anlatımları için otomatik yazılandırma ve geri getirim sistemi geliştirilmesi", *Signal Processing and Communications Applications Conference(SIU),* 2017

[26] R. Aşlıyan, K. Günel, T. Yakhno, "Dinamik zaman bükmesi ve çok katmanlı algılayıcı kullanarak hece tabanlı türkçe konuşma tanıma", *Signal Processing and Communications Applications Conference(SIU),* 2008

[27] K. Kuru, K. Arda, N. Baykal, "Konuşma tanıma sistemi kullanılarak bilgisayar ile etkileşimli, yapısal tıbbi rapor hazırlanması", *Signal Processing and Communications Applications Conference(SIU)*, 2007

[28] O. Büyük, H. Erdoğan, K. Oflazer, "Konuşma tanımada karma dil birimleri kullanımı ve dil kısıtlarının gerçeklenmesi", *Dede Korkut Türk Dili ve Edebiyatı Araştırmaları Dergisi*, 2006

[29] D. Can, M. Saraçlar, "Türkçe haber bültenlerinin açık kaynak yazılımları ile yazılandırılması", *Signal Processing and Communications Applications Conference(SIU),* 2009

[30] R. Edizkan, B. Tiryaki, T. Büyükcan, İ. Uzun, "Ses komut tanıma ile gezgin araç kontrolü", *9. Akademik bilişim konferansı*, 2007

[31] İ. Uzun, R. Edizkan, "Dağıtılmış sistemlerde internet protokolü üzerinden yalıtık ses tanıma", *Signal Processing and Communications Applications Conference(SIU),* 2007

[32] M. Bayğın, M. Karaköse, "Gerçek zamanlı ses tanıma tabanlı akıllı ev uygulaması", *Signal Processing and Communications Applications Conference(SIU),* 2012

[33] M. K. Bayrakçeken, M.A. Çay, A. Barkana, "Ortak yöney yaklaşımı ile anahtar sözcük yakalama", *Signal Processing and Communications Applications Conference(SIU),* 2007

[34] O. Türk, L. M. Arslan, "Konuşma terapisine yönelik konuşma tanıma yöntemleri", *Signal Processing and Communications Applications Conference(SIU),* 2004

[35] M. Caner, S. V. Üstün, "Yapay sinir ağları ile konuşmacı kimliğini tanıma uygulaması", *Journal of Engineering Sciences,* 2006

[36] E. Arısoy, L. M. Arslan, "Türkçe gazete haberleri dikte sistemi", *Signal Processing and Communications Applications Conference(SIU),* 2005

[37] T. Aksungurlu, S. Parlak, H. Sak, M. Saraçlar, "Türkçe haber bültenleri için dil modelleme yaklaşımlarının karşılaştırılması", *Signal Processing and Communications Applications Conference(SIU),* 2016

[38] T. Çiloğlu, M. Çömez, S. Şahin, "Takılı bir dil olarak türkçe için dil modelleme", *Signal Processing and Communications Applications Conference(SIU)*, 2004

[39] Ç. Oflazoğlu, S. Yıldırım, "Türkçe duygusal konuşma için duygu sınıflarının ikili sınıflandırma performansları", *Signal Processing and Communications Applications Conference(SIU)*, 2015

[40] N. Yalçın, Ö. F. Bay, "Htk ile konuşmacıdan bağımsız türkçe konuşma tanıma sistem oluşturma", *Gazi üniversitesi Endüstriyel sanatlar eğitim fakültesi dergisi*, 79-97, 2007

[41] O. Türk, Ö. Şayli, A.S.Özsoy, L.M. Arslan, "Türkçe'de ünlülerin formant incelemesi", *18. Ulusal dilbilim kurultayı*, 2004

[42]R. Aşlıyan, K. Günel, T. Yakhno, "Dinamik zaman bükmesi yöntemiyle hece tabanlı konuşma tanıma sistemi", *akademik bilişim*, 2008

