T.C.
KASTAMONU ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
MALZEME BİLİMİ VE MÜHENDİSLİĞİ ANA BİLİM DALI

TEKRARLAYAN SİNİR AĞLARINI KULLANARAK KONUŞMA TANIMA

MUSTAFA AHMED ELBERRI

YÜKSEK LİSANS TEZİ

DR. ÖĞR. ÜYESİ ÜMİT TOKEŞER

HAZIRAN - 2020
KASTAMONU

TEZ ONAYI

Mustafa Ahmed ELBERRI tarafından hazırlanan "**TEKRARLAYAN SİNİR AĞLARINI KULLANARAK KONUŞMA TANIMA**" adlı tez çalışmasının savunma sınavı **26.06.2020** tarihinde yapılmış olup aşağıda verilen jüri tarafından oy birliği / oy çokluğu ile Kastamonu Üniversitesi Fen Bilimleri Enstitüsü **MATEMATİK BÖLÜMÜ YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir

Danışman	Dr. Öğr. Üyesi Ümit TOKEŞER Kastamonu Üniversitesi	.....
Jüri Üyesi	Dr. Öğr. Üyesi Cevat RAHEBİ Altınbaş Üniversitesi	.....
Jüri Üyesi	Doç Dr. Üyesi Oytun Emre SAKICI Kastamonu Üniversitesi	.....

Jüri üyeleri tarafından kabul edilmiş olan bu tez Kastamonu Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulunca onanmıştır.

Enstitüsü Müdürü Prof. Dr. İzzet ŞENER .....

TAAHHÜTNAME

Bu tezin tasarımı, hazırlanması, yürütülmesi, araştırmalarının yapılması ve bulgularının analizlerinde bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu; ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını, bilimsel etiğe uygun olarak kaynak gösterildiğini bildirir ve taahhüt ederim.

Mustafa Ahmed ELBERRI

ÖZET

YÜKSEK LİSANS TEZI

TEKRARLAYAN SİNİR AĞLARINI KULLANARAK KONUŞMA TANIMA

MUSTAFA AHMED ELBERRI

KASTAMONU ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ
MALZEME BİLİMİ VE MÜHENDİSLİĞİ ANA BİLİM DALI
DANIŞMAN:DR. ÖĞR. ÜYESİ ÜMİT TOKEŞER

Konuşma tanıma yada otomatik konuşma tanıma bilgisayar biliminin, mühendisliğin ve bilişimsel dilbilimin ortak bir uygulama alanıdır. Bu alan, özellikle bilgisayarlar için, konuşulan dilden otomatik olarak veri elde etmeyi sağlayan süreçlerin araştırılmasıyla ve geliştirilmesiyle ilgilenir. Konuşma tanımanın bir biyometrik kişisel tanıma yöntemi olan ses veya konuşmacıyı tanıma ile karıştırılmaması gerekir. Ancak her iki uygulamanın realizasyonunda benzerlikler söz konusudur. Bu tezde insanların konuşma sinyallerinden kelimeleri tanımada derin öğrenme yöntemi kullanılmıştır. Önerilen yöntemin yürütülmesinde Python kullanılır ve bu adım adım gerçekleştirilir. Çalışmamız neticesinde, konuşma tanıma için CNN'in (evrişimli sinir ağları) oldukça başarılı bir araç olduğu görülmüştür, bu durum daha fazla etiketin söz olması halinde daha sofistike bir mimari inşası yönünde cesaret verici olmuştur. Daha farklı çözümleri denemek için Kaggle'ın TensorFlow Speech Recognition Challenge platform kullanılmıştır.

ANAHTAR KELİMELER: Konuşma tanıma, yapay sinir ağları, derin öğrenme.

Haziran 2020, 58 Sayfa,
Bilim Dalı: 91

ABSTRACT

MSC. THESIS

SPEECH RECOGNİTİON WİTH DEEP LEARNİNG NEURAL NETWORKS

MUSTAFA AHMED ELBERRI

KASTAMONU UNIVERSITY INSTITUTE OF SCIENCE
DEPAERTMENT OF ENVIRONMENTAL ENGINEERING
SUPERVISOR: DR. ÜYESİ ÜMİT TOKEŞER

The speech recognition or automatic speech recognition is a branch of applied computer science, the engineering and computational linguistics. It is concerned with the investigation and development of processes that make automatic, especially computers, the spoken language available for automatic data acquisition. The speech recognition is to be distinguished from the voice or speaker identification, a biometric method of personal identification. However, the realizations of these procedures are similar. In this thesis the deep learning method is used to recognize words from human speech signals. For implementation of proposed method Python is used and presented in a step-by-step manner. The CNN turned out to be a really good tool in speech recognition and encourage to build more complex architecture in case of more labels. The TensorFlow Speech Recognition Challenge Kaggle is used to check out more solutions.

KEYWORDS: Speech recognition, artificial neural network, deep learning

June 2020, 58 Page,
Science Code:91

TEŞEKKÜR

Öncelikle danışmanım Dr. Ümit TOKEŞER'e bu araştırma boyu süren rehberliği ve yol göstericiliği için teşekkür ederim. Matematik Bölümü'ndeki diğer hocalarıma ve asistanlara da, bu araştırma bağlamında ortaya çıkan pek çok gereksinime sundukları çözümler için minnettarım. Destekleri için Kastamonu Üniversitesi Master Programındaki diğer arkadaşlarıma ve Kastamonu'da bulunan Libya toplumuna da teşekkür etmek isterim. Son olarak; bu çalışmayı tamamlayabilmek için ihtiyaç duyduğum özgüveni borçlu olduğum eşime ve aileme benden hiç bir zaman esirgemedikleri manevi destekleri için minnettarım. Bu çalışmada elde ettiğimiz sonuçların bu alana ilgili insanlara faydalı olmasını ve gelecekte yapılacak yeni çalışmalara katkı sağlamasını ümit ediyorum.

MUSTAFA AHMED ELBERRI

Kastamonu, 2020

İÇİNDEKİLER

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

TABLOLAR DİZİNİ

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

Simgeler

ZCR	**:** Sıfırdan Geçiş Sayısı (Zero Crossing Rate)
AAC	**:** İleri Düzeyde Ses Kodlaması (Advanced Audio Coding)
DSP	**:** Dijital Sinyal İşleme (Digital Signal Processing)
ASR	**:** Otomatik Konuşma Tanıma (Automatic Speech Recognition)
SNR	**:** Sinyal-Gürültü Oranı (Signal-To-Noise Ratio)
DFT	**:** Ayrık Fourier Dönüşümü (Discrete Fourier Transform)
FFT	**:** Hızlı Fourier Dönüşümü (Fast Fourier Transform)
FFT	**:** Lineer/Doğrusal Tahmin (Linear Prediction)
DWT	**:** Kesikli Dalgacık Dönüşümü (Discrete Wavelet Transform)
MRMR	**:** Minimum Artıklık Maksimum İlgililik (Minimum Redundancy Maximum Relevance)
DTW	**:** Dinamik Zaman Bükmesi (Dynamic Time Warping)
HMM	**:** Saklı Markov Modeli (Hidden Markov Modeling)
NN	**:** Sinir Ağı/Ağları (Neural Networks)
SVM	**:** Destek Vektör Makinesi (Support Vector Machine)
RF	**:** Rastfele Orman (Random Forest)
MFCC	**:** Mel-Frekansı Cepstral Katsayıları (Mel-Frequency Cepstral Coefficients)
STFT	**:** Kısa Süreli Fourier Dönüşümü (Short-time Fourier Transform)
CNN	**:** Evrişimli Sinir Ağı/Ağları (Convolutional Neural Network)
DCNN	**:** Derin Evrişimli Sinir Ağı/Ağları (Deep Convolutional Neural Network)

1. GİRİŞ

1.1 Arka Plan

Bir ses dosyası, en yaygın türü konuşma ve müzik olmak üzere birçok kaynaktan gelen pek çok ses bileşeninden oluşur. Bu iki ana tür sadece birincil bileşenlerdir ve buna hemen her zaman var olan çoklu gürültü kaynakları eklenmelidir.

Daha fazla ayrıntıya girmeden önce, özellikle konuşma ve müzik üzerine belirli kavram ve tanımların netleştirilmesi gereklidir.

1.2 Tarihsel Gelişim

Konuşma tanıma sistemleri üzerine araştırmalar 1960'larda başladı, ancak o zaman büyük ölçüde başarısız oldu: farklı özel şirketler tarafından geliştirilen sistemler, ancak birkaç düzine bireysel kelimenin tanınmasını laboratuvar koşullarında olmak şartıyla imkan veriyordu. Bunun bir nedeni, oldukça yeni sayılabileck bu araştırma alanındaki sınırlı bilgiyse, bir diğer nedeni sınırlı teknik imkanlardı.

Bu alanda gelişme ancak 1980'lerin ortalarında sağlanabildi. Bu sıralarda, sesteş sözcüklerin bağlam kontrolü ile ayırt edilebileceği keşfedilmişti. Belirli kelime kombinasyonlarının sıklığı hakkında istatistikler oluşturarak ve bu istatistikler çerçevesinde bir kısım değerlendirmeler yaparak, hangi kelimelerin seslerinin birbirine yakın olduğuna veya bu benzer kelimelerin hangisinin kastedildiğine karar verilebilir. Bu istatistikler zamanla tüm konuşma tanıma sistemlerinin önemli bir parçası haline geldi. IBM yaklaşık 5000 ayrı İngilizce kelimeyi tanıyabilen ilk konuşma tanıma sistemini 1984 yılında tanıttı. Ancak, sistemin tanıma işlemini gerçekleştirebilmesi için bir ana bilgisayarda birkaç dakika hesaplama süresine ihtiyacı vardı. Bununla birlikte, Dragon Systems tarafından geliştirilen bir sistem daha gelişmişti: taşınabilir bir kişisel bilgisayarda kullanılabiliyordu.

1988-1993 yılları arasında yürütülen SUNDIAL Avrupa projesi (Peckham, 1991), Almanca tren tarifelerinin sesli tanınmasını sağladı. SUNDIAL'da ayrıca

konuşma tanıma değerlendirme göstergeleri de incelendi (Danieli ve Gerbino, 1995)(Zeqiri vd., 2003)(Charpentier vd., 1995).

IBM 1991 yılında, CeBIT fuarında 20 000 ila 30 000 Almanca kelime tanıyabilen bir konuşma tanıma sistemi tanıttı. Ancak, TANGORA 4 adı verilen bu sistemin sunumu özel olarak yalıtılmış bir odada gerçekleşmek zorundaydı, aksi takdirde fuardan gelen gürültü sistemi işlevsiz hale getirebilirdi.

1993 yılı sonunda IBM, piyasa için geliştirilen ilk konuşma tanıma sistemini duyurdu: IBM Kişisel Dikte Sistemi olarak adlandırılan sistem normal PC'lerde çalışıyordu ve maliyeti 1000 dolardan daha azdı. CeBIT 1994'te IBM VoiceType dikte sistemi adı altında tanıtıldığında, ziyaretçiler ve basından büyük ilgi gördü.

1997'de PC son kullanıcısı için hem ViaVoice (IBM VoiceType'ın devamı) hem de Dragon NaturallySpeaking yazılımının 1.0 sürümü ortaya çıktı. 1998 yılında Philips Konuşma Tanıma Sistemleri, PC son kullanıcıları için bir konuşma tanıma sistemi olan FreeSpeech 98'i piyasaya sürdü, bu sistemin kontrollü şirket içi dijital dikte aygıtı SpeechMike'a uyarlanabiliyordu, ancak bu sistem ikinci sürüm olan FreeSpeech 2000'den sonra üretimden kaldırıldı. 2004 yılında IBM'in, konuşma tanıma uygulamalarının bir kısmını açık kaynak olarak yayınlaması ortalığı ayağa kaldırdı. Sektör uzmanları, bu açılımın aynı alanda aktif olan Microsoft'a karşı bir taktik olabileceğini dile getirdiler; Microsoft 2007'den bu yana PC işletim sistemi Windows'un ayrılmaz bir parçası olarak konuşma tanıma işlevleri sunmaktadır ve aynı uygulama Windows 10 ile daha da gelişmiştir.

IBM ViaVoice'in geliştirilmesine devam edilmezken, Dragon NaturallySpeaking, üçüncü taraf olarak Windows PC'ler için konuşmacıya bağımlı konuşma tanımada şu anda en yaygın yazılım konumuna geldi, bu yazılım 2005'ten beri Nuance Communucations tarafından üretilmekte ve pazarlanmaktadır.

Nuance 2008 yılında Philips Speech Recognition Systems-Vienna'yı satın alınca, sağlık sektöründe özellikle popüler hale gelen SpeechMagic yazılım geliştirme kitinin (SDK) de haklarını almış oldu. Apple'ın iMac kişisel bilgisayarları için MacSpeech, 2006'dan beri Philips bileşenlerine dayanan iListen adı altında üçüncü taraf konuşma

tanıma yazılımı satmaktaydı. 2008'de MacSpeech'in Dragon Dictate'de Nuance Communications 2010 tarafından satın alınmasından sonra bu durum değişti, MacSpeech Dictate tarafından Dragon NaturallySpeaking'in temel bileşenleri kullanılarak yapıldı ve yeniden adlandırıldı (Sürüm 2.0 - Sürüm 3.0 ise 2012'den beri mevcuttur).

Siri Inc. 2007 yılında kurulmuş ve Nisan 2010'da Apple tarafından satın alınmıştır. Ekim 2011'de Apple, iPhone 4s için Siri konuşma tanıma yazılımını tanıtmıştır; bu program, doğal olarak konuşulan dili (Apple sunucularını kullanarak) tanımak ve işlemek için kullanılır ve kişisel asistanın işlevlerini yerine getirir.

1.3 Güncel Durum

Günümüz itibariyle, konuşma tanıma sistemleri arasında kabaca şöyle bir ayrım yapılabilir:

- Konuşan kişiden bağımsız olarak konuşma tanıma
- Konuşan kişiye bağlı olarak konuşma tanıma

“Konuşan kişiden bağımsız” konuşma tanımanın karakteristik özelliği, kullanıcının bir eğitim aşamasına gerek olmadan konuşma tanımayı kullanmaya hemen başlayabilmesidir. Ancak, kelime haznesi birkaç bin kelimeyle sınırlıdır.

“Konuşan kişiye bağlı” konuşma tanıyıcıların, kullanımdan önce veya kullanım sırasında kullanıcı tarafından kendi telaffuz özellikleri konusunda eğitilmesi gereklidir. Bu tanıyıcıların işlevselliğinde önemli bir faktör, konuşmacıya bağlı olarak optimum bir sonuç elde etmek için sistemle kullanıcının bireysel etkileşim olasılığıdır (kullanıcının kendi terminolojisi, kısaltmalar, vb.). Bu nedenle, sık değişen kullanıcılara sahip uygulamalarda (ör. çağrı merkezleri) kullanımı mantıklı değildir. Buna karşılık, kelime haznesi, konuşan kişiden bağımsız tanıyıcılardan çok daha büyüktür. Mevcut sistemler 300 000'den fazla kelime içermektedir. Konuşma tanıyıclar arasında aşağıdaki kritere göre de bir ayrım yapılabilir:

- Ön-uç sistemleri

- Geri-uç sistemleri.

Ön uç sistemlerinde, konuşma dilinin işlenmesi ve metne dönüştürülmesi hemen gerçekleşir, böylece kullanıcı belirgin bir zaman gecikmesi olmaksızın metni okuyabilir. Uygulama, kullanıcının bilgisayarında veya bulut tabanlı olarak yürütülebilir. En yüksek düzeyde tanıma, kullanıcı ve sistem arasındaki doğrudan etkileşim yoluyla elde edilir. Sistemin komutlar aracılığıyla kontrolü ve gerçek zamanlı yardım sistemleri gibi diğer bileşenlerin eklemlenmesi de mümkündür. *Arka uç sistemlerde* ise uygulama bir gecikme ile yürütülür. Bu genellikle uzak bir sunucuda meydana gelir. Metin ancak gecikmeli olarak elde edilmektedir. Bu sistemler tıp alanında hala yaygındır. Konuşmacı ve tanıma sonucu arasında doğrudan bir etkileşim olmadığından, yalnızca kullanıcının konuşma tanıma deneyimi varsa, mükemmel bir kalite elde edilebilir.

Konuşma tanımanın gelişimi çok hızlı bir seyir izlemektedir. Günümüz itibariyle konuşma tanıma sistemleri akıllı telefonlarda da yaygın şekilde kullanılmaktadır; örneğin Siri, Google now, Cortana ve Samsung'un Voice uygulamaları. Mevcut konuşma tanıma sistemleri artık eğitilmeye ihtiyaç duymamaktadır. Sistemin esnekliği, günlük dilin haricinde de yüksek derecede doğruluk için çok önemlidir. Yüksek düzeyli taleplere de cevap verebilmek açısından ve kullanıcıya bağlı sonuçları daha da geliştirebilmek amacıyla profesyonel sistemler belirli bir listeye göre ses testleri yapma imkanı sunmaktadır (Arnold vd., 1994).

1.3.1 Dudak Okuma

Tanımanın doğruluğunu daha da arttırmak için, bazen bir video kamera ile konuşmacının yüz ifadesi kayda alınmaya ve dudak hareketleri okunmaya çalışılır. Bu sonuçlar akustik algılama sonuçlarıyla birleştirildiğinde, özellikle gürültünün mevcut olduğu kayıtlarda çok daha yüksek bir algılama oranı elde etmek mümkündür.

Bu, konuşma tanımasına dair önemli bir gözleme karşılık gelir: Harry McGurk 1976'da insanların konuşma dilini anlamlandırmada karşılarındakii kişinin dudak hareketlerini de kullandıklarını keşfetti (McGurk etkisi).

1.3.2 Konuşma

İnsanlar arası iletişim genellikle iki kişi arasında bir diyalog olduğundan, konuşma tanıma ile çoğunlukla konuşma sentezi bağlantılıdır. Bu şekilde, sistem kullanıcısına konuşma tanıma işleminin başarısı ve gerçekleştirilmiş olabilecek eylemler hakkında akustik geri bildirim verilebilir. Aynı şekilde, kullanıcıdan seslendirmeyi tekrar yapması da istenebilir.

1.4 Problemin Ortaya Konulması

Bir konuşma tanıma sisteminin nasıl çalıştığını anlayabilmek için, ilk olarak üstesinden gelinmesi gereken zorlukların anlaşılması gerekir.

1.4.1 Ayrık ve Sürekli Konuşma

Günlük konuşmalarda, sözcükler cümle içinde aralarında belirgin bir duraklama olmaksızın telaffuz edilirler. Kişisel anlamda, kendiniz sezgisel yolla kelimeler arasındaki geçişleri kavrayabilirsiniz—ilk konuşma tanıma sistemleri bunu yapamıyordu. Bu sistemler kelimeler arasında yapay duraklamalar içeren ayrık (kesintiye uğramış) bir dile ihtiyaç duyuyorlardı.

Buna karşın, modern sistemler sürekli (akıcı) dili anlayabilecek durumdadır.

1.4.2 Ayrık Konuşma

Örnek olarak, “Özgür Ansiklopedi”, cümlesi ayrık olarak telaffuz edilecek olursa:

Ayrık konuşmada kelimeler arası duraklamalar net bir şekilde gösterilir, öyleki, bu duraklamalar ansiklopedi kellimesindeki heceler arası duraklamalardan çok daha net ve uzundur.

1.4.3 Sürekli Konuşma

Aynı şekilde, örnek olarak, "Özgür Ansiklopedi", cümlesi sürekli olarak telaffuz edilecek olursa:

Sürekli konuşmada her bir kelime bir diğeriyle bütünleşik gibidir, herhangi bir duraksama neredeyse tanımlanamaz.

1.4.4 Kelime Haznesi

Gramer fonksiyonuna bağlı olarak hem kelimelere eklenen ekler, hem de zaman çekimleri yoluyla, aynı kelime köklerinden pek çok kelime formunun türetilmesi mümkündür. Bu durum kelime dağarcığı için önemlidir, çünkü tüm kelime formları konuşma tanımada bağımsız kelimeler olarak düşünülmelidir.

Sözlüğün boyutu büyük ölçüde söz konusu dile bağlıdır. Örneğin, ortalama yaklaşık 4 000 kelimeyle konuşulan Almanca'nın, ortalama yaklaşık 800 kelimeyle konuşulan İngilizce'ye göre çok daha büyük bir kelime haznesi vardır. Buna ek olarak, Almanca'daki türetmeler (ekler, çekimler) ortalama yaklaşık on farklı kelime formunu netice verirken İngilizce'deki türetmeler ortalama yaklaşık dört farklı kelime formunu netice verir.

1.4.5 Eşsesler

Birçok dilde, farklı anlamları olan ancak aynı şekilde telaffuz edilen kelimeler veya kelime formları vardır. Örneğin İngilizce'de "sea" ve "see" sözcüklerinin telaffuzu aynı olmakla beraber aralarında hiçbir anlam ilişkisi yoktur. Bu tür kelimelere sesteş veya eşsesli kelimeler (İng. homophone) denir. Bir konuşma tanıma sisteminin, insanların aksine, dünya hakkında hiçbir bilgisi yoktur, ve anlam bütünlüğü gözeterek çeşitli seçenekler arasında ayrım yapamaz. Büyük veya küçük harf kullanımından kaynaklanacak sorunlar da bu kapsama girer.

1.4.6 Biçimleyiciler (Formant)

İngilizce'de formant olarak tanımlanan biçimleyicilerin konumu akustik seviyede önemli bir rol oynar: konuşulan sesli harflerin frekans bileşenleri genellikle formant (biçimleyici) olarak adlandırılan belirli farklı frekanslara odaklanır. En düşük iki formant, sesli harflerin ayırt edilmesi için özellikle önemlidir: düşük frekans 200 ila 800 Hertz aralığındadır, yüksek frekans ise 800 ila 2 400 Hertz aralığındadır. Tek tek ünlüler bu frekansların yeri ile ayırt edilebilir.

Ünsüzlerin tanınması nispeten daha zordur; bireysel ünsüzler (patlamalı ünsüzler), ancak aşağıdaki örnekte gösterildiği gibi, komşu seslere geçişle belirlenebilir:

İngilizce'de konuşmak anlamına gelen *speaking* sözcüğü içinde, ünsüz p (daha açıkça söylemek gerekirse: fonem p'nin kapanma aşaması) aslında sessizdir ve ancak diğer sesli harflere geçişle tanındığı görülebilir—bu nedenle kaldırılması sesli bir fark yaratmaz.

Diğer ünsüzler karakteristik spektral kalıplarla, yani bahsettiğimiz frekans aralıklarıyla tanınabilir. Örneğin, *s* ve *f* sesi, daha yüksek frekans bantlarında yüksek oranda enerji ile karakterize edilir. Bu iki sesin ayrımının yapılabilmesi için gerekli olan frekans aralığının telefon şebekelerinde iletilen spektral aralığın (yaklaşık 3,4 kHz'e kadar) dışında olması dikkat çekicidir. Bu durum, harfleri kodlamaksızın telefonda hecelemenin neden iki kişi arasındaki iletişimde son derece yanlış anlamaya ve hataya açık olduğunu açıklar.

1.4.7 Lehçeler ve Topluluk Dilleri

Bir konuşma tanıma programı üst düzey bir dil için son derece iyi şekilde ayarlanmış olsa bile, bu durum tanıma programının söz konusu dilin her biçimini anlayabileceği anlamına gelmez. Özellikle lehçeler ve topluluk dilleri (sociolect) söz konusu olduğunda, bu tür programların sınırlarının aşıldığı, yani konuşma tanımanın doğru şekilde gerçekleştirilemediği durumlar ortaya çıkabilir. İnsanlar genellikle karşılarındaki kişinin öncesinde aşina olmadıkları lehçesine hızlı bir şekilde uyum

sağlayabilirler—tanıma yazılımları bunu kolayca yapamaz. Öncesinde karmaşık süreçler izlenerek lehçelerin programa öğretilmesi gerekir.

Ayrıca, kelime anlamları zamana bağlı olarak ve bölgesel olarak değişebilir. Örneğin Bavyera ve Berlin bölgelerinde, pankek söz konusu olduğunda farklı tatlılar akla gelir. Tanıma programlarının şu an yapamadığı bir diğer iş olarak; bir kişinin kültürel arka planına dair bilgi, bu tür yanlış anlamaları önlemeyi daha kolay hale getirir.

1.4.8 Stratejiler ve Komünikasyon Problemleri

Komünikasyon esnasında anlaşılırlık sorunu varsa, insanlar doğal olarak özellikle yüksek sesle konuşma veya yanlış anlaşılmış terimleri daha ayrıntılı olarak tanımlama eğilimindedir. Ancak bu durumun, bağlamı anlamaktan ziyade anahtar kelimelere odaklanma stratejisi güden ve normal konuşma tonuna göre eğitilmiş bilgisayar yazılımı üzerinde ters etkisi olabilir.

1.5 Realizasyon

Bir konuşma tanıma sisteminde şu bileşen de bulunur: analog olan konuşma sinyallerini ayrı frekanslara ayıran bir önişleme safhası. Gerçek tanıma ise, akustik modeller, sözlükler ve dil modelleri yardımıyla gerçekleşir

1.5.1 Önişleme

Önişleme safhası, örnekleme, filtreleme, frekans alanına dönüştürme ve öznitelik vektörünün elde edilmesi gibi alt adımlardan oluşur.

1.5.1.1 Örnekleme

Tarama sırasında, analog (sürekli) olan sinyal dijitalize edilir, yani daha kolay işlenebilmesi için elektrok olarak işlenebilecek bit dizilerine ayrıştırılır.

1.5.1.2 Filtreleme

Filtrelemenin başlıca amacı, asıl sesle ortamdaki gürültüyü birbirinden ayrıştırmaktır. Bunun için sinyalin enerjisi veya sfırdan geçiş sayısı kullanılabilir.

1.5.1.3 Dönüşümler

Konuşma tanıma için elverişli olan sinyal zaman düzleminde değil, frekans alanında olan sinyaldir. Bunun için FFT (Hızlı Fourier) dönüşümü uygulanır. Sinyalde var olan frekans bileşenleri sonuçtan yani frekans spektrumundan okunabilir.

1.5.1.4 Öznitelik Vektörü

Gerçek konuşma tanıma için bir öznitelik vektörü oluşturulur. Bu vektör, dijital konuşma sinyalinden üretilen bağımsız veya birbirine bağımlı özelliklerden oluşur. Daha önce bahsedilen spektruma ek olarak, içerdiği en önemli kısımlardan biri cepstrum'dur. Öznitelik vektörleri örneğin daha önce tanımlanmış bir *metrik* kullanarak karşılaştırılabilir.

1.5.1.5 Cepstrum

Cepstrum, logaritmik spektrumun FFT'sini hesaplayarak elde edilir. Bu şekilde, spectrum içinde yer alan periyodik özellikler tanınabilir. Bunlar insanların ses yolunda ses tellerinin uyarılmasıyla ortaya çıkarlar. Ses tellerinin uyarılmasının neticesi olan periyodiklikler üst kısımlarda baskındırlar ve bu nedenle cepstrumun üst kısmında bulunurlar, alt kısım ise ses yolunun konumunu gösterir. Bu konuşma tanıma için önemlidir, bu nedenle sadece cepstrumun alt kısımları öznitelik vektörüne dahil edilir. Transfer fonksiyonu—yani z sinyalinde değişiklik, duvarlardaki yansımalar sebebiyle—zamanla değişmez, bu cepstrumun ortalama değeri ile temsil edilebilir. Bu nedenle yankıları hariç tutmak için cepstrumdan çıkarılır genellikle. Benzer şekilde, öznitelik vektörüne dahil edilebilecek olan cepstrumun ilk türevi, transfer fonksiyonunun etkisini nötürlemek için kullanılmalıdır.

1.6 Algılama

1.6.1 Saklı Markov Modelleri

Saklı Markov Modelleri (HMM) sonraki aşamalarda önemli bir rol oynayacaktır. Bu modeller giriş sinyallerine en uygun fonemleri bulmayı mümkün kılar. Bu amaçla, bir fonemin akustik modeli farklı parçalara ayrılır: başlangıç, uzunluğa ve merkez parçasının sayısına bağlı olarak ara parçalar ve son. Girdi sinyalleri bu kaydedilmiş bölümlerle karşılaştırılır ve Viterbi algoritması kullanılarak olası kombinasyonlar aranır.

Kesikli (her kelimeden sonra bir duraklama yapıldığı ayrık) dilin algılanması için, HMM içindeki bir duraklama modeliyle birlikte bir seferde bir kelimeyi hesaplamak yeterliydi. Bununla birlikte, modern bilgisayarların hesaplama kapasitesi önemli ölçüde arttığından, akıcı (sürekli) dil artık birkaç kelimeden ve aralarındaki geçişlerden oluşan daha büyük saklı Markov modelleri oluşturularak tanınabilir.

1.6.2 Sinir Ağları

Akustik model için sinir ağları bir alternatif olarak kullanılmaya çalışılmıştır. Zaman içerisinde frekans spektrumundaki değişiklikleri tespit için Zaman Gecikmeli Sinir Ağları kullanılmalıdır. Bu girişimin başlangıçta olumlu sonuçları vardı, ancak daha sonra HMM'ler lehine olarak vazgeçildi. Ancak son yıllarda bu kavram Derin Sinir Ağlarının bir parçası olarak yeniden keşfedildi. Derin öğrenmeye dayanan konuşma tanıma sistemleri, neredeyse insanlar seviyesinde tanıma oranları sağlamaktadır.

Bununla beraber, ön işlemeden elde edilen verilerin bir sinir ağı tarafından önceden sınıflandırıldığı ve ağ çıktısının saklı Markov modelleri için bir parametre olarak kullanıldığı hibrit bir yaklaşım da vardır. Bunun avantajı, HMM'lerin karmaşıklığını arttırmadan, henüz düzenlenmiş olan süreden kısa bir süre önce ve kısa bir süre sonra kullanabilmenizdir. Ek olarak, verilerin sınıflandırması ve bağlama duyarlı kompozisyon (anlamlı sözcüklerin/cümlelerin oluşumu) ayrılabilir.

1.6.3 Dil modeli

Daha sonraki aşamada dil modeli belirli kelime kombinasyonlarının olasılığını belirlemeye çalışır ve böylece yanlış veya olanaksız hipotezleri eler. Biçimsel dilbilgisi kullanan bir dilbilgisi modeli veya N-gram kullanılan bir istatistiksel model kullanılabilir.

Bir bigram veya trigram istatistiği ile, iki veya daha fazla kelimenin bir araya geldikleri kombinasyonların olasılıkları hesaplanır. Bu istatistikler çok büyük metinlerden (örnek metinler) elde edilir. Konuşma tanıma ile belirlenen her hipotez daha sonra kontrol edilir ve gerekirse olasılığının çok düşük olması durumunda reddedilir. Bu, sesteş sözcüklerin, yani özdeş telaffuza sahip farklı sözcüklerin ayırt edilmesini sağlar. Bu nedenle, her ikisi de eşit olarak söylense de "çok teşekkür ederim" kombinasyonunun olasılığı "teşekkür ederim" kombinasyonunun olasılığından daha fazladır.

Trigramlar, bigramlara kıyasla kelime kombinasyonlarının meydana gelme olasılığına ilişkin teorik olarak daha doğru tahminlere izin verir. Bununla birlikte, trigramların çıkarıldığı örnek metin veritabanları, bigramlardan önemli ölçüde daha büyük olmalıdır, çünkü üç kelimeden oluşan izin verilen tüm kelime kombinasyonlarının istatistiksel olarak anlamlı sayıda olması gerekir (yani, her biri bir kereden çok daha fazla). Dört veya daha fazla kelimeden oluşan kombinasyonlar uzun süredir kullanılmamaktadır, çünkü genel olarak yeterli sayıda tüm kelime kombinasyonlarını içeren örnek metin veritabanlarını bulmak pek mümkün değildir. Bunun bir istisnası Dragon'dur, sürüm 12'de sistemin tanıma doğruluğunu artıran pentagramlar dahi kullanılmaktadır.

Dilbilgisi kullanıldığında, çoğunlukla bağlamdan, içerikten bağımsız olarak kullanılır. Ancak, her sözcüğe dilbilgisi içindeki işlevi atanmalıdır. Bu nedenle, bu tür sistemler çoğunlukla sadece sınırlı bir kelime haznesi ve özel uygulamalar için kullanılır, ancak PC'ler için yaygın konuşma tanıma yazılımında kullanılmaz.

1.6.4 Değerlendirme

Bir konuşma tanıma sisteminin kalitesi çeşitli ölçümlerle veya farklı kriterlerle belirlenebilir. Tanıma hızına ek olarak—genellikle gerçek zamanlı bir faktör (EZF) olarak verilir—tanıma kalitesi kelime doğruluğu veya kelime tanıma oranı olarak ölçülebilir.

1.7 Kelime Hazneleri

Konuşma tanıma ile çalışmayı kolaylaştıracak profesyonel konuşma tanıma sistemlerine entegre edilme amaçlı önceden tanımlanmış kelime hazneleri vardır. Bu kelime hazneleri SpeechMagic *ConText* ve Dragon *Datapack* olarak belirlenmiştir. Kelime haznesi, konuşmacı tarafından kullanılan kelime haznesine ve dikte stiline ne kadar iyi uyarlanırsa (kelime dizilerinin sıklığı), tanıma doğruluğu o kadar yüksek olur. Konuşmacıdan bağımsız sözlüğün (teknik ve temel kelime bilgisi) yanı sıra, bir kelime haznesi ayrıca bireysel bir kelime dizisi modeli (dil modeli) içerir. Kelime haznesinde, yazılım tarafından bilinen tüm kelimeler fonetik ve yazımdır. Dolayısıyla, konuşulan bir kelime sesiyle sistem tarafından tanınır. Eğer kelimeler anlam ve yazım açısından farklılık gösterip aynı telafuza sahipse, yazılım kelime dizisi modelini kullanır. Yazılım, belirli bir kullanıcı için bir kelimenin diğerini takip etme olasılığını tanımlar. Akıllı telefonlardaki konuşma tanıma aynı metodu kullanır, ancak kullanıcının önceden tanımlanmış kelime haznesi üzerinde herhangi bir etkisi bulunmamaktadır. Yeni teknolojiler katı, depolanmış bir kelime listesi fikrinden uzaklaşmaktadır çünkü bileşik kelimeler oluşturulabilmektedir. Tüm sistemlerin ortak özelliği, sadece tek tek kelimeleri ve kelime dizilerini öğrenmesidir, bu da ilgili kullanıcıyı düzelterek elde edilir.

1.8 Tezin Amaçları

Bu tezin temel amacı özellikle DNN'yi incelemek ve Türkçe bir otomatik konuşma tanıma (ASR) programına uyarlandığında ortaya çıkan uygulama ve çalışma prensiplerini anlamaktır.

Buna ek olarak, GMM ve HMM gibi geleneksel sistemlerle DNN'in başarılarını bir araya getirmek de bir diğer hedefimizdir.

1.9 Tezden Elde Edilmesi Hedeflenen Sonuçlar

Tez çalışmamızın ilk hedeflenen sonucu, DNN algoritmalarını baz alan ve Türkçe konuşmaları tanıyan bir sistemin hayata geçirilmesidir.

Hedeflenen daha genel sonuç ise, geleneksel konuşma tanıma sistemlerine nazaran DNN'nin biraz daha iyi bir performans göstermesidir. Burada özellikle belirtilmelidir ki, her iki sistemin performanslarının da büyük ölçüde çeşitli ağ ve uygulama parametrelerine bağlı olması beklenmektedir. Bu parametrelerin analizi yapılacaktır.

Elde edilen verilere dayalı tartışmalar ve sonuçlar sunulacaktır.

1.10 Tezin Organizasyonu

Tez şu şekilde yapılandırılmıştır:

Birinci bölümde konuşma tanıma sistemleri üzerine ayrıntılı bir genel bakış sunulur ve problem ortaya konulur.

İkinci bölümde, örüntü tanıma, otomatik konuşma tanıma, sinir ağları ve derin sinir ağları gibi farklı kavramlarla ilgili literatür taramasına yer verilir.

Üçüncü bölümde, önişlemeye, öznitelik çıkarmaya, veri bölümüne ve ağ eğitimi algoritmasına dair açıklamalar verilerek önerilen yöntem ortaya konulur.

Dördüncü bölümde deneyler ve uygulama ayrıntıları sunulur. Bu bölüm, kullanılan verilerin tanımını, derin sinir ağının uygulanmasını ve ayrıca belirli uygulama parametrelerinin ağ performansları üzerindeki etkisini detaylandırır.

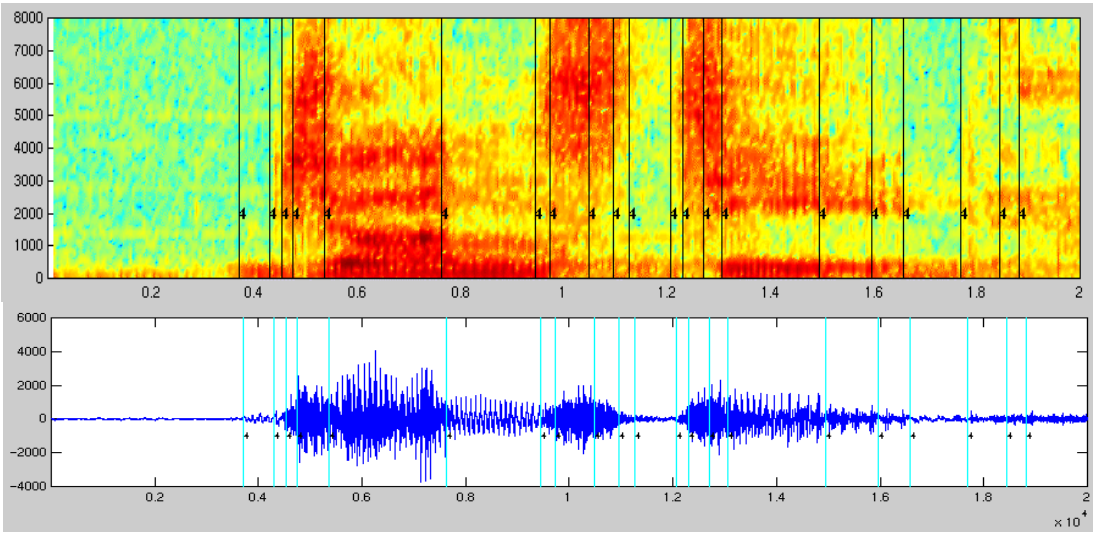
Son olarak, beşinci bölümde ise gelecekte yapılabilecek araştırmalara dönük olarak sonuç ve öneriler açıklanmıştır.

2. LİTERATÜR TARAMASI

2.1 Konuşma

Makine öğrenmesi, uzun yıllardır insan ve makinenin doğal etkileşimine zemin hazırlamak için algoritmalar sağlayan ve böylece dünyada açıkça devrim yaratan bir alan olmuştur ve bu alan sürekli olarak gelişmektedir. Bu algoritmaların tasarımı genellikle doğadan ilham alır, tıpkı sinir ağlarının insan beyninin işleyişinden belirli ölçülerde ilham alması gibi. Yapay bir sinir ağı her biri birkaç nöron içeren birçok katmandan oluşur ve çok boyutlu ve doğrusal olmayan verileri analiz etme yetenekleri nedeniyle sınıflandırma amaçları olarak kullanılır. Bu sinir ağlarının kullanımının makine çevirisi (Lawrence, 2008), konuşma tanıma (LeCun vd., 1989), el yazısı çıktısı (Fadhilah vd., 2018), karakteristik özelliklerde metin çıktısı (Fu, 1976), görüntü tanıma (Saunders, 1996) gibi birçok makine öğrenmesi görevi için yararlı olduğu gösterilmiştir.

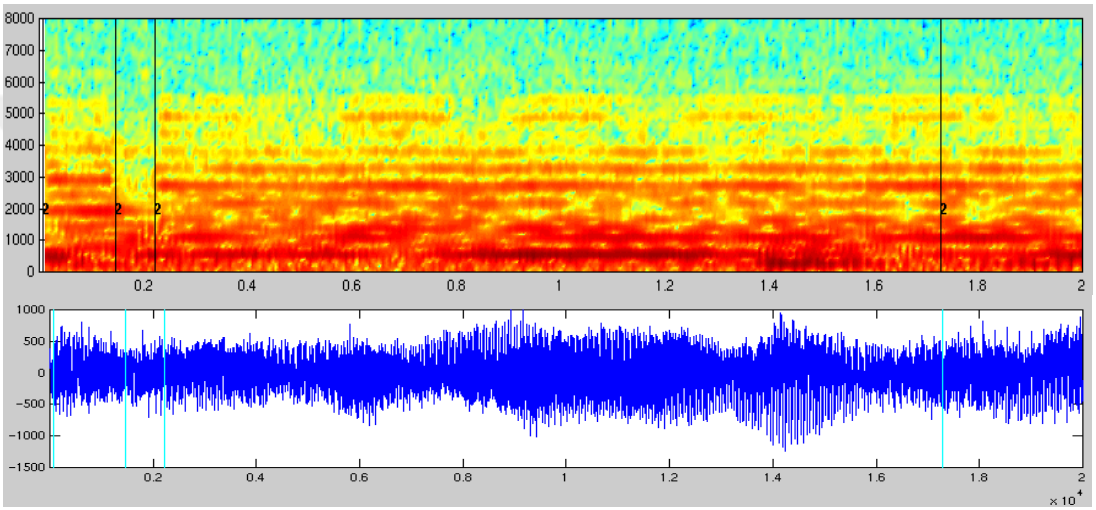
Konuşma, ses tellerinin titreşimleri ile (neredeyse periyodik ses kaynağı), ses yollarında akan hava tarafından veya bu yolun tıkanıklığının serbest bırakılması sırasında oluşan bir türbülans tarafından üretilen bir dizi sesten oluşur (gürültü kaynağı görünmezdir) (Lawrence, 2008). Bu seslerin biçimsel (formantik) bir yapısı vardır. Bir sesin süresi yaklaşık 60 ila 100 ms'dir (bkz. Şekil 1.1'deki spektrogram).



Şekil 1.1 Konuşma spektrogramı: 1.2s boyunca değişmeli olarak ötümlü veya ötümsüz sesler. Ötümlü durumda, formantik bir yapı mevcuttur.

2.2 Müzik

Müziği tanımlamak çok zordur çünkü müzik çok farklı şekillerde üretilebilir. Bu nedenle, birçok araştırmacı, (bu bileşenin elde edilmesi amaçlandığında), amaçlarını "geleneksel enstrümantal" müzik çerçevesi içine sınırlarlar, bu müzik bir ahenkli sesler bileşkesidir (klasik anlamda notalardan oluşur), muhtemelen çok seslidir; o zaman mesele bir (veya daha fazla) frekans temel perdesinin bulunması sorunudur (bkz. Şekil 1.2, spektrogram).



Şekil 1.2 Alışılagelmiş müziğin spektrogramı: ahenkli yapı ve 1.2 sn boyunca temel frekansın veya perdenin varlığı.

2.3 Parametreler (veya karakteristikler)

Bir ses dosyasında yukarıda açıklanan iki temel bileşenin, konuşma ve müziğin bulunabilmesi için iki adım gereklidir:

- Parametrik hale getirme
- Karar

O zaman yapılmak istenen bir örüntü tanıma tipi analizine karşılık gelir. Gerçekten de, bir sinyalin özelliklerinin çıkarılması, bir sesin ötümlü/ötümsüz olarak farklılıkları gibi karşıtlıklarının ortaya çıkmasını mümkün kılar.

Mevcut sistemlerde birçok öznitelik kullanılmaktadır, burada sadece en sık kullanılanlar listelenmiştir. Bunlar dört gruba ayrılmıştır:

- Zaman parametreleri (Scheirer ve Slaney, 1997)(Zhang ve Kuo, 1998)(Foote, 1997).
- Frekans parametreleri (Zhang ve Kuo, 1998)(Scheirer ve Slaney, 1997).
- Karışık parametreler (Zhang ve Kuo, 1998).
- MFCC (Hinton vd., 2012).

Son zamanlarda, derin öğrenme olarak bilinen çok ilginç ve umut verici bir teknik yapay sinir ağlarında (ANN) esinlenerek geliştirilmiştir. Bu tekniğin derin olarak isimlendirilmesinin sebebi, katman sayısının bir gizli katmana sahip sığ ağlara kıyasla fazlaca olmasıdır. Daha derin ağlar, makine öğrenmesi uygulamalarında daha iyi performans sergileyebilir çünkü, girdiler ve çıktılar arasındaki doğrusal olmayan ilişkileri araştırmaktadırlar.

Günümüzde daha derin ağları eğitmek için iyi organize edilmiş yeni yöntemler bulunmaktadır ve dolayısıyla bu yöntemlerin otomatik dil tanıma programlarına (ASR) uygulanması son derece doğaldır.

Son on yıllık süreçte, ASR probleminin çözümü için GMM-HMM yaklaşımının baskın olduğu söylenebilir. Günümüz itibariyle DNN yaklaşımının ise, İngilizce, Fransızca ve Mandarin (Çince) gibi yaygın olan bazı dillerde geleneksel GMM-HMM yaklaşımın performansını geride bıraktığı görülmektedir. 2014'ten beri Baidu, Apple, Google, Microsoft ve Amazon gibi birçok teknoloji şirketi ASR sistemlerine derin öğrenmeyi dahil etmektedir (Avci vd., 2005).

Farklı diller için bir ASR sisteminin adaptasyonu potansiyel olarak yeni zorlukları haizdir, çünkü her dilin kendine has bir yapısı ve kuralları vardır. Türkçe üzerine ASR çalışmalarının yirmi yıllık bir geçmişi bulunmaktadır. Bu alanda yapılan araştırmalardan bazıları aşağıda listelenmiştir:

2005 yılında Fırat Üniversitesi'ndeki araştırmacılar dalgacık paket uyarlamalı ağ tabanlı bulanık çıkarım sistemi ile bir Türkçe ASR sistemi tasarladılar. Tasarlanan sistemde iki katman bulunuyordu: bir katman dalgacık paketinden ve diğer katman ise uyarlanabilir ağdan oluşuyordu. Dalgacık paket katmanının temel amacı, zaman-frekans düzleminden uyarlanabilir öznitelik çıkarımıydı; bu bir dalgacık paket

ayrışması ve bir dalgacık paket entropisi yoluyla yapıldı. Elde edilen doğru sınıflandırma oranı kelime tanıma için yaklaşık %90 seviyesindeydi (Arısoy, 2004).

2006 yılında Boğaziçi Üniversitesi'nde uyarlanabilir dil modeline sahip bir Türkçe konuşma tanıma yazılımı geliştirilmiştir. Sistem, Türkçe'de yaşanan kelime haznesi dışı (Out Of Vocabulary-OOV) sorunundan kaçınmak için alt sözcük birimleri kullanmıştır. Geliştirilen tanıma sistemi, radyoloji alanı terminolojisi ile sınırlandırıldı ve %90'lık kelime tanıma doğruluğu elde edildi, ancak daha geniş bir kelime haznesi için aynı oran %40'tı (Arısoy vd., 2007).

2007 yılında yine Boğaziçi Üniversitesi'ndeki araştırmacılar Türkçe haberlere dönük olarak bir ASR sistemi geliştirdiler. Geniş bir Türkçe haber veri tabanı oluşturdular, daha sonra GMM-HMM gibi istatistiksel modelleri kullanarak sistemi eğittiler, ancak dil modellemesi için çeşitli alt sözcükleri (kelime kökleri) inceleniyordu. Bu çalışma %20'den daha az hata veren bir sistem ortaya koymuştur (Can ve Artuner, 2013).

2007 yılında Hacettepe Üniversitesi araştırmacıları, sinir ağlarının kullanıldığı, hece tabanlı bir Türkçe ASR sistemi tasarladılar. Algoritmaları Türkçe heceleri tespit etmede %44'lük bir başarı oranı yakalamıştır (Avci, 2007). 2007 yılında Fırat Üniversitesi'ndeki araştırmacılar sinir ağlarınının kullanıldığı bir ASR sistemi tasarladılar. Çalışmada uyarlamalı entropiye dayanan ayrık dalgacık kullanıldı. Sistemin testleri gürültülü koşullar altında Türkçe konuşulan kelimeler üzerinde yürütüldü. Sistem sadece 15 kelime için %92 kelime tanıma oranı Verdi (Aşlıyan, 2008).

2008 yılında Dokuz Eylül Üniversitesi'ndeki araştırmacılar, konuşmacıya bağımlı hece tabanlı bir konuşma tanıma motoru tasarladılar. Bu motor MFCC, LPC, PARCOR ve RASTA gibi öznitelik çıkarma yöntemlerini kullanıyordu. Yukarıda belirtilen öznitelik çıkarma yöntemleriyle NN, SVM, HMM ve DTW sınıflandırma yöntemlerini beraber kullanarak bir karşılaştırma yaptılar. Karşılaştırmalarının neticesinde bu teknikler arasında en iyi performansın %94.2'lik bir oranla DTW-MFCC kombinasyonundan elde edildiği bulunmuştur (Arisoy vd., 2009).

2009 yılında bir araştırma ekibi Türkçe haberleri transkripte eden (yazılı hale getiren) ve ulaşım sağlayan bir sistem kurdular. Bu çalışma, kelime tabanlı ve alt kelime tabanlı Türkçe ASR sistemi üzerine bir araştırmaydı. Bu araştırmaya göre, çeşitli alt-kelime tabanlı tanıma birimlerinin kullanılması, Türkçe'deki çok sayıda kelime-haznesi-dışı (OOV) kelime problemini hafifletmektedir. Tanıma birimleri, kelimeler ve alt kelimeler arasındaki etkileşim de incelendi. Sistem eğitim amacıyla, HMM'nin ayırt edici eğitimini kullandı ve bu da kelime tanıma için %65.1'lik, alt kelime tanıma içinse %68.6'lık bir doğruluk oranı vermiştir (Karakaş, 2010).

2010 yılında Dokuz Eylül Üniversitesi araştırmacıları ses girdisini kullanan bilgisayar tabanlı bir sistem kontrolü tasarladılar. Öznitelik çıkarma ve sınıflandırma için sırasıyla MFCC ve dinamik zaman sıçraması (time warping) yöntemlerini kullandılar. Performans testleri hem eğitilmiş hem de eğitilmemiş konuşmacılar üzerinde uygulandı. Bunun sonucunda eğitimli konuşmacılar için %90'lık, eğitilmemiş konuşmacılar içinse %84'lük bir performans elde edilmiştir (Rabiner vd., 1985).

2010 yılında Osmangazi Üniversitesi araştırmacıları bir sürekli sayı zamanı dizisi tanıma sistemi tasarladılar. Bu araştırma konuşmacıdan bağımsız ASR sistemine odaklanmıştır. Çalışmada sınıflandırma yöntemi olarak HMM kullanıldı. Sistem %74'lük bir kelime tanıma oranı verdi (Phadke vd., 2004).

2012 yılında Atılım Üniversitesi'ndeki araştırmacılar, izole hece bazlı bir ASR'yi test ettiler. Sistemleri 50 kelimeyle sınırlıydı ve sadece hece yaklaşımına odaklandı. Bu hecelerin sınıflandırılmasında yapay sinir ağları kullanılmıştır. 50 farklı Türkçe kelime için 5 farklı kullanıcılı 10 grupla test sonucunda, sistem konuşmanın tanınmasında %85 performansa ulaştı. Ayrıca sistem, eğitimsiz bir kullanıcı için de test edilmiş ve konuşmanın tanınmasında %75 performans elde edilmiştir (Arisoy ve Saraçlar, 2015).

2015 yılında MEF Üniversitesi ve Boğaziçi Üniversitesi araştırmacıları tarafından yürütülen ortak bir projede, çoklu akışlı uzun-kısa süre bellekli (LSTM) sinir ağı dil modeli tasarlandı. Bu LSTM sinir ağı, dizideki önceki girdileri göz önüne alarak bir sonraki kelimenin olasılığını hesaplayabiliyordu. Bu çalışma maalesef sadece genel Türkçe dil modelinin geliştirilmesine odaklanmıştı ve Türkçe bir ASR sisteminin geliştirilmesi veya iyileştirilmesi amacını taşımıyordu. Çalışma neticesinde, modern

ngram dil modeline göre %0.5'lik bir iyileşme ortaya çıkmıştır (Arisoy ve Saraglar, 2018).

2018 yılında MEF Üniversitesi ve Boğaziçi Üniversitesi araştırmacıları, sinir ağlarına dayanan güncel bir Türkçe haberler transkripsiyon sistemi kurdu. Bu çalışmada, Türkçe ASR sisteminin odak bileşenleri olarak sinir ağına dayalı akustik ve dil modelleri kullanılmıştır. Akustik modelleme için, optimizasyon çapraz entropi ve sekans ayırt edici objektif fonksiyonlar kullanılarak yapılmıştır. Dil modeli içinse tekrarlayan bir dil modeli kullanıldı. Bu çalışmada %10'un altında bir hata oranı elde edilmiştir (Deng, 2016).

Yukarıda bahsedilen araştırmalar üzerinden, yapay sinir ağları tabanlı olarak tasarlanan Türkçe konuşma tanıma sistemlerinin evrimi net bir şekilde görülmektedir. Bu araştırmalardan da anlaşılacağı üzere, Türkçe konuşma tanıma sistemlerinde derin sinir ağlarının tam anlamıyla her yönüyle kullanılmadığı ortadadır. Türkçenin sondan eklemeli olan yapısının doğal bir neticesi olarak; derin öğrenme tekniklerini bu alanda kullanabilmek için çok büyük miktarda konuşma verisi gereklidir, mevcut Türkçe konuşma veritabanları bu anlamda yetersiz kalmaktadır.

Bu tezde, mobil cihazlardan toplanan bir veri tabanı kullanılarak Türkçe bir ASR (otomatik konuşma tanıma) için derin öğrenme tekniği uygulanmıştır.

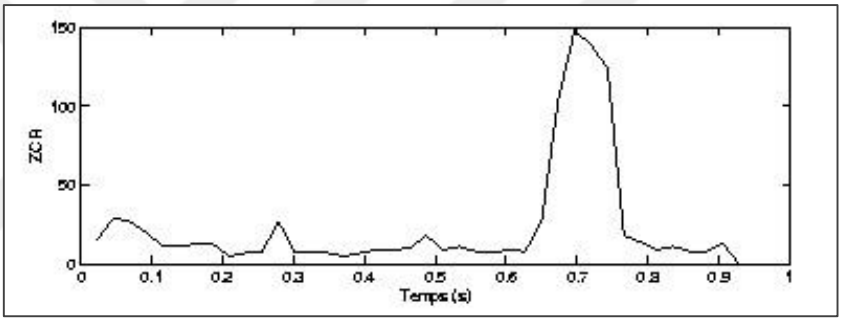
Geçtiğimiz sekiz yıl içinde, DNN'in konuşma sinyali işleme, dil modelleme, konuşma tanıma, bilgi toplama, ayrıştırma, konuşma çevirisi, konuşma sentezi, oyun vb. gibi birçok araştırma alanında şüpheye yer bırakmayacak şekilde iyi bir performans gösterdiği açıktır. DNN söz konusu alanlara uygulandığında, elde edilen performans sonuçları, o andaki en modern teknikleri geride bırakmıştır (Hu, 2005). DNN'nin bu yüksek performansının nedeni, büyük miktarda veri kullanarak öğrenme ve sinyallerde bileşik ve karmaşık yapıları bulma yeteneğidir. Bu DNN'ler ASR'de derin öğrenme mimarisinden türetilen akustik modeller geliştirmek için kullanılmıştır.

2.4 Zaman parametleri

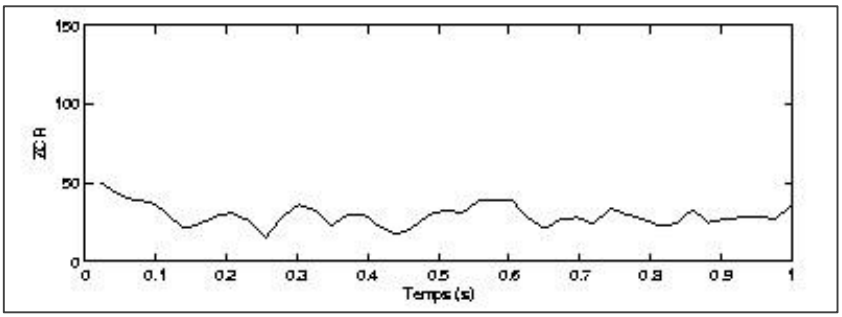
İki ana zaman parametresi ZCR ve enerjidir.

2.4.1 ZCR

ZCR (zero crossing rate) sıfırdan geçiş sayısıdır. Bu öznitelik konuşma/müzik sınıflandırması için sıklıkla kullanılır. Aslında, ZCR'deki ani değişimler ötümlü/ötümsüz değişimler için belirgindir, dolayısıyla konuşmanın varlığı için de belirleyicidir. Bir konuşma söz konusuysa ZCR, konuşmanın olduğu anlarda düşükken, sessiz yani konuşmanın olmadığı anlar içinse çok yüksektir (Şekil 2.1); müzik söz konusuysa ZCR'deki farklılaşma çok küçüktür (Şekil 2.2).



Şekil 2.1 Bir saniyelik konuşma sinyalinin ZCR'si (tepeler, sessiz anlara karşılık gelir)

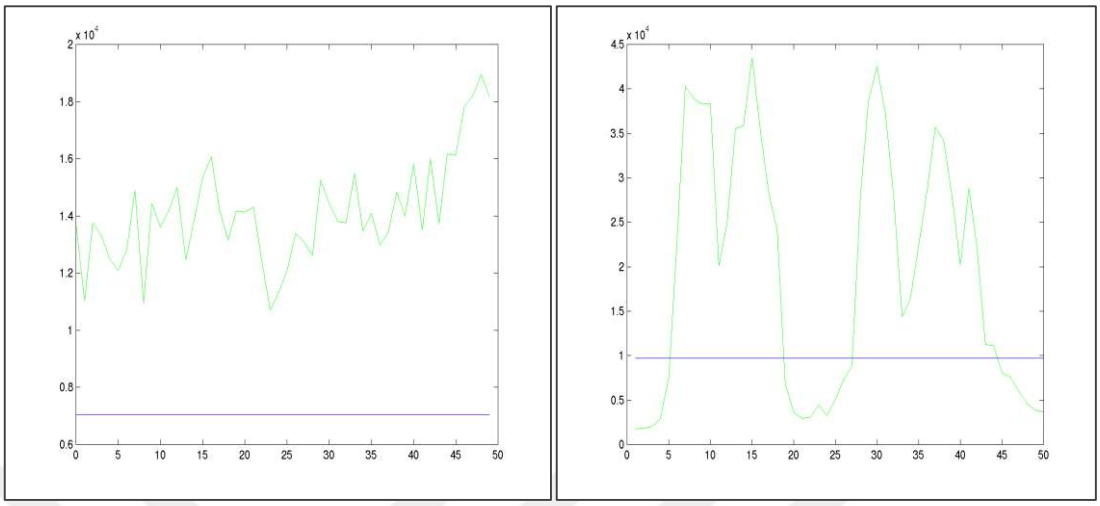


Şekil 2.2 Bir saniyelik müzik sinyalinin ZCR'si (çok az farklılık)

2.4.2 Enerji

Enerji, ZCR ile yakından ilişkilidir. Aslında, enerji konuşma için çok değişkendir ve müzik içinse oldukça durağandır. Bu nedenle, ZCR gibi, hem konuşma ve müziği ayırt etmeyi (Şekil 2.3) hem de sessizlikleri tespit etmeyi sağlar. Genellikle enerji ZCR'ye

bağlaşıktır: düşük ZCR ve güçlü bir enerji sesin varlığıyla eşanlamlıdır, yüksek ZCR ise sessiz bir alanı işaret eder.



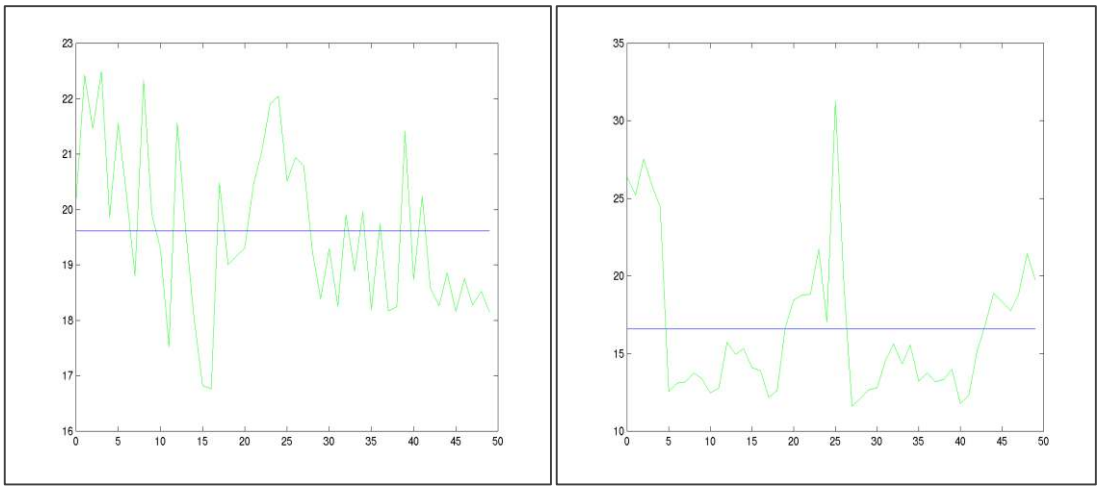
Şekil 2.3 Bir saniye boyunca müzik (solda) ve konuşma (sağda) için enerjinin değişimi

2.5 Frekans parametreleri

Bu parametreler PSD'den (Power Spectral Density-Güç Spektral Yoğunluğu) gelir. Başlıca üç parametre spektral sentroid, spektral akı ve "spektral rollof (spectral etekler) noktası"dır.

2.5.1 Spektral sentroid

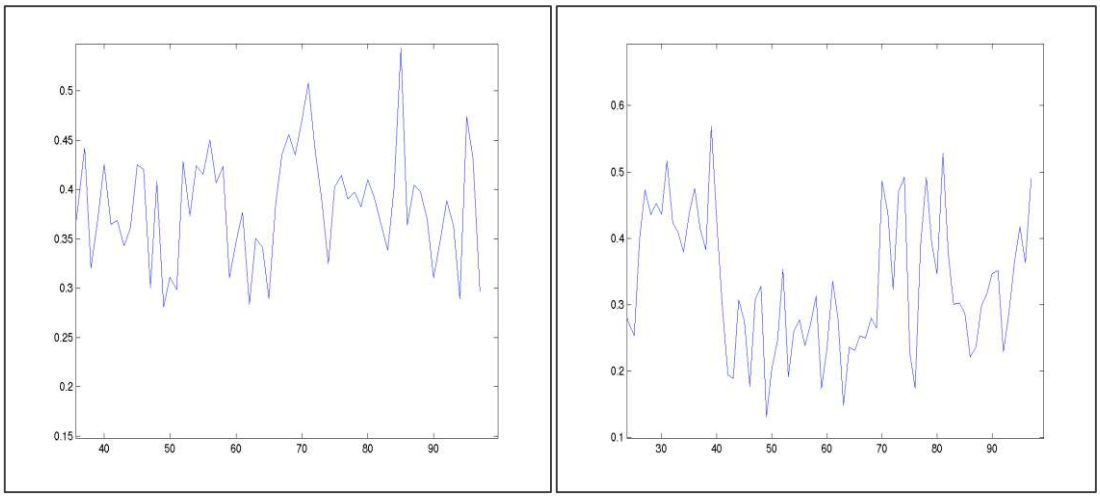
Spektral sentroid, bir PSD'nin frekans ağırlık merkezidir. Müzikler için daha yüksektir, çünkü konuşmaya nazaran perdeler daha geniş bir alana dağılır (genellikle müzik için 6 oktav, konuşma için 3 oktav). Spektral sentroidin varyasyonu da belirleyici olabilir: eğer bu varyasyon çok belirginse, sesli/sessiz değişim karakterize edilebilir. Spektral sentroid ayrıca bir sesin parlaklığına delil olarak kullanılır (Şekil 2.4).



Şekil 2.4 Müzik (solda) ve konuşma (sağda) için spektral sentroid

2.5.2 Spektral Akı

Spektral akı, zaman aralığının 20 ms kadar olduğunu göz önüne alarak iki ardışık analizin spektrumundaki değişimi ölçer. Konuşma için, varyasyonlar belirgindir ve spektral akının değeri düşüktür; müzik içinse profil daha durağan ve seviye yüksektir. Dolayısıyla, yüksek bir spektral akı müziksel bir içeriği karakterize eder ve fazlaca varyasyon, yada değişim konuşmanın varlığına işaret eder (Şekil 2.5).

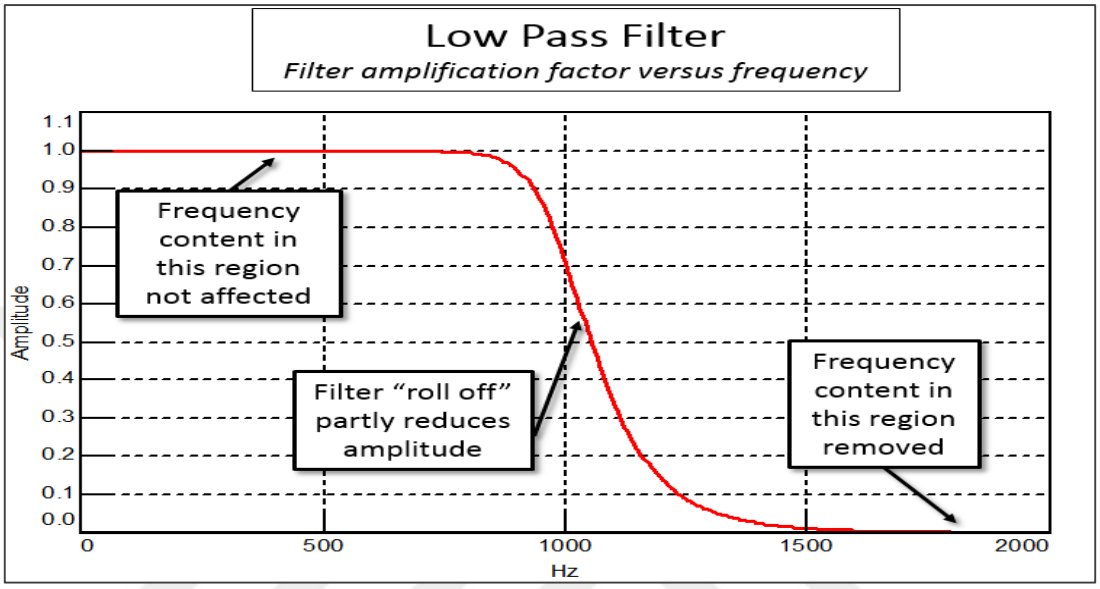


Şekil 2.5 Bir saniyelik müzik (solda) ve 1s'lik konuşma (sağda) üzerinde spektral akı

2.6 Spektral Etekler (Rolof) Noktası

"Spektral etekler noktası," DSP'nin gücünün %95'inin konsantre olduğu eşik frekanstır (Şekil 2.6). Spektral etekler noktası, ötümlü bir ses için (gücün daha düşük

frekanslarda konsantre olduğu ses), ötümsüz bir sesten (yüksek frekanslar yönünden zengin ses) daha yüksektir. Dolayısıyla bu ölçüm konuşmadaki sesli/sessiz değişimlerini karakterize etmeyi mümkün kılar. Müzik için bu ölçüm hemen hemen her zaman aynı değeri verir.



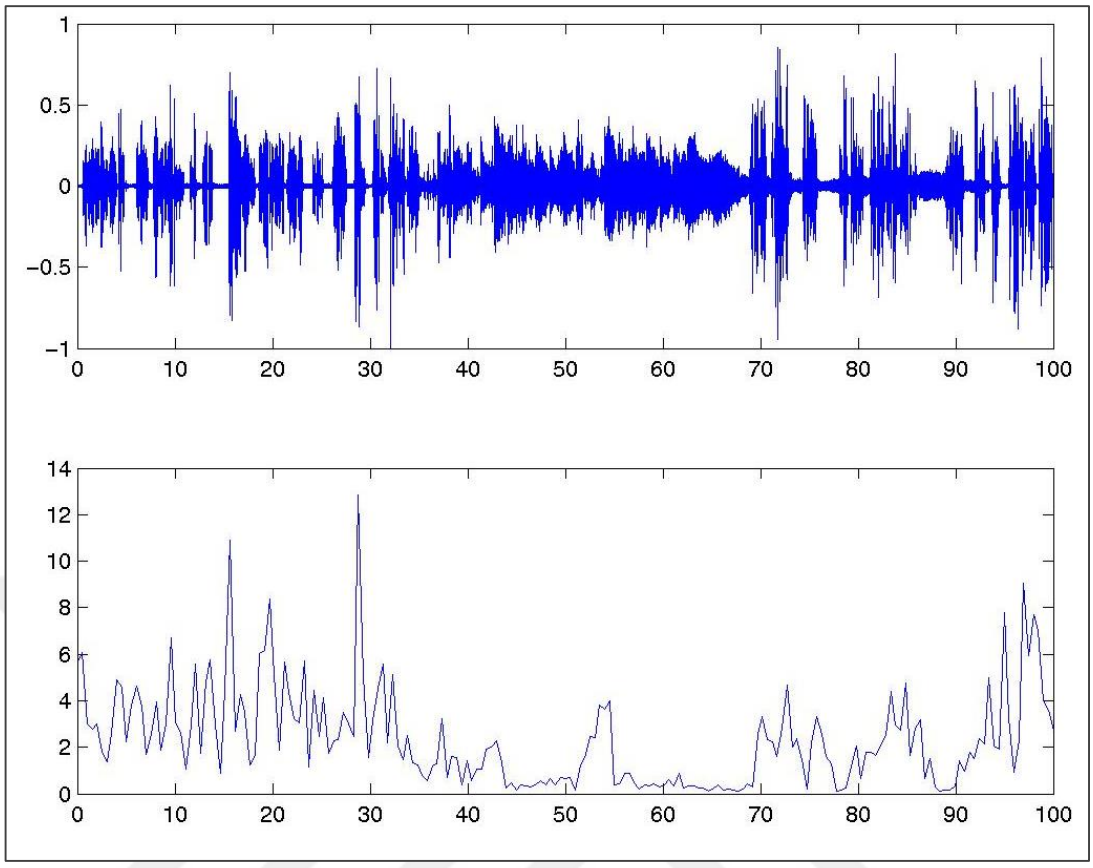
Şekil 2.6 Spektral Etekler Noktasının Tanımı

2.7 Karışık parametreler

Enerjinin 4Hz'deki modülasyonu karışık bir parametredir, yani hem frekans analizinden hem de zaman analizinden parametre çıkar.

Konuşmanın hecelerinde, 4 Hz civarında enerjinin davranışının ve modülasyonunun gözlemi ilginç sonuçlar verir, çünkü değişiklikler bu frekansın etrafında ortaya çıkar. Gerçekten de, bir hece düşük enerji bölgesi (ünlü harf) ve yüksek enerji bölgesinin (ünsüz harf) birleşimidir.

Yöntem, 20 frekans bandı üzerindeki sinyalin filtrelenmesinden oluşur. Ardından enerjiler gözlemlenebilir bir değer elde etmek amacıyla toplanır. Daha sonra, bu değerin mutlak değerinin logaritması alınır. Son olarak, varyans hesaplanır. Şekil 2.7'de görüldüğü gibi, müzik yaklaşık 37. ve 70. saniyeler arasındadır (varyasyon oldukça küçüktür). Büyük varyasyonlar konuşmaya karşılık gelir.



Şekil 2.7 Yaklaşık 100 saniyelik bir sinyale (üst) karşılık gelen enerjinin 4Hz'de modülasyonu (altta)

3. KONUŞMA SİNYALİNİ İŞLEME VE YÖNTEMLER

3.1 Sinyal İşleme

Günlük yaşamımızda kullandığımız hatta hayatımızın vazgeçilmez birer parçası haline gelmiş olan radio, bilgisayar, video, cep telefonu gibi teknoloji ürünleri sinyal işleme üzerine kuruludur, sinyal işleme elektrik mühendisliğinin bir dalıdır ve fiziksel olayları temsil eden verilerin modellenmesi ve analiziyle ilgilenir. Sinyal işleme, modern dünyamızın merkezindedir; günümüzün eğlencesine ve yarının teknolojisine güç vermektedir. Sinyal işleme, biyoteknolojinin ve sosyal etkileşimlerin kesişim noktasındadır. İletişim kurma ve bilgi paylaşma gibi faaliyetlerimiz büyük ölçüde bu kanal üzerinden gerçekleşmektedir. Sinyal işleme, dijital yaşamlarımızın arkasındaki bilimdir.

3.1.1 Konuşma ve Ses İşleme

Gerek sabit hatlar, gerek mobil gerekse akıllı telegonlar, iki (veya daha fazla) kişi arasında sesli iletişimi mümkün kılmak için büyük ölçüde konuşma işleme tekniklerine dayalıdır. Bu iletişimde yer alan işlemlerin analog-dijital dönüşümünden ses kalitesini geliştirmeye (filtreleme, yankı-gürültü baskılama ve otomatik kazanım kontrolü), bir uçta konuşmayı kodlamaktan diğer uçta çözümlemeye kadar farklı aşamaları vardır. Bu işlemlerin her aşamasında sinyal işleme tercih edilen araçtır. Sinyal işleme olmadan Siri, Google Asistan ve Cortana gibi modern dijital asistanlar bir kullanıcının sesini tanıyamazlardı. MP3 ve AAC (Gelişmiş Ses Kodlayıcı) gibi ses sıkıştırma teknikleri, müzik dinleme biçimimizde devrim yarattı. Artık dünyanın her bölgesinden müzik kataloğunu avucumuzda tutarken hatta hareket halindeyken, Bluetooth aracılığıyla gözle görülür bir bağlantınız olmaksızın müzik dinlemenin tadını çıkarmak mümkündür. Yine, sinyal işleme sayesinde bu olay mümkün hale gelmiştir (Aadit vd., 2016).

Konuşma işleme, konuşma sinyallerinin ve bu sinyalleri işleme yöntemlerinin incelenmesidir. Sinyaller genellikle bir dijital gösterim yani temsil üzerinde işlenir. Dolayısıyla konuşma işlemeyi, dijital sinyal işlemenin konuşma sinyallerine

uygulaması olarak düşünebiliriz. Otomatik Konuşma Tanıma teknolojisi son yıllarda büyük aşamalar kaydetmiştir. Konuşma tanıma ilgiye mazhar olan çok büyük bir alandır ve karmaşık bir problem olarak görülür. Pratik anlamda, konuşma tanıma sorunları çözer, üretkenliği artırır ve hayat tarzımızı değiştirir. Güvenilir konuşma tanımanın elde edilmesi zor bir problemdir, çözümü birçok tekniğin kombinasyonunu gerektirir; ancak modern yöntemler etkileyici bir doğruluk derecesine ulaşmıştır. Gerçek zamanlı dijital sinyal işleme, gelişmiş DSP (Dijital Sinyal Processing) işlemcilerinin piyasaya sürülmesinden sonra önemli ilerlemeler kaydetmiştir (Chi vd., 2004).

Bütün konuşmacı tanıma sistemleri iki ana modül içerir: öznitelik çıkarma ve öznitelik eşleştirme. Öznitelik çıkarma, ses sinyalinden daha sonra her bir konuşmacıyı temsil etmek için kullanılabilecek az miktarda veri çıkaran işlemdir. Öznitelik eşleme ise, ses girdisinden çıkarılan öznitelikleri bilinen bir grup konuşmacının özniteliklerıyle karşılaştırarak bilinmeyen konuşmacıyı tanımlamak için izlenen prosedürü içerir (Kalia vd., 2020).

3.2 Konuşma tanıma

Konuşma tanıma sinyal işlemenin hayati bir uygulamasıdır ve muhtemelen anlaşılması en kolay olanıdır. Sinyal işleme, otomatik konuşma tanımayı (ASR) mümkün hale getirebilmek için sinyallerdeki bilgi içeriğini değiştirir. Konuşma sinyallerinden bilgi çıkarılmasına yardımcı olur ve daha sonra bu bilgiyi tanınabilir kelimelere dönüştürür. Konuşma tanıma teknolojisi savaş uçaklarında, akıllı telefonlarda konuşmaları metne dönüştüren uygulamalarda, iyileştirici (terapötik) uygulamalarda, tercüme ve dil öğreniminde ve engelli insanlar için tanıma programlarında bulunur (Raj ve Shaji, 2016).

Analog konuşma sinyalinin parametrize edilmesi, konuşma tanıma işleminin ilk adımıdır. Literatürde birçok popüler sinyal analiz tekniği adeta de facto standart olarak ortaya çıkmıştır. Bu algoritmalar, konuşma sinyalinin “algısal olarak anlamlı” parametrik temsilini üretmeyi amaçlamaktadır: bu parametreler insan işitsel ve algısal sistemlerinde gözlemlenen bazı davranışları taklit eden parametreler olmalıdır.

Elbette, bu algoritmalar nihayetinde tanıma performansını en üst düzeye çıkarmak için tasarlanmıştır.

Bu tekniklerin çoğunun kökleri, konuşmacıya bağlı konuşma tanıma üzerine yapılan ilk araştırmalara kadar uzanır. Bugün, konuşma tanıma araştırmalarının önemli bir kısmı artık konuşmacıdan bağımsız tanıma sorununa odaklanmış olsa da, bu geçmiş parametrizasyonların birçoğu halen yararlı olmaya devam etmektedir. Konuşmacıdan bağımsız konuşma tanımada, başlıca gaye konuşmacının değişmesiyle değişmeyecek hususların geliştirilmesidir. Belirli bir konuşmacının sesine bağlı ayrıntılar yerine, sesin belirgin spektral enerjilerini temsil eden parametrelerin tanımlanması amaçlanır.

3.2.1 Sinyal Modelleme

Sinyal modelleme dört temel işleme ayrılabilir: spektral şekillendirme, spektral analiz, parametrik dönüşüm ve istatistiksel modelleme. İlk üç işlem, dijital sinyal işlemenin sıradan problemleridir. Ancak, son işlem genellikle sinyal modelleme sistemi ile konuşma tanıma sistemi arasında bölünür.

3.2.2 Spektral Şekillendirme

Spektral şekillendirme iki temel işlemi içerir: A/D dönüşümü—sinyalin ses basınç dalgasından dijital sinyale dönüşümü; ve sinyaldeki önemli frekans bileşenlerini ayıklayan dijital filtreleme.

Dijitalleştirme işleminin temel amacı, konuşma sinyalinin sinyal/gürültü oranı (SNR) olabildiğince yüksek olan örneklenmiş bir veri temsilini üretmektir.

3.2.3 Spektral Analiz

Konuşma tanıma sistemlerinde kullanılan spektral ölçüm türleri iki sınıfa ayrılır: sinyalin brüt spektral (veya zamansal) gücünün ölçümü; veya spektrumdaki belirli frekans aralıklarında gücün ölçümü—spektral genlik.

Son zamanlarda, prosodik bir öznitelik olarak kullanmak amacıyla temel frekansa yeniden ilgi duyulmaktadır, bunun ton dillerinde (ör. Çince) veya bazı ton bileşenlerine sahip dillerde (ör. Japonca) konuşma tanınmasında ve konuşmacı kimliğinin veya özgünlüğünün tepitinde kullanımı mümkündür.

Temel frekans, bir seslendirme sırasında ses tellerinin titreşim frekansı olarak tanımlanır. Konuşma sinyalinden güvenilir bir çıkarımı yapma konusunda temel frekans, uzun süredir zor bir parametre olagelmiştir. Daha önce, konuşma tanıma sistemlerinde çok sayıda nedenden dolayı ihmal edilmiştir bu parametre, bu ihmalin sebepleri doğru tahmin için gereken büyük hesaplama yükü, güvenilir olmayan tahminin yüksek performans elde etmenin önünde bir engel olacağı endişesi ve parçaüstü olgular arasındaki karmaşık etkileşimleri karakterize etme zorluğu şeklinde sıralanabilir.

Konuşma tanımada bir tür güç ölçümünün kullanılması bugün oldukça standart hale gelmiştir. Birçok konuşma tanıma sistemi, gücü doğrudan kullanmak yerine, gücün logaritmasını 10 ile çarparak kullanır, bu gücün desibel cinsinden ifadesidir; bu tanımın ortaya çıkmasının sebebi insan işitme sisteminin logaritmik olan tepkisini taklit etmektir. Konuşma tanıma sistemindeki çoğu parametre gibi, güç de çerçeve çerçeve, yani anlık küçük parçalar için hesaplanır.

Bugün konuşma tanıma sistemlerinde kullanılan altı ana spektral analiz algoritması sınıfı vardır. Filtre bankası yöntemleri (analog devrelerde uygulanmıştır) tarihsel olarak ilk ortaya atılan yöntemlerden biridir. Doğrusal tahmin yöntemleri 1970'lerde kullanılmaya başlanmış ve 1980'lerin başında baskın teknik haline gelmiştir. Şu anda, hem Fourier dönüşümü hem de doğrusal tahmin teknikleri çeşitli konuşma işleme uygulamalarında yaygın olarak kullanılmaktadır.

3.3 Dijital Filtre Bankası

Dijital filtre bankası, konuşma işlemenin en temel kavramlarından biridir. Bir filtre bankası, insan işitme sistemindeki transdüksiyonun (ses dalgalarının dönüştürülüp iletimi) ilk aşamalarının ham bir modeli olarak görülebilir. Filtre bankasıyla temsilin arkasında iki temel motivasyon kaynağı vardır (Xu vd., 2011).

İlk olarak, saf tonlar gibi uyaranlar için baziler zardaki maksimum yer değiştirme pozisyonu, tonun frekansının logaritması ile orantılıdır. Bu hipotez, bir işitme teorisi olan "yer teorisi"nin bir parçasıdır.

İkincisi, insan algısındaki deneyler, karmaşık bir sesin bazı nominal frekansların bant genişliği içindeki frekanslarının ayrı ayrı tanımlanamayacağını göstermiştir. Bu sesin bileşenlerinden biri bu bant genişliğinin dışına çıktığında, ayrı ayrı ayırt edilebilir.

3.3.1 Fourier Dönüşüm Filtre Bankası

Bir sinyalin düzgün olmayan aralıklı bir filtre bankası modelini hesaplamanın en kolay ve en etkili yollarından biri, sinyal üzerinde basit bir Ayrık Fourier Dönüşümü (DFT) gerçekleştirmek ve dönüşüm çıktısını istenen frekanslarda örneklemektir (Logan, 2000).

Hızlı Fourier Dönüşümü (FFT) de sinyalin spektrumunu hesaplamak için alternatif bir yöntem olarak kullanılabilir. FFT, spektrumun ayrık bir frekans kümesinde değerlendirileceği kısıtlaması altında DFT'nin hesap yükü yönünden verimli olan bir uygulamasıdır.

3.3.2 Cepstral Katsayılar

Cepstrum, doğrusal olmayan bir fonksiyon yani logaritma fonksiyonu kullanılarak hesaplandığından, genellikle belirli türdeki gürültü ve sinyal bozulmalarına duyarlı olduğu düşünülmektedir. Çarpımsal gürültü süreçlerinin genel olarak üstesinden gelmek mümkünken, toplanır gürültü (arka plandaki akustik gürültü gibi) problemlere yol açabilir. Yüksek çözünürlüklü spektral tahmincilerden veya spektrumun parametrik uyumlarından türetilen cepstral parametreler, gürültülü ortamlardaki uygulamalar için sıklıkla tercih edilir (McCree, 2005).

3.3.3 Doğrusal Tahmin Katsayıları

Şimdi doğrusal spektral analize dayanan Fourier dönüşüm yöntemlerinden spektrumu özbağlanımlı (autoregressive) bir süreç olarak en uygun şekilde modellemeye çalışan

parametrik modelleme teknikleri sınıfına dönüyoruz. En küçük ortalama kare hata hesabına dayalı olarak bir parametrik modelin hesaplanması doğrusal tahmin (LP) olarak bilinir. Yordayıcı katsayılarını hesaplamanın başlıca üç yolu vardır: kovaryans matrisine dayanan kovaryans yöntemleri (en küçük kareler yönteminin en saf hali olarak da bilinir), otokorelasyon yöntemleri ve latis (kafes veya harmonik) yöntemi (Mantha vd., 1999).

LP katsayılarına güç ve temel frekans bilgilerini ekleyerek, konuşma sinyalinin bir ses versiyonunu yeniden oluşturmak mümkündür. Konuşma sinyalinin, özellikle konuşma tanıma modellerinin parametrik açıklamalarını dinlemek, sorunları teşhis etmek için çok yararlıdır. Cepstral katsayılar gibi bazı parametrik dönüşümlerin orijinal LP verileriyle bire bir eşleşmesi yoktur. Bu nedenle, bir parametre kümesinin geçerliliğini ölçmek biraz daha zordur. Konuşma tanıma sistemleri tarihsel olarak ilk başta LP parametrelerini tanıma sürecinde doğrudan kullanılmıştır. O zamandan beri, bu parametrelerin daha karmaşık dönüşümleri tasarlanmıştır. Bununla birlikte, doğru bir LP modeli oluşturmanın spektral analizde önemli bir ilk adım olduğunu hatırlamak önemlidir. LP analizi doğrusal olmayan bir işlem olduğundan, gürültülü ortamlardaki performans bazen sorunludur. Bu nedenle, bazı sistemlerde hala Fourier dönüşüm tabanlı filtre bankası analizi kullanılmaktadır.

3.3.4 LP-Türetimli Filtre Bankası Genlikleri

Doğrusal kestirimin türetilen filtre bankası genlikleri, (sinyal spektrumu yerine) uygun filtre bankası frekanslarında LP spektral modelinin örneklenmesinden kaynaklanan filtre bankası genlikleri olarak tanımlanır. LP modelinin veya sıklıkla bahsedildiği gibi yüksek çözünürlüklü modelin kullanımının daha sağlam spektral tahminler sağladığı öne sürülmüştür. LP modelinin doğasında bulunan spektral düzeltme, işlemenin sonraki aşamaları için genellikle daha kararlı parametreler sağlar. Bununla birlikte, konuşma tanıma ve DSP teknolojileri ilerledikçe, bu yaklaşımlardaki farklılıkların eskisi kadar büyük olmadığı söylenebilir (Assaleh ve Mammone, 1994).

3.3.5 LP-Türetimli Cepstral Katsayıları

LP'den türetilen filtre bankası genliklerini hesaplamak için LP modelini sonuna kadar kullandık. Bu yöndeki bir diğer mantıklı adım, cepstral katsayılarını hesaplamak için LP modelini kullanmak olacaktır (Loizou, 2013).

LP'den türetilmiş cepstral katsayıların bir dezavantajı, doğrusal olmayan bir frekans ölçeği kavramını tanıtmak için biraz daha fazla çalışmamız gerektiğidir. Tercih edilen yaklaşım, dijital filtre tasarımındaki frekansları esnetmek için kullanılan bir yönteme dayanmaktadır. Bu yöntem dijital sinyal işlemede çok önemli bir dönüşüm kullanır: bilinear (çiftdoğrusal) dönüşüm.

3.3.6 Parametrik Dönüşüm

Sinyal parametreleri, sinyal ölçümlerinden iki temel işlemle üretilir: farklılaştırma ve ardarda bağlama. İşlemin bu aşamasının çıktısı, sinyale ilişkin ham tahminlerimizi içeren bir parametre vektörüdür.

3.3.7 İstatistiksel Modelleme

Eğer dikkatimizi istatistiksel parametreler sorununa çevirecek olursak; Sinyal parametrelerinin, altta yatan çok değişkenli bazı rastgele işlemlerden üretildiğini varsayalım. Süreç hakkında sahip olacağımız tek bilgi, gözlemlenen çıktılarıdır, yani hesaplanan sinyal parametreleridir. Bu nedenle, bu işlem aşamasında ortaya çıkan parametre vektörü çıktısı genellikle sinyal gözlemleri olarak adlandırılır. Bu vektörlerin tüm sinyal için toplanmasına sinyal gözlem matrisi denir. Konuşma tanıma sistemlerinde son derece sofistike istatistiksel modeller kullanılır—bu bir konuşma tanıyıcısının temel işlevlerinden biridir. Bununla birlikte, burada sunulan tekniklerin çok çeşitli konuşma işleme uygulamalarında yararlı olduğu görülmüştür ve bu teknikler daha karmaşık algoritmaların temelini oluşturmaktadır (Arısoy vd., 2007).

3.4 Konuşma İyileştirilme

Konuşma iyileştirme, çeşitli algoritmalar kullanarak konuşma kalitesini artırmayı amaçlamaktadır. Çok sık kullanılan bir kelime olmakla birlikte, kalite kelimesinden tam olarak ne kastedilmektedir? En azından netlik ve anlaşılabilirlik, tatmin edicilik veya konuşma işlemedeki diğer bazı yöntemlerle uyumluluk kalite kapsamında sayılabilir.

Anlaşılabilirliği ve tatmin ediciliği herhangi bir matematiksel algoritma ile ölçmek zordur. Bu amaçla genellikle dinleme testleri kullanılır. Bununla birlikte, dinleme testlerinin organizasyonu pahalı olabileceğinden, dinleme testi sonuçlarının nasıl bir netice vereceği üzerine fazlasıyla çalışma yapılmıştır. Şimdiye kadar herhangi bir felsefe taşı veya minimizasyon kriteri keşfedilememiştir ve keşfedilmesi de oldukça şüphelidir.

Konuşmayı iyileştirmenin öncelikli yöntemleri; arka plan gürültüsünün giderilmesi, yankının bastırılması ve belirli frekansların konuşma sinyaline yapay olarak dahil edilmesi işlemidir. Diğer yöntemlere kısaca değindikten sonra arka plan gürültüsünün giderilmesine odaklanacağız.

Her şeyden önce, doğal ortamında yapılan her konuşma ölçümü bir miktar yankı içerir. Yankının olmadığı, yankısız olarak tasarlanan bir odada ölçülen konuşma insan kulağına kuru ve donuk gelmektedir. Yankı bastırma büyük salonlarda, özellikle de mikrofon ile hoparlör arasındaki mesafe büyükse, konuşma sinyalinin kalitesini artırmak için gereklidir.

Arkaplan gürültüsü bastırılırken, konuşma sinyaline zarar vermemek veya bu sinyali bozmamak çok önemlidir; ya da en azından bu yan etkilerin çok kötü olmaması gerekir. Hatırlanması gereken bir başka şey de, sessiz sayılabilecek doğal arka plan gürültüsünün, daha sessiz ancak doğal olmayan çarpık gürültüden az rahatsız edici olduğudur. Eğer konuşma sinyalinin insanlar tarafından dinlenmesi amaçlanmıyor, fakat örneğin bir konuşma tanıyıcıya yönlendiriliyorsa, o zaman rahatsız edici olma durumu tabii ki söz konusu olamaz. Bu durumda arka plan gürültüsünü düşük tutmak çok önemlidir.

Arka plan gürültüsünü bastırmanın birçok örneği bulunmaktadır. Arabalarla dolu bir sokak gibi gürültülü bir ortamda, örneğin telefonla konuşmak bunun bir örneğidir. Geleneksel olarak, bir uçağın kokpitinden yere veya kabine konuşma gönderirken arka plan gürültüsü bastırılır. Benzer örnekler bulmak kolaydır (Hira ve Gillies, 2015).

3.5 Konuşma Segmentasyonu (bölümlendirme)

Konuşma sinyallerinin otomatik segmentasyonu üzerine 30 yılı aşkın bir süredir araştırmalar yapılmaktadır. Birçok konuşma işleme sistemi, konuşma dalga formunun ana akustik birimlere ayrılmasını gerektirir. Segmentasyon, bir konuşma sinyalini daha küçük birimlere ayırma işlemidir. Segmentasyon, konuşma tanıma sistemleri ve konuşma sentez sistemlerinin eğitimi gibi sesli herhangi bir aktif sistemdeki en önemli adımdır. Konuşma segmentasyonu Dalgacık Dönüşümü, Bulanık Mantık, Yapay Sinir Ağları ve Gizli Markov Modeli gibi yöntemler kullanılarak yapılır.

Konuşma segmentasyonu, konuşma dilinin doğal akışı içinde kelimeleri, heceleri veya fonemleri birbirinden ayrıştırma, aralarındaki sınırları (belirli karakteristiklerle) bulma süreci olarak tanımlanabilir. Konuşma segmentasyonunun temel amacı, konuşma sentezi, konuşma tanıyıcılar için veri eğitimi gibi diğer konuşma analizi problemlerine hizmet etmek veya prosodik veritabanları üretmek ve etiketlemektir. Bu nedenle, konuşma analizi ve araştırmasında segmentasyon çeşitli alanlar için hayati bir alt birim olarak görülebilir. Bu konudaki geleneksel yaklaşım, genellikle uzman fonetikçiler tarafından icra edilmek üzere, konuşmanın manuel olarak ayrıştırılmasıdır. Ancak bu yöntemin gerekli sınırlar üzerinde dinleme ve görsel yargıya dayanması bu yaklaşımı tutarsız ve zaman alıcı hale getirmektedir. Çok kullanışlı olduğu düşünülen bir diğer yöntem de otomatik segmentasyondur. Konuşma otomatik olarak akustik şeklinde tanımlanan alt sözcük birimlerine bölünebilir. ASR (Otomatik Konuşma Tanıma) sistemlerinde segmentasyon şu aşamalarda uygulanabilir:

1) Sistemin eğitimi aşamasında, eğitim kümesi kayıtlarına segmentasyon uygulandığında.
2) Tanıma aşamasında.

Konuşma sinyalini bölümlere ayırmak için iki tür sinyal özniteliği kullanılır: zaman alanı öznitelikleri ve frekans alanı öznitelikleri.

3.5.1 Zaman Alanı Sinyal Öznitelikleri

Zaman alanı özenitelikleri, konuşma segmenti çıkarmak için yaygın olarak kullanılmaktadır. Bu öznitelikler, basit bir şekilde uygulanabilecek ve hesaplamalarda verimlilik sağlayacak algoritmaya ihtiyaç duyulduğunda faydalıdır.

3.5.2 Frekans Alanı Sinyal Öznitelikleri

En çok konuşma bilgisi 250Hz-6800Hz frekans aralığında toplanır. Frekans alanı özniteliklerini çıkarmak için ayrık Fourier dönüşümü kullanılabilir. Bir sinyalin Fourier gösterimi, sinyalin spektral kompozisyonunu verir. Otomatik segmentasyonda kullanılan yöntemlerin bazıları aşağıda açıklanmaktadır.

3.6 Dalgacık Metodu

Dalgacık metodu, fonemlerin başlangıç ve bitişlerini tanımak için uygulanan DWT (Ayrık Dalgacık Dönüşümüne) ve spektral analize dayanır; bu metot son derece etkili bir teknik olmuştur.

3.6.1 Siyah Alanı Bloklama Yöntemi

Ostu'nun dinamik eşik için var olan yöntemi kullanılarak sürekli konuşmadaki sesli alanlar bloklar haline getirilir, yani bloklanır ve böylece sessizlerden ayrılır ve bu yönteme siyah alanı bloklama denir. Sürekli konuşmadaki blokların kenarları kelimelerin sınırları olarak kullanılır. Sistemin performansı, 3280 kelimeden oluşan 500 Bangla cümlesinden oluşan bir veritabanı ile test edilir. Bu araştırmada kullanılan tüm algoritmalar ve yöntemler MATLAB'da uygulanmış ve önerilen sistem %90,58'lik bir ortalama segmentasyon doğruluğu yakalamıştır.

3.6.2 Kısa Süreli Enerji

Kısa Süreli Enerji, konuşmanın hecelere segmentasyonu için kullanılır. Yine aynı şekilde bu yöntem de Matlab'da uygulanmıştır. Çeşitli konuşma sinyalleri kaydedilmiş ve segmentasyona tabi tutulmuştur. Önerilen yöntem farklı konuşma sinyalleri için uygulanmış ve analiz edilmiştir.

3.6.3 Konuşma Segmentasyonu için Hibrit Algoritması

Otomatik Konuşma Tanıma için önerilen hibrit bir segmentasyon yöntemi mevcuttur. Segmentasyon yöntemleri zaman alanı özniteliklerine veya frekans alanı özniteliklerine dayanmaktadır. Zaman alanı öznitelikleri, kısa zaman enerjisi, kısa zaman sıfır geçiş oranıdır. Frekans alanı öznitelikleri ise spektral sentroid ve spektral akıdır. Öznitelikler çıkarıldıktan sonra kelimenin sınırlarını tespit etmek için basit bir eşik metodolojisi kullanılır. Bu nedenle segmentasyon, sürekli konuşmanın bir dizi kelimeye veya alt kelimeye bölünmesi ile karakterize edilir (Avci, 2007).

3.7 Öznitelik Çıkarımı ve Seçimi

Makine öğrenmesinde verilerin boyutsallığı arttıkça, yani çok boyutlu hale geldikçe, güvenilir bir analiz yapabilmek için gerekli olan veri miktarı adeta üstel bir fonksiyon şeklinde artar. Boyutsallık veya çok boyutluluk arttıkça, hesaplama yükü yada maliyeti de genellikle artar, üstelik yine bu artış hemen hemen üsteldir. Bu sorunun üstesinden gelmek için, dikkate alınan öznitelik sayısını azaltmanın bir yolu bulunmalıdır. Bu doğrultuda kullanılan genellikle iki teknik vardır:

1) Öznitelik altkümesi seçimi.
2) Öznitelik çıkarma.

Öznitelik altkümesi seçimi, kısaca ilgili olmayan veya gereksiz olan özniteliklerin kaldırılmasıdır. Öznitelik seçim algoritmaları üç kategoriye ayrılır:

1) Herhangi bir öğrenmeye/eğitime gerek kalmadan verilerden öznitelikler çıkaran filtreler.

2) Hangi özniteliklerin yararlı olduğunu değerlendirmek için öğrenme tekniklerini kullanan sarmalayıcılar (wrapper).
3) Öznitelik seçme adımını ve sınıflandırıcı inşasını birleştiren gömülü teknikler.

Filtreler, sınıflandırıcıyı dikkate almadan çalışır. Bu durum filtreleri hesaplama açısından çok verimli kılar. Filtreler çok değişkenli ve tek değişkenli yöntemlere ayrılırlar. Çok değişkenli yöntemler öznitelikler arasında ilişki bulabilirken, tek değişkenli yöntemler her bir özniteliği ayrı ayrı ele alır (Amin ve Mahmood, 2008).

Konuşma özniteliği çıkarmanın amacı, konuşma dalga formunu bir tür parametrik temsile (oldukça az bir bilgi içeriğiyle) dönüştürmektir, böylece daha fazla analiz ve işlem mümkün olacaktır. Bu genellikle sinyal işlemenin ön ucu olarak adlandırılır. Konuşma sinyali yavaş zamanlanan değişken bir sinyaldir.

3.8 Mel-frekanslı Cepstrum Katsayıları İşlemcisi

MFCC'ler, insan kulağının kritik bant genişliklerinin bilinen frekans varyasyonuna dayanır, düşük frekanslarda filtre aralıkları doğrusaldır; yüksek frekanslarda ise filtre aralıkları logaritmiktir, böylece konuşmanın fonetik olarak önemli özellikleri yakalanır. Bu, 1 000 Hz'nin altında doğrusal frekans aralığına ve 1 000 Hz'nin üzerinde logaritmik frekans aralığına sahip olan mel-frekans ölçeğinde (MFCC) ifade edilir. MFCC işlemcinin temel amacı, insan kulağının davranışını taklit etmektir (Peker vd., 2015).

MFCC'ler, temel olarak insan kulağının kritik bant genişlikleri ile frekans arasındaki ilişkisini tanımlamak için kullanılan bir algoritma türüdür. Bu yöntem temel olarak adım vektörlerinin analizi ve çıkarımı için kullanılır (Peker vd., 2015).

3.9 (Çok Değişkenli) Minimum Artıklık Maksimum İlgililik (mRMR)

mRMR temel olarak sınıf etiketleri arasında sadece en ilgili olan öznitelikleri seçmeye çalışan bir filtreleme algoritmasıdır. Algoritma bir yandan en ilgili öznitelikleri tanımlarken, bir yandan da seçili/ilgili öznitelikler arasındaki artıklığı en aza indirmeye çalışır. Spesifik olarak, mRMR algoritması her özniteliği ve sınıf vektörünü

(bağımlı değişken veya çıktı değişkeni) ayrık/süreksiz rastgele/rassal değişken olarak ele alır. İki öznitelik arasındaki veya bir öznitelik ile sınıf vektörü arasındaki benzerliği ölçmek için karşılıklı bilgi ölçümünü ($I(x,y)$) kullanır. İki rastgele değişken birbirinden tamamen bağımsız ise karşılıklı bilgi (MI) değeri 0'dır; bu değer simetriktir ve negatif olamaz (Mueen ve Keogh, 2016).

3.10 Sınıflandırma

Konuşma sınıflandırması için yaygın olarak kullanılan başlıca teknikler şunlardır:

3.10.1 Dinamik Zaman Bükmesi (DTW)

Bu teknikte, çıkarımı yapılan kelimeler referans kelimelerle karşılaştırır. Her referans kelimenin bir dizi spektrası vardır; ancak kelime içindeki ayrı sesler arasında bir ayrım yoktur. Bir kelime farklı hızlarda telaffuz edilebildiğinden, zamanın normalize edilmesi gerekecektir. Dinamik Zaman Bükmesi, bilinmeyen sözcüğün zaman boyutunun, herhangi bir referans kelimeyle benzerliği oluncaya kadar değiştirildiği (uzatılıp kısaltıldığı) bir programlama tekniğidir (Muhammad vd., 2018).

3.10.2 Saklı Markov Modellemesi (HMM)

Günümüz itibariyle, konuşma tanımada en çok kullanılan ve en başarılı olan örüntü tanıma yöntemi Saklı Markov Modellemesidir (HMM). Bu yöntem Markov Modelinden türetilmiş olan matematiksel bir modeldir. Konuşma tanımada hafif değişikliklere uğramış bir Markov Modeli kullanılır. Konuşma en küçük işitilebilir birimlere bölünür (sadece ünlüler ve ünsüzler olarak değil, aynı zamanda birleşik sesler olarak da). Tüm bu birimler Markov Modelinde birer durum olarak temsil edilirler. Bir kelime Saklı Markov Modeline girdiğinde, en uygun modelle (birim) karşılaştırılır. Bir durumdan diğerine geçiş