T.C.
RADYO VE TELEVİZYON ÜST KURULU

KONUŞMA TANIMA SİSTEMLERİNDE UYGULANAN ÖZNİTELİK ÇIKARIM ALGORİTMALARININ KARŞILAŞTIRILMASI VE RTÜK İNCELEMESİ

UZMANLIK TEZİ

MURATCAN İTAP

ANKARA
KASIM, 2022


RTÜK
RADYO VE TELEVİZYON ÜST KURULU

T.C.

RADYO VE TELEVİZYON ÜST KURULU

KONUŞMA TANIMA SİSTEMLERİNDE UYGULANAN ÖZNİTELİK ÇIKARIM ALGORİTMALARININ KARŞILAŞTIRILMASI VE RTÜK İNCELEMESİ

UZMANLIK TEZİ

MURATCAN İTAP

DANIŞMAN: ÜST KURUL UZMANI CEM ULUOĞLAKÇI

ANKARA

KASIM, 2022

Radyo ve Televizyon Üst Kurulu Başkanlığına,

Muratcan İtap adlı Üst Kurul Uzman Yardımcısının hazırladığı "Konuşma Tanıma Sistemlerinde Uygulanan Öznitelik Çıkarım Algoritmalarının Karşılaştırılması ve RTÜK İncelemesi" başlıklı bu çalışma Tez Değerlendirme Komisyonu üyelerince oy birliği / oy çokluğu ile **Uzmanlık Tezi** olarak kabul edilmiştir.

	Komisyon Üyeleri	İmza
Başkan	Ebubekir ŞAHİN Üst Kurul Başkanı	……..
Üye	Deniz GÜÇER Üst Kurul Üyesi	……..
Üye	Dr. İbrahim USLU Üst Kurul Başkan Yard.	……..
Üye	Deniz GÜLER Daire Başkanı	……..
Üye	Cem ULUOĞLAKÇI Üst Kurul Uzmanı	……..

ONAY

Bu tez Radyo ve Televizyon Üst Kurulu Uzmanlığı Yönetmeliği'nin ilgili maddeleri uyarınca …/…/20... tarihinde kabul edilmiştir.

……

Ebubekir ŞAHİN

Üst Kurul Başkanı

ETİK İLKELERE UYGUNLUK BEYANI

Bu tez çalışmasında;

Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,

Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,

Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,

Kullanılan verilerde ve ortaya çıkan sonuçlarda herhangi bir değişiklik yapmadığımı,

Sunduğum çalışmanın özgün olduğunu bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim. …./…./20….

(İmza)

Muratcan İtap

ÖZET

KONUŞMA TANIMA SİSTEMLERİNDE KULLANILAN ÖZNİTELİK ÇIKARIM ALGORİTMALARININ KARŞILAŞTIRILMASI VE RTÜK İNCELEMESİ

İTAP, Muratcan

Uzmanlık Tezi

Tez Danışmanı: Üst Kurul Uzmanı Cem ULUOĞLAKÇI

Kasım, 2022, 85

50'ler ve 60'larda ilk basit örneklerini veren konuşma tanıma sistemleri, 70'lerde öznitelik çıkarım tekniklerinin, 80'lerde ise istatistiksel modellerin kullanılmasıyla, büyük gelişim kaydetti. 2000'lerde gelişen yazılım ve donanım teknolojileri ile birlikte yapay zekâ teknolojilerinin de bu alanda kullanılmasıyla, konuşma tanıma sistemleri yeni bir boyut kazandı.

Konuşma tanıma sistemleri günümüzde sanal asistanlardan müşteri hizmetlerine kadar pek çok alanda kullanılan bir teknolojidir. Konuşma tanıma sisteminin adımlarından birisi olan öznitelik çıkarımı, sistemin verimliliği ve tanıma oranı bakımından çok önemlidir. Bu sebeple günümüzde yaygın olarak kullanılan üç önemli öznitelik çıkarım tekniğinin karşılaştırılması bu tezin konusu olarak seçilmiştir.

Bu çalışmanın ilk bölümünde konuşma tanıma sistemlerinin genel yapısı verilmiştir. Bu bölümde amaç sistemin genel yapısını ve öznitelik çıkarımının sistemin neresinde yer aldığını anlamaktır. Ayrıca konuşma sinyal işleme konusu öznitelik çıkarım tekniklerinin temelini oluşturduğundan bu bölümde detaylıca ele alınmıştır. İkinci bölümde öznitelik çıkarımı için seçilen üç teknik olan LPC, MFCC, PLP-RASTA detaylıca açıklanmış ve literatürde yapılan çalışmalar ışığında bu tekniklerin çeşitli yönlerden karşılaştırması yapılmıştır. Üçüncü bölümde ise ele alınan üç öznitelik çıkarım yöntemi için Türkçe ve İngilizce komutlar için bir saniyelik ses dosyalarından oluşan veri seti üzerinden elde edilen öznitelik vektörleriyle, her yöntem ve dil için ayrı olacak şekilde aynı özelliklere sahip evrişimsel sinir ağı eğitilmiştir. Öznitelik çıkarım tekniklerinin Türkçe ve İngilizce için yakaladığı tanıma oranlarının ve hesaplama sürelerinin karşılaştırılması yapılmıştır.

Son olarak, bulgular ve sonuçlar bölümlerinde, elde edilen çıkarımlar, yorum ve öneriler eklenerek verilmiştir. Ayrıca, RTÜK bünyesinde yayınların kaydedilmesini sağlayan SKAAS ile konuşma tanıma sistemlerinin birlikte kullanılması bu bölümlerde incelenecektir.

Anahtar Kelimeler: Konuşma Tanıma, Öznitelik Çıkarım, LPC, MFCC, PLP-RASTA

ABSTRACT

COMPARISON OF FEATURE EXTRACTION ALGORITHMS USED IN SPEECH RECOGNITION SYSTEMS

İTAP, Muratcan

Specialization Thesis

Thesis Advisor: Supreme Council Specialist Cem ULUOĞLAKÇI

November, 2022, 85

Speech recognition systems, which gave their first simple examples in the 50s and 60s, made great progress with the use of feature extraction techniques in the 70s and statistical models in the 80s. In 2000s, with the use of artificial intelligence technologies in this field along with the advanced software and hardware technologies, speech recognition systems gained a new dimension.

At the present time speech recognition systems are a technology that used in many areas, from virtual assistants to customer service. Feature extraction, which is one of the steps of the speech recognition system, is very important in terms of the efficiency and success rate of the system. For this reason, the comparison of three important feature extraction techniques, which are widely used today, has been chosen as the subject of this thesis.

In the first part of this study, the general structure of speech recognition systems is given. The purpose of this section is important to understand the general structure of the system and where the feature extraction takes place in the system. In addition, in this section, speech signal processing, which is necessary to understand the feature extraction methods, is discussed in detail. In the second part, the three selected techniques for feature extraction, LPC, MFCC and PLP-RASTA, are explained in detail and a comparison of these techniques in various aspects is made in the light of the studies in the literature. In the third part, for the three feature extraction methods discussed, a convolutional neural network with the same features is trained with the feature vectors obtained from the dataset consisting of one-second audio files for Turkish and English commands, separately for each method and language.

Comparison of feature extraction methods for Turkish and English commands are made for their recognition rate and calculation time.

Finally, in the findings and results sections, the inferences obtained are given by adding comments and suggestions. In addition, the use of SKAAS, which ensures the recording of broadcasts within the body of RTÜK, and speech recognition systems together will be examined in these sections.

Key Words: Speech Recognition, Feature Extraction, LPC, MFCC, PLP-RASTA

İÇİNDEKİLER

Sayfa

ETİK İLKELERE UYGUNLUK BEYANI iii
ÖZET iv
ABSTRACT vi
İÇİNDEKİLER viii
TABLOLAR DİZİNİ xi
ŞEKİLLER DİZİNİ xii
KISALTMALAR/SİMGELER xiii
GİRİŞ 1
Literatür Taraması 4
Problem 8
Amaç 8
Önem 9
Sınırlılıklar 9
Tanımlar 10
Öznitelik Çıkarımı 10
Fonem 11
Alofon 11
Yapay Sinir Ağları 11
Uygulama Programlama Ara yüzü (API) 12
Gizli Markov Modeli (HMM) 12
Yöntem 12
Araştırmanın Modeli 12
Evren ve Örneklem 13
Verilerin Toplanması 14
Verilerin Çözümlenmesi 15
BÖLÜM 1 16
KONUŞMA TANIMA TEKNOLOJİSİNE GENEL BAKIŞ 16
1.1. Konuşma Sinyal İşleme 18
1.1.1. Ses & Ses Dalga Formu 19
1.1.2. Ses Dalgasında Genlik, Frekans, Güç, Tını 21
1.1.3. Analog Sinyalden Dijital Sinyale Geçiş 21

1.1.4. Nyquist Teoremi ve Duyma Aralığı 22
1.1.5. Ses Sinyal Alanı 24
1.1.6. Spektral Sızıntı ve Pencereleme 31
1.2. Öznitelik Çıkarımı 32
1.3. İstatistiksel ve Yapay Zekâ Modelleri 33
1.3.1. HMM-GMM Modeli 33
1.3.2. Yapay Sinir Ağı Modelleri 36
1.4. Akustik Modelleme 38
1.5. Dil Modelleme 39
1.6. Konuşma Tanıma Kullanım Alanları 40
BÖLÜM 2 43
ÖZNİTELİK ÇIKARIM TEKNİKLERİ 43
2.1. Önişlem 44
2.1.1. Filtreleme 44
2.1.2. Sinyalin Bloklara Ayrılması 46
2.1.3. Pencereleme 47
2.2. Doğrusal Tahmin Kodlaması (LPC) 47
2.2.1. İnsanda Konuşma Üretimi ve Kaynak Filtre Modeli 48
2.2.2. LPC Matematiksel Modeli 49
2.2.3. Ses Sinyali Sıkıştırmasında Kullanımı 51
2.3. Mel-Frekans Kepstral Katsayıları (MFCC) 54
2.3.1. Log-Büyüklük Spektrumu 55
2.3.2. Mel Ölçeği ve Mel Spektogramları 55
2.3.3. Ayrık Cosine Dönüşümü (DCT) Uygulanması 58
2.3.4. MFCC Üzerine Literatürde Yapılan Çalışmalar 59
2.4. Algısal Doğrusal Tahmin – Bağıl Spektrum (PLP-RASTA) 59
2.4.1. Eşit Ses Yüksekliği Eğrisi 61
2.4.2. Yoğunluk-Ses Şiddeti Kanunu 62
2.5. LPC, MFCC ve PLP-RASTA Karşılaştırması 63
BÖLÜM 3 65
ÖZNİTELİK VEKTÖRLERİYLE CNN EĞİTİMİ 65
3.1. Kullanılan Veri Seti 65

3.2. Kullanılan Hesaplama Yöntemleri 66
3.3. Evrişimsel Sinir Ağı Modeli 67
3.4. Geliştirme Süreci 68
3.5. Sonuçlar 69
BULGULAR VE YORUMLAR 71
SONUÇLAR VE ÖNERİLER 74
Sonuçlar 74
Öneriler 76
KAYNAKLAR 77
ÖZGEÇMİŞ 85

TABLOLAR DİZİNİ

Tablo **Sayfa**

Tablo 1. Kriterlerin Literatüre veya Geliştirilen Uygulamaya Göre Değerlendirilmesi……9

Tablo 2. Türkçe Veri Setindeki Komutlar ve Ses Dosyası Sayıları……..……13

Tablo 3. İngilizce Veri Setindeki Komutlar ve Ses Dosyası Sayıları...…….…….……..14

Tablo 4. Oluşturulan Modellerin Komut Komutu Tanıma Oranları ……...…….……69

Tablo 5. Tablo 2 ve Tablo 3'teki Veri Setleri için Öznitelik Çıkarma Süreleri…….…..70

Tablo 6. Üç Teknik İçin Literatüre Göre Elde Edilen Bulgular ……...…….……73

Tablo 7. Üç Teknik İçin Geliştirilen Uygulamaya Göre Elde Edilen Bulgular …….…..73

ŞEKİLLER DİZİNİ

Şekil **Sayfa**

Şekil 1. Ses Tanıma Sistemine İlişkin Genel Blok Yapı ……17

Şekil 2. Titreşen Hava Moleküllerinin Oluşturduğu Basınç Grafiği……19

Şekil 3. "Radyo ve Televizyon Üst Kurulu" İfadesinin Dalga Formu…….20

Şekil 4. İnsanlarda ve Bazı Hayvanlarda Duyma Aralığı ……23

Şekil 5. Örnek Genlik Zarfı……..25

Şekil 6. "Radyo ve Televizyon Üst Kurulu" İfadesine Ait Spektrum …….28

Şekil 7. "Radyo ve Televizyon Üst Kurulu" İfadesine Ait Spektogram..……31

Şekil 8. HMM Üç Adet Durum ve Bu Durumlar Arasındaki Geçiş Olasılıkları ……....34

Şekil 9. Kısmi Bir Durum ve Gözlemler Dizisi ……..35

Şekil 10. Yapay Sinir Ağları Genel Yapısı …….37

Şekil 11. "Radyo ve Televizyon Üst Kurulu" İfadesine Ait Filtrelenmiş Sinyal……46

Şekil 12. İnsanda Ses Yolu……..49

Şekil 13. Kaynak Filtre Modeli ……..……..50

Şekil 14. Sesli Harf Sinyali …….53

Şekil 15. Sessiz Harf Sinyali ……...53

Şekil 16. Altı Adet Mel Filtre Bandından Oluşan Mel Filtre Bankası…….57

Şekil 17. "Radyo Televizyon Üst Kurulu" İfadesine Ait Mel Spektogramı……58

Şekil 18. PLP-RASTA Adımları……..61

Şekil 19. Eşit Ses Yüksekliği Eğrisi…….62

Şekil 20. Uygulama İçin Kullanılan Evrişimsel Sinir Ağı Modeli……...68

KISALTMALAR/SİMGELER

Kısaltma	Açıklama
AMDF	Average Magnitude Difference Function (Ortalama Büyüklük Farkı Fonksiyonu)
API	Uygulama Programlama Arayüzü (Application Programming Interface)
ASR	Otomatik Konuşma Tanıma (Automatic Speech Recognition)
CELP	Kod Uyarımlı Doğrusal Tahmin (Code-Excited Linear Predictive)
CPS	Saniye Başına Döngü (Cycle Per Second)
CTC	Bağlantıcı Zamansal Sınıflandırma (Connectionist Temporal Classification)
DARPA	Savunma İleri Araştırma Projeleri Ajansı (Defense Advanced Research Projects Agency)
DCT	Ayrık Cosine Dönüşümü (Descrete Cosine Transform)
DFT	Ayrık Fourier Dönüşümü (Descrete Fourier Transform)
DNN	Derin Sinir Ağları (Deep Neural Network)
FFT	Hızlı Fourier Dönüşümü (Fast Fourier Transform)
FIR	Sonlu Darbe Tepkisi (Finite Impulse Response)
GMM	Gauss Karma Modeli (Gauss Mixture Model)
HMM	Saklı Markov Modeli (Hidden Markov Model)
HTK	Saklı Markov Modeli Araç Seti (Hidden Markov Model Toolkit)
IEEE	Elektrik ve Elektronik Mühendisleri Enstitüsü (Institute of Electrical and Electronics Engineers)
IJERT	Uluslararası Mühendislik Araştırma ve Teknoloji Gazetesi (International Journal of Engineering Research&Technology)
LPC	Doğrusal Tahmin Kodlama (Linear Predictive Coding)
LSTM	Uzun Kısa-Süreli Bellek (Long Short-Time Memory)
MFCC	Mel Frekansı Kepstral Katsayıları (Mel Frequency Cepstral Coefficients)
MIT	Massachusetts Teknoloji Enstitüsü (Massachusetts Institute of Technology)
NDLTD	Tez Dijital Kütüphane Ağı (The Networked Digital Library of Theses and Dissertations)
OATD	Açık Erişim Tezler (Open Access Theses and Dissertations)
PLP-RASTA	Algısal Doğrusal Tahmin – Bağıl Spektrum (Perceptual Linear Prediction – Relative Spectra)
ReLU	Düzeltilmiş Lineer Birim (Rectified Linear Unit)

RMS	Karekök Ortalama (Root Mean Square)
RNN	Tekrar Eden Sinir Ağı (Recurrent Neural Networks)
RTÜK	Radyo ve Televizyon Üst Kurulu
SKAAS	Sayısal Kayıt Arşiv Analiz Sistemi
SRI	Stanford Araştırma Enstitüsü (Stanford Research Institute)
STFT	Kısa Zamanlı Fourier Dönüşümü (Short Time Fourier Transform)
VODER	Sesle Çalıştırma Göstericisi (Voice Operating Demonstrator)
VSELP	Vektör Toplamı Uyarılmış Doğrusal Kodlama (Vector Sum Excited Linear Prediction)
YSA	Yapay Sinir Ağları
ZCR	Sıfır Geçme Oranı (Zero Crossing Rate)
W	Watt
WER	Kelime Hata Oranı (Word Error Rate)

GİRİŞ

Bu bölümde ilk olarak, genel olarak konuşma tanıma sistemlerine ve bu çalışma kapsamında incelenen üç önemli öznitelik çıkarım tekniğine yönelik olarak yapılan literatür taraması verilmiştir. Ardından problem, amaç, önem, sınırlılıklar, tanımlar ve yöntem sırasıyla verilmiştir.

İnsan davranışını, özellikle doğal konuşma ve konuşulan dile uygun şekilde yanıt verme yeteneğini taklit eden bir makine tasarlamak, yüzyıllardır mühendislerin ve bilim adamlarının ilgisini çekmiştir. Temelde insanların birbirleri ile iletişimini destekleyen teknolojiler zaman içerisinde yapısal olarak değişmiş ve insan-makine etkileşimi kavramı ortaya çıkmıştır. Erken evrelerde basit tuşlu sistemler ile gerçekleştirilen insan makine veya elektronik cihaz etkileşimi günümüzde yerini dokunmatik sistemlere bırakmıştır. Dokunmatik ve mekanik sistemler hala makine insan etkileşimi en çok kullanılan yöntemlerdir. Bunun yanı sıra özellikle son yıllarda bu alanda yeni bir yöntem olarak konuşma tanıma sistemlerinden de bahsedebiliriz. Sesli arama, konuşmadan metne dönüşüm, akıllı ev aletlerine sesli komutlar, eğitim, hukuk, müşteri hizmetleri vb. gibi alanlardaki kullanımı nedeniyle konuşma tanıma sistemleri, insan-makine etkileşiminde önemli bir yere sahiptir.

Bilgisayarlar tarafından konuşma üretimi ve tanıma, araştırmacılar tarafından kapsamlı bir şekilde ele alınan karmaşık bir süreçtir. Bir konuşma tanıma sistemi hem donanımsal açıdan önemli bir kaynak gerektirir hem de yazılımsal açıdan sistemi geliştirmek için ciddi çaba gerekir. Sistemin bu kadar maliyetli olması nedeniyle, bu sisteme ihtiyaç duyan ticari firmalar, akademik araştırmacılar, öğrenciler için sıfırdan bu sistemi üretmek çok zordur. Ancak sistemi bu maliyetlere katlanmadan elde etmek mümkündür.

Günümüzde büyük teknoloji firmaları kendi konuşma tanıma sistemlerini çeşitli API'lar (Uygulama Programlama Ara yüzü) aracılığıyla, belirlenen süre veya kullanım limitleri dahilinde ücretsiz olarak, yoğun veya uzun süreli kullanımda ise belirli bir ücret karşılığı isteyen kişi ve kurumlara sunulmaktadırlar. "Google ASR", "IBM Watsons",

“Amazon Transcribe”, “Microsoft Azure Cognitive Services” bu alandaki başlıca servisler veya ürünler olarak gösterilebilirler.

Bu çalışmanın ilk bölümünde konuşma tanıma sistemleri genel olarak ele alınacaktır. Konuşma tanıma sistemlerinin mimarisini oluşturan alt sistemler olan, konuşma sinyal işleme, öznitelik çıkarımı, akustik modelleme, dil modelleme açıklanıp, günümüzde teknolojik olarak geldiği noktalar belirtilecektir. Ayrıca literatürde bu sistemler üzerine yapılan bazı çalışmalara değinilecektir. Bu bölümde konuşma sinyal işleme konusu, öznitelik çıkarım işleminin temelini oluşturduğundan, detaylıca ele alınmıştır.

İkinci bölümde ilk olarak konuşma tanıma sistemlerinin en önemli adımlarından biri olan öznitelik çıkarımı açıklanmıştır. Sonrasında öznitelik çıkarımının öncesinde, sinyali bu işleme hazırlayan önişlem aşamasında yapılanlar açıklanmıştır. Ardından öznitelik çıkarımı için literatürde öne çıkan üç teknik (LPC, MFCC, PLP-RASTA) detaylıca ele alınıp, literatürde bu teknikler için yapılmış çalışmalar incelenmiştir.

Üçüncü bölümde ele alınan öznitelik çıkarım teknikleri kullanılarak, İngilizce ve Türkçe veri seti üzerinden elde edilen öznitelik vektörleri ile evrişimsel sinir ağı (CNN) eğitilmiştir. Ayrıca verilen bir ses dosyasının hangi komuta ait olduğunu tahmin eden bir servis geliştirilmiştir. Her teknik ve dil için elde edilen bulgular karşılaştırılmalı olarak incelenmiştir.

Konuşma tanıma teknolojilerinin tarihindeki kırılma noktalarına bakacak olursak; 50’li ve 60’lı yıllarda Bell laboratuvarlarının geliştirdiği sistemler ve 70’li yıllarda Amerikan Savunma Bakanlığı’na bağlı DARPA’nın (İleri Savunma Araştırma Projeleri Ajansı) desteklediği projeler, sistemin ilk örneklerini oluşturmaktadır. 80'lerde istatistiksel modelleme olan HMM’nin (Gizli Markov Modeli) konuşma tanıma sistemlerinde uygulanmaya başlanması bu teknoloji için dönüm noktalarından biri olmuştur. 2000’li yıllardan itibaren de yapay zekâ ve makine öğrenmesi teknolojilerinin bu alanda kullanılmaya başlaması ise konuşma tanıma teknolojisinin bugünkü halini almasında büyük katkılar sağlamıştır. Konuşma tanıma sistemleri teknolojilerinin detaylı tarihi literatür taraması bölümünde verilmiştir.

Çalışma kapsamında incelenen LPC, MFCC ve PLP-RASTA teknikleri 70’lerden itibaren kullanılmaya başlanmıştır. Konuşma sinyalinden bu teknikleri kullanarak çıkarılan öznitelik vektörleri, 2000’li yıllara kadar daha çok HMM-GMM’ye (Gizli Markov Modeli-

Gauss Karma Modeli) dayalı istatistiksel modellemelerle birlikte kullanılarak konuşma tanıma sistemleri gerçekleştiriliyordu (Wang ve ark., 2019, s. 1).

HMM-GMM'nin konuşma tanıma sistemlerinde kullanımı için üç rolden bahsedebiliriz; konuşma sinyalinin küçük blokları için çıkarılan öznitelik vektörlerini kullanarak, sinyalin en olası hangi fonemlere karşılık geldiğinin belirlenmesi (bkz. 1.4), fonemlerin yan yana gelerek en olası kelimelerin belirlenmesi ve kelimelerin bir araya gelerek en olası cümlelerin belirlenmesi (bkz. 1.5) (cümle yapısına göre olası kelimeler değişebilmektedir.). HMM-GMM olasılıksal olarak en uygun durumları belirlemek için bir veri seti tarafından eğitilmelidir. Bu veri seti çeşitli koşullarda (gürültü, yüksek ses, alçak ses, farklı kişilere ait sesler vb.) üretilen ses dosyaları (akustik modelleme için), gazete metinleri, elektronik kitaplar (dil modelleme için) veya konuşma tanıma sistemleri için özel oluşturulmuş veri setleridir (Koçer ve Ahmed, 2019, s. 42; Oyucu ve ark., 2019, s. 845, para. 4; Wang ve ark., 2019, s. 3, para. 8).

2000'li yıllardan sonra bilgisayar ve yazılım teknolojilerinin olağanüstü bir hızda gelişmesi, konuşma tanıma sistemlerinin de aynı hızda gelişmesine ve kullanımının yaygınlaşmasına neden olmuştur. Kullanılan istatistiksel modeller yerini yapay zekâ ve makine öğrenmesinin alt dalı olan derin öğrenme ya da bunları istatistiksel modellerle beraber kullanan hibrit modellere bırakmıştır (Wang ve ark., 2019, s.1).

Yapay zekâ, insan zekâsını taklit etmeye çalışan sistemlerin genel adıdır. Yapay zekânın dallarından biri olan makine öğrenmesi, insanların öğrenme şekillerini taklit etmek için veri ve algoritmaların kullanımına odaklanıp doğruluğunu kademeli olarak artırır. Derin öğrenme ise, yapay sinir ağlarını temel alan bir makine öğrenmesi alt kümesidir. Başlıca derin öğrenme mimarileri olarak evrişimsel sinir ağları, tekrarlayan sinir ağları, uzun-kısa vadeli hafıza ağları ve sınırlı boltzmann makinelerini gösterebiliriz (Şeker ve ark., 2017). Konuşma tanıma sistemlerinde, 2000'li yıllarla beraber istatistiksel modellerin yerini alan yapay zekâ sistemleri, yapay sinir ağlarını temel alan derin öğrenme mimarileridir.

Konuşma tanıma sistemlerinde derin öğrenme mimarilerinin kullanımı için, dile ait veri setleriyle bu mimarilerin eğitilmesi gerekmektedir. Bu eğitme aşamasında yine konuşma sinyallerinden çıkarılan öznitelik vektörleri eğitim girdisi olarak kullanılmaktadır. Bu iş için özel olarak oluşturulmuş veri setleri mevcuttur (Kurtkaya, 2021; Warden, 2017). Üçüncü bölümde geliştirilen uygulamada kullanılan evrişimsel sinir ağı benzer bir veri seti ile eğitilmiştir.

Literatür Taraması

Bu bölümde kökleri 1950'lere uzanan konuşma tanıma sistemleri teknolojisinde ve bu konu özelinde öznitelik çıkarım tekniklerinde, günümüze kadar yapılan çalışmalara değinilecektir. 80'lere kadar yapılan basit tanıma sistemleri 80'lerde istatiksel modellemenin kullanımı ile büyük atılım yaşamıştır. Konuşma tanıma teknolojisinde bir diğer büyük atılım 2000'li yıllardan sonra yapay zekâ ve makine öğrenmesi tekniklerinin bu alanda kullanılması ile olmuştur.

20. yüzyılın ilk yarısında, Fletcher (1922) ve Bell Laboratuvarları'ndaki araştırmacılar tarafından yapılan çalışma, belirli bir konuşma spektrumu (bkz. 1.1.5.2) ile ses özellikleri arasındaki ilişkiyi belgelemiştir. Dudley ve ark. (1939), Fletcher'ın araştırmasından büyük ölçüde etkilenerek, mekanik kontrollü ancak elektriksel bir konuşma makinesi olan VODER'i geliştirdi.

Harvery Fletcher ve Homer Dudley gibi bu alanın öncüleri, bir konuşma sesinin fonetik doğasının güvenilir bir şekilde tanımlanması için sinyal spektrumunun önemini kesin olarak ortaya koydu. Bu iki seçkin bilim adamı tarafından kurulan sistemi takiben, konuşma tanıma için tasarlanan pek çok sistem ve algoritma, kısmen konuşma gücü spektrumunun ölçümü kavramına dayanmaktadır. Günümüzde bir sinyalden güç spektrumunun ölçülmesinin modern dijital sinyal işleme teknikleri ile başarılması oldukça kolaydır (Juang ve Rabiner, 2005, s. 5, para. 2).

Bell Laboratuvarları'ndan Davis ve ark. (1952), tek bir konuşmacı için izole edilmiş rakam tanıma için bir sistem kurdu. Ayrıca yine 1950'lerin başlarında RCA Laboratuvarları'ndan Olson ve Belar (1956), tek bir konuşmacının 10 hecesini tanımak için bir sistem kurdu. Denes (1959), İngiltere'deki Üniversite Koleji'nde dört sesli harf ve dokuz sessiz harfi tanımak için bir ses tanıyıcı geliştirdi.

1960'larda, birkaç Japon laboratuvarı, konuşma tanıma görevini gerçekleştirmek için özel bir donanım oluşturdu. Bunlardan dikkate değer olanları, Tokyo'daki Radyo Araştırma Laboratuvarı'ndaki Suzuki (1961) tarafından geliştirilen sesli harf tanıyıcısı ve Kyoto Üniversitesi'ndeki Sakai (1962) tarafından geliştirilen fonem tanıyıcısıdır. Sakai'nin çalışmaları, girdi konuşmasının farklı bölümlerindeki konuşmanın analizi ve tanınması için bir konuşma bölümleyicinin ilk kullanımını içeriyordu.

Atal ve Hanauer (1971), konuşma sinyalinden insan ses yolunu formüle eden, günümüzde de öznitelik çıkarım tekniği ve konuşma sinyali sıkıştırma alanında kullanılan, LPC'nin temel kavramlarını formüle etti. LPC bu çalışma kapsamında incelenen öznitelik çıkarım tekniklerinden biridir. Itakura (1975) ve Rabiner ve ark. (1979), LPC yöntemlerine dayalı olarak temel örüntü tanıma teknolojisinin konuşma tanımaya uygulanmasına ilişkin temel fikirler önerdi.

1984'te Amerikan Savunma Bakanlığı, bu çalışmada incelenen öznitelik çıkarım tekniklerinden biri olan, LPC'nin detaylarını bir taslak halinde sunan Federal Standart 1015'i yayınladı. 1980'lerde müzik sinyallerinin sıkıştırılması amacıyla LPC'nin genişletilmiş versiyonları olan CELP ve VSELP kullanıldı. 90'larda bu algoritmalar daha da geliştirilerek sıkıştırma oranları ve ses kaliteleri arttırıldı (Bradbury, 2000, s.5, para. 1).

LPC'nin konuşma tanıma uygulamaları için en yaygın öznitelik çıkarım tekniği olarak kullanılması 80'lere kadar sürdü. 80'lerde LPC konuşma tanıma sistemlerindeki yerini MFCC'ye bıraktı. LPC 2000'li yılların başında, konuşmanın telefon hatlarında iletimi için, konuşma sinyalini sıkıştırması sebebiyle kullanılan bir tekniktir (O'Shaughnessy, 2008, s. 2967, para. 1).

MFCC (Mel-Frekans Kepstral Katsayıları) bu çalışmada incelenen bir diğer öznitelik çıkarım tekniğidir. Mermelstein (1976) MFCC öznitelik çıkarım tekniğini geliştirdi. Mermelstein MFCC'nin temel kavramları için Bridle ve Brown'dan (1974) esinlenmiştir.

IBM'den Jelinek ve ark. (1975), ana işlevi konuşulan bir cümleyi bir ekranda gösterilebilecek veya bir kâğıda yazabilecek bir sistem oluşturmayı amaçlıyordu. IBM bu çalışmalar neticesinde "Tangora" isimli konuşmacıdan bağımsız (daktilo her bir kullanıcı tarafından eğitilmelidir), konuşulanı yazan bir daktilo sistemi geliştirdi. Bu sistemin teknik olarak odaklandığı noktalar, kelime dağarcığının arttırılması ve dil modelinin oluşturulmasıydı.

AT&T Bell Laboratuvarlarında, araştırma programının amacı, sesli arama ve telefon aramalarının yönlendirilmesi için komuta ve kontrol gibi halka otomatik telekomünikasyon hizmetleri sağlamaktı. Bu otomatik sistemlerin, bireysel konuşmacı eğitimine ihtiyaç duymadan çok sayıda konuşmacı için iyi çalışması bekleniyordu. Bell Laboratuvarları'ndaki odak noktası, çoğu zaman belirgin şekilde farklı aksanlara sahip birçok farklı konuşmacıdan gelen konuşma sinyallerindeki içsel akustik değişkenlikle başa çıkabilen, konuşmacıdan bağımsız bir sistemin tasarımındaydı. Bu, çok çeşitli konuşmacılar ve aksanlarda

kullanılabilecek kelime ve ses referans kalıpları (başlangıçta şablonlar, ancak nihayetinde istatistiksel modeller) oluşturmak için bir dizi konuşma kümeleme algoritmasının oluşturulmasına yol açtı. Ayrıca, konuşmacılar arasında çeşitli konuşma örneklerinin akustik değişkenliğini anlamak ve kontrol etmek için yapılan araştırmalar, çok sayıda kişiden alınan zengin konuşma ifadeleri üreten bir dizi spektral mesafe ölçümü (örneğin, Itakura mesafesi Itakura, 1975) ve istatistiksel modelleme tekniklerinin araştırılmasına yol açtı (Juang ve ark., 1986, s. 11, para. 1).

IBM ve AT&T Bell Laboratuvarları'ndaki konuşma tanıma yaklaşımlarının her ikisi de insan-makine sesli iletişim teknolojisinin evriminde derin bir etkiye sahip oldu. Bu çabalar arasındaki farklılıklara rağmen ortak bir yön ise, matematiksel formüllemenin konuşma tanıma araştırmalarının farklı ve önemli yönü olarak ortaya çıkmaya başlamasıydı. Bu iki yaklaşımın hedeflerindeki farklılık, teknolojinin çeşitli uygulamalarda farklı şekilde gerçekleştirilmesine yol açarken, 1980'lerde istatistiksel yöntemlerin hızlı gelişimi, özellikle de Gizli Markov modeli (HMM), belirli bir derecede sistem tasarımında benzeşmeye neden oldu (Juang ve Rabiner 2005, s. 12, para. 1).

1980'ler ve 1990'lardaki istatistiksel yöntemlerin başarısı sebebiyle DARPA konuşma tanıma sistemlerine yeniden ilgi gösterdi ve verdiği desteklerle Carneige Mellon üniversitesinden Lee (1988) "Sphinx" isimli, Schwartz ve ark. (1989) "Byblos" isimli, SRI'dan Murveit ve ark. (1989) "Depicher" isimli konuşma tanıma sistemlerini geliştirdiler.

Hermanski (1990) konuşma sinyali analizi için PLP'yi geliştirildi. Bu teknik, işitsel spektrumun (bkz. 1.1.5.2) bir tahminini elde etmek için işitmenin psikofiziğinden üç kavram kullanır: kritik bant spektral çözünürlüğü, eşit ses yüksekliği eğrisi (bkz. 2.4.1) ve yoğunluk-ses şiddeti yasası (bkz. 2.4.2). PLP-RASTA Hermanski ve ark. (1991) tarafından doğrusal spektral bozulmalara karşı PLP'yi daha sağlam yapmak için geliştirildi. PLP-RASTA bu çalışma kapsamında incelenen öznitelik çıkarım tekniklerinden biridir.

1990'larda yapay sinir ağları (YSA) konuşma tanıma için kullanılmaya başlandı. İlk girişimler, birkaç fonemi veya birkaç kelimeyi (örneğin, izole edilmiş rakamlar) tanımak gibi basit görevlere odaklandı ve başarılı oldu (Lippmann, 1990). Patil (1998) konuşma tanıma uygulaması için ileri beslemeli, çok katmanlı, yapay sinir ağını LPC vektörleriyle eğitti.

1990'larda, tüm dünyada birçok bireysel araştırma programını mümkün kılan yazılım araçlarının geliştirilmesinde büyük ilerleme kaydedildi. Sistemler daha karmaşık hale geldikçe, yeni kavramları ve algoritmaları birleştirmek için daha fazla araştırma ve geliştirme için bir yazılım kütüphanesine gerek duyuldu ve bu doğrultuda Cambridge Üniversitesinden Young ve ark. (1999) liderliğindeki araştırma ekibi HTK (Gizli Markov Model Araç Kiti) sistemini kullanıma sundu. Bu sistem HMM için bir yazılım kütüphanesidir ve HMM için bazı standartlar belirlemiştir. HTK günümüzde hala kullanılmaktadır.

2000'li yıllara kadar, HMM-GMM gibi istatistiksel sistemler konuşma tanıma sistemlerinde ana model olmuştur (Wang ve ark., 2019, s. 1). İstatistiksel modeller, akustik ve dil modeli (bkz. 1.4, 1.5) için, konuşulan ifadeyi istatistiksel olarak en olası metne dönüştürmek için, olasılık parametrelerinden oluşurlar. Bu olasılık parametrelerini ise geniş bir veri seti ile eğitildikten sonra elde ederler (Koçer ve Ahmed, 2019, s. 42; Oyucu ve ark., 2019, s. 844, 845).

2000'li yıllarda ise yapay zekânın konuşma tanıma sistemlerinde kullanılmaya başlaması ağırlık kazandı. Yapay zekânın alt dallarından biri olan makine öğrenmesi ve makine öğrenmesinin alt dallarından biri olan derin öğrenme, konuşma tanıma sistemlerinde kullanılan yapay zekâ teknolojisidir. Makine öğrenmesi bu yıllarda yüksek hesaplama gerektiren sistemler için hesaplama maliyetini düşürecek çözümler sundu. Konuşma tanıma sistemleri de yoğun hesaplama gerektiren sistemler olduğundan, derin öğrenme mimarileri konuşma tanıma sistemlerinde kullanılmaya başlandı.

2000'li yıllardan günümüze gelene kadar ise, HMM-DNN (Gizli Markov Modeli – Derin Sinir Ağı) gibi hem istatistiksel yöntemleri hem de YSA (Yapay Sinir Ağı) modellerini bir arada kullanan hibrit modeller veya sadece derin öğrenme kullanan uçtan uca modeller, sadece HMM-GMM'ye dayanan istatiksel modellerin ötesinde bir performans elde etti. Hibrit modeller ve YSA'ya dayanan modeller, karşılaştırılabilir performanslara sahiptir (Wang ve ark., 2019, s. 1). Hibrit veya sadece YSA'ya dayalı modellerde akustik, telaffuz ve dil modeli olarak, geniş veri setlerinden çıkarın öznitelik vektörleriyle eğitilen yapay sinir ağları modelleri, konuşulan ifadenin tahminini yapar.

Problem

Konuşma sinyali üzerinden öznitelik çıkarımı, konuşma tanıma sistemlerinin en önemli aşamalarından biridir. Konuşma tanıma sistemlerinde kullanılan istatistiksel modeller ve Derin öğrenme yöntemleri bu alanda çok isabetli tahminlemelerde bulunabilmektedir. Ancak yapısal olarak bu modellerin girdileri ve çıktıları sadece rakamlardır.

Öznitelik çıkarım tekniklerinin amacı, ses verisini, konuşma tanıma sistemlerinin diğer adımlarında kullanılmak üzere, sinyali temsil eden, orijinal sinyalden daha küçük boyuta sahip, sayılara (vektörlere) dönüştürmektir. Ses verisinin, onu en iyi şekilde temsil edecek vektörlere dönüştürülmesi bu araştırmada irdelenecek ana sorundur.

Ayrıca konuşma sinyalinden çıkarılan öznitelik vektörlerinin, istenen uygulama için gereken tüm özellikleri olabilecek en uygun şekilde, donanımsal hesaplama yükünü de hesaba katarak, temsil etmesi gerekmektedir. Öznitelik çıkarımı sorununun ne şekilde aşıldığı, sistemin verilen ifadeyi tanıma oranını doğrudan etkileyecektir.

Amaç

Bu tez çalışmasında amaç çeşitli konuşma tanıma sistemlerinde yaygın olarak kullanılan MFCC, LPC, PLP-RASTA öznitelik çıkarım tekniklerini aşağıdaki yönleriyle inceleyerek karşılaştırılmasının yapılmasıdır. Bu karşılaştırma yapılırken literatürde yapılan benzer çalışmalar ve Bölüm 3'te geliştirilen, bu teknikleri Türkçe ve İngilizce veri seti üzerinde uygulayarak elde ettiği öznitelik vektörleriyle, yapay sinir ağı eğiten uygulama temel alınmıştır (bkz. Tablo 1).

- Verilen ifadeyi tanıma oranı
- Hesaplama süresi
- Gürültüye karşı dayanıklılık
- Türkçe ve İngilizce veri seti için tanıma oranları

Tablo 1

Kriterlerin Literatüre veya Geliştirilen Uygulamaya Göre Değerlendirilmesi

	Literatüre Göre	Uygulamaya Göre
Verilen İfadeyi Tanıma Oranı	✓	✓
Hesaplama Süresi	✓	✓
Gürültüye Karşı Dayanıklılık	✓	
Türkçe ve İngilizce Tanıma Oranları		✓

Önem

Bu tez çalışmasıyla, konuşma tanıma sistemlerinin genel yapısı ve üç önemli öznitelik çıkarım tekniğinin açıklaması yapılıp, bu tekniklerin artı ve eksi yönleri ortaya konulmuştur. Ayrıca örnek bir uygulama geliştirilerek, İngilizce ve Türkçe için tanıma oranları kıyaslanmıştır.

Bu çalışmanın önemi, farklı öznitelik çıkarım tekniklerinin, konuşma tanıma sistemlerine etkilerinin tespit edilmesi, Türkçe ve İngilizce için performanslarının ortaya konulması ve öznitelik çıkarım tekniklerini kıyaslayan ilk Türkçe çalışma olmasıdır. Ayrıca bu çalışmanın, Üst Kurulun yayın takip sistemlerinde kullanılabilecek teorik ve pratik bir altyapı oluşturması öngörülmektedir.

Sınırlılıklar

Bu çalışmada konuşma tanıma sistemlerinin genel yapısı içinde öznitelik çıkarımı ve öznitelik çıkarımının temelini oluşturan sinyal işleme konularına detaylandırılarak incelenecektir. Konuşma tanıma sistemlerinin diğer bölümleri olan akustik ve dil modelleme ise genel olarak açıklanacaktır.

Öznitelik çıkarımı kısmında ise sadece yaygın olarak kullanılan üç tekniğin (MFCC, LPC, PLP-RASTA) teorik olarak açıklaması yapılıp, literatürde bu teknikler üzerine yapılan çalışmalar çeşitli yönleriyle incelenecektir. Bu tekniklerin seçilme sebebi kullanımlarının diğer tekniklere göre daha yaygın olmasıdır.

Uygulama kısmında ise öznitelik çıkarımı için kullanılan veri setleri, İngilizcede 30 komut Türkçede 14 komut olarak sınırlandırılmıştır. Her komut için 1700 ile 2400 adet arasında, bir saniyelik örnek ses dosyaları kullanılmıştır.

Tanımlar

Öznitelik Çıkarımı

Konuşma sinyalinin ilk hali, işlenmesi için çok fazla hesaplama kaynağı gerektiren çok sayıda değişkenden oluşmaktadır. Öznitelik çıkarma, konuşma sinyalinin ilk ham veri kümesinin işlenmek üzere daha yönetilebilir gruplara indirgendiği bir boyutsallık azaltma işlemidir. Öznitelik çıkarma, değişkenleri seçen ve/veya özelliklerde birleştiren, işlenmesi gereken veri miktarını etkili bir şekilde azaltan ve yine de orijinal veri kümesini doğru ve tam olarak tanımlayan yöntemdir. Öznitelik çıkarma, belirli bir analiz için gerekli olmayan veri miktarını da azaltabilir. Ayrıca, verilerin azaltılması ve makinenin değişken kombinasyonlar (özellikler) oluşturma çabaları, makine öğrenme sürecinde öğrenme ve genelleme adımlarının hızını arttırır. Kısacası konuşma tanıma sistemlerinde kullanılan modeller (istatistiksel ve yapay zekâ modelleri), sinyalden konuşma metnini üretmek için, sinyalin tamamına bakmak yerine, sinyali küçük bloklara ayırıp her bir blok için çıkarılan öznitelik vektörlerini kullanırlar (Alim ve Rashid, 2018, s. 4; Dave, 2013, s. 1).

Fonem

Modern fonem teorisinde fonem anlam ayırt edici sestir. Bir kelimede anlamı değiştiren en küçük ses birimine denir. Fonem, genel anlamda insana ait olan konuşmanın ve bilimsel anlamda fonolojinin en küçük birimidir. Örneğin, yay sözcüğünün başındaki y sesi onu bay ve ray sözcüklerinden ayırır. Bu durumda y-ay, b-ay, r-ay sözcüklerinde /y/, /b/, /r/ farklı ses birimlerdir. Türkiye Türkçesinin resmî alfabesinde 29 fonem vardır (Demirci, 2011, s. 361, para. 2).

Türkçe yazıldığı gibi okunup okunduğu gibi yazıldığı için, fonem harfe karşılık gelmektedir. Ancak her dil için bu geçerli değildir. Örneğin İngilizcede harflerin telaffuzunun farklı olması nedeniyle fonem sayısı harf sayısından fazladır (Bett, 2002, s. 11).

Alofon

Fonemler konuşmacının aksanına, yaşına, cinsiyetine, duygu durumuna ve fonemin kelime içindeki konumuna göre değişiklik göstermektedir. Fonemin bu farklı varyasyonlarına alofon denir. Alofon fonemin alt birimidir (Demirci, 2011, s. 361, para. 2).

Yapay Sinir Ağları

Yapay Sinir Ağları (YSA) isminden de anlaşılabileceği gibi beyni oluşturan sinir hücrelerini (nöron) matematiksel olarak taklit ederek akıllı bir sistem oluşturmaya çalışan bir yapay zekâ yöntemidir. YSA başka bir deyişle biyolojik sinir ağlarını taklit eden bilgisayar programlarıdır (Taşova, 2011, s. 18)

Uygulama Programlama Ara yüzü (API)

API herhangi bir uygulamanın belli işlevlerinin diğer uygulamalar tarafından kullanılabilir modülleridir. Günümüzde pek çok kapsamlı uygulama API kullanarak kendi işlevselliğini diğer uygulamalara entegre etmekle beraber, bünyesinde bulunan pek çok veriyi de API'lar aracılığıyla üçüncü parti kişi veya uygulamalara sunmaktadır.

Gizli Markov Modeli (HMM)

Gizli Markov modeli, konuşma sinyalinin içsel değişkenliğini ve ayrıca konuşulan dilin yapısını bütünleşmiş ve tutarlı bir istatistiksel modelleme çerçevesinde modeller. Gerçekçi bir konuşma sinyali doğası gereği telaffuz ve aksan farklılıklarının yanı sıra yankı ve gürültü gibi çevresel faktörlerden dolayı oldukça değişkendir. İnsanlar aynı kelimeyi konuştuğunda, altta yatan dil yapısı, telaffuz, sözdizimi ve dilbilgisi açısından aynı kalsa bile ses sinyalleri aynı değildir. HMM, konuşulan ifadedeki akustik oluşumlardaki değişkenlik için bir olasılık dağılımı dizisi ve dil yapısının temsili için Markov zincirini kullanan bir olasılık ölçümüdür (Juang ve Rabiner, 2005, s. 12, 13).

Yöntem

Araştırmanın Modeli

Bu tez çalışmasında model olarak "Karşılaştırmalı Analiz" ve "Deney" yöntemi kullanılmıştır. Tez konusuna giren öznitelik çıkarım teknikleri hesaplama süresi, verilen ifadeyi tanıma oranları ve çevredeki gürültüye karşı dayanıklılıkları yönleriyle incelenmiştir. Bu teknikleri kullanarak, Türkçe ve İngilizce komutlar için evrişimsel sinir ağı eğiten, konuşma tanıma sistemi geliştirilmiştir. Bulgular bilimsel etik ve kurallara uygun biçimde aktarılmıştır.

Evren ve Örneklem

Çalışmanın evrenini konuşma tanıma sistemlerinde kullanılan öznitelik çıkarım teknikleri oluşturmaktadır. Bu bağlamda belirlenen üç öznitelik çıkarım tekniği bu çalışmanın örneklemini oluşturmaktadır. Bu tekniklerin konuşma tanıma sistemlerindeki kullanım oranlarının, diğer tekniklere göre daha yüksek olması bu tekniklerin seçilme nedenidir.

Çalışma kapsamında geliştirilen uygulamada, Türkçe ve İngilizce'ye ait kelimeler çalışmanın evrenini, Tablo 2 ve Tablo 3'teki komutlar ise çalışmanın örnekleminin oluşturmaktadır. Bu komutların seçilme nedeni; dijital ortamda istenen niteliklere sahip veri setlerinin genellikle bu komutlardan oluşmasıdır.

Tablo 2

Türkçe Veri Setindeki Komutlar ve Ses Dosyası Sayıları

Komut	Ses Dosyası Sayısı
Aç	1.955
Aşağı	1.870
Başlat	1.879
Devam	1.880
Dur	1.887
Evet	1.885
Geri	1.880
Hayır	1.843
İleri	1.886
İptal	1.952
Kapa	1.882
Sağ	1.883
Sol	1.910
Yukarı	1.892

Tablo 3

İngilizce Veri Setindeki Komutlar ve Ses Dosyası Sayıları

Komut	Ses Dosyası Sayısı
Bad	1.713
Bird	1.731
Cat	1.733
Dog	1.746
Down	2.359
Eight	2.352
Five	2.357
Four	2.372
Go	2.372
Happy	1.742
House	1.750
Left	2.353
Marvin	1.746
Nine	2.364
No	2.375
Off	2.357
On	2.367
One	2.370
Right	2.367
Seven	2.377
Sheila	1.734
Six	2.369
Stop	2.380
Three	2.356
Tree	1.733
Two	2.373
Up	2.375
Wow	1.745
Yes	2.377
Zero	2.376

Verilerin Toplanması

Bu tez çalışmasında için kullanılan kaynaklar; dijital makale ve tezleri içeren kütüphanelerden (IEEE, Academia, Dergipark, NDLTD, OATD, Springer, IJERT), YÖK Tez Merkezinden, Üniversitelerin dijital ortamda paylaştığı ders notlarından, bazı araştırmacıların kendi web sitelerinden ve Google Akademi'den edinilmiştir.

Çalışma kapsamında geliştirilen uygulama için veri seti olarak; Türkçe için Kurtkaya (2021) tarafından oluşturulan ve Tablo 2'de detayları verilen veri seti, İngilizce için Google bünyesinde "TensorFlow" ve "AIY" (Google tarafından desteklenen yapay zekâ donanım kiti) ekipleri tarafından oluşturulan (Warden, 2019), detayları Tablo 3'te verilen veri seti kullanılmıştır.

Verilerin Çözümlenmesi

Çalışmada detaylı kaynak taraması yapılmasının ardından araştırmanın amaçlarının ve sınırlılıklarının dışında kalan veriler elendikten sonra geriye kalan veriler, odaklandıkları konulara göre sınıflandırılmıştır. Elde edilen verilerin çözümlenmesi için literatürden hareketle incelemeler ve karşılaştırmalar yapılmıştır.

BÖLÜM 1

KONUŞMA TANIMA TEKNOLOJİSİNE GENEL BAKIŞ

Bu bölümde konuşma tanıma sistemlerinin alt sistemleri olan konuşma sinyal işleme, öznitelik çıkarımı, akustik modelleme ve dil modellemesi ele alınacaktır. Ayrıca akustik ve dil modeli için kullanılan iki temel model olan HMM-GMM ve yapay sinir ağı modelleri açıklanacaktır. Bu bölüm öznitelik çıkarım işleminin konuşma tanıma sistemindeki yerini anlamamız açısından önem arz etmektedir.

Konuşma sinyal işleme, konuşma sinyallerinin bilgisayarda ilk oluşumunu ve çeşitli gösterim ve yöntemlerle analinizi yapmaktadır. Konuşma sinyal işleme öznitelik çıkarımının temelini oluşturmaktadır. Öznitelik çıkarımı, konuşma sinyalinden çeşitli özelikleri çıkarabilmek için, konuşma sinyalini temsil eden katsayılardan oluşan vektör ya da vektörlerin elde edilmesi işlemidir. Akustik ve dil modelleme, dile özgü kelime ve cümle yapılarının modellenmesi işlemidir. Akustik modellemede istatistiksel (HMM-GMM) ve yapay zekâ modellemeleri (yapay sinir ağları) kullanılmaktadır (bkz. 1.3). Bu modellemelerde hedef en olası harf kelime ve cümle dizilimini üretmektir. Konuşma tanıma sistemleri fonetik (dilbilim), bilgisayar bilimi ve elektrik ve elektronik mühendisliği alanlarında uzmanlık gerektiren bir teknolojidir.

Fonetik, insan konuşmasının ürettiği seslerin incelenmesine odaklanan, bu seslerin oluşumlarını, akustik özelliklerini ve algılanma biçimlerini kapsayan, dilbilimin bir parçasıdır. Fonetik temelde 3 ana bölümde incelenir ve bu bölümlerin hepsi konuşma tanıma sistemleri alanına girer (Hoque, 2020, s. 2).

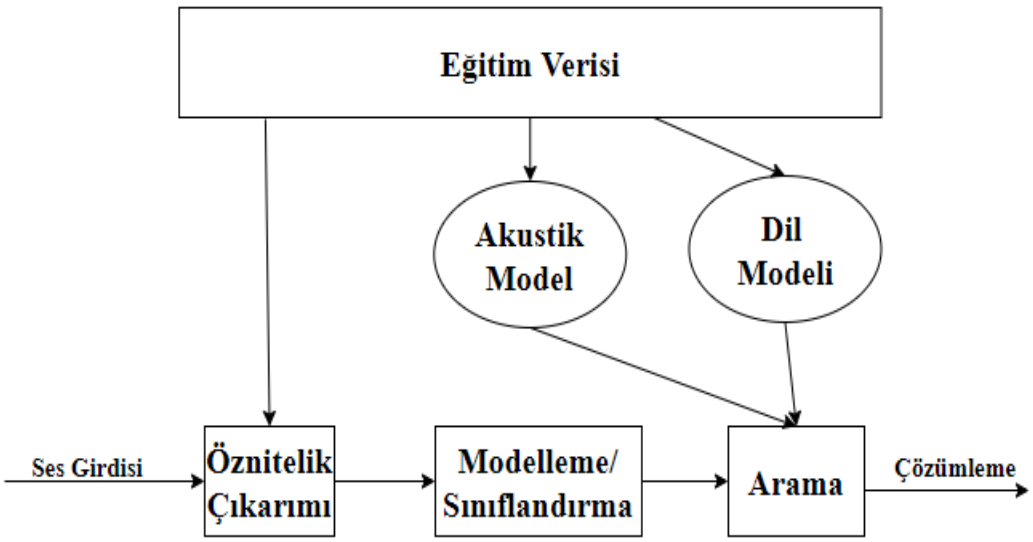
- Artikülasyon fonetik, ses yolu ve çeşitli ses yolu elemanlarının (dudak, çene, dil, yutak, yumuşak damak) aracılığıyla konuşma seslerinin üretimine odaklanır.
- Akustik fonetik, konuşma seslerinin bir konuşmacıdan bir dinleyiciye iletilmesine odaklanır.
- İşitsel fonetik, dinleyici tarafından konuşma seslerinin alınmasına ve algılanmasına odaklanır.

Konuşma sesinin en küçük birimine fonem denir. Kelimeler sırayla bir veya daha fazla fonemden oluşur. Bir fonemin akustik olarak gerçekleşmesine fon denir. Fonemlerin önemli bir yönü, oluşumlarının arkası ve önündeki fonlara bağlı olarak değişebilmesidir, buna fonetik bağlam denir. Tüm son teknoloji konuşma tanıma sistemleri, fonemlerin bu bağlama bağlı değişen hallerini kullanırlar (Demirci, 2011, s. 362, para. 1).

Konuşma tanıma sistemleri Şekil 1'deki gibi gösterebiliriz. Akustik ve dil modeli gerçekleştirmek için geliştirilen modellerde iki farklı yaklaşım vardır. Bu yaklaşımların biri istatistiksel yaklaşım diğeri ise yapay zekâya dayalı yaklaşımdır. Şekil 1'deki eğitim verisi istatistiksel modellerde olasılık parametrelerini elde etmek için, yapay zekâya dayalı modellerde ise derin öğrenme mimarilerini eğitmek için kullanılır.

Şekil 1

Ses Tanıma Sistemine İlişkin Genel Blok Yapı (Cole ve ark., 1997, s. 5)



1.1. Konuşma Sinyal İşleme

Bu bölümde ilk olarak konuşma sinyaline ilişkin tanımlar verilmiştir. Bu tanımlar konuşulan bir ifadenin, mikrofon ile yakalanıp, konuşma sinyaline dönüşümünü ve bu sinyalin özelliklerini anlamak için önemlidir. Ardından analog sinyalden dijital sinyale dönüşüm ve bu dönüşüm sırasında ortaya çıkan örneklem sayısı kavramı açıklanıp, insanların duyma aralığı bakımından ilişkisine yer verilecektir. Örneklem sayısını konuşma sinyalinin pikseli gibi düşünebiliriz, bu değer arttıkça o konuşmaya ait daha fazla özellik konuşma sinyalinde temsil edilmiş olur. Sonrasında sinyalin farklı amaçlar için gösterim türlerine ve bu gösterim türlerinden hangi özelliklerin analizinin yapılabileceğine değinilmiştir. Bu gösterimlerden, sinyalin zamana, frekansa veya hem zaman hem frekansa göre değişimlerini görebiliriz. Son olarak sinyalin, konuşma tanıma sistemleri açısından önem arz eden spektral sızıntı ve pencereleme kavramları açıklanacaktır. Bu kavramları bilmek, konuşma tanıma sistemlerinde, öznitelik çıkarımının bir aşaması olan önişlemi (bkz. 2.1) anlamak için gereklidir.

Konuşma sinyal işleme, konuşma tanıma sistemleri için bilinmesi gereken en temel konudur. Konuşma sinyalinin analizi, bilgisayarda konuşma ile ilgili hangi işlem yapılırsa yapılsın gerçekleştirilmesi gereken bir adımdır. Burada konuşma sinyali ilk olarak zaman alanında, zamana bağlı havadaki basıncı gösterecek şekilde oluşur. Bu gösterim, konuşma sinyali için en temel gösterimdir. Sonrasında bu gösterim, sinyalinin farklı incelemeleri için, farklı Fouirer dönüşümleri vasıtasıyla, frekans alanı ve frekans-zaman alanı olarak gösterilebilir. Özellikle frekans-zaman alanı gösterimi (spektogram), farklı analizler için pek çok veri sağladığından önemlidir. Bu çalışma kapsamında incelenen tekniklerden biri olan MFCC spektogramı insanın ses algılamasını yansıtacak şekilde dönüştürerek kullanmaktadır. Öznitelik çıkarım teknikleri nihayetinde konuşma sinyali üzerinde analizler yaparak bu sinyali temsil eden katsayılardan oluşan, daha küçük boyutlu öznitelik vektörleri elde etmek olduğundan, konuşma sinyal işlemeye dair tanımları ve sinyalin özelliklerini kavramak önemlidir.

Konuşma, ses telleri tarafından oluşturulan, ses yolunda ilerleyen, konuşanın ağzında üretilen ve dinleyici kulağı tarafından bir basınç dalgası olarak algılanan bir ses sinyalidir. Ses sinyali sesin bir temsilidir ve sesi yeniden üretmek için gerekli olan tüm bilgileri içerir. Ses sinyallerinin oluşum süreci ise şöyledir; ses dalgaları havada yayılır ve havadaki ses

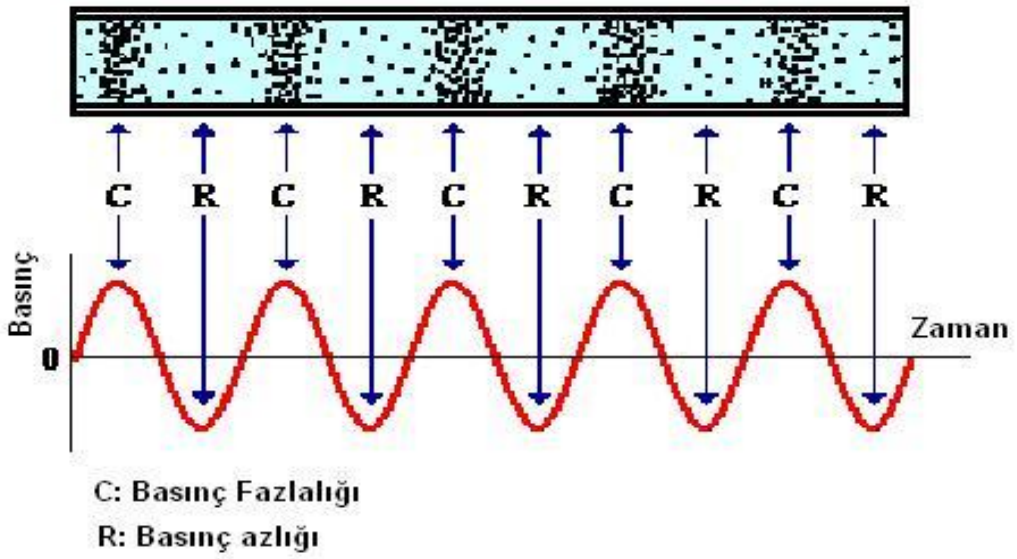
dalgasını elektriksel aktiviteye dönüştüren bir mikrofon tarafından yakalanır, sonrasında mikrofonun oluşturduğu elektriksel aktivite, sinyali tanımlayan bir dizi dalga formu örneği oluşturur. Analog ve sürekli bir yapıya sahip olan bu sinyal, üzerinde analiz ve işlem yapmak için dijital hale dönüştürülür (bkz. 1.1.3) (Peleg, 2010).

1.1.1. Ses & Ses Dalga Formu

Ses titreşen nesneler tarafından oluşturulan basınç dalgasıdır. Bu titreşimler, havadaki molekülleri hareket ettirir, harekete geçen moleküller sayesinde enerji havada bir yerden bir yere taşınır. Hareket eden moleküllerin havada oluşturduğu basınç dalgası Şekil 2'de ve örnek bir ses basınç dalgasının bilgisayarda oluşturduğu dalga formu Şekil 3'te görülebilir (Penn, 2010).

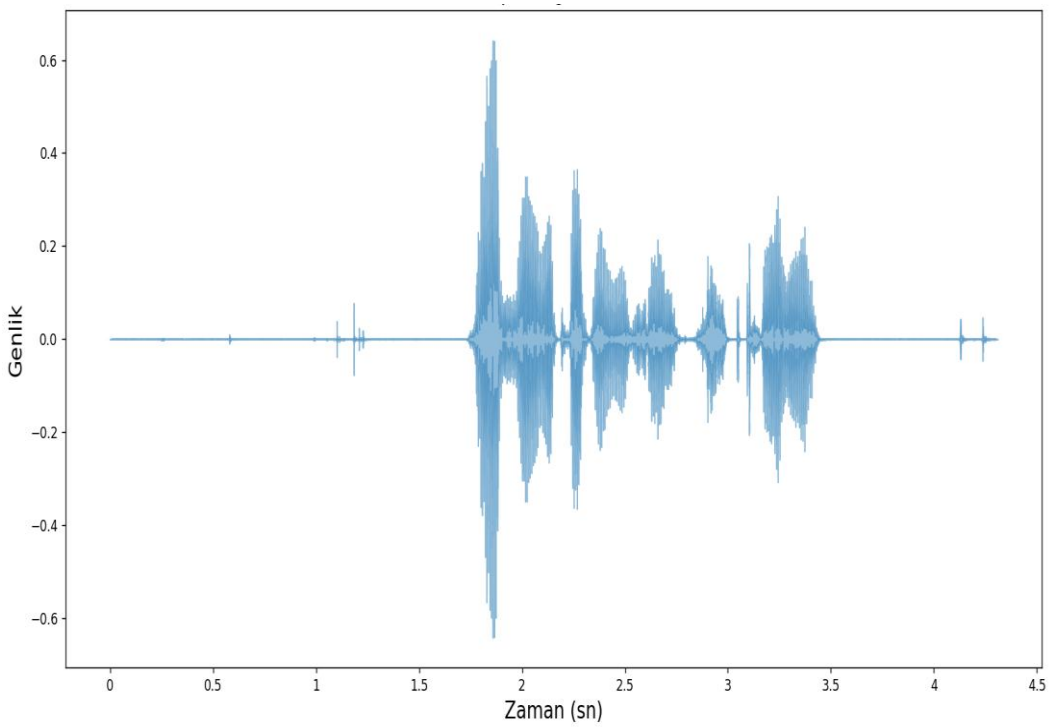
Şekil 2

Titreşen Hava Moleküllerinin Oluşturduğu Basınç Grafiği (Mikayilova, s.14)



Şekil 3

"Radyo ve Televizyon Üst Kurulu" İfadesinin Dalga Formu



Ses dalga formu, bir sesin iki boyutlu bir temsilidir. Bir ses dalga formu iki boyutludur. Dikey boyut basınç dalgası, yatay boyut ise zamandır. Dalga formları ses basıncının zaman içindeki değişimlerini gösterdiği için gösterime zaman alanı gösterimi denir (bkz. 1.1.5.1).

Ses basıncı genellikle desiBel (dB) cinsinden belirtilir. DesiBeller, ses basıncının veya yoğunluğunun logaritmik bir ölçeklendirmesidir ve kulak ve beynin ses genliğini yeniden ölçeklendirme şeklini yaklaşık olarak verir. Bu çalışma kapsamında incelenen MFCC ve PLP_RASTA öznitelik çıkarım teknikleri de insanın sesi logaritmik olarak algılamasına dayanırlar. Şekil 2 ve Şekil 3'te dalga formlarında dB ölçeği kullanılmaz. Bu dalga formlarındaki dikey eksen ses basıncının basit bir doğrusal ölçeklendirmesidir (Mannell, 2020).

1.1.2. Ses Dalgasında Genlik, Frekans, Güç, Tını

Ses sinyalini incelemek için ses dalga formuna ilişkin bazı tanımları bilmek gerekir. Ayrıca bu tanımlar öznitelik çıkarım tekniklerini anlamak için de gereklidir. Sesin genliği, havadaki molekülleri harekete geçirmek için gereken enerjiyi üretmek için yapılan iş miktarı olarak ölçülür. Genlik, dalganın ortalama konumundan maksimum dikey yer değiştirmesinin mesafesini ifade eder. Genlik ne kadar büyük olursa, enerji o kadar yüksek olur. Seste genlik, ses dalgasının içinden geçtiği ortamın (genellikle hava) yaşadığı sıkıştırma ve genişlemenin büyüklüğünü ifade eder. Bu genlik, kulaklarımız tarafından gürlük olarak algılanır. Yüksek genlik, yüksek seslere eşdeğerdir (Hillenbrand, 2016, s. 3). Şekil 3'te konuşma dalga formunun zamana bağlı değişen genliği verilmiştir.

Bir ses dalgasındaki frekans, havada dolaşan sesin titreşim oranını ifade eder. Frekans, bir sesin yüksek perdeli mi yoksa alçak perdeli olarak mı algılanacağına karar verir. Seste, frekans sesin perdesi olarak da bilinir. Şekil 2'deki basınç grafiğini düşündüğümüzde iki tepe noktası veya iki dip nokta arasındaki mesafe periyodu verir. Frekans ise 1/Periyot olarak tanımlanır. Ölçü birimi Hertz'dir ve 1 Hertz 1 saniyedeki döngünün frekansına eşittir. Hertz'den önce bir saniyedeki döngüye eşit olan CPS (Saniye Başına Döngü) kullanılırdı (Hillenbrand, 2016, s. 2).

Bir sesin gücü, o ses kaynağı tarafından her yöne yayılan birim zaman başına enerjidir, ölçü birimi Watt'dır. Birim alandaki ses gücüne sesin yoğunluğu denir birimi W/m^2'dir. Tını ise bir sesin aynı zaman aralığı, frekans ve yoğunluğa sahip diğer bir ses ile arasındaki fark olarak tanımlanır. Yani bir sesin aynı fiziksel şartlar altında diğer seslerden ayıran karakteristiğine o sesin tınısı denir (Hillenbrand, 2016, s. 17).

1.1.3. Analog Sinyalden Dijital Sinyale Geçiş

Bir analog sinyal, sürekli olan bilgiyi ifade eder. Analog konuşma sinyali, basınç ve zaman özelliklerinde sürekli olarak değişir. Örneğin kasetler, vinil kayıtlar, FM ve AM radyo yayınları. Manyetik bant kaydında saklanan analog sinyalin avantajı, dinleme ve analiz için oynatılabilmesidir. Oynatma, manyetik enerjinin tekrar elektrik enerjisine

dönüştürülmesiyle gerçekleştirilir ve bu da hoparlörler veya kulaklıklar tarafından akustik enerjiye dönüştürülür. Analog sinyalde zaman ve genlik için devamlı değerler dijital sinyalde ayrı değerler dizisine dönüştürülür, dijital sinyalde devamlılık yoktur, belli bir veri noktasında sınırlı sayıda değer vardır (Koni, 2022).

Ses, bir bilgisayarda dijital (ayrık) olarak temsil edilmesi gereken, analog bir olgudur. Bir analog sinyali dijitale dönüştürmek için ayrık örnekler elde etmemiz gerekir ve bunun için iki aşamalı işlem gerçekleştirilir; örnekleme ve niceleme.

Örnekleme, sinyali zaman ekseninde örneklere ayırma işidir. Saniye başına elde ettiğimiz örneklem sayısı o ses sinyali için örnekleme oranını verir ve birimi Hertz'dir. Eğer saniye başına çok az örneklemimiz varsa, sinyali doğru bir şekilde temsil edemeyiz. Eğer saniye başına çok fazla örneklemimiz varsa bu da daha fazla depolama alanı kullanmamıza neden olur. Örnek olarak, CD kalitesinde sesin örnekleme hızı 44,1 kHz, yani saniyede 44.100 örneklemdir, telefon için örnekleme oranı 8 kHz'dir. Örnekleme oranı seçmek için Nyquist teoremi bize önemli bir ölçüt vermektedir (bkz. 1.1.4).

Bir analog sinyale niceleme işlemi yapıldığında dikey eksendeki genlik ayrık değerlere veya artışlara dönüştürülür. Niceleme seviyelerinin sayısı ne kadar yüksekse, nicelenmiş sinyal analog sinyali o kadar doğru temsil eder. Konuşma sinyalinde, en az 12 bitlik bir dönüşümle niceleme yapılmalıdır. Çok az niceleme seviyesi kullanılırsa, sinyal niceleme gürültüsü adı verilen bozulmaya sahip olacaktır (Koni, 2022).

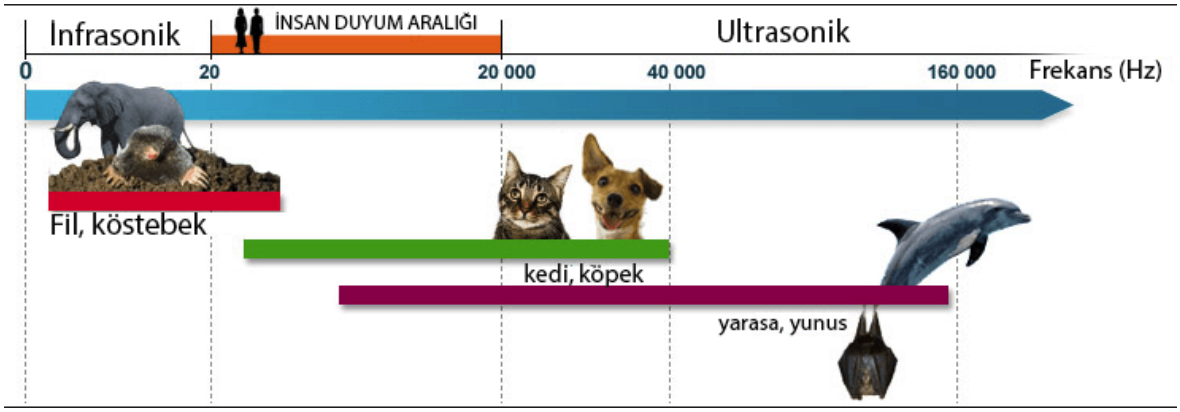
1.1.4. Nyquist Teoremi ve Duyma Aralığı

Nyquist teoremine göre örnekleme oranı, analog sinyaldeki en yüksek frekans bileşeninin iki katına eşit veya iki katından büyük olmalıdır. Günümüzde konuşma sinyalinin aslına uygun olarak yakalanabilmesi için frekansı en fazla 22.050 Hz. olmalıdır. Normal konuşma sinyalleri genelde 8.000 Hz. civarındadır bu nedenle normal konuşmalar için genellikle 16.000 Hz. örnekleme oranı kullanılır. Geleneksel telefonlarda frekans 3.400 Hz. ile sınırlandırılmıştır bu sebeple de telefon konuşmaları için Nyquist teoremi gereğince 8.000 Hz. örneklem oranı kullanılır (Landau, 1967).

Duyma aralığı, insanda 20 ile 20.000 Hz. aralığındadır. Frekansı 20 Hz'den küçük olan ses dalgalarına infrasonic ses dalgaları denir. Frekansı 20000 Hz'den büyük olan ses dalgalarına ultrasonic ses dalgaları denir. Bu aralık fillerde 14 ile 14.000 Hz, kedilerde 48 ile 75.000 Hz, köpeklerde 64 ile 45.000 Hz, farelerde 1.000 ile 70.000 Hz, yarasalarda 7.000 ile 200.000 Hz. aralığında gerçekleşmektedir (Heffner ve Heffner, 2007).

Şekil 4

İnsanlarda ve Bazı Hayvanlarda Duyma Aralığı (Solmaz, 2017)



Çalışma kapsamında incelenen MFCC tekniği insan kulağının kritik frekans bant genişliğini kullanan filtreler kullanmaktadır (bkz. 2.3.2). Üçüncü bölümde geliştirilen uygulama için de insan duyma aralığı baz alınarak veri setleri saniye başına 22.050 örneklem olacak şekilde örneklendirilmiştir.

1.1.5. Ses Sinyal Alanı

Çoğu konuşma işleme uygulaması, görevlerini yerine getirirken konuşma sinyallerinin belirli özelliklerini kullanır. Bu özelliklerin çıkarılması ve bunların bir konuşma sinyalinden nasıl elde edileceği, konuşma analizi olarak bilinir. Bu analiz zaman alanında veya frekans alanında yapılabilir. Konuşmayı zaman alanında analiz etmek genellikle basit hesaplama ve yorumlama gerektirir. Frekans alanı, konuşma analizinde en faydalı parametreleri elde etmek için kullanılır (Joy ve Savitha, 2015, s. 1, para. 2).

Zaman ve frekans alanına ek olarak hem zaman hem frekansı temsil eden zaman-frekans alanı da konuşma analizinde önemlidir. Bu gösterime spektogram denilmektedir. MFCC öznitelik çıkarım tekniğinde özel bir spektogram türü kullanılmaktadır (bkz. 2.3.2).

1.1.5.1. Zaman Alanı

Havadaki titreşim hareketini veya bu hareketin yarattığı hava basıncı bozulmasını temsil etmek için kullandığımız grafiklere zaman alanı gösterimi denir. Konuşma sinyalinin zaman alanında gösterimi, konuşmanın havada yarattığı basıncın zamana bağlı en basit gösterimidir (Hillenbrand, 2016, s. 8, para. 1).

Zaman alanında x ekseni zaman y ekseni genlik olacak şekilde bir gösterim söz konusudur (bkz. Şekil 3). Zaman alanındaki gösterimden konuşma sinyaline yönelik pek çok analiz yapılabiliriz bunlardan bazıları genlik zarfı, karekök ortalama, sıfır geçiş oranıdır (Joy ve Savitha, 2015).

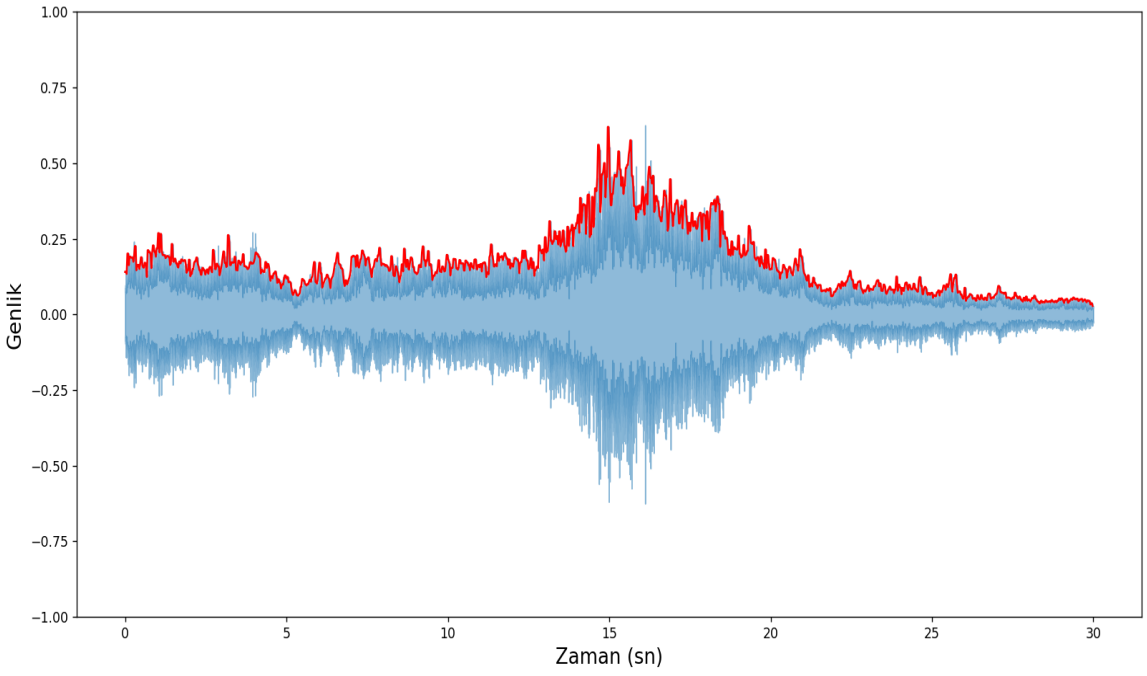
Genlik Zarfı

Genlik zarfında sinyali belli sayıda örneklem içeren bloklara ayırdıktan sonra her bir bloktaki maximum genliğe sahip örneklem bulunur. Denklem 1'de *GZ* genlik zarfı fonksiyonu olmak üzere *t* blok numarasını, *s(k)* blok içerisindeki k numaralı örneklemi, *K* bloktaki örneklem sayısını, *t.K* bloktaki ilk örneklemi *(t+1).K-1* ise bloktaki son örneklemi temsil etmektedir. Genlik zarfı bize kabaca ses seviyesinin yüksekliği hakkında bilgi verir ve müzik türü sınıflandırılmasında kullanılabilir (Goswami ve ark., 2002). Şekil 5'te 30 saniyelik bir ses dosyası için genlik zarfı hesaplanıp kırmızı noktalarla gösterilmiştir. Konuşma sinyaline genlik zarfı uygulayarak sinyali daha düşük boyutta, belli miktarda veri kaybını göze alarak, temsil edebiliriz, bu yönüyle öznitelik çıkarım teknikleriyle benzerdir.

$$GZ_t = max_{k=t.K}^{(t+1).K-1} s(k)$$ **Denklem 1** (Goswami ve ark., 2002)

Şekil 5

Örnek Genlik Zarfı



Karekök Ortalama (RMS)

RMS'de sinyali belli sayıda örneklem içeren bloklara ayırdıktan sonra her bir bloktaki maksimum genliğe sahip örneklem bulunur. Denklem 2'de *RMS* karekök ortalama fonksiyonu olmak üzere *t* blok numarasını, *k* numaralı *s(k)* örnekleminin karesi örneklemin enerjisini, *K* bloktaki örneklem sayısını, *t.K* bloktaki ilk örneklemi, *(t+1).K-1* ise bloktaki son örneklemi temsil etmektedir. RMS her bir blok için örneklemlerine bakarak ortalama enerjilerini hesaplar. RMS de genlik zarfı gibi kabaca ses seviyesinin yüksekliğini gösterir. Ayrıca müzik türü sınıflandırması ve ses sınıflandırmasında (konuşma, müzik, gürültü, sessizlik vb.) kullanılır (Gray ve Markel, 1976).

$$RMS_t = \sqrt{\frac{1}{K}\sum_{k=t.K}^{(t+1).K-1} s(k)^2}$$ **Denklem 2** (Gray ve Markel, 1976)

Sıfır Geçiş Oranı (ZCR)

Ardışık örneklerin farklı cebirsel işaretleri varsa, sıfır geçişinin meydana geldiği söylenir. Sıfır geçişlerinin meydana gelme hızı, bir sinyalin frekans içeriğinin basit bir ölçüsüdür. Sıfır geçiş oranı, sinyali belli sayıda örneklem içeren bloklara ayırdıktan sonra, her bir blokta sinyalin x eksenini geçme sayısının toplamıdır. Denklem 3'te *ZCR* sıfır geçiş oranı fonksiyonu olmak üzere *t* blok numarasını, *sgn(s(k))* örneklemin x eksenine konumunu (üzerindeyse +1 altındaysa -1 kesiyorsa 0 değerini alır.), *K* bloktaki örneklem sayısını, *t.K* bloktaki ilk örneklemi, *(t+1).K-1* ise bloktaki son örneklemi temsil etmektedir. ZCR değeri sesli harfler için düşük, sessiz harfler için yüksektir (Bachu ve ark., 2008). Ayrıca ZCR sinyaldeki herhangi bir ifade temsil etmeyen yerleri sinyalden çıkarmak veya sinyalin sonunu tespit etmek için de kullanılabilir (Joy ve Savitha, 2015, s. 2).

$$ZCR_t = \frac{1}{2} \cdot \sum_{k=t.K}^{(t+1).K-1} |sgn\big(s(k)\big) - sgn\big(s(k+1)\big)|$$ **Denklem 3** (Joy ve Savitha, 2015, s. 2).

1.1.5.2. Frekans Alanı

Frekans alanı gösterimine spektrum denir. Bu gösterimi elde etmek için zaman alanındaki sinyale hızlı fourier dönüşümü (FFT) veya ayrık fourier dönüşümü (DFT) uygulanır. Bu gösterimde x ekseni boyunca frekans, y ekseni boyunca genlik ölçüsü verilir (Hillenbrand, 2016, s. 8, para. 1).

Hızlı Fourier Dönüşümü (FFT)

FFT zaman bazlı (x ekseni zaman, y ekseni genlik) devamlı bir konuşma sinyalini frekans bazlı (x ekseni frekans, y ekseni büyüklük) bir sinyale(spektrum) dönüştürür. *g(f)* frekans bazlı sinyali, *g(t)* zaman bazlı sinyali temsil etmek üzere Denklem 4'deki gibi ifade edilir (Nussbaumer, 1981).

$$\hat{g}(f) = \int g(t) \, . \, e^{-i\,2\pi\,ft} \; dt$$ **Denklem 4** (Nussbaumer, 1981)

Ayrık Fourier Dönüşümü (DFT)

DFT'de zamana bağlı sinyale olduğu gibi FFT uygulayıp spektruma dönüştürmek yerine sinyal *T* zaman aralıklarında örneklemlere ayrılır, sonrasında ise belli sayıda örneklem içeren bloklara ayrılır ve her bir bloğa DFT (Ayrık Fourier Dönüşümü) ayrı ayrı uygulanır. Bunun sonucunda blok sayısı kadar kompleks sayıdan oluşan Fourier katsayıları elde edilir. Bu kompleks katsayılar frekans büyüklüğünü temsil eder. Dönüşüm uygulanırken orijinal sinyaldeki blok sayısı spektrumdaki frekans sayısına eşit olarak hesaplanır bunun nedeni hesaplama verimliliğini arttırmak ve spektrumdan orijinal sinyale tersine dönüşümü kolaylaştırmaktır.

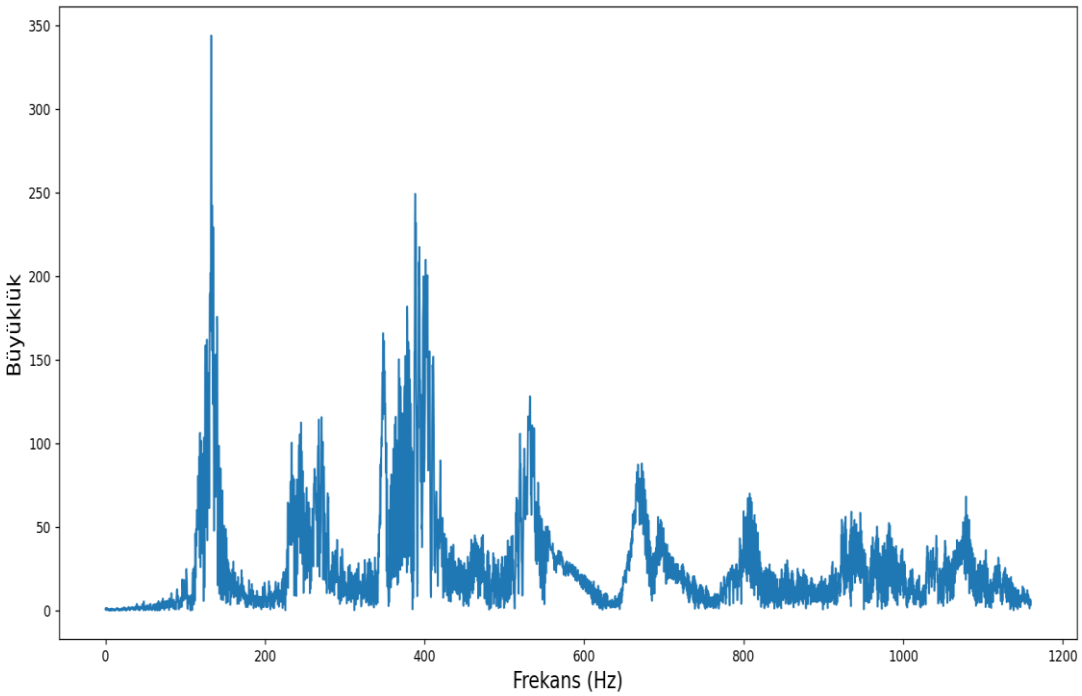
DFT Denklem 5'teki gibi ifade edilir. Denklem 5'te *x(n)* her bir örneklemi *N* ise örneklem sayısını temsil etmektedir. Spektrumda x ekseni frekans y ekseni büyüklük olacak şekildedir. Şekil 3'te zaman alanında gösterimi verilen ses sinyalinin, Şekil 6'da FFT uygulandıktan sonra oluşan spektrumu görülebilir (Winograd, 1978).

$k = [0, N-1]$ Olmak üzere

$$\hat{x}(k/N) = \sum_{n=0}^{N-1} \left(\mathrm{x(n)}.\ e^{-i\,2\pi\,n\frac{k}{N}}\right)$$ **Denklem 5** (Winograd, 1978).

Şekil 6

"Radyo ve Televizyon Üst Kurulu" İfadesine Ait Spektrum



Spektrumun İncelenmesi

Seslerin oluşturduğu dalga biçimlerini gözlemleyerek, dalga biçiminin özelliklerine iki faktörün katkıda bulunduğunu görürüz. Bu faktörlerden birincisi havayı ses yolundan ve ağızdan dışarı iten ses tellerinden gelen uyarı ikincisi ise ses yolunun bir sesi çıkarırken ki şekli. Örneğin, hem 's' hem de 'f' seslerinin, ses tellerindeki sessiz uyarılma nedeniyle gürültüye benzer şekilde ses dalga formunda yerlerini alırlar ancak farklı sesler oldukları için, ses dalga formunda, genel itibariyle farklı şekillere sahip olduklarını görebiliriz. Aynı şey sesli harfler için de geçerlidir, aynı konuşmacıdan gelen farklı bir sesli harf benzer bir periyodikliğe ancak farklı bir genel şekle sahip olacaktır çünkü aynı ses telleri sesli harf için aynı uyarıyı oluşturur, ancak farklı bir ses üretirken ses yolunun şekli farklıdır. Örnek olarak çıkardığımız bir sesin oluşturduğu sinyal dalgasını incelediğimizde çoğu zaman sessiz harflerin rastgele ve gürültülü, sesli harflerin ise periyodik ve düzenli gibi göründüğünü görebiliriz bunun nedeni, ses tellerimizin bu sesleri çıkarırken farklı şekilde titreşmesidir (Bradbury, 2000, s. 6, 7).

1.1.5.3. Zaman ve Frekans Alanı

Bu gösterime spektogram denir. Frekans-zaman gösterimini elde etmek için zaman gösterimindeki sinyale kısa zamanlı fourier dönüşümü (STFT) uygulanır. MFCC öznitelik çıkarım tekniği, spektogramı insan duyma aralığını yansıtacak şekilde temsil eden Mel spektogramını hesaplamasında kullanır (bkz. 2.3.2).

Kısa Zamanlı Fourier Dönüşümü (STFT)

STFT zaman bazlı sinyali zaman-frekans sinyaline (spektrogram) dönüştürmek için kullanılır. STFT gerçekleştirilirken ilk olarak zaman bazlı sinyal *T* zaman aralıklarında ve belli sayıda örneklem içeren örtüşen bloklara ayrılır. Örtüşen bloklara ayrılmasının sebebi

pencereleme işleminden sonra oluşabilecek sinyal kayıplarını gidermektir. Sonrasında pencereleme işlemi yapılır. Pencereleme işlemi için zil şeklindeki kıvrıma sahip olan Hann fonksiyonu kullanılır. Buraya kadar yapılan sinyalin bloklara ayrılması ve pencereleme işlemleri öznitelik çıkarımının önişlem kısmında da uygulanmaktadır ve bu bölümde detaylıca açıklanmıştır (bkz. 2.1.2, 2.1.3). Ardından *N* blok sayısı olmak üzere her bir bloğa STFT uygulanır (Lu ve Zhang, 2009).

STFT sonucunda, her bir blok için frekans vektörleri ayrı ayrı oluştuğundan, her blok için kompleks Fourier katsayılarından oluşan vektörlerin birleşimi ile bir matris oluşur. STFT uygulanırken bir bloktaki örneklem sayısı arttıkça yani bloğun temsil ettiği zaman miktarı arttıkça spektogramda frekans çözünürlüğü artarken zaman çözünürlüğü azalır. Bu sebeple bloğun büyüklüğü (ne kadar örneklem içerdiği) için yapılan uygulamanın zaman veya frekans hassasiyetine göre optimal bir değer seçilir. Ayrıca hesaplama verimliliği arttırmak için bu değer ikinin kuvvetleri olacak şekilde seçilir. STFT Denklem 6 ile ifade edilir (Lu ve Zhang, 2009).

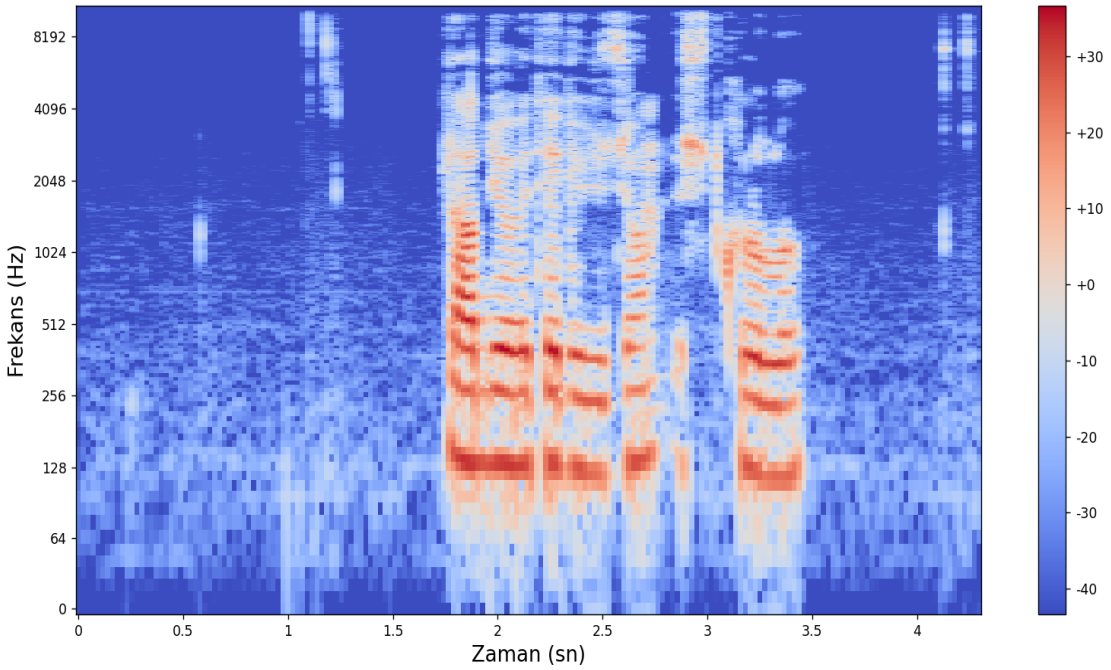
$k = [0, N-1]$ Olmak üzere

$$S(m,k) = \sum_{n=0}^{N-1} \left(\mathrm{x(n + mH)} \,.\, \mathrm{w(n)} \; e^{-i\,2\pi\, n \frac{k}{N}} \right)$$ **Denklem 6** (Griffin ve Lim, 1984)

Denklem 6'da $\hat{x}(k/N)$ dönüşüm fonksiyonunu temsil etmek üzere, *x(n)* pencerelenmiş örtüşen blokları, *N* ise blok sayısını temsil etmektedir. Spektogram sese ait hem zaman hem de frekans bilgilerini aynı anda temsil ettiği için pek çok uygulamada sıklıkla kullanılır (Griffin ve Lim, 1984). Şekil 7'de örnek bir ses sinyaline STFT uygulandıktan sonra oluşan spektogram gösterilmiştir.

Şekil 7

"Radyo ve Televizyon Üst Kurulu" İfadesine Ait Spektogram



1.1.6. Spektral Sızıntı ve Pencereleme

Öznitelik çıkarım tekniklerinin önişlem aşamasında, pencereleme, spektral sızıntıyı önlemek ve diğer bazı amaçlar için kullanılır (bkz. 2.1.3). Spektral sızıntı, zaman bazlı analog bir sinyali frekans bazı bir dijital sinyale çevirmek için Fourier dönüşümünü uyguladığımızda, sinyalde devamlı olmayan yerler (çoğu zaman sinyalin başı ve sonu), frekans bazlı sinyalde gereksiz yere yüksek frekans varmış gibi görünmektedir. Bunun önüne geçmek için her bir sinyal bloğuna pencereleme dediğimiz bir işlem uygulanır.

Pencereleme temel olarak zaman bazlı orijinal sinyalin başını ve sonunu yok etmektedir. Pencereleme işlemi her bir sinyal bloğuna uygulandıktan sonra her bloğun başı ve sonu için veri kaybı oluşabilmektedir. Bu sebeple sinyal bloklara ayrılırken pencerelemeden kaynaklanacak veri kaybını önlemek için blok ayrımı genellikle 10 ms'lik örtüşecek şekilde ayrılır (Alim ve Rashid, 2018, s. 5, para. 1, 2).

1.2. Öznitelik Çıkarımı

Öznitelik çıkarımı Otomatik konuşma tanıma sistemlerinin en önemli bölümlerinden biridir. Burada yapılan işlem konuşma sinyalinden, sinyali temsil için gereken özelliklerin çıkarımıdır. Öznitelik çıkarma, konuşma sinyalinin gücüne herhangi bir zarar vermeden giriş sinyalinin (vektörün) büyüklüğünü sıkıştırır (Narang ve Gupta, 2015, s. 109, para. 2).

Bu konuşma sinyalleri, yavaş zamanlamalı değişen sinyallerdir (yarı durağan). Yeterince kısa bir sürede (5-100 ms) incelendiğinde, özellikleri oldukça durağandır. Bu kısa süre için sinyal özellikleri değişirse, konuşulan sesin farklılaşmasına sebep olur. Konuşma sinyalindeki bilgi aslında konuşma dalga formunun kısa vadeli genlik spektrumu ile temsil edilir. Bu sebeple konuşmadan kısa vadeli genlik spektrumuna dayalı öznitelik vektörleri çıkarabiliriz. Konuşma tanımanın temel zorluğu, farklı konuşmacılar, konuşma hızları, içerikler ve akustik koşullar nedeniyle konuşma sinyalinin oldukça değişken olmasıdır. Bir Konuşma tanıma sisteminin öznitelik çıkarımı bileşeni, sistemin genel performansında çok önemli bir rol oynar (Alim ve Rashid, 2018, s.1).

Üç önemli öznitelik çıkarım tekniği "Doğrusal Tahmin Kodlaması (LPC)", "Mel-Frekans Kepstral Katsayıları (MFCC)", "Algısal Doğrusal Tahmin – Bağıl Spektrum (PLP-RASTA)", literatürde öne çıkmaktadır. Bu teknikler öznitelikleri çıkarırken farklı yaklaşımlar kullanırlar. Örneğin MFCC ve PLP-RASTA insan işitsel algı sistemini baz alırken, farklı hesaplamalarla bu sistemi modellemektedirler. LPC ise geçmiş konuşma örneklerinin doğrusal kombinasyonunu kullanarak sonraki konuşma örneğini yaklaşık olarak tahmin etmeye dayanır (Kurzekar ve ark., 2014, s. 18006).

Bu tekniklerin dışında kullanılan bazı öznitelik çıkarım teknikleri; "Doğrusal Tahmin Kepstral Katsayıları (LPCC)", "Çizgi Spektral Frekansları (LSF)", "Ayrık Wavelet Dönüşümü (DTW)", "Dalgacık Paket Ayrıştırma (WPD)" olarak gösterilebilirler. Bu tekniklerin kullanımı çalışma kapsamında seçilen teknikler kadar yaygın olmasa da bazı özel uygulanma alanlarına sahiptirler (Alim ve Rashid, 2018, s.1, Ghule ve Deshmukh, 2015).

Öznitelik çıkarımıyla, sinyali temsil eden sinyalden daha küçük boyutta öznitelik vektörleri üretilir. Öznitelik vektörleri, konuşma tanıma sistemlerinin diğer aşamaları olan akustik ve dil modellerinin, konuşma sinyalini temsil eden, girdilerini oluşturmaktadır. Öznitelik çıkarımının ardından akustik ve dil modelleme için iki temel yaklaşım olan istatistiksel ve yapay zekâ modelleri ele alınacaktır.

1.3. İstatistiksel ve Yapay Zekâ Modelleri

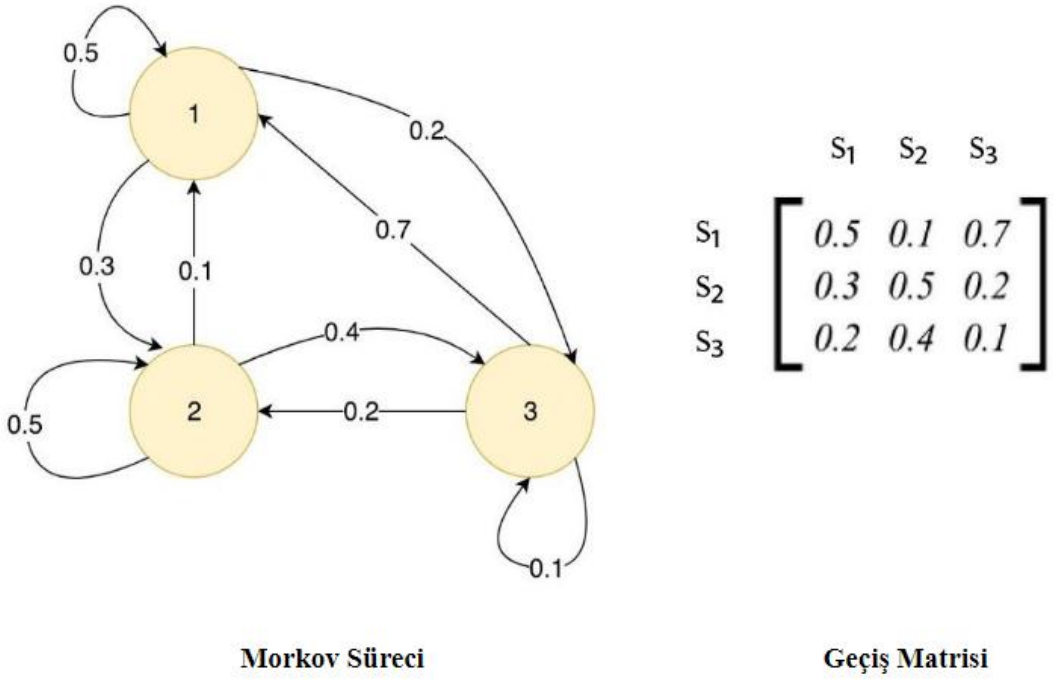
Konuşma tanıma sistemlerinde 2000'li yıllara kadar akustik ve dil modeli için istatistiksel modeller kullanılıyordu. 2000'li yıllardan sonra ise yapay zekâ uygulamaları konuşma tanıma sistemlerinde istatistiksel modellerin yerini almaya başladı. Günümüzde kullanılan son teknoloji konuşma tanıma sistemlerinde tamamen yapay zekâ modelleri veya hem yapay zekâyı hem de istatistiksel modelleri bir arada kullanan hibrit modeller kullanılmaktadır. Bu iki modeli kullanan sistemler, ilgili konuşmayı tanımada karşılaştırılabilir başarım oranlarına sahiptir. Konuşma tanıma sistemlerinde kullanılan istatistiksel modeller ağırlıklı olarak HMM-GMM'dir. Yapay zekâya dayalı modellerde ise genellikle derin öğrenme mimarileri olan yapay sinir ağı modelleri kullanılmaktadır.

1.3.1. HMM-GMM Modeli

HMM için bir Markov zinciri, bir sistemin tüm olası durumlarını ve bir durumdan diğerine geçiş olasılığını içerir. Birinci dereceden bir Markov zinciri, bir sonraki durumun yalnızca mevcut duruma bağlı olduğunu varsayar. Ancak birçok sistemde tüm durumlar gözlemlenebilir değildir, bu durumlara gizli durumlar veya içsel durumlar denir. Bunlar girdiler için gizli faktörler olarak değerlendirebilir. Örneğin, bir kişinin mutlu mu yoksa üzgün mü olduğunu bilmek kolay olmayabilir. Bu durumda kişinin gizli durumu m (mutlu) ve (ü) üzgün olur. Ancak gözlemlediklerimizden bazı ipuçları alabiliriz. Örneğin, mutlu olduğumda film izleme şansım 0,2, üzgün olduğumda ise bu şans 0,4'e çıkıyor. Buradan hareketle, bir iç durumdaki gözlenebilir bir veriyi gözlemleme olasılığına emisyon olasılığı denir. Bir iç durumdan diğerine geçiş olasılığına ise geçiş olasılığı denir. HMM modeli geçiş ve emisyon olasılığından oluşur (Hui, 2019; Juang ve Rabiner, 1991). Şekil 8'de üç durum ve bu üç durum için geçiş olasılıklarını içeren geçiş matrisi verilmiştir.

Şekil 8

Üç Adet Durum ve Bu Durumlar Arasındaki Geçiş Olasılıkları (Hui, 2019)



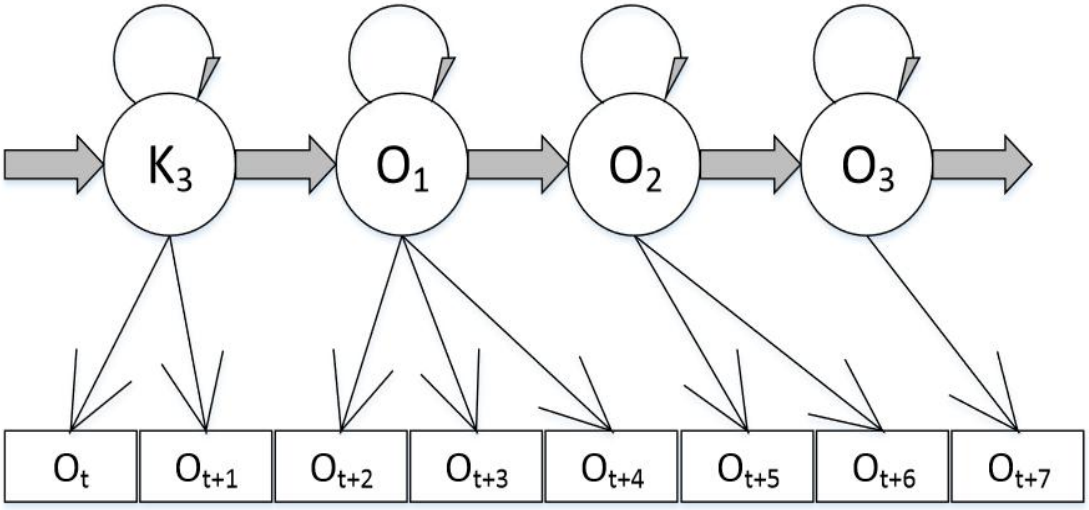
Morkov Süreci **Geçiş Matrisi**

Yukarıda verilen tanımı, konuşma tanıma sistemleri için düşündüğümüzde gözlemlenebilir olan, her bir ses bloğunun öznitelik vektörleridir, emisyon olasılığı öznitelik vektörünün hangi foneme karşılık gelme olasılığına, geçiş olasılığı ise fonemlerin ya da kelimelerin sıralanması olasılığına karşılık gelir. HMM modeli geçiş ve emisyon olasılığından oluşur. HMM emisyon olasılığı için, yani öznitelik vektörleri ile fonem eşleştirme olasılığı için GMM'yi kullanır (Hui, 2019; Juang ve Rabiner, 1991).

Şekil 9'de kısmi bir durum dizisi (daireler; K_3 durumu, *K* foneminin son durumudur ve O_1, O_2 O_3 O fonemine ait durumlar) ve kısmi bir gözlemler dizisi (dikdörtgenler; O_t t anındaki sinyal bloğu için çıkarılmış öznitelik vektörü olarak düşünülebilir).

Şekil 9

Kısmi Bir Durum ve Gözlemler Dizisi (Zelasko, 2016)



Konuşma tanıma sistemlerinde her bir sinyal bloğu için çıkarılmış öznitelik vektörünün, eğer bir foneme veya bir fonemin alofonuna karşılık geliyorsa, bu fonemi belirlemek için GMM kullanılır. GMM hangi fonemin en yüksek olasılıkla verilen öznitelik vektörüne karşılık geldiğini bulur. GMM ile bulunan fonem zamandan ve bağlamdan bağımsızdır. HMM zamana ve bağlama göre GMM ile bulunan fonemi (eğer bulunmuşsa) değiştirebilir.

Özet olarak HMM-GMM modelinde dil ve akustik model için, dile ait geniş veri setlerinden elde edilen olasılıksal parametreler, konuşma sinyalinden konuşma metnini elde etmek için kullanılır. YSA modellerinde ise aynı veri seti olasılıksal parametreleri elde etmek yerine, sinir ağını eğitmek için kullanılır.

1.3.2. Yapay Sinir Ağı Modelleri

YSA (Yapay Sinir Ağı) modellerinin konuşma tanıma sistemlerinde kullanımı 1990'lara dayanmaktadır (Lippmann, 1990; Patil, 1998). YSA modellerinin, akustik ve dil modeli için, dile özgü bir veri seti tarafından eğitilmesi gerekir. Bu eğitim için yine konuşma sinyalinin tamamı yerine, sinyali belli küçüklükte bloklara ayırdıktan sonra, bu bloklar için çıkarılan öznitelik vektörleri kullanılır. Ayrıca tahminleme aşamasında yine sinyalden çıkarılan öznitelik vektörleri YSA'ya verilerek bu vektörlerin hangi ifadeye karşılık geldiğinin eğitilmiş YSA tarafından tahmin edilmesi beklenir.

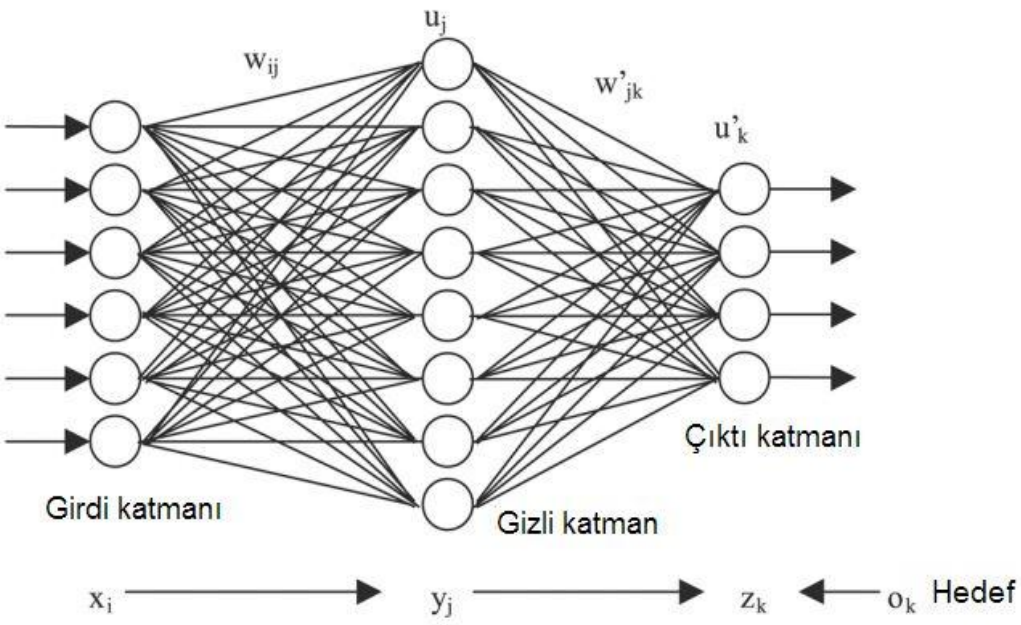
Yapay sinir ağları insan beynini örnekleyerek geliştirilmiş bir teknolojidir. Düşünme, öğrenme, hatırlama gibi insan davranışları beyindeki sinir hücreleriyle gerçekleştirilmektedir. İnsan beyninde yaklaşık 10^{11} adet sinir hücresi olduğu düşünülmektedir. Bu sinir hücreler, neredeyse sonsuz sayıda sinaptik bağlarla birbirlerine bağlanmaktadırlar. İnsan beyninin bu kapasitesine bakarak günümüz teknolojisinin insan beyninin gerisinde olduğunu söyleyebiliriz (Ergezer ve ark., 2003, s. 14).

YSA büyük verileri işleme gücünü, öğrenebilme ve genelleme yapabilme yeteneği ve paralel dağılmış yapısından alır. YSA genelleme yeteneği sayesinde eğitim aşamasında verilmeyen girdiler için de uygun sonuçları üretebilir. YSA hücrelerin birleşiminden oluşmaktadır ve bu hücrelerin doğrusal olmaması YSA'yı doğrusal olmayan karmaşık problemlerin çözümünde önemli bir araç yapmıştır. YSA'da öğrenme aşamasında hücreler arası bağlantıların ağırlıkları belirlenir. Şekil 10'da w_{ij}, w_{jk} hücreler arası bağlantıların ağırlıklarını, u_j, u_k ise hücreleri temsil etmektedir. YSA ilgili problemlerdeki değişikliklere göre yeniden uyarlanabilir. Yani problem her değiştiğinde yeni girdilerle tekrar eğitilebilir ya da değişim gerçek zamanlı ise eğitime sürekli olarak devam edilerek hücreler arası bağlantıların ağırlıkları gerçek zamanlı güncellenir. (Ergezer ve ark., 2003, s. 14; Yalçın, 2008, s. 257, 258).

YSA çok sayıda farklı ağırlıklarla birbirine bağlanan hücrelerden oluştuğu için, şekil 10'da da görülebileceği üzere, paralel bir yapıya sahiptir. YSA'da bilgi bu yapı üzerinde dağılmış durumdadır. Eğitilmiş bir YSA'da bazı bağlantıların ve hücrelerin etkisiz olması sistemin ürettiği sonuçları önemli ölçüde etkilemez.

Şekil 10

Yapay Sinir Ağları Genel Yapısı (Türk ve Kiani, 2019, s. 34)



YSA'nın genelleme yeteneği onu konuşma tanıma sistemleri için çok değerli kılmaktadır çünkü konuşma tanıma sistemlerinde eğitim verisi ne kadar büyük olursa olsun, bir dil için tüm olası ifadeleri içermesi imkânsızdır. Örneğin YSA eğitim sırasında hiç karşılaşmadığı ya da düzün ifade edilmeyen bir kelimeyi tanıyabilir. YSA öğrenme için ilgili problemden aldığı eğitim verisiyle eğitilir, bu konuşma tanıma sistemleri için öznitelik vektörleridir.

YSA'nın kullanım alanları konuşma tanıma sistemleriyle sınırlı değildir. Görüntü işleme, yüz tanıma, finansal tahminler (makroekonomik tahminler, döviz kuru tahminleri), tıp (kanserli hücre analizi), savunma sanayi (hedef izleme, uçakların uçuş yörüngelerini belirleme) gibi pek çok alanda kullanılmaktadırlar (Ağyar, 2015, s. 22, para. 5; Ergezer ve ark., 2003, s. 15, para. 2).

İstatistiksel ve yapay zekâ modelleri konuşma tanıma sistemlerinde akustik ve dil modellemenin temel yapılarını oluştururlar. Akustik ve dil modellemede, sistem önceden kaydedilmiş seslerle ve metinlerle eğitilerek dile özgü akustiğin ve dilbilgisinin öğrenilmesi sağlanır.

1.4. Akustik Modelleme

Bir konuşma işleminin akustik modellenmesi, tipik olarak konuşma dalga formundan hesaplanan öznitelik vektör dizileri için istatistiksel bilgilerin oluşturulması işlemidir. Akustik modellemede ilk olarak belirli zaman sinyali bloğundaki foneme ait sonsal olasılık hesaplanmaktadır. (Oyucu ve ark., 2019, s. 844).

Bu aşamada amaç farklı seslerin spektrumlarındaki farkların matematiksel olarak modellenmesidir. Örneğin /a/ sesini /e/ sesinden ayırt edebilmemiz için iki sesin birçok kişiden alınan örneklerinin frekans incelemesi yapılır ve her bir ses için ortalama spektrum elde edilir. Seslerin birbirinden ayırt edilmesi için sadece akustik olasılıkların hesap edilmesi yeterince yüksek konuşma tanıma performansı vermemektedir (Gürel ve Arslan, 2008, s. 84). Burada dil modellemesi devreye girerek akustik modellemeyi bir nevi tamamlayıcı rol üslenmektedir.

Akustik modelleme için geleneksel konuşma tanıma uygulamalarında, konuşulan ifadenin zamansal ve bağlamsal değişkenliğini ele almak için Gizli Markov Modeli (HMM) kullanılır ve her bir HMM durumunun konuşma sinyali bloğuna ne kadar uyduğuna karar vermek için Gauss Karma Modeli (GMM) kullanılır. Ancak son zamanlarda, HMM-derin sinir ağı (DNN) modeli ve derin öğrenme kullanan uçtan uca model, HMM-GMM'nin ötesinde bir performans elde etti. Her ikisi de derin öğrenme tekniklerini kullanan bu iki model, karşılaştırılabilir performanslara sahiptir (Wang ve ark., 2019, s. 1).

Hinton ve ark. (2012) birçok gizli katmana sahip ve yeni metotlarla eğitilmiş derin sinir ağlarının GMM'den daha iyi sonuçlar verdiğini çeşitli konuşma tanıma uygulaması testlerinde göstermiştir. Sak ve ark., (2015) derin Uzun Kısa Vadeli Bellek (LSTM) tekrarlayan sinir ağlarının (RNN), konuşma tanıma için akustik modeller olarak, ileri beslemeli derin sinir ağlarından (DNN) daha iyi performans verdiğini gösterdi ve LTSM RNN tekniğinin performansını daha da ileriye taşıdı.

1.5. Dil Modelleme

Dil modeli hangi kelimenin, önceden tanınan kelimeleri takip edebileceğini tanımlar ve olası olmayan kelimeleri çıkararak eşleştirme sürecini önemli ölçüde kısıtlamaya yardımcı olur. En yaygın dil modelleri n-gram dil modelleridir. N-gram dil modelleri kelime dizilerinin istatistiklerini içerir. Bir dil modeli, genellikle dikkate alınan kelime dağarcığını, içerdiği kelimelerle sınırlar. Bu, isim tanıma için bir sorundur. Bununla başa çıkmak için bir dil modeli, alt kelimeler ve hatta fonemler gibi daha küçük parçalar içerebilir. Bu durum arama alanı kısıtlaması açısından iyi değildir (Buyuk ve ark., 2005, s. 112, 113; Kutlugün ve Şirin, 2018, s. 3).

Dil modelleme, dildeki kelime dizilerinin ve cümlelerin, yapı ve dizilişlerini modelleyip bunların olasılıklarını belirtir. Burada gerçek bir istatistik elde etmek için büyük metin verileri ile dil modeli eğitilmelidir. Bu veriler çevrimiçi gazetelerden, elektronik kitaplardan ve bu tür sistemler için özellikle oluşturulmuş çevrimiçi veri setlerinden elde edilebilirler (Oyucu ve ark., 2019, s. 845).

Dil modellemede yaygın olarak kullanılan en basit teknik n-gram modelidir. N-gram modelleri, ardışık bir dizinin bir sonraki elemanını (n-1) formatlı Markov zincirinden yararlanarak tahmin etmekte kullanılan bir istatistiksel dil modelidir. N-gram modelleri, tekrarlayan sinir ağlarına (RNN) ve dönüşüme dayalı dil modellerine göre oldukça basit olsa da dil modellemenin temel kavramlarını anlamak için önemli bir temel araçtır (Jurafsky ve Martin, 2021, s. 2, 3).

Dönüşüme dayalı dil modeli mimarilerinin avantajı, tekrarlayan sinir ağlarında (RNN) olduğu gibi sıralı bir işlem olmadığı için eğitim aşamasında hızlı bir yineleme hızına sahip olmasıdır. Bununla birlikte, bir RNN, etkili ortak eğitim ve kod çözme yöntemleri nedeniyle genel eğitim hızı ve kelime hata oranı (WER) açısından uçtan uca otomatik konuşma tanıma görevleri için daha en iyi seçenektir (Nakatani, 2019, s. 1408).

Chen ve ark. (2015) Dil modelinde RNN kullanırken, donanımsal hesaplama maliyeti açısından yük getiren, RNN'nin çıkış katmanında bulunan, softmax fonksiyonundaki normalleştirme işlemini gerektirmeyen, gürültü karşıtı tahmin (NCE) sistemini RNN'de kullandı. Böylece Eğitim ve test aşamalarında önemli ölçüde hız artışı sağladı.

Nakatani (2019) daha hızlı ve daha doğru bir konuşma tanıma sistemi için dil modelinde dönüşüm tabanlı mimarilere, öğrenme ve çözümleme aşamaları aşamalarında iyileşme sağlamak adına, bağlantıcı zamansal sınıflandırma (CTC) entegre etti. Oluşturdukları modelle çeşitli veri setlerinde kelime hata oranlarında yaklaşık yüzde %5 iyileştirme sağladı.

1.6. Konuşma Tanıma Kullanım Alanları

Günümüzde kullandığımız pek çok teknoloji konuşma tanıma özelliğini barındırmaktadır. Yapay zekâ, makine öğrenmesi gibi teknolojilerle doğrudan bağlantılı olan konuşma tanıma teknolojisinin, bu teknolojilerin gelişmesi ile kullanım alanının da artacağını söylemek mümkündür. Konuşma tanıma teknolojisinin günümüzdeki en popüler kullanım alanının sanal asistan uygulamaları olduğunu söyleyebiliriz, dünyanın önde gelen teknoloji firmalarının ürünleri olan Alexa (Amazon), Cortana (Microsoft), Google Asistan (Google), Siri (Apple) bu alanda en çok bilinen ve kullanılan uygulamalardır.

Ancak ses tanıma teknolojisinin sanal asistanlardan daha geniş kullanım alanları vardır. Bu bağlamda, pazarlama, müşteri hizmetleri, sağlık ve daha pek çok alanda ses tanıma teknolojisinden faydalanıldığını aşağıdaki örneklerle gösterebiliriz.

Sesli arama ses tanıma teknolojilerinin en yaygın kullanımlarındandır. 2018'de PWC tarafından Amerika'da yapılan ankete göre, Amerikan vatandaşlarının %72'si sesli asistanı kullanmış ve özellikle, katılımcıların %71'i internette arama yaparken sesli aramayı tercih ettiklerini söylüyorlar (PwC, 2018).

Bir diğer konuşma tanıma kullanım alanı konuşmadan metne dönüşümdür. Sesten metne dönüştürme için aşağıdaki kullanım alanları için aşağıdaki örnekler verilip bu örnekler artırılabilir. Ayrıca RTÜK bünyesinde kullanılan SKAAS sisteminde de belli yayın bölümlerinde, konuşulanların otomatik olarak metne dönüştürülmesi, sistemi kullananlar uzmanlar için kolaylık sağlayacak bir özelliktir.

- Eposta yazma
- Google dokümanlarda belge oluşturma
- Ses tanıma ile otomatik altyazı ekleme (Youtube)
- Otomatik çeviri
- Mesaj Gönderme

Ayrıca akıllı ev aletlerinde de konuşma tanıma teknolojisinden faydalanılmaktadır. Akıllı ev cihazları, ışıkları açmak, su kaynatmak, termostatı ayarlamak vb. gibi ev işlerini gerçekleştirmek için konuşma tanıma teknolojisinden yararlanır. İlerleyen yıllarda akıllı ev aletlerinin kullanım alanının artacağı ve teknolojisinin gelişeceği düşünüldüğünde, sesli komutlar da akıllı ev aletlerinin gelişen teknolojisinin önemli bir parçası olacağını söyleyebiliriz.

Ses biyometrisi, bir kişinin kimliğini doğrulamak için kişinin sesini kullanır, telefonunuzun parmak izlerinizle açmanıza izin vermesi gibi. Kullanım durumu açısından, ses biyometrisi hem çalışanlara hem de müşterilere fayda sağlayabilir. Örneğin, oturum açma işlemleri için, kullanıcılardan parola girmek yerine kendi benzersiz sesleriyle adlarını söylemeleri istenebilir.

Günümüzde modern müşteri hizmetleri servislerinin çoğu ses tanıma teknolojilerinden yararlanır. Ses tanıma teknolojileri etkileşimli sesli yanıt sistemlerinde kullanılarak, müşterinin doğru temsilcilere ulaşmasını veya sorunlarını sesli komutlarla çözmesini sağlar. Ayrıca müşteriler ve temsilciler arasındaki binlerce telefon görüşmesinin dökümü, yaygın arama kalıplarının ve sorunlarının belirlenmesine yardımcı olur.

Konuşma tanıma sistemleri kör veya az gören öğrencilerin öğrenmede yaşadığı güçlükleri azaltabilecek potansiyele sahiptir. Ayrıca kullanıcının dil telaffuzunu değerlendirmek için konuşma tanımayı kullanan “Duolingo” gibi dil öğrenme mobil uygulamaları da vardır. Telaffuz değerlendirmesi, pratik bir bilgisayar destekli dil öğrenme uygulamasıdır.

Dikte yazılımı gibi araçlar, insanların daha kısa sürede daha fazla kelime yazmasını sağlayabilir. Makaleler, konuşmalar, kitaplar, notlar veya e-postalar yazan içerik oluşturucular bu yazılımları kullanarak daha hızlı içerik oluşturabilirler. Bu araçlar hala yüzde yüz doğru sonuçlar vermese de ilk taslaklar için faydalıdır.

Tıbbi transkripsiyon yazılımı sayesinde, hasta muayeneleri sırasında, doktorların hastaların belirtilerini not alma konusunda endişelenmelerine gerek kalmaz. Bu tür yazılımı kullanmak doktorların zamanını daha verimli kullanmasını sağlar. Konuşma tanıma teknolojisi, depresyonun alt tonlarının varlığını veya eksikliğini belirlemek için bir hastanın sesini dinleyebilir. Örneğin “Sonde Health” isimli şirketin geliştirdiği, diğer değişkenlerin yanı sıra seslerinin tonuna, enerjisine, dalgalanmalarına ve ritmine dayalı olarak kullanıcılara bir "zihinsel uygunluk" puanı veren temel algoritmalarla mobil uygulaması vardır.

Hukuk alanında “DoNotPay” isimli firmanın geliştirdiği robot avukat ses tanıma teknolojilerinden faydalanarak hukuk danışmanlık hizmeti vermektedir. Ayrıca mahkeme raporlaması (Gerçek Zamanlı Konuşma Yazma), ifadelerde ve sorgulamalarda otomatik transkriptler gibi alanlarda yine konuşma tanıma teknolojisinden yararlanılmaktadır.

Konuşma tanıma sistemleri çoğu modern araç için standart bir özellik haline geldi. Bu teknolojinin en büyük faydası, sürücünün sürüş sırasında önündeki yol dışında başka bir yere bakma ihtiyacını ortadan kaldırarak, sesini kullanarak çoklu görev yapabilmelerine olanak sağlamasıdır. Bu sistemler sayesinde sürücüler, telefon görüşmeleri başlatmak, radyo istasyonlarını seçmek veya müzik çalmak için basit sesli komutları kullanabilir (Dilmegani, 2021).

BÖLÜM 2

ÖZNİTELİK ÇIKARIM TEKNİKLERİ

Bu bölümde ilk olarak konuşma sinyalinde öznitelik çıkarımı yapılmadan önce yapılan önişlem aşamasından ve bu aşamada farklı araştırmacıların kullandığı farklı tekniklerden bahsedilmiştir. Ardından bu çalışmanın konusu olarak seçilen LPC, MFCC, PLP-RASTA öznitelik çıkarım teknikleri açıklanıp hesaplama metotları verilmiştir. Son olarak literatürde bu tekniklere yönelik yapılan karşılaştırmalı çalışmalar verilmiştir.

Öznitelik çıkarımı otomatik konuşma tanıma sistemlerinin en önemli bölümlerinden biridir. Literatürde özellik çıkarımı olarak da geçmektedir. Öznitelik çıkarım tekniği genellikle uygulandığı konuşma sinyalinden o sinyali temsil eden katsayılardan oluşan çok boyutlu bir vektör üretir (Ravikumar ve ark. 2009). Öznitelik vektörleri, konuşma tanıma sistemlerinin yanında, konuşmacı tanımlama, konuşma sinyalini sıkıştırma gibi alanlarda da kullanılmaktadır (Narang ve Gupta, 2015, s. 109, para. 2).

Konuşma sinyalinde öznitelik çıkarımı için iki baskın akustik ölçümü vardır. Biri, konuşma sesini üreten insan ses yolunun rezonans yapısını yakından eşleştirmek için geliştirilen yaklaşımdır örneğin LPC. Diğeri, insan işitsel algı sistemini modelleyen yöntemdir örneğin MFCC, PLP-RASTA (Wang ve ark., 2002, s. 111).

Konuşma sinyalinden öznitelik çıkarımı işlemi yapılmadan önce birkaç adımdan oluşan ön işlemler dizisi vardır. Ön işlem aşamasında hedef konuşma sinyalini netleştirmek, sinyali spektral olarak düzleştirmek, sinyalin içinde konuşmayı temsil etmeyen yerleri elimine etmektir. Günümüze kadar araştırmacılar konuşma sinyalinin öznitelik çıkarımından önce bu işleme daha hazır hale getirmek için farklı algoritmalar, filtreler kullanmışlardır. Bunlardan bazıları burada ele alınmıştır.

2.1. Önişlem

Gürültü, yüksek frekanslar üzerinde düşük frekanslara göre daha büyük bir etkiye sahiptir. Bu sebeple, gürültünün sinyale olan etkisini azaltmak için, önişlem aşamasında sinyaldeki yüksek frekanslar filtreler kullanılarak yapay olarak arttırılır. FIR (Sonlu Darbe Tepkisi) bu aşamada yaygın olarak kullanılmaktadır. Ayrıca çeşitli frekans aralıklarını sinyalden eleyen filtreler sıklıkla kullanılmaktadır.

Ardından sinyal eşit süreli örtüşen bloklara ayrılır. Genellikle blok uzunluğu 20 ile 40 ms arasındadır. Örtüşme uzunluğu ise genellikle 10 ms'dir (Gupta ve Gupta, 2016, s. 499). Sonrasından bloklara ayrılan sinyale pencereleme işlemi uygulanır. Her sinyal bloğunun başı ve sonundaki kopukluğu azaltmak için, pencereleme işlemi olarak Hamming ve dikdörtgen pencereleme teknikleri oldukça sık kullanılan pencereleme yöntemlerindendir. Pencereleme işlemi harmoniklerin keskinliğini arttırır, sinyal bloklarının başlangıç ve bitişini sıfıra yakınsatarak sinyalin süreksizliğini ortadan kaldırır. Ayrıca sinyalin bloklara ayrılması aşamasındaki örtüşmenin oluşturduğu spektral bozulmayı da azaltır (Alim ve Rashid, 2018, s.5, para. 1, 2).

2.1.1. Filtreleme

Bu bölümde ön işlem aşamasında kullanılan filtrelerin ne olduğuna ve neden kullanıldığına, hangi çeşit filtrelerin kullanıldığı açıklanacaktır. Bazı araştırmacıların hangi filtreleme yöntemlerini seçtiğine değinilecektir.

Sesi herhangi bir şekilde değiştirdiği sürece, sinyalinin içinden geçtiği herhangi bir ortam, biçimi ne olursa olsun bir filtre olarak kabul edilebilir. Örneğin, hoparlör kablosu bir filtre olarak kabul edilmez, ancak hoparlör kabul edilir. İnsan ses yolu şekil değiştirerek farklı sesler çıkardığı için bir filtre olarak kabul edilir. Diğer taraftan konuşma tanıma sistemlerinde kullandığımız filtre, konuşma sinyalini analog değil dijital olarak ele aldığımız için dijital bir filtredir. Yani burada kullandığımız manasıyla filtre, bir sayı dizisini (giriş sinyali) alan ve yeni bir sayı dizisi (filtrelenmiş sayı dizisi) üreten bir hesaplamadır (Smith, 2007).

Konuşma sinyalinde önişlem aşamasında, sinyali gürültüye karşı dayanıklı hale getirmek için, çeşitli filtreleme fonksiyonları kullanılabilir ancak çoğunlukla FIR filtresi kullanılmaktadır. FIR filtresi ile düşük frekanslara göre yüksek frekansların boyutunu arttırılmış olur böylece konuşma sinyalinde konuşmayı temsil etmeyen yerleri elimine edilmiş olur (Alim ve Rashid, 2018, s. 5, para. 1; Joshi ve Cheeran, 2014, s. 1). FIR filtresi, *z* sinyal değerlerini, *a* sabit bir katsayı olmak üzere Denklem 7'de gibi tanımlanır. HTK'ya göre *a* katsayısı için varsayılan değer 0,97'dir (Young ve ark., 2002). Bazı araştırmacılar FIR filtresi dışında farklı filtreler kullanmışlardır.

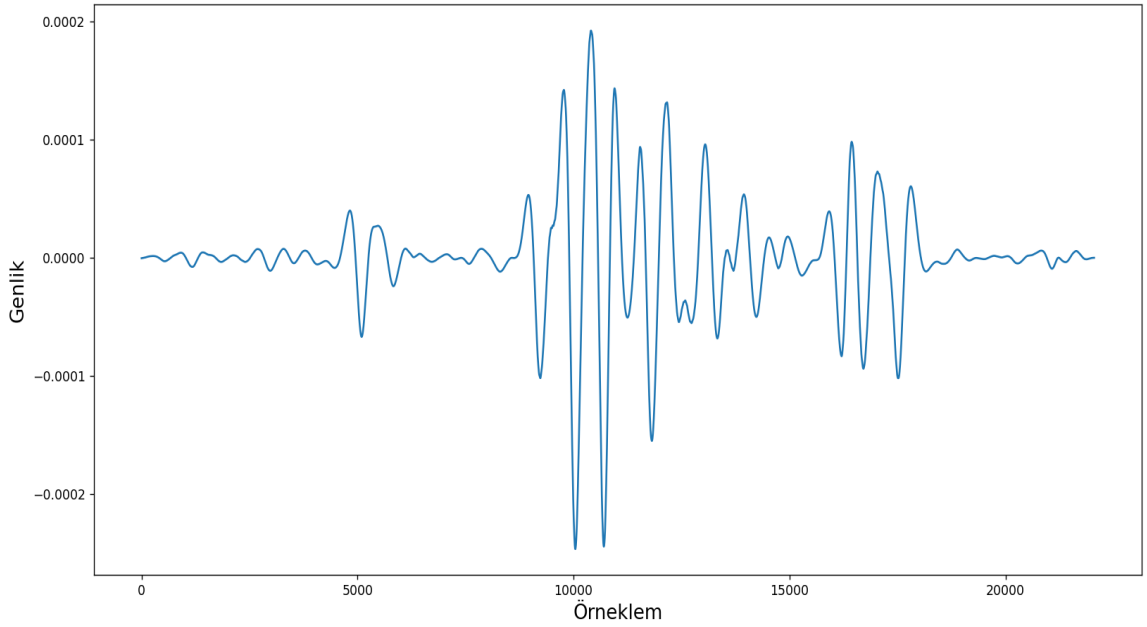
$$H(z) = 1 - az^{-1}\,, 0.9 \leq a \leq 1.0$$ **Denklem 7** (Joshi ve Cheeran, 2014, s. 1)

Barua ve ark. (2014) ve Ittichaichareon ve ark. (2012) 16 üçgen bant geçiren filtresini kullandı. Özseven (2019) belli sadece frekans aralıklarını geçiren, yüksek geçiren, alçak geçiren filtreleri kullandı. Magsi ve ark. (2018) alçak geçiren filtresi, hareketli ortalama filtresi ve FIR filtresini karşılartırdığı çalışmada, gürültüyü azaltma bakımından FIR filtresinin diğerlerinden daha iyi olduğunu belirtti. Mohan (2014) ön işlem aşamasında filtreleme işleminin yanında ses kapısı denilen bir metodu kullanarak belli bir eşik değeri altındaki genlikleri sinyalden çıkardı ve sonraki bölümlerde hesaplama süresini azaltmak için sinyali sıfırdan başlayacak şekilde hizaladı.

"Radyo ve Televizyon Üst Kurulu" ifadesinin orijinal zaman alanında sinyali Şekil 3'te verilmişti. Şekil 11'de ise aynı sinyalin bant geçiren filtresinden geçirildikten sonraki durumu, örneklem-genlik gösterilmiştir. Bu iki şekli kıyaslarsak gösterimlerin ana hatlarıyla aynı olduğunu, orijinal sinyalin bant geçiren filtreden geçtikten sonra spektral olarak daha düzleştiğini görebiliriz.

Şekil 11

"Radyo ve Televizyon Üst Kurulu" İfadesine Ait Filtrelenmiş Sinyal



2.1.2. Sinyalin Bloklara Ayrılması

Konuşmanın sinyalinin sürekliliğe sahip olması istatistiksel özelliklerinin zamanla değiştiği anlamına gelir. Bu sebeple konuşma sinyalini düzgün bir şekilde analiz etmek için, sinyalin bu noktalarda durağan olacağını varsayabileceğimiz kadar küçük olacak şekilde bloklara ayırmamız gerekir. Böylece analizi bu kısa ve örtüşen sıralı bloklar üzerinde yapabiliriz. Konuşma tanımada, genellikle 0,01 sn (10 ms) örtüşme ile 0,025 sn (25 ms) uzunluğunda bloklar kullanırız. Blok uzunlukları 5 ms aşağı veya yukarı olarak değişebilir. Barua ve ark. (2014) ve Han ve ark. (2006) 20 ms'lik blok uzunluğu ve 10 ms'lik örtüşme kullandı.

1970'lerde DARPA'nın LPC-10 standardına göre konuşma sinyali bir saniyede 8.000 örneklem ve her örneklem 8 bit boyutundadır. Ardından her saniyedeki 8.000 örneklem 180 ayrı bloğa ayrılır. Yani her blok 22,5 milisaniyelik sinyal içerir (Bradbury, 2000, s. 10). Günümüzde saniye başına örneklem sayısı artmıştır ama blok uzunluğu aşağı yukarı LPC-10 standardına yakındır.

Örtüşen blokların kullanılma sebebi bir sonraki adımdaki pencereleme işleminde blok sinyallerin başlangıç ve bitişlerinin zayıflamasıdır. Pencereleme nedeniyle sinyal bloğunun başında ve sonunda veri kaybı yaşanabileceğinden bloklara ayırma işleminde bir miktar örtüşme payı bırakılır. (Barua ve ark., 2014, s. 3, para. 3; Dhingra, 2013, s. 4088, 4089).

2.1.3. Pencereleme

Sinyalin bloklara ayrılmasından sonra, her sinyal bloğuna pencereleme işlemi uygulanır. Pencereleme işlemiyle sinyal bloğunun harmonik keskinliği arttırılır, bloğun başı sonu sıfıra indirgenir. Ayrıca örtüşmenin oluşturduğu spektral bozulmayı da azaltır (Alim ve Rashid, 2018, s.5, para. 2).

Pencereleme işlemi için Hamming, Hann ya da Dikdörtgen pencereleme teknikleri kullanılır. Genellik pencereleme işlemi için Hamming Pencerelemesi, iyi frekans çözünürlüğü sağladığı için tercih edilir (Barua ve ark. 2014). Kunigami (2021) LPC yöntemi için Hann pencerelemesi kullanmıştır. Han ve ark. (2006) Hamming penceresini kullandı. Hamming penceresi Denklem 8'de verilmiştir. Denklemde *N* bir bloktaki örneklem sayısını, *n* ise 1'den *N*'ye kadar olan sayıları temsil etmektedir.

$$Ham(N) = 0{,}54 - 0{,}46\cos(2\pi \frac{n-1}{N-1}) \qquad \textbf{Denklem 8} \text{ (Han ve ark., 2006)}$$

2.2. Doğrusal Tahmin Kodlaması (LPC)

LPC konuşma sinyalini modellerken insanda konuşma üretimini esas alır. Bu üretimi modellerken de kaynak-filtre modelini kullanır. Bu sebeple, bu bölümde ilk olarak insanda konuşma üretimi ve kaynak-filtre modeli ele alınmıştır. Sonrasında LPC'nin bu modele göre

matematiksel olarak hesaplanması verilmiştir. Son olarak LPC'nin yaygın olarak kullanıldığı konuşma sinyali sıkıştırma konusu ele alınmıştır.

LPC konuşma tanıma uygulamalarında yaygın olarak kullanılan bir öznitelik çıkarım tekniğidir. Bu çalışma kapsamından geliştirilen uygulamada LPC vektörleri CNN ağını eğitmek için kullanılmıştır. Ayrıca LPC salt olarak ses sinyali sıkıştırmasında da kullanılmaktadır ve oldukça yüksek sıkıştırma oranına sahiptir. LPC ses sinyalini sıkıştırmak için kullanıldığında, sıkıştırılan sesin yeniden oluşturulması için, LPC vektörlerinden fazlasına ihtiyaç vardır (bkz. 2.2.3) (Kim, s. 1)

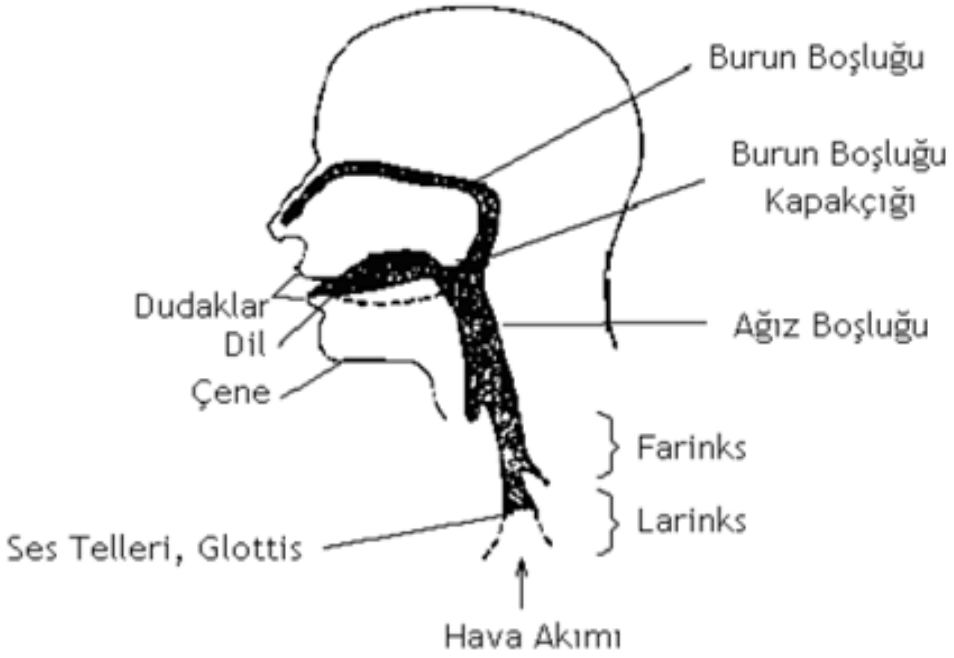
LPC'nin gürültü içeren ses dosyalarında doğruluk oranı yüksektir ve hesaplama maliyeti açısından verimlidir. LPC insan ses yolundan ilham alınarak tasarlanmıştır. Konuşmacıların kısa konuştuğu ve aynı kelimeleri söylediği durumlarda daha iyi performans gösterir. Bu teknikte ses yolları, rezonans frekanslarına göre ayrılır ve buna formant denir. Bu tekniği kullanarak, bir ses sinyalindeki formantların konumları, ses sinyal blokları lineer tahmin katsayıları hesaplanarak tahmin edilebilir. LPC, hesaplama hızı, doğruluğu ve sesin akustik bilgilerinin elde edilmesi açısından yüksek performansa sahiptir (Ayrancı ve ark., 2021, s. 19, para. 2).

2.2.1. İnsanda Konuşma Üretimi ve Kaynak Filtre Modeli

Konuşulan dilden bağımsız olarak hemen hemen her insan aynı anatomiyi kullanarak ses üretirler. İnsanların üretebileceği sesler bu anatomiye dayanan fizik yasalarıyla sınırlıdır. İnsanlarda konuşma üretme süreci, havanın akciğerlerden çıkıp ses yolundan dışarı itilmesi ve konuşma oluşturmak için ağızdan dışarı çıkması olarak özetlenebilir. Bu tür bir tanımlamada akciğerler sesin kaynağı, ses yolu ise konuşmayı oluşturan çeşitli ses türlerini üreten bir filtre olarak düşünülebilir. Bu tanımlamaya sesin üretimi için kaynak-filtre modeli denir (Bradbury, 2000, s. 6, 7).

Şekil 12

İnsanda Ses Yolu (Bilmes ve ark., 2001, s. 60)

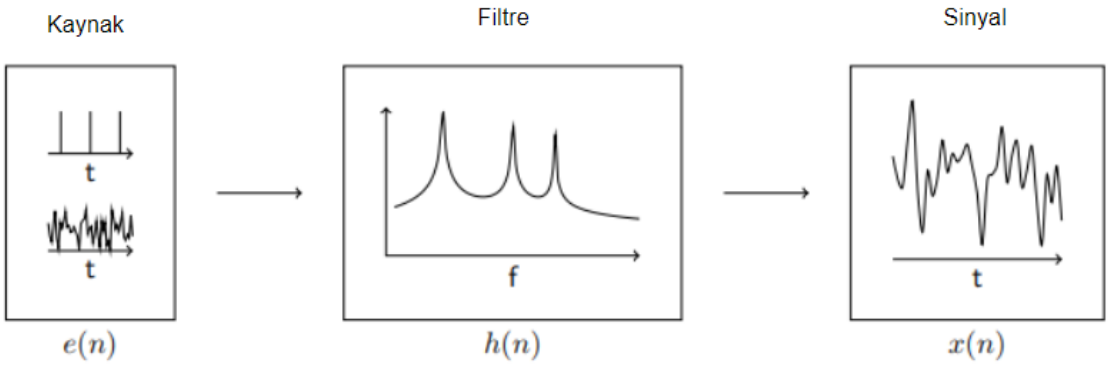


2.2.2. LPC Matematiksel Modeli

LPC insanda konuşma üretimini modellerken yukarıda da belirtilen kaynak filtre modeline göre matematiksel olarak modeller. LPC *e(n)* kaynak (akciğerlerden gelen havanın ses tellerinde ürettiği titreşim), *h(n)* filtre (ses yolu; ses tellerinden çıkan titreşimin ağızdan çıkana kadar geçtiği bölüm), *x(n)* ise ortaya çıkan sinyal olmak üzere (bkz. Şekil 13) Denklem 9'daki gibi tanımlar (Kim, s. 2).

Şekil 13

Kaynak Filtre Modeli (Kim, s. 2)



$$x(n) = h(n) * e(n)$$ **Denklem 9** (Kim, s. 2)

LPC filtrenin (*h(n)*) p dereceden tüm kutuplu bir filtre olduğunu varsayar. Yani *n* anındaki bir *x(n)* sinyali için filtre kendisinden önceki *p* adet sinyal ile Denklem 9'deki gibi ilişkili olduğunu varsayar. Bu insanda ses üretimini düşündüğümüzde fizyolojik açıdan kesin değildir ancak rezonansları modellemek için genişletilebilir bir yöntem sağlar. Bu aynı zamanda *x(n)*'den *h(n)*'i hesaplamak için bir çözüm sunar. Başlangıçta konuşma sinyalleri için geliştirilmiş olsa da spektral olarak düz bir kaynak sinyali ve bir rezonans filtresi varsayımı, doğal olarak oluşan birçok sesin yanı sıra çoğu tonal enstrümandan gelen sinyallerin modellenmesi için de geçerlidir.

$$x_n = \sum_{k=1}^{p} a_k x_{n-k} + e_n$$ **Denklem 10** (Kim, s. 3)

Denklem 10'u elimizdeki *N* adet sinyal örneği için matris formuna çevirmek istersek; $x_i = [x_{n-1+i} \dots x_{n-p+i}]$ ve $a = [a_1, \dots, a_p]^T$ olmak üzere Denklem 11'deki *N* adet eşitliği elde ederiz.

$$x(n) = x_0.a + e(n)$$

$$x(n+1) = x_1.a + e(n+1)$$

. **Denklem 11** (Kim, s. 3)

.

$$x(n+N) = x_N.a + e(n+N)$$

Denklem 11'de ise değişkenleri birleştirerek yeni değişkenler oluşturursak $b = [x(n) \dots x(n+N)]^T \in R_{Nx1}$, $e = [e(n) \dots e(n+N)]^T \in R_{Nx1}$, $A = [x_0 \dots x_N]^T \in R_{Nxp}$ olmak üzere Denklem 12'i elde ederiz. Denklem 12'de ise $\hat{b} \equiv A.a$ olmak üzere Denklem 13'ü elde ederiz.

$e = b - A.a$ **Denklem 12** (Kim, s. 4)

$e = b - \hat{b}$ **Denklem 13** (Kim, s. 4)

Denklem 13 lineer regresyon olarak bilinir. Burada $\hat{b}$ A'nın a ağırlıklarıyla bir b tahminidir ve e ise tahmin hatasıdır. Önceden de belirtildiği gibi kesin bir sonuç yoktur ancak LPC bu durum için üzerinde oldukça çalışılmış olan en küçük kare hataları yöntemini kullanır. Yani $||e(n)||^2$'yi minimize edecek olan a vektörü seçilir. Bu LPC vektörüdür (Kim, s. 4).

2.2.3. Ses Sinyali Sıkıştırmasında Kullanımı

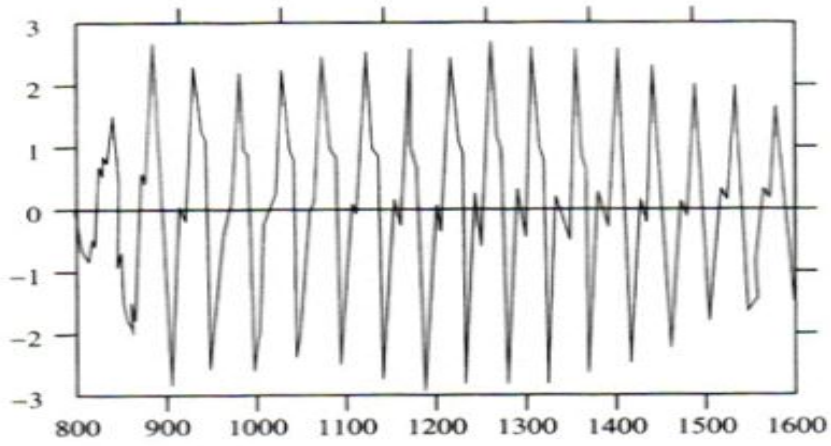
LPC konuşma tanıma sitemlerinin yanı sıra ses sinyalinin sıkıştırılmasında da yaygın olarak kullanılır. LPC ile ses sinyali sıkıştırılması uygulamasını analiz ve sentez olarak iki aşamada düşünebiliriz. Analiz kısmında sinyal LPC ile kodlanır. Bu aşamada LPC

vektörlerinin yanı sıra, sinyal bloğunun sesli veya sessiz olması ve ton periyodu tahmini yapılmaktadır. Daha sonra LPC vektörü ve analiz aşamasında edinilen diğer parametreler kullanılarak sentez aşamasında ses sinyali yeniden üretilmeye çalışılır (Kamil, 2017).

Konuşma sinyalinde sessiz ve sesli harflerin bölgeleri belirgin şekilde farklıdır. Sesli harfler için konuşma sinyalini incelediğimizde sesli bölgelerde yüksek enerjili (geniş genlik), periyodik ve belirgin frekansa sahip dalga formlarıyla karşılaşırız (bkz. Şekil 14). Sessiz harflerin olduğu bölgeler ise az enerjili (düşük genlik) ve yüksek frekanslı olduğundan daha çok karmaşık ve rastgele bir dalga formuna sahiptir (bkz. Şekil 15). LPC kodlama aşamasında her bir bloğun sesli veya sessiz olduğunu belirtmek için 1 bit kullanır, bu bit kod çözme aşamasında kullanılır. Bir bloğu sesli veya sessiz olarak sınıflandırmak için iki aşamalı bir işlem uygulanır. İlk aşamada bloğun dalga formunun enerjisinin (genliğine) farklılığından yararlanılır. Enerji yüksekse sesli, enerji düşükse sessiz olarak belirlenir. Burada enerjinin neye göre yüksek veya düşük olduğu hakkında ön tanımlı bir eşik gerekir. Birinci aşamada tam olarak sınıflandırılamayan bloklar için ikinci aşamada sesli bölgeler yüksek enerjili olurken sessiz bölgelerin yüksek frekansa sahip olmasından yararlanılır. Burada bloğun x eksenini kaç defa geçtiğine bakılır ve bu değer, diğer sesli veya sessiz blokların kaç defa x eksenini kestiğiyle karşılaştırılarak bloğun sesli veya sessiz olduğuna karar verilir. Ayrıca bu sınıflandırmayı etkileyen bir başka faktör komşu bloklardır, genellikle sesli harflerin arasında sessiz harf ve sessiz harflerin arasında sesli harf olması muhtemeldir (Bradbury, 2000, s. 10).

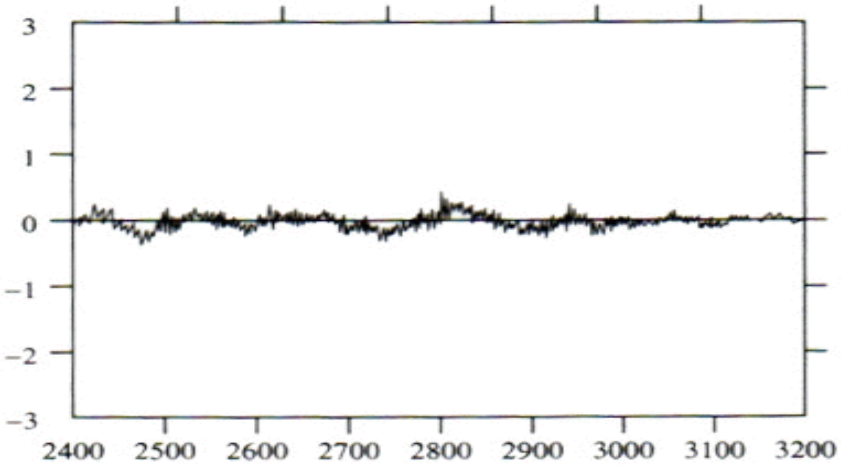
Şekil 14

Sesli Harf Sinyali (Bradbury, 2000, s. 10)



Şekil 15

Sessiz Harf Sinyali (Bradbury, 2000, s. 10)



LPC sıkıştırılan sinyali yeniden üretmek için, bir sinyal bloğunun sesli veya sessiz olmasının yanında, o bloğun ton periyoduna da ihtiyaç duymaktadır. Bir dalga formunun ton periyodu, bir dalga döngüsünün sabit bir pozisyonu tamamen geçmesi için gereken süre olarak tanımlanabilir. Konuşma sinyallerinde ton periyodu, ses üretilirken titreyen ses

tellerinin periyodudur. İnsanda konuşma üretimi bölümünde de belirtildiği gibi ses telleri yalnızca sesli harfleri üretirken titrer bu sebeple ton periyodu sadece sesli bloklar için kod çözme aşamasında LPC filtresi tarafından ihtiyaç duyulur ve kodlama aşamasında hesaplanır. LPC ton periyodu tahmini için AMDF (Ortalama Büyüklük Farkı Fonksiyonu) algoritmasını kullanır. (Bradbury, 2000, s. 10-12).

Kunigami (2021) 44,1 kHz örneklem oranına sahip (saniyede 44100 örneklem) bir ses sinyalinin örnekleme oranını 8 kHz'ye düşürüp ardından 30 ms'lik bloklara ayırdı (Hann pencerelemesini kullanarak.) ve LPC (altı kutuplu filtre kullanarak) tekniğini kullanarak sinyali 17 kat sıkıştırdı.

2.3. Mel-Frekans Kepstral Katsayıları (MFCC)

Bu bölümde ilk olarak MFCC'de geçen kepstral kavramının dayandığı log büyüklük spektrumu açıklanmıştır. Ardından Mel ölçeği ve Mel filtre bankası kullanılarak Mel spektogramlarının elde edilme süreci anlatılmıştır. Son olarak Mel spektogramlarından MFCC öznitelik vektörlerini elde etmek için uygulanması gereken DCT (Ayrık Cosine Dönüşümü) verilmiştir.

MFCC günümüzde en yaygın olarak kullanılan öznitelik çıkarma tekniğidir. MFCC 1 kHz üzerindeki frekansları algılayamayan, insan işitme algılarına dayanmaktadır. MFCC ile elde edilen özellikler kulak salyangozunun (kohlea) kritik frekans bant genişliğine benzerdir. Bu yöntem gürültüye karşı oldukça hassas olmasına rağmen doğruluk oranı oldukça yüksektir. MFCC yöntemini kullandığımızda gürültü arttıkça performans açısından azalma olur (Koçer ve Ahmed, 2019, s. 40, para. 1).

Mel-frekans kepstral katsayıları hesaplamak için öncelikle, ses dalga formuna DFT (Ayrık Foruier Dönüşümü) uygulanarak frekans bazlı sinyal (x ekseni frekans, y ekseni büyüklük) elde edilir. DFT konuşma sinyal işleme bölümünde açıklanmıştır (bkz. 1.1.5.2). Ardından Log-büyüklük spektrumu hesaplanır. Sonrasında log büyüklük spektrumuna Mel ölçeği uygulanıp Mel spektogramı elde edilir. Son olarak Mel spektogramlarına DCT (Ayrık Cosine Dönüşümü) uygulanarak MFCC vektörü hesaplanır (Alim ve Rashid, s. 5, 6).

MFCC'nin genel formülü, *k* mel kepstrum katsayı sayısı, $\widehat{S_k}$ filtre bank çıktısı olmak üzere Denklem 14'te verilmiştir.

$$\widehat{C_n} = \sum_{n=1}^{k}\left(log\widehat{S_k}\right)\cos[n(k-\frac{1}{2})\frac{\pi}{k}] \qquad \textbf{Denklem 14} \text{ (Alim ve Rashid, s. 6)}$$

2.3.1. Log-Büyüklük Spektrumu

Log büyüklük spektrumuna kepstrum denmektedir. MFCC'deki kepstral sözcüğü bu kepstrumdan gelmektedir. Kepstrum 1960'larda MIT tarafından sismik sinyallerin yankılarını araştırırken geliştirildi. 1970'lerde kepstrum konuşma tanıma uygulamaları için sesin özniteliği olarak kullanıldı. 2000'lerde ise kepstrum müzik sinyal işlemede ve MFCC için kullanılmaya başlandı (Oppenheim ve Schafer, 2004). Kepstrum Denklem 15 ile ifade edilmektedir.

$$C\big(x(t)\big) = F^{-1}[\log(F[x(t)])] \qquad \textbf{Denklem 15} \text{ (Oppenheim ve Schafer, 2004)}$$

Denklem 15'te *x(t)* zaman bazlı konuşma sinyalini, *C(x(t))* zaman bazlı konuşma sinyalinin kepstrumunu, *F[x(t)]* zaman bazlı sinyalin Fourier dönüşümünden sonra elde edilen frekans bazlı sinyali (spektrum) ifade etmektedir. Denklemi kabaca özetleyecek olursak zaman bazlı bir konuşma sinyalinin Fourier ile frekans bazlı sinyale dönüştürüp logaritmasını alıp sonra ters Fourier'ini alarak kepstrumu elde ediyoruz. Kepstrum ile spektrum kelimelerinin benzerliği de bu denklemdeki ilişkilerinden dolayıdır.

2.3.2. Mel Ölçeği ve Mel Spektogramları

İnsanlar sesleri algılarken lineer değil logaritmik algılarlar. Normal bir spektogram, frekansı lineer olarak gösterdiğinden (bkz. 1.1.5.3 STFT) insan algılamasını yansıtmaz.

MFCC tekniğinde amaç frekansı logaritmik olarak ifade eden ve insanın konuşma algılamasını yansıtan spektogramlar oluşturmaktır. Bunlara Mel spektogramları denir. Mel spektogramını elde etmek için Mel filtre bankası orjinal spektograma uygulanır. (Tak ve ark. 2017, s. 318, 319)

Mel Ölçeği ve Mel Filtre Bankasının Oluşturulması

Mel ölçeği, Hertz ile Mel frekansı arasındaki ilişkiyi tanımlar. Mel frekansı ile Hertz'in birbiri aralarındaki dönüşüm, *m* Mel frekansı *f* Hertz olarak olmak üzere; Denklem 16 ve Denklem 17'deki gibi tanımlanır (Shannon ve Paliwal, 2003, s. 2).

$$m = 2595 \,.\log(1 + \frac{f}{500})$$ **Denklem 16** (Shannon ve Paliwal, 2003, s. 2)

$$f = 700(10^{\frac{m}{2595}} - 1)$$ **Denklem 17** (Shannon ve Paliwal, 2003, s. 2)

Frekanslar üç adımda Mel ölçeğine dönüştürülür. İlk adım Mel frekans bandı sayısını seçmektir, ikinci adım mel filtre bankasını oluşturmaktır, üçüncü adım ise oluşturulan Mel filtre bankasını spektograma uygulamaktır.

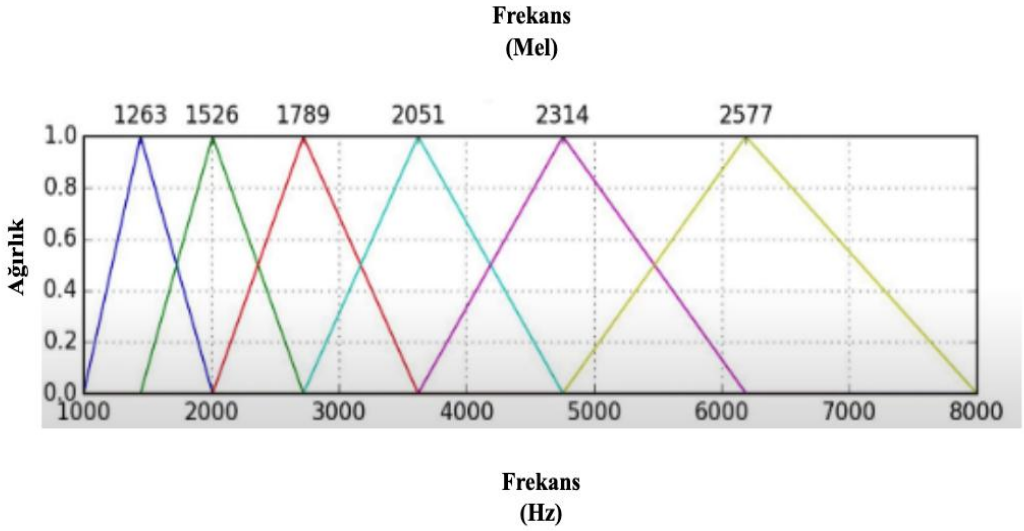
Mel frekans bandını sayısını insanın algılayabileceği farklı frekans sayısı olarak düşünebiliriz. Genellikle 40 ila 130 çeşit arasında Mel frekans bandı kullanılır. Kaç tane kullanılacağı probleme göre değişiklik gösterir. Bir problem için farklı Mel frekans bandı sayıları denenerek optimal sayı bulunabilir.

Mel frekans bankasını oluşturmak için ilk olarak en yüksek ve en düşük frekans değerleri Denklem 16'ya göre Mel frekansına çevrilir. Sonra çevrilen frekanslar ilk ve son noktayı oluşturmak üzere bu noktaların arasına eşit aralıklarla belirlenen sayıda Mel frekans bantları yerleştirilir. Sonrasında bu noktalar Denklem 17 kullanılarak tekrardan Hertz'e çevrilir. Ardından elde edilen bu frekans noktaları üzerinde elimizde olan orijinal frekanslara

göre yuvarlama işlemi yapılır. Son olarak elde edilen frekans noktalarından Şekil 16'daki gibi üçgensel filtreler oluşturulur (Molau ve ark., 2001).

Şekil 16

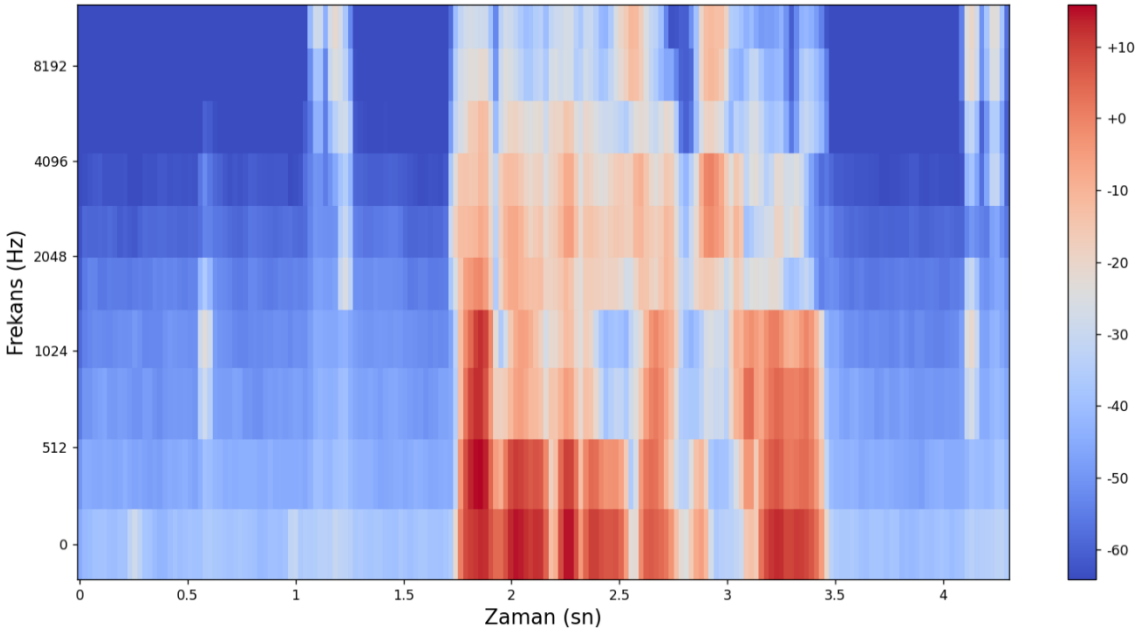
Altı Adet Mel Filtre Bandından Oluşan Mel Filtre Bankası



Oluşturulan Mel filtre bankası orijinal spektograma uygulanarak mel spektogramı elde edilir. Uygulama işleminde matris ile ifade edilen Mel filtre bankası ile STFT ile elde edilen orijinal spektogramı ifade eden matris ile çarpılır. Mel spektogramları insan algılamasını orijinal spektrogramdan daha iyi yansıtır. Mel spektogramları ses sınıflandırması, müzik sınıflandırması, ruh hali tanıma, müzik enstrümanı sınıflandırma gibi alanlarda kullanılır (Molau ve ark., 2001). Şekil 17'de örnek bir ifadenin Mel spektogramı verilmiştir.

Şekil 17

"Radyo Televizyon Üst Kurulu" İfadesine Ait Mel Spektogramı



2.3.3. Ayrık Cosine Dönüşümü (DCT) Uygulanması

Mel spektogramlarından MFCC öznitelik vektörlerini elde etmek için DCT uygulanır. DCT bir sinyali uzamsal alandan frekans alanına dönüştürür. DCT *N* blok sayısı ve *k* ise 0 ile *N-1* arasındaki sayılar olmak üzere Denklem 18'deki gibi ifade edilir (Marshall, 2001).

$$X_k = \sum_{n=0}^{N-1} x_n \cos\frac{\pi}{N}(n + \frac{1}{2})k$$ **Denklem 18** (Rao ve Yip, 2014)

2.3.4. MFCC Üzerine Literatürde Yapılan Çalışmalar

Lee ve ark. (2003) Mel kritik bant filtreleri için, filtrelerin şekil, merkez frekansı ve bant genişliği gibi parametrelerini değiştirerek Mel kepstrumun performansını analiz etti. Ardından, simpleks yöntemini kullanarak filtrelerin parametrelerini optimize etmek için bir algoritma öne sürdü ve Korece kelimelerle yapılan deneylerle, tanıma oranının yaklaşık %4-7 oranında arttığını gösterdi.

MFCC'nın konuşma tanıma alanındaki doğruluğunu arttırmak için pek çok çalışma yapılmıştır ancak bu çalışmalar genellikle hesaplaması çok yoğun olduğu için maliyeti arttırıp ve performansı azaltmışlardır. Bunun yanı sıra 2006 yılında Hong Kong Üniversitesinde MFCC için geliştirilen yeni bir algoritmayla donanımsal hesaplama maliyetini %53 düşürerek doğruluk oranında ise yalnızca %1,5'lik bir gerileme (Doğruluk oranı %94,43'ten %92,93'e düşmüştür.) sağlayarak MFCC metodunun hesaplama verimliliğinin önemli bir şekilde arttırılabileceğini göstermiştir (Han ve ark. 2006).

Wang ve ark. (2002) MFCC algoritmasının hesaplama karmaşıklığı ve bellek gereksinimi büyük ölçüde iyileştiren bir yonga tasarladı. Önerilen yonga, MFCC algoritmasına dayalı konuşma öznitelik çıkarımı için özel olarak oluşturulmuş ilk yongadır.

2.4. Algısal Doğrusal Tahmin – Bağıl Spektrum (PLP-RASTA)

PLP yaklaşımında, işitmenin birkaç belirgin özelliği kopyalanır ve bunun sonucunda ortaya çıkan işitsel benzeri konuşma spektrumu, otoregresif bir yaklaşım ile tahmin edilmeye çalışılır. PLP-RASTA PLP öznitelik çıkarım tekniğinin iyileştirilmiş versiyonudur. PLP-RASTA, Hermanski ve ark. (1991) tarafından doğrusal spektral bozulmalara karşı PLP'yi daha sağlam yapmak için geliştirildi ve PLP-RASTA için gereken adımları Şekil 18'deki gibi tanımladı. Şekil 18'de verilen logaritma ve ters logaritma adımları PLP'de yoktur. PLP ile PLP-RASTA'nin en büyük farkı adından da anlaşılacağı üzere, uygulanan RASTA filtresidir (Këpuska ve Elharati, 2015, s. 5). Ayrıca önişlem adımları yine bu yöntem öncesinde de uygulanır.

RASTA, kısa süreli gürültü değişimlerini yumuşatmak ve konuşma kanalındaki statik spektral renklenmeden kaynaklanan herhangi bir sabit ofseti ortadan kaldırmak için her frekans alt bandındaki enerjiye bir bant geçiren filtre uygulayan ayrı bir tekniktir (Shrawankar ve Thakare, 2013, s. 6). Hermanski ve ark. (1991) RASTA filtresini *z* sinyal değeri olmak üzere Denklem 19'daki gibi tanımladı.

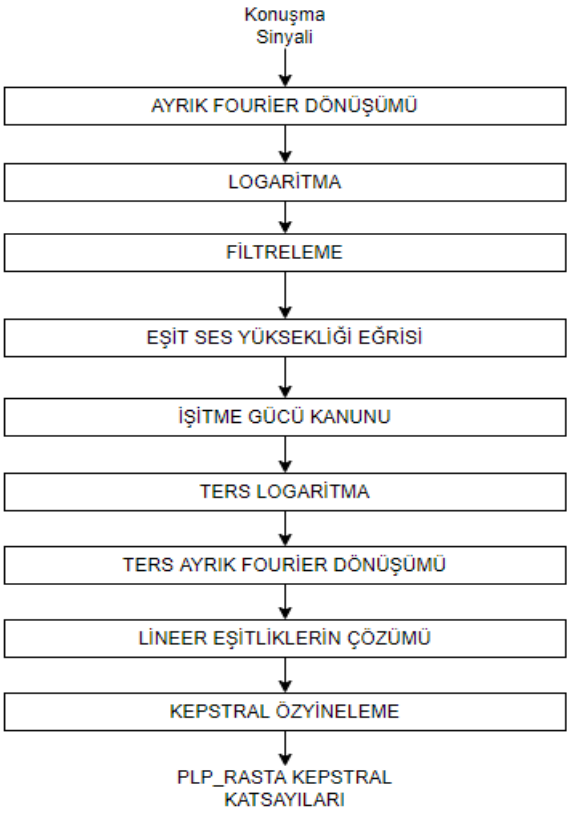
$$H(z) = 0.1 \times \frac{2+z^{-1}-z^{-3}-2z^{-4}}{z^{-4}\times(1-0.98z^{-1})}$$ **Denklem 19** (Hermanski ve ark.,1991)

Şekil 18'deki ilk adım olan DFT ve yedinci adım olan ters DFT birinci bölümde konuşma sinyal işleme başlığı altında açıklanmıştır (bkz. 1.1.5.2). İkinci ve altıncı adımlar basitçe sinyal değerlerinin log ve ters log'larının alınmasıdır. Üçüncü adımda kullanılan RASTA filtrelemesi konuşma sinyalinin belli frekans bandındaki enerjilere bant-geçiren filtre uygulayarak, sinyaldeki kısa zamanlı gürültü değişimlerini biraz düzleştirir ve sinyalinin geçtiği fiziksel hattan kaynaklı etkileri sinyalden kaldırır (Ellis, 2012).

Eşit ses yüksekliği ve işitme gücü kanunu ayrı başlıklar altında açıklanmıştır. Sekizinci adımdaki lineer eşitliklerin çözümü LPC tekniğindeki gibi geriye doğru yapılan lineer tahminden oluşan eşitliklerin çözümüdür (bkz. 2.2.2).

Şekil 18

PLP-RASTA Adımları (Hermanski ve ark., 1991)

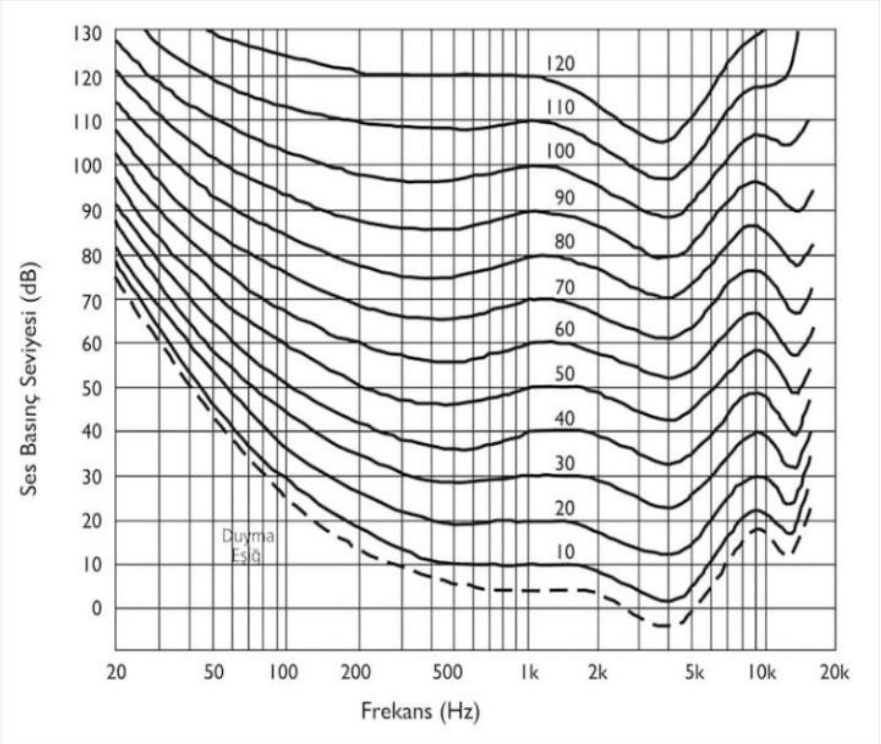


2.4.1. Eşit Ses Yüksekliği Eğrisi

Eşit ses yüksekliği eğrisi 1930'larda Fletcher ve Munson'un öncülüğünü yaptığı çalışmalara dayanır. Fletcher ve Munson, insan kulağının hangi frekanslara daha çok, hangi frekanslara daha az duyarlı olduğu üzerinde araştırmalar yapmış ve eşit ses yüksekliği eğrisi adı verilen bir çizelge geliştirmişlerdir (bkz. Şekil 19) (Nave, 2017)

Şekil 19

Eşit Ses Yüksekliği Eğrisi (Nave, 2017)



2.4.2. Yoğunluk-Ses Şiddeti Kanunu

Bazen bir uyaran güçlendikçe, uyaran değişikliklerine karşı giderek daha az duyarlı hale geliriz. Bazen uyaran güçlendikçe hassasiyetimiz kabaca sabit kalır. Bazen de uyaran yoğunluğu arttıkça duyarlılığımız da artar. Stevens, bu olası durumların tamamını kapsamak amacıyla bir denklem geliştirdi. Buna Stevens'ın Güç Yasası denir ve bunun insanın sesi frekans olarak algılamasına uygulanmasına da işitme gücü kanunu denir (Schwartz ve Krantz, 2015).

$P = cI^b$ **Denklem 20** (Schwartz ve Krantz, 2015).

Denklem 20'de *P*, algılanan büyüklüğe eşittir. *I*, gerçek uyaranların şiddetine eşittir. *C* harfi, her sensor modeli için farklı olacak bir sabite eşittir örneğin fahrenhayt ve santigrat'ın sıcaklığı farklı ölçeklerde ölçmesi gibi. Üs b, uyaranın şiddetinin yükseltildiği güce eşittir. Örneğin acı uyaranı için bu muhtemelen birden büyüktür. (Schwartz ve Krantz, 2015).

2.5. LPC, MFCC ve PLP-RASTA Karşılaştırması

Bu bölümde, çalışma kapsamında geliştirilen, öznitelik vektörleriyle eğitilen CNN'de verilen veri setleri için (Türkçe ve İngilizce) modelin tanıma oranlarının karşılaştırılması yapılacaktır. Ayrıca literatürde bu teknikler için yapılan karşılaştırmalı çalışmalara değinilerek bu tekniklerin birbirlerine göre avantajları ve dezavantajları açıklanacaktır.

Uygulama kapsamında geliştirilen üç katmanlı modelin, üç teknik için, (LPC, MFCC, PLP-RASTA) İngilizce ve Türkçe veri setleri ile eğitilmesi ve test edilmesi sonucunda edinilen tanıma oranları sonuçlar bölümünde verilmiştir. Geliştirilen uygulamada türkçe veri setinde tanıma oranının İngilizce veri setine göre daha yüksek olduğu görülmüştür. Bunun sebebinin Türkçenin yazıldığı gibi okunan bir dil olmasından kaynaklandığı düşünülmektedir.

Shrawankar ve Thakare (2013) öznitelik çıkarım tekniklerine yönelik yaptığı araştarmalarda; konuşma sinyalinde gürültüyü bastıran metodlar kullanılmazsa MFCC'nin performansının düşük olacağını belirtmektedir. LPC'nin kelime hata oranının zor koşullarda (çevresel gürültü) PLP'den daha iyi olduğunu saptamıştır. PLP'nin bazı özel koşullarda MFCC'den daha başarılı tanıma oranları yakaladığını, bunun haricinde genel olarak MFCC kullanan uygulamaların tanıma oranının daha yüksek olduğu sonucuna vardı.

Narang ve Gupta (2015) LPC'nin diğer tekniklere göre kıyasla hesaplama hızının iyi olması ve örneklem oranı düşük konuşma sinyalleri için kullanışlı olmasının başlıca avantajları olduğunu belirtir. LPC'nin dezavantajları olarak ise artık hatası sebebiyle önemli konuşma bölümlerini temsil etmeyebileceğini belirtiyor. MFCC için ise avantaj olarak yüksek tanıma oranını ve fonemlerin başlıca karakteristiklerini yakalaması olarak belirtti.

MFCC'nin dezavantajı olarak ise arka plandaki gürültüye karşı hassas olması ve performansının kullanılan filtre sayısına göre kolay etkilenebilmesi olarak belirtti. Barua ve ark. (2014) da MFCC'nin arka plandaki gürültüye karşı hassas olduğunu yaptıkları uygulamayla gösterdi.

Dave (2013) bu çalışmada ele alınan üç teknik için (MFCC, LPC, PLP) -PLP çalışmada ele alınan PLP-RASTA tekniğinin RASTA filtresi kullanılmamış halidir bu sebeple PLP ile ilgili yapılan çalışmalar PLP-RASTA için de geçerlidir diyebiliriz.- karşılaştırmaya dayalı bir çalışma yapmıştır. Dave bu çalışmasında konuşma tanıma için MFCC ve PLP'nin LPC'den daha başarılı olduğu sonucuna varmıştır. Bunun sebebi olarak ise insan sesinin doğası gereği doğrusal olmamasını ve MFCC ve PLP'deki, logaritmik olarak aralıklı frekans filtre bankası kavramının insan sesini daha iyi modellemesine bağlamıştır. Literatürde bu yöntemlerin kullanım alanlarına yönelik yapılan araştırma da bu sonucu güçlendirmektedir.

Alim ve Rashid (2018) öznitelik çıkarım tekniklerine dair yaptıkları karşılaştırmada hesaplama hızı bakımından MFCC ve LPC'nin PLP'den daha iyi sonuçlar verdiğini buldular. Ayrıca LPC'nin gürültüye dayanıklılık bakımından MFCC ve PLP'den daha iyi olduğu sonucuna vardılar. (Alim ve Rashid, 2018, s. 15) MFCC'nin düşük frekanstaki ses sinyallerini göstermede etkili olduğunu vurguladı. Ayrıca LPC için ise hesaplama hızının yüksek olduğunu ve konuşma kaynağının (ses yolu, bkz. Şekil 12) stabil ve tutarlı olduğu durumlarda daha iyi sonuçlar verdiğini belirtti.

PLP-RASTA de MFCC gibi insanın işitme algısına dayanır. MFCC'deki gibi log sıkıştırması yerine eşit ses yüksekliği ön vurgusu ve küp-kök sıkıştırması kullanır. PLP-RASTA yaklaşımında, işitmenin birkaç belirgin özelliği kopyalanır ve bunun sonucunda ortaya çıkan işitsel benzeri konuşma spektrumu, otoregresif bir yaklaşım ile tahmin edilmeye çalışılır. PLP doğruluk ve gürültüye karşı dayanıklılık açısından MFCC'den biraz daha iyidir ancak yine de MFCC'nin, öznitelik çıkarımı için daha güvenilir bir seçenek olduğuna inanılıyor. (Alim ve Rashid, 2018, s. 14).

BÖLÜM 3

ÖZNİTELİK VEKTÖRLERİYLE CNN EĞİTİMİ

Bu çalışma kapsamında incelenen üç öznitelik çıkarım tekniği ile İngilizce ve Türkçe ses dosyalarından oluşan veri seti için öznitelik vektörleri elde edilip, elde edilen öznitelik vektörleri ile CNN (Evrişimsel Sinir Ağı) eğitilip test edilmiş, oluşan model kaydedilmiştir. Test sonucunda elde edilen sonuçlar Tablo 4 ve Tablo 5'te verilmiştir. Bu uygulamanın geliştirilmesindeki amaç üç farklı yöntemle elde edilen öznitelik vektörlerinin aynı CNN'de elde edeceği öğrenme oranlarını tespit etmektir.

Uygulama geliştirilirken ilk olarak kullanılacak olan veri setleri belirlenmiştir. Sonrasında öznitelik çıkarım teknikleri için kullanılacak hesaplama yöntemlerine karar verilmiştir. Ardından eğitilecek olan modelin özellikleri belirlenmiştir. Ardında kodlama aşamasına geçilmiştir.

3.1. Kullanılan Veri Seti

Çalışma kapsamında İngilizce ve Türkçe komutlar için 1 saniyelik ses dosyalarından oluşan veri seti kullanılmıştır. Komutlar tek kelimelik olup, her bir komutun binlerce farklı tonlamaya sahip örneği mevcuttur.

Türkçe veri seti için Murat Kurtkaya tarafından oluşturulan, 14 komuttan oluşan, 26.485 ses dosyasına sahip veri seti kullanılmıştır. Çalışma kapsamında incelenen üç öznitelik çıkarım tekniği, yine bu dosyaların hepsi için ayrı ayrı uygulanıp öznitelik vektörleri elde edilmiştir (Kurtkaya, 2021). Kullanılan Türkçe komutların listesi ve her bir komut için örnek ses dosyası sayısı Tablo 2'de verilmiştir.

İngilizce veri seti için Google bünyesinde "TensorFlow" ve "AIY" (Google tarafından desteklenen yapay zekâ donanım kiti) ekipleri tarafından oluşturulan, konuşma tanıma alanında çalışan araştırmacıların kullanımına sunulan, veri seti kullanılmıştır. Bu veri setinde 30 komut için, binlerce kişinin seslendirdiği, 64.721 adet bir saniyelik ses dosyaları mevcuttur (Warden, 2017). Çalışma kapsamında incelenen üç öznitelik çıkarım tekniği, 64.721 ses dosyası için ayrı ayrı uygulanıp öznitelik vektörleri elde edilmiştir. Kullanılan İngilizce komutların listesi ve her bir komut için örnek ses dosyası sayısı Tablo 3'te verilmiştir.

Bu veri setleri üzerinde, uygulamanın sonraki aşamalarında, LPC, MFCC, PLP-RASTA öznitelik çıkarım algoritmaları kullanılarak her bir ses dosyası için öznitelik vektörleri elde edilmiştir. Sonrasında bu öznitelik vektörleriyle evrişimsel sinir ağı eğitilip, bu komutları tanıyan model oluşturulacaktır.

3.2. Kullanılan Hesaplama Yöntemleri

MFCC için Librosa kütüphanesine ait hesaplama tekniği kullanılmıştır (Librosa). Sonuç olarak her bir saniyelik ses dosyası, 44 adet 22,5 ms'lik bloklara bölünmüş ve her blok için 13 tane öznitelik katsayı üretilmiştir.

LPC hesaplaması için Kunigami'nin (2021) izlediği yöntem ("lpc_encode" metodu) temel alınmıştır. Ancak hesaplamanın diğer öznitelik çıkarım teknikleriyle aynı olması amacıyla LPC (katsayı sayısı) 13, örneklem oranı 22.050 olarak ayarlanmıştır. Ayrıca yine her bir saniyelik ses dosyası, 44 adet 22,5 ms'lik bloklara bölünmüş ve her blok için 13 tane öznitelik katsayı üretilmiştir.

PLP-RASTA hesaplaması için Lee'nin (2018) izlediği "RastaPlp" methodu baz alınmıştır. Ancak yine hesaplamanın diğer öznitelik çıkarım teknikleriyle aynı olması amacıyla PLP-RASTA (katsayı sayısı) 13, örneklem oranı 22.050 olarak ayarlanmıştır. Ayrıca yine her bir saniyelik ses dosyası, 44 adet 22,5 ms'lik bloklara bölünmüş ve her blok için 13 tane öznitelik katsayı üretilmiştir.

3.3. Evrişimsel Sinir Ağı Modeli

Derin öğrenmede verileri temsil eden öznitelik vektörlerini öğrenmek için bir yapı vardır. Derin öğrenmede ilgili veriden çıkarılan öznitelikler, hiyerarşik bir yapı oluşturur ve bu sayede soyutlamanın farklı seviyelerini öğrenilir. Örneğin bir görüntü için piksel başına yoğunluk, görüntünün kenar kümeleri veya özel şekiller o görüntü için çıkarılan özellik vektörleri için temsil edilebilir. Derin öğrenme veriyi en iyi temsil eden ve hiyerarşik öznitelik çıkarımı algoritmaları kullanmaktadır. Bir derin öğrenme mimarisi olan CNN (Evrişimsel Sinir Ağı) bir veya daha fazla evrişimsel katmanlardan, alt örnekleme katmanından ve standart çok katmanlı bir sinir ağı gibi bir veya daha fazla bağlı katmanlardan oluşur. Görüntü ve ses işleme alanlarında sıkça kullanılmaktadır (Şeker ve ark., 2017, s. 48, 50).

Uygulamayı geliştirirken CNN mimarisinin seçilmesinin nedenlerinden biri, ele aldığımız öznitelik çıkarım tekniklerinde amaç konuşma sinyalini en iyi temsil eden öznitelik vektörlerini çıkarmak olduğu için ve bu işlem gerçekleştirilirken konuşma sinyaldeki hiyerarşiye dikkat edilmesidir (önemsiz sinyal verilerinin çıkarılması, filtreden geçirilmesi).

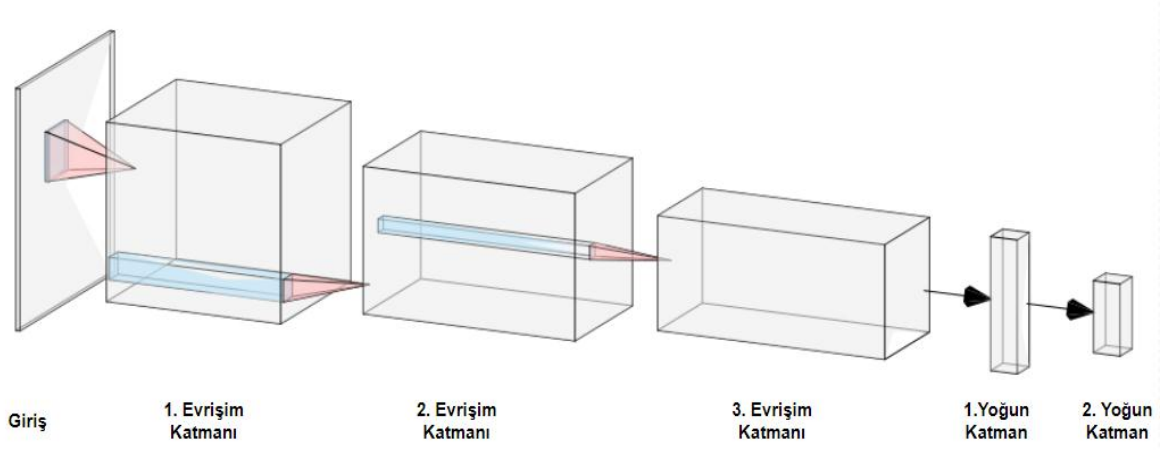
CNN seçilmesinin bir başka nedeni, CNN'nin diğer yapay sinir ağlarından farklı olarak evrişim ve sınıflandırma katmanları içermesidir. Evrişim, bir matrise uygulanan kayan pencere işlemidir. CNN'ler esasen sonuçlara uygulanan doğrusal olmayan aktivasyon fonksiyonlarına (ReLU veya tanh) sahip birkaç evrişim katmanıdır. Geleneksel bir ileri beslemeli sinir ağında, her bir girdi nöronunu bir sonraki katmandaki her bir çıktı nöronuna bağlanır. CNN'lerde çıktıyı hesaplamak için bunun yerine girdi katmanı üzerinde evrişimler kullanılır (Britz, 2015).

CNN (Evrişimsel Sinir Ağı) için TensorFlow altyapısı ve python için yapay sinir ağlarında kullanılan Keras kütüphanesinden faydalanılmıştır. Oluşturulan CNN beş katmanlı bir yapıya sahiptir. Bu katmanların ilk üçü evrişim katmanı son ikisi yoğun katmandır (bkz. Şekil 20).

CNN Türkçe ve İngilizce veri setleri ve üç ayrı öznitelik vektörleri için (MFCC, LPC, PLP-RASTA) için ayrı ayrı eğitilmiştir (toplam altı defa). Veri setinden elde edilen öznitelik vektörlerinin % 90'ı eğitim için %10'u da test için ayrılmıştır. Eğitim için ayrılan veri setinin de %10'u validasyon için kullanılmıştır.

Şekil 20

Uygulama İçin Kullanılan Evrişimsel Sinr Ağı Modeli



Öznitelik vektörleriyle oluşturulan CNN eğitilirken, vektörler 32'şerli kümeler halinde 40 defa modelden geçirilmiştir (batch_size = 32, epoch = 40). Ayrıca tahminlemedeki kayıp oranı hesaplanarak modelin ağırlıkları güncellenmiştir. Veri seti, hesaplama teknikleri ve CNN modeli belirlendikten sonra uygulama geliştirilmeye başlanmıştır.

3.4. Geliştirme Süreci

Öznitelik vektörleriyle CNN eğitimi uygulaması python programlama dili kullanılarak geliştirilmiştir. Bu çalışma kapsamında geliştirilen uygulama ve hesaplamalara ait kodlar açık kaynak olarak internetten erişilebilir durumdadır (İtap, 2022). Spesifik olarak python versiyon 3.9 kullanılmıştır. Ses ve müzik analizi için kullanılan "Librosa" isimli python kütüphanesinin 0.9 versiyonundan faydalanılmıştır. Evrişimsel sinir ağı için "TensorFlow" ve "Keras" yazılım kütüphanelerinden faydalanılmıştır.

Uygulama geliştirilirken ilk olarak veri seti eğitim için hazırlanmıştır. Veri seti eğitim için hazırlanırken, her bir veri (bir saniyelik komut ses dosyası) etiketlenip, üç ayrı

yöntemle (MFCC, LPC, PLP-RASTA) öznitelik vektörleri elde edilmiştir. Elde edilen vektörler için, üç yöntem için İngilizce ve Türkçe ayrı olacak şekilde, altı farklı dosyaya etiketleri ile yazılmıştır. Üç yöntem için, Tablo 2 ve Tablo 3'teki veriler üzerinde yapılan öznitelik çıkarım hesaplama süreleri Tablo 5'te verilmiştir.

Ardından Veri setlerine ait öznitelik vektörlerini içeren dosyalar kullanılarak özellikleri 4.2'de verilen bir CNN modeli oluşturulup eğitilmiş ve test edilmiştir. Modelin test sonucu ulaştığı öğrenme oranları Tablo 4'te verilmiştir. Ardından bu model H5 formatında bir dosyaya kaydedilmiştir. Son olarak girdi olarak verilen ses dosyasının, hangi komut olduğunu tahmin eden (kaydedilen CNN modelini kullanarak) bir servis geliştirilmiştir.

3.5. Sonuçlar

Bu bölümde geliştirilen uygulama ile elde edilen sonuçlar paylaşılacaktır. Bu sonuçlar, ele alınan öznitelik çıkarım teknikleri ve Türkçe ve İngilizce veri setleri için elde edilen tanıma oranları ve Tablo 2 ve Tablo 3'teki veri setleri için öznitelik çıkarım sürelerinden oluşmaktadır. Uygulamada evrişimsel sinir ağı eğitilirken elimizdeki veri setinin %10'u test için ayrılmıştır. Bu veri setiyle yapılan testlerde elde edilen tanıma oranları Tablo 4'te verilmiştir.

Tablo 4

Oluşturulan Modellerin Komut Tanıma Oranları

	Türkçe Veri Seti	İngilizce Veri Seti
LPC	% 88,8	% 73,9
MFCC	% 97,4	% 91,2
PLP-RASTA	% 97,0	% 90,2

Uygulamada Tablo 2 ve Tablo 3'teki veri setleri için öznitelik çıkarma işleminin süreleri Tablo 5'te verilmiştir. Bu tablo incelenirken kullanılan hesaplama yöntemlerine bakılması faydalı olacaktır.

Tablo 5

Tablo 2 ve Tablo 3'teki Veri Setleri İçin Öznitelik Çıkarma Süreleri

	Türkçe Veri Seti	İngilizce Veri Seti
LPC	2656 sn.	4286 sn.
MFCC	1196 sn.	3304 sn.
PLP-RASTA	1345 sn.	3636 sn.

BULGULAR VE YORUMLAR

Bu çalışmanın ilk bölümünde konuşma tanıma sistemleri genel olarak açıklanmıştır. Bu sistemi kavramak için gerekli alt sistemler olan, konuşma sinyal işleme, öznitelik çıkarımı, akustik ve dil modelinin temel yaklaşımları olan HMM-GMM ve YSA, akustik ve dil modelleme sistemleri için geçmişten günümüze kullanılan teknikler açıklanmıştır. Öznitelik vektörleri akustik ve dil modellerinin temel girdisini oluşturmaktadır.

İkinci bölümde tez kapsamında incelenen üç öznitelik çıkarım tekniği olan LPC, MFCC, PLP-RASTA tekniklerinin hesaplama yöntemleri açıklanmıştır. Öznitelik çıkarımı işleminden önce yapılması gereken işlemler olan filtreleme, sinyalin bloklara ayrılması ve pencereleme işlemlerine yine bu bölümde yer verilmiştir. Ayrıca bu teknikler üzerine literatürde yapılan karşılaştırmalar incelenerek bir özet halinde sunulmuştur.

Üçüncü bölümde Tablo 2 ve Tablo 3'teki komutları tanımak için bir uygulama geliştirilmiştir. Bu bölümde eğitim verisi hazırlanıp, öznitelik vektörleri çıkarılmış ve çıkarılan öznitelik vektörleriyle evrişimsel sinir ağı eğitilmiştir. Ayrıca verilen bir ses dosyasının hangi komut olduğunu tahmin eden bir servis geliştirilmiştir. Bu serviste tahmin için eğitilen evrişimsel sinir ağının oluşturduğu model kullanılmıştır.

LPC tekniği diğer tekniklere göre en önce kullanılan öznitelik çıkarım tekniğidir. Literatürde yapılan araştırmalara göre LPC; konuşma tanıma sistemlerinde yakaladığı tanıma oranı MFCC ve PLP-RASTA'ya göre daha düşüktür ve bu sebeple bu sistemlerde kullanım oranı da daha düşüktür, konuşma sinyalini sıkıştırmada ise diğer tekniklere göre daha başarılıdır dolayısıyla bu alandaki kullanım oranı oldukça yüksektir. Diğer taraftan gürültüye karşı dayanıklılık ve hesaplama süresi olarak LPC, literatürde yapılan araştırmalara göre, diğer tekniklerden daha başarılıdır.

Çalışma kapsamında geliştirilen uygulama da literatürde yapılan araştırmalara paralel olarak, göstermiştir ki konuşma tanıma sistemlerindeki tanıma oranı, LPC için diğer tekniklere göre belirgin bir oranda daha düşüktür (bkz. Tablo 4). Yine geliştirilen uygulamaya göre hesaplama süresi bakımından LPC, diğer tekniklere göre daha yüksek hesaplama süresine sahiptir ve bu bulgu literatürde yapılan araştırmalarla ters düşmektedir.

MFCC, günümüzde konuşma tanıma sistemlerinde kullanılan en yaygın öznitelik çıkarım tekniğidir. Kronolojik olarak LPC'den sonra, PLP-RASTA'dan önce geliştirilmiş ve bu sistemlerde kullanılmaya başlanmıştır. Literatürde yapılan araştırmalara göre MFCC; konuşma tanımadaki tanıma oranı olarak LPC'den daha iyidir, PLP-RASTA'ya göre ise karşılaştırılabilir sonuçlara (birbirine yakın tanıma oranları) sahiptir, gürültüye dayanıklılık bakımından diğer tekniklere göre daha kötü performans sergilemektedir, MFCC'nin hesaplama süresi ise PLP-RASTA'dan daha kısa, LPC'den ise daha uzun sürmektedir.

Tez için geliştirilen uygulamaya göre MFCC, konuşma tanıma sistemindeki tanıma oranı bakımından LPC'ye göre oldukça iyidir, PLP-RASTA'dan ise küçük bir farkla daha ileridedir (bkz. Tablo 4), Hesaplama süresi olarak, geliştirilen uygulamaya göre, diğer tekniklerden daha kısa sürmüştür (bkz. Tablo 5).

PLP-RASTA diğer tekniklere göre daha sonra geliştirilmiş ve konuşma tanıma sistemlerinde kullanılmıştır. Literatürde yapılan araştırmalara göre PLP-RASTA; konuşma tanıma sistemlerindeki tanıma oranı bakımından LPC'den daha iyidir, MFCC ile benzer tanıma oranlarına sahiptir, hesaplama süresi olarak, diğer tekniklerden daha uzun sürmektedir, gürültülü sesler için LPC'den daha kötü, MFCC'ye göre ise daha iyi performans göstermektedir.

Çalışma kapsamında geliştirilen uygulama da PLP-RASTA için, literatürde yapılan araştırmalara çoğunlukla benzer sonuçlar vermektedir. Konuşma tanıma sistemlerindeki başarım oranları MFCC ile çok yakın, LPC'den ise daha iyidir (bkz. Tablo 4). Hesaplama süresi bakımından MFCC'den biraz daha uzun sürmektedir, LPC'den ise daha kısa sürmektedir (bkz. Tablo 5). Literatürde yapılan çalışmalar LPC'nin hesaplama süresinin PLP-RASTA'dan kısa olduğu yönündedir ancak geliştirilen uygulamada bunun tam tersi bir sonuç görülmüştür.

Çalışma kapsamında incelenen üç öznitelik çıkarım tekniği de geliştirilen uygulamada, her iki dilde de eğitim için komut başına veri sayısı birbirine yakın olmasına rağmen (bkz. Tablo 2, Tablo 3), Türkçede daha yüksek tanıma oranlarına ulaşmıştır (bkz. Tablo 4). Türkçenin yapısı itibariyle yazıldığı gibi okunan bir dil olmasının, bunun başlıca sebebi olduğu düşünülmektedir.

Elde edilen bulgular, Tablo 1'deki karşılaştırma kıstaslarına göre, Tablo 6 ve Tablo 7'de verilmiştir. Ayrıca Türkçe ve İngilizce veri seti için tanıma oralarının karşılaştırması Tablo 4'te verilmiştir.

Tablo 6

Üç Teknik İçin Literatüre Göre Elde Edilen Bulgular

	LPC	MFCC	PLP-RASTA
Verilen ifadeyi tanıma oranı	Orta	İyi	İyi
Hesaplama Süresi	İyi	İyi	Orta
Gürültüye Karşı Dayanıklılık	İyi	Kötü	Orta

Tablo 7

Üç Teknik İçin Geliştirilen Uygulamaya Göre Elde Edilen Bulgular

	LPC	MFCC	PLP-RASTA
Verilen ifadeyi tanıma oranı	Orta	İyi	İyi
Hesaplama Süresi	Kötü	İyi	İyi

SONUÇLAR VE ÖNERİLER

Bu tez çalışmasında, ilk bölümde konuşma tanıma sistemlerinin genel yapısı ele alınmış, bu yapıyı oluşturan alt sistemler ve bu sistemleri gerçekleştirmek için kullanılan modeller ve yaklaşımlar incelenmiştir. İkinci bölümde literatürde öne çıkan üç önemli öznitelik çıkarım tekniği incelenmiş, hesaplama yöntemleri ve konuşma sinyalini temsil etmek için ortaya koydukları yaklaşımlar açıklanıp, bu teknikler üzerinde literatürde yapılan karşılaştırmalı araştırmalara değinilmiştir. Üçüncü bölümde ise ikinci bölümde ele alınan öznitelik çıkarım tekniklerini kullanarak Türkçe ve İngilizce veri setleri için öznitelik vektörleri elde edilmiş, bu vektörler kullanılarak her teknik için ve dil için ayrı olacak şekilde aynı özelliklere sahip evrişimsel sinir ağı eğitilmiştir. Ardından bu sinir ağı test edilerek, elde sonuçlar Tablo 4 ve Tablo 5'te verilmiştir.

Sonuçlar

Konuşma tanıma sistemlerinin en temel adımı olan konuşma sinyal işleme konusuna bakacak olursak; Fourier dönüşümünün çeşitli varyasyonları sayesinde konuşma sinyali çeşitli formlarda gösterip, ihtiyaç duyduğumuz analizi yapabilmekteyiz (spektrum ve spektogram analizi). Bu sebeple Fourier dönüşümleri konuşma sinyal işlemede oldukça önemli bir yere sahiptir.

Konuşma sinyal işlemede spektogramların insan duyusunu yansıtacak şekilde geliştirilmiş versiyonu olan Mel spektogramları, MFCC öznitelik çıkarım tekniğindeki kullanımı dışında, ses sinyaline ait insan duyusu ölçeğinde frekans zaman bilgisini vermesi açısından, konuşma sinyallerinin çeşitli analizleri için büyük fayda sağlamaktadır.

Öznitelik çıkarımı konuşma tanıma sistemlerinde en önemli adımlardan biridir. Çünkü tanıma işlemini gerçekleştirecek olan dil ve akustik modeller konuşma sinyalinin kendisi yerine, bu sinyali temsil eden, sinyalin küçük blokları için elde edilen öznitelik vektörlerini kullanmaktadır. Öznitelik çıkarımı için çeşitli teknikler kullanılmaktadır. Bu teknikler konuşma sinyalini modellerken çoğu zaman insanda ses oluşumunu veya insan duyma yapısını temel almışlardır. Bunlar arasından MFCC günümüzde en çok kullanılan

öznitelik çıkarım tekniğidir. MFCC'nin yaygın kullanımı ile birlikte araştırmacılar bu tekniği iyileştirmek için pek çok çalışma yapmışlardır ve gelinen noktada MFCC konuşma tanıma sistemlerinde çok yüksek tanıma oranlarına ulaşmıştır. MFCC'nin en büyük dezavantajı gürültüye karşı hassasiyetidir. Yapılan araştırmalara ve geliştirilen uygulama sonuçlarına göre MFCC normal şartlarda bir konuşma tanıma sistemi için en ideal öznitelik çıkarım tekniğidir.

Çalışma kapsamında incelenen PLP-RASTA; RASTA filtresinin kullanılması sebebiyle gürültüye karşı dayanıklılığı yüksektir. Konuşma tanıma sistemlerindeki tanıma oranı bakımından MFCC'ye çok yakın sonuçlar vermektedir. Çalışma kapsamında geliştirilen uygulamaya göre hesaplama süresi MFCC'den daha uzundur. LPC ile karşılaştırıldığında ise tanıma oranı olarak daha iyidir. Hesaplama süresi olarak yine geliştirilen uygulama göz önüne alındığında LPC'den daha iyi bir performans göstermiştir. PLP-RASTA arkaplanda gürültü olan konuşmalar için MFCC'nin yerine kullanılabilir.

LPC öznitelik çıkarım tekniği'nin konuşma tanımadaki tanıma oranı diğer tekniklere göre daha düşüktür. LPC konuşma sinyalini sıkıştırmadaki başarısı sebebiyle bu alanda kullanımı oldukça yaygındır. LPC'nin hesaplama süresi literatürdeki araştırmalara göre diğerlerinden düşüktür ancak bu çalışma kapsamında geliştirilen uygulamada LPC hesaplaması diğerlerinden daha fazla zaman almıştır.

Konuşma sinyalinden fonemlerin kelimelerin ve cümlelerin üretilmesinden sorumlu olan akustik ve dil modelleri günümüzde özellikle yapay sinir ağları teknolojilerinin kullanımıyla beraber büyük gelişmeler kaydetmiştir. Özellikle YSA'da bulunan genelleme özelliği, bu yapıları konuşma tanıma sistemleri için uygun kılmaktadır çünkü bir YSA genelleme özelliği sayesinde eğitim sırasında verilmeyen ifadeleri de tanıyabilme yeteneğine sahiptir.

Öneriler

Gerçekleştirilecek bir konuşma tanıma sisteminde oluşturulan dil ve akustik modellerin gerçek zamanlı olarak eğitilmesine dikkat edilmelidir. Yapay sinir ağlarının gerçek zamanlı eğitilmesi yani eğitimin sürekli olarak ilgili verilerle devam etmesi, konuşma tanıma sisteminin tanıma oranını oldukça artırabilmektedir. Yine dil ve akustik model için istatistiksel metotlar (HMM-GMM) kullanılacaksa yine bu modellerinde parametrelerinin sürekli olarak yeni verilerle güncellenmesi tanıma oranını yükseltecektir.

RTÜK bünyesinde SKAAS sistemi ile yüksek miktarda veri kurum sunucularında barınmaktadır. Bu veriler, oluşturulacak bir konuşma tanıma sisteminde kullanılacak yapay sinir ağını eğitmek adına büyük önem arz etmektedir. Yine bu veriler hem YSA hem istatistiksel metotları bir arada kullanan hibrit modellerde istatistiksel parametreleri elde etmek için de kullanılabilir. Ayrıca bu veriler konuşan kişiyi tanımlama, konuşma duygu analizi gibi uygulamalarda da eğitim verisi olarak kullanılabilir.

SKAAS sistemi ile anlık olarak yayınlar RTÜK'e gelmektedir. Oluşturulacak bir konuşma tanıma sisteminde, bu verilerin gerçek zamanlı olarak sınıflandırılıp bir ön işlemden geçirilmesi akabinde YSA veya istatistiksel modelleri eğitmek için kullanılması, oldukça başarılı modeller oluşturmamıza olanak sağlayacaktır. Böyle bir sistemde, MFCC öznitelik çıkarım tekniği olarak kullanılmalıdır çünkü hem tanıma oranı hem de hesaplama süresi olarak diğer tekniklerden daha başarılı sonuçlar vermektedir. Diğer taraftan elimizdeki verileri gürültü içerme oranlarına göre sınıflandırabilirsek, gürültü oranı fazla olan veriler için PLP-RASTA tekniğinin kullanılması daha doğru olacaktır çünkü yapılan araştırmalar göstermiştir ki MFCC gürültü içeren verilerde düşük performans göstermektedir.

KAYNAKLAR

Ağyar, Z. (2015). Yapay sinir ağlarının kullanım alanları ve bir uygulama. Mühendis ve Makine, 56(662), 22-23.

Alim, S. A., & Rashid, N. K. A. (2018). Some commonly used speech feature extraction algorithms, 2-19.

Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. The journal of the acoustical society of America, 50(2B), 637-655.

Ayrancı, A. A., Atay, S., & Yıldırım, T. (2021). Speaker Accent Recognition Using MFCC Feature Extraction and Machine Learning Algorithms. International Journal of Advances in Engineering and Pure Sciences, 33, 17-27.

Bachu, R. G., Kopparthi, S., Adapa, B., & Barkana, B. D. (2008, June). Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In American Society for Engineering Education (ASEE) zone conference proceedings.

Barua, P., Ahmad, K., Khan, A. A. S., & Sanaullah, M. (2014, May). Neural network based recognition of speech using MFCC features. In 2014 international conference on informatics, electronics & vision (ICIEV).

Bett, S. (2002). The number of phonemes in English. Memory of Ken Ives (1917–2002), 30(1).

Bilmes, J. A., Zweig, G., Filali, K., Livescu, K., Xu, P., Jackson, K., ... & Byrne, B. (2001). Discriminatively Structured Graphical Models for Speech Recognition The Graphical Models Team.

Bradbury, J. (2000). Linear predictive coding. Mc G. Hill.

Bridle, J. S., & Brown, M. D. (1974). An experimental automatic word recognition system. JSRU report, 1003(5), 33.

Britz, D. (2015, November). Understanding Convolutional Neural Networks for NLP. Retrieved September 19, 2022, from https://dennybritz.com/posts/wildml/understanding-convolutional-neural-networks-for-nlp/#what-areconvolutional-neural-networks

Buyuk, O., Erdogan, H., & Oflazer, K. (2005, May). Using hybrid lexicon units and incorporating language constraints in speech recognition. In Proceedings of the IEEE 13th Signal Processing and Communications Applications Conference, 111-114.

Chen, X., Liu, X., Gales, M. J., & Woodland, P. C. (2015, April). Recurrent neural network language model training with noise contrastive estimation for speech recognition. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5411-5415.

Cole R., Mariani J., Uszkoreit H., Zaenen A., & Zue V. (1997). Survey of the state of the art in human language technology, Cambridge University Pres

Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. International journal for advance research in engineering and technology, 1(6).

Davis, K. H., Biddulph, R., & Balashek, S. (1952). Automatic recognition of spoken digits. The Journal of the Acoustical Society of America, 24(6), 637-642.

Demirci, K. (2011). Fonem Teorisinin Önemi Hakkında. Electronic Turkish Studies, 6(2).

Denes, P. (1959). The design and operation of the mechanical speech recognizer at University College London. Journal of the British Institution of Radio Engineers, 19(4), 219-229.

Dhingra, S. D., Nijhawan, G., & Pandit, P. (2013). Isolated speech recognition using MFCC and DTW. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2(8), 4085-4092.

Dilmegani, C. (2021, February), Top 11 Speech Recognition Applications. Retrieved May 04, 2022, from https://research.aimultiple.com/voice-recognition-applications/

Dudley, H., Riesz, R. R., & Watkins, S. S. (1939). A synthetic speaker. Journal of the Franklin Institute, 227(6), 739-764.

Ellis, D. (2012, September). PLP and RASTA (and MFCC, and inversion) in Matlab using melfcc.m and invmelfcc.m. Retrieved June 06, 2022, from https://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

Ergezer, H., Dikmen, M., & Özdemir, E. (2003). Yapay sinir ağları ve tanıma sistemleri. PiVOLKA, 2(6), 14-17.

Fletcher, H. (1922). The nature of speech and its interpretation. Journal of the franklin institute, 193(6), 729-747.

Goyal, S. (2018, May). What are the characteristics of Sound Waves. Retrieved May 02, 2022, from https://www.jagranjosh.com/general-knowledge/what-are-the-characteristics-of-sound-waves-1525678871-1

Gupta, H., & Gupta, D. (2016, January). LPC and LPCC method of feature extraction in Speech Recognition System. In 2016 6th international conference-cloud system and big data engineering (confluence), 498-502.

Warden, P. (2017, August). Launching the Speech Commands Dataset. Retrieved June 07, 2022, from https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html

Ghule, K., & Deshmukh, R. (2015). Feature Extraction Techniques for Speech Recognition: A Review. International Journal of Scientific and Engineering Research. 6, 143-147

Goswami, U., Thomson, J., Richardson, U., Stainthorp, R., Hughes, D., Rosen, S., & Scott, S. K. (2002). Amplitude envelope onsets and developmental dyslexia: A new hypothesis. Proceedings of the National Academy of Sciences, 99(16), 10911-10916.

Graham, T. (2015, October). Artificial neural networks are changing the world. What are they? Retrieved August 12, 2022, from https://www.extremetech.com/extreme/215170-artificial-neural-networks-are-changing-the-world-what-are-they

Gray, A., & Markel, J. (1976). Distance measures for speech processing. IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(5), 380-391.

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time Fourier transform. IEEE Transactions on acoustics, speech, and signal processing, 32(2), 236-243.

Gürel, A., & Arslan, L. M. (2008). Konuşma tanıma için insan-makine karşılaştırması. Dilbilim Araştırmaları Dergisi.

Kamil, H. O. (2017). Speech sound coding using linear predictive coding (LPC). Journal of Information, Communication, and Intelligence Systems (JICIS), 3(1).

Han, W., Chan, C. F., Choy, C. S., & Pun, K. P. (2006, May). An efficient MFCC extraction method in speech recognition. In 2006 IEEE International Symposium on Circuits and Systems (ISCAS).

Han, Z., & Wang, X. (2019). A signal period detection algorithm based on morphological self-complementary Top-Hat transform and AMDF. Information, 10(1), 24.

Heffner, H. E., & Heffner, R. S. (2007). Hearing ranges of laboratory animals. Journal of the American Association for Laboratory Animal Science, 46(1), 20-22.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. the Journal of the Acoustical Society of America, 87(4), 1738-1752.

Hermansky, H., Morgan, N., Bayya, A., & Kohn, P. (1991, December). RASTA-PLP speech analysis. In Proc. IEEE Int'l Conf. Acoustics, speech and signal processing (Vol. 1), 121-124.

Hillenbrand, J. (2016). The Physics of Sound. Retrieved June 28, 2022, from https://homepages.wmich.edu/~hillenbr/206/ac.pdf

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine, 29(6), 82-97.

Hoque, M. (2020). Phonetics,Branches of Phonetics. Language and Linguistics. 1.

Hui, J. (2019, September). Speech Recognition GMM, HMM. Retrieved June 28, 2022, from https://jonathan-hui.medium.com/speech-recognition-gmm-hmm-8bb5eff8b196

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. IEEE Transactions on acoustics, speech, and signal processing, 23(1), 57-72.

Ittichaichareon, C., Suksri, S., & Yingthawornsuk, T. (2012, July). Speech recognition using MFCC. In International conference on computer graphics, simulation and modeling (Vol. 9), 135-138.

İtap, M. (2022). Speech recognition application. https://github.com/mitap45/speech-reconition, Erişim tarihi: 09.10.2022

Jelinek, F., Bahl, L., & Mercer, R. (1975). Design of a linguistic statistical decoder for the recognition of continuous speech. IEEE Transactions on Information Theory, 21(3), 250-256.

Joshi, S. C., & Cheeran, A. N. (2014). MATLAB based feature extraction using Mel frequency cepstrum coefficients for automatic speech recognition. International Journal of Science, Engineering and Technology Research (IJSETR), 3(6), 1820-1823.

Joy, S., & Savitha, U., (2015). Speech Analysis in Time and Frequency Domain, Internatıonal Journal of Engıneerıng Research & Technology (IJERT) ICNTE, 3(1).

Juang, B. H., Levinson, S., & Sondhi, M. (1986). Maximum likelihood estimation for multivariate mixture observations of markov chains (corresp.). IEEE Transactions on Information Theory, 32(2), 307-309.

Juang, B. H., & Rabiner, L. R. (1991). Hidden Markov models for speech recognition. Technometrics, 33(3), 251-272.

Juang, B. H., & Rabiner, L. R. (2005). Automatic speech recognition–a brief history of the technology development. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara, 1, 67.

Jurafsky, D. & Martin, J. H. (2021, Aralık). N-gram language models. Stanford University. Retrieved June 10, 2022, from https://web.stanford.edu/~jurafsky/slp3/3.pdf

Këpuska, V. Z., & Elharati, H. A. (2015). Robust speech recognition system using conventional and hybrid features of MFCC, LPCC, PLP, RASTA-PLP and hidden Markov model classifier in noisy conditions. Journal of Computer and Communications, 3(06), 1.

Kim, H. S. (n.d.). Linear predictive coding is all-pole resonance modeling. Center for Computer Research in Music and Acoustics, Stanford University.

Koçer, H. E., & Ahmed, M. C. (2019). Turkish Speech recognition using Mel-frequency cepstral coefficients (MFCC) and Hidden Markov Model (HMM). Veri Bilimi, 2(2), 39-44.

Koni, S. S. (2020, Haziran). Analog to Digital Conversion of Speech Signal. Retrieved May 13, 2022, from https://courseware.cutm.ac.in/wp-content/uploads/2020/06/ADC.pdf

Kunigami, G. (2021, May). Linear Predictive Coding in Python. Retrieved August 14, 2022, from https://www.kuniga.me/blog/2021/05/13/lpc-in-python.html

Kurtkaya, M. (2021, March). Turkish speech command dataset. Retrieved August 07, 2022, from https://www.kaggle.com/datasets/muratkurtkaya/turkish-speech-command-dataset

Kurzekar, P. K., Deshmukh, R. R., Waghmare, V. B., & Shrishrimal, P. P. (2014). A comparative study of feature extraction techniques for speech recognition system. International Journal of Innovative Research in Science, Engineering and Technology, 3(12), 18006-18016.

Kutlugün, M. A., & Şirin, Y. (2018, May). Turkish meaningful text generation with class based n-gram model. In 2018 26th Signal Processing and Communications Applications Conference (SIU).

Landau, H. J. (1967). Sampling, data transmission, and the Nyquist rate. Proceedings of the IEEE, 55(10), 1701-1706.

Lee, C., Hyun, D., Choi, E., Go, J., & Lee, C. (2003). Optimizing feature extraction for speech recognition. IEEE Transactions on Speech and audio processing, 11(1), 80-87.

Lee, K. F. (1988). On large-vocabulary speaker-independent continuous speech recognition. Speech communication, 7(4), 375-379.

Lee, W. G. (2018, June). RASTA-PLP and MFCC tool based rasta-mat. Retrieved September 15, 2022, from https://github.com/mystlee/rasta_py

Levinson, S. E., Rabiner, L. R., & Sondhi, M. M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. Bell System Technical Journal, 62(4), 1035-1074.

Librosa. (n.d.). MFCC Feature Extraction Method. Retrieved August 21, 2022, from https://librosa.org/doc/main/generated/librosa.feature.mfcc.html

Lippmann, R. P. (1990). Review of Neural Networks for Speech Recognition, Readings in Speech Recognition, A. Waibel and K. F. Lee, Editors, Morgan Kaufmann Publishers, 374-392

Lu, W. K., & Zhang, Q. (2009). Deconvolutive short-time Fourier transform spectrogram. IEEE Signal Processing Letters, 16(7), 576-579.

Magsi, H., Sodhro, A. H., Chachar, F. A., & Abro, S. A. K. (2018, March). Analysis of signal noise reduction by using filters. In 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET).

Mannell, R. (2020, June). The Waveforms of Speech. Retrieved August 11, 2022, from https://www.mq.edu.au/about/about-the-university/our-faculties/medicine-and-health-sciences/departments-and-centres/department-of-linguistics/our-research/phonetics-and-phonology/speech/acoustics/speech-waveforms/the-waveforms-of-speech

Marshall, D. (2001). The Discrete Cosine Transform (DCT). Retrieved September 07, 2022, from https://users.cs.cf.ac.uk/Dave.Marshall/Multimedia/node231.html

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. Pattern recognition and artificial intelligence, 116, 374-388.

Mikayilova, M. (t.y.). Gestalt Etkisi ve Akustik. https://avys.omu.edu.tr/storage/app/public/mihriban.mikayilova/132784/DERS%20NOTU%20Hafta%208.%20GESTALT%20ETK%C4%B0S%C4%B0%20VE%20AKUST%C4%B0K%20.pdf, Erişim tarihi: 16.06.2022

Mohan, B. J. (2014, January). Speech recognition using MFCC and DTW. In 2014 international conference on advances in electrical engineering (ICAEE).

Molau, S., Pitz, M., Schluter, R., & Ney, H. (2001, May). Computing mel-frequency cepstral coefficients on the power spectrum. In 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (cat. No. 01CH37221) (Vol. 1), 73-76.

Murveit, H., Cohen, M., Price, P., Baldwin, G., Weintraub, M., & Bernstein, J. (1989). SRI's DECIPHER system. In Speech and Natural Language: Proceedings of a Workshop Held at Philadelphia, Pennsylvania, February 21-23, 1989, 238-242

Nakatani, T. (2019). Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. In Proc. Interspeech, 1408-1412.

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

Nave, C. R. (2017). Equal Loudness Curves. Retrieved August 13, 2022, from http://hyperphysics.phy-astr.gsu.edu/hbase/Sound/eqloud.html

Nussbaumer, H. J. (1981). The fast Fourier transform. In Fast Fourier Transform and Convolution Algorithms, 80-111.

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

Oppenheim, A. V., & Schafer, R. W. (2004). From frequency to quefrency: A history of the cepstrum. IEEE signal processing Magazine, 21(5), 95-106.

O'Shaughnessy, D. (2008). Automatic speech recognition: History, methods and challenges. Pattern Recognition, 41(10), 2965-2979.

Oyucu, S., Sever, H., & Polat, H. (2019). Otomatik Konuşma Tanımaya Genel Bakış, Yaklaşımlar ve Zorluklar: Türkçe Konuşma Tanımanın Gelecekteki Yolu. Gazi University Journal of Science Part C: Design and Technology, 7(4), 834-854.

Özseven, T. (2019). Konuşma Tabanlı Duygu Tanımada Ön İşleme ve Öznitelik Seçim Yöntemlerinin Etkisi. Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, 10(1), 99-112.

Patil, P. B. (1998). Multilayered network for LPC based speech recognition. IEEE Transactions on Consumer Electronics, 44(2), 435-438.

Peleg, N. (2010, February). Speech Signal Basics. Retrieved August 21, 2022, from http://cs.haifa.ac.il/~nimrod/Compression/Speech/S1Basics2010.pdf

Penn, G. (2010). What is sound. Retrieved August 28, 2022, from http://www.cs.toronto.edu/~gpenn/csc401/soundASR.pdf, Erişim tarihi:15.07.2022

PwC. (2018). Consumer Inteligence Series voice assistants survey. https://www.pwc.com/us/en/advisory-services/publications/consumer-intelligence-series/voice-assistants.pdf, Erişim tarihi: 03.05.2022

Rabiner, L., Levinson, S., Rosenberg, A., & Wilpon, J. A. Y. G. (1979). Speaker-independent recognition of isolated words using clustering techniques. IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(4), 336-349.

Ravikumar, K. M., Rajagopal, R., & Nagaraj, H. C. (2009, June). An approach for objective assessment of stuttered speech using MFCC. In The international congress for global science and technology, 19-24.

Sak, H., Senior, A., Rao, K., & Beaufays, F. (2015). Fast and accurate recurrent neural network acoustic models for speech recognition. arXiv preprint arXiv:1507.06947.

Sakai, T. (1962). The phonetic typewriter. In Proc. IFIP Congress 62, 445-450.

Schwartz, B. L. & Krantz, J. H. (2015, January). Steven's Power Law. Sensation & Perception Book. Retrieved October 13, 2022, from https://isle.hanover.edu/Ch02Methods/Ch02PowerLaw_evt.html

Schwartz, R., Barry, C., Chow, Y. L., Deft, A., Feng, M. W., Kimball, O., ... & Vandegrift, J. (1989). The BBN BYBLOS continuous speech recognition system. In Speech and Natural Language: Proceedings of a Workshop Held at Philadelphia, Pennsylvania, February 21-23, 1989, 94-99

Shannon, B. J., & Paliwal, K. K. (2003, November). A comparative study of filter bank spacing for speech recognition. In Microelectronic engineering research conference (Vol. 41).

Shrawankar, U., & Thakare, V. M. (2013). Techniques for feature extraction in speech recognition system: A comparative study. arXiv preprint arXiv:1305.1145.

Smith, J. O. (2007). Introduction to Digital Filters with Audio Applications. Retrieved October 04, 2022, from https://ccrma.stanford.edu/~jos/filters/What_Filter.html

Solmaz, S. (2017, Ağustos). Canlıların Duyduğu Ses Frekans Aralıkları. https://www.sercansolmaz.com/canlıların-duyduğu-ses-frekans-aralıkları-9cfe6c11d12a, Erişim tarihi: 26.07.2022

Suzuki, J. (1961). Recognition of Japanese vowels. Journal of the Radio Research Laboratory, 8(37).

Şeker, A., Diri, B., & Balık, H. H. (2017). Derin öğrenme yöntemleri ve uygulamaları hakkında bir inceleme. Gazi Mühendislik Bilimleri Dergisi, 3(3), 47-64.

Tak, R. N., Agrawal, D. M., & Patil, H. A. (2017, December). Novel phase encoded mel filterbank energies for environmental sound classification. In International Conference on Pattern Recognition and Machine Intelligence, 317-325.

Taşova, O. (2011). Yapay sinir ağları ile yüz tanıma (Doctoral dissertation, DEÜ Fen Bilimleri Enstitüsü).

Türk, E. & Kiani, F. (2019). Yapay Sinir Ağları ile Talep Tahmini Yapma: Beyaz Eşya Üretim Planlama Örneği. İstanbul Sabahattin Zaim Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 1(1), 30-37.

Velardo, V. (2020). Deep Learning Audio Application From Design to Deployment. Retrieved August 04, 2022, from https://github.com/musikalkemist/Deep-Learning-Audio-Application-From-Design-to-Deployment

Wang, D., Wang, X., & Lv, S. (2019). An overview of end-to-end automatic speech recognition. Symmetry, 11(8), 1018.

Wang, J. C., Wang, J. F., & Weng, Y. S. (2002). Chip design of MFCC extraction for speech recognition. Integration, 32(1-2), 111-131.

Winograd, S. (1978). On computing the discrete Fourier transform. Mathematics of computation, 32(141), 175-199.

Yalçın, N. (2008). Konuşma Tanima Teorisi ve Teknikleri. Kastamonu Eğitim Dergisi, 16(1), 249-266.

Yang, C. H. H., Qi, J., Chen, S. Y. C., Chen, P. Y., Siniscalchi, S. M., Ma, X., & Lee, C. H. (2021, June). Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6523-6527.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., ... & Woodland, P. (2002). The HTK book. Cambridge university engineering department, 3(175), 12.

Zelasko, P. (2016). How does GMM-HMM model work in ASR? Retrieved August 24, 2022, from https://www.quora.com/How-does-GMM-HMM-model-work-in-ASR/answer/Piotr-%C5%BBelasko

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı ve Soyadı : Muratcan İtap

E-Posta Adresi : mitap45@gmail.com

İş Deneyimi :

Unvan	**Görev Yeri**	**Yıl**
Yazılım Mühendisi	Epam Sistemleri, Türkiye	2017
Yazılım Mühendisi	Enuygun, enuygun.com	2018
Üst Kurul Uzman Yardımcısı	Radyo ve Televizyon Üst Kurulu	2019

Akademik Bilgiler

Öğrenim Durumu:

Derece	**Bölüm/Program**	**Üniversite**	**Yıl**
Lisans	Bilgisayar Mühendisliği	Pamukkale Üniversitesi	2012-2017
Yüksek lisans	Bilgisayar Mühendisliği	Hacettepe Üniversitesi	2022

Yayınlar:

RTÜK