[43]İ. Gelegin, B. Bolat, "Ayrık kelime tabanlı bir konuşma tanıma sistemiyle bilgisayar kontrolü", *Fırat üniversitesi dergisi*, 2010

[44] E. Köklükaya, İ. Çoşkun, "Endüktif öğrenmeyi kullanarak konuşma tanıma", *SAU Fen bililmler enstitüsü dergisi*, 2003

[45] B. Asefisaray, E. Mengüşoğlu, M. Hacıömeroğlu, H. Sever, "Türkçe ses tanıma sistemlerinde dil modeli boyutunun doğruluk oranına etkisi", *PAU Mühendislik Bilimleri Dergisi*, 100-105, 2016

[46] C. Özbey, S. Bayar, "Otomatik ses tanıma: türkçe için genel dağarcıklı akustik model oluşturulması ve test edilmesi", *19. Akademik bilişim konferansı,* 2017

[47] A. Gürel, L.M. Arslan, "Konuşma tanıma için insan-makine karşılaştırılması", *Signal Processing and Communications Applications Conference(SIU),* 2008

[48] H. Çakır, B. Okutan, "Ses kontrollü web tarayıcı", *Bilişim teknolojileri dergisi*, 2011

[49] H. Sak, M. Saraçlar, T. Güngör, "Morpholexical and discriminative language models for turkish automatic speech recognition", *Signal Processing and Communications Applications Conference(SIU)E transactions on audio, speech and language processing*, 2012

[50] B. Can, H. Artuner, "A syllable-based turkish speech recognition system by using time delay neural network(tdnn), *Signal Processing and Communications Applications Conference(SIU),* 2013

[51] A. Akyol, H. Erdoğan, "Filler model based confidence measures for spoken dialogue systems: a case study for turkish", *Signal Processing and Communications Applications Conference(SIU)* 2004

[52] İ. Uzun, R. Edizkan, "Performance improvement in distributed turkish continuous speech recognition system using packet loss concealment techniques", *Signal Processing and Communications Applications Conference(SIU)*, 2011

[53] H. Sak, M. Saraçlar, T. Güngör, "Morphology-based and sub-word language modelling for turkish speech recognition", *Signal Processing and Communications Applications Conference(SIU)*, 2010

[54] N. Yalçın, Y. Altun, "A sample work developed using speech recognition technology about basic language education: Speak and learn- learn with spech recognition", *2.world conference on educational technology researches*, 2012

[55] E. Arısoy, L.M. Arslan, “Turkish radiology dictation system”, 9. *Conference speech and computer,* St. Petersburg, 2004

[56] Ö. Salor, B. L. Pellom, T. Çiloğlu, M. Demirekler, “turkish speech corpora and recognition tools developed by porting SONIC: towards multilingual speech recognition”, *computer speech and language*, 580-593, 2007

[57] H. Erdoğan, O. Büyük, K. Oflazer, “Incorporating language constraints in sub-word based speech recognition, *Signal Processing and Communications Applications Conference(SIU)*, 2005

[58] E. Arısoy, L.M. Arslan, “Turkish dictation system for broadcast news applications”, *Signal Processing and Communications Applications Conference(SIU),* 2009

[59] K. Çarkı, P. Geutner, T. Schultz, “Turkish lvcsr: towards better speech recognition for agglutinative language”, *Signal Processing and Communications Applications Conference(SIU),* 2000

[60] Ö. Yakar, R. Aşlıyan, “Saklı markov modeli kullnarak türkçe konuşma tanıma”, *Signal Processing and Communications Applications Conference(SIU)*, 2014

[61] D. Z. H. Tür, K. Oflazer, G. Tür, “Statistial morphological disambiguation for agglutinative languages”, *computers and the humanities*, 381-410, 2002

[62] E. Arısoy, H. Dutağacı, L. M. Arslan, “A unified model for large vocabulary continuous speech recognition of Turkish”, *Signal processing,* 2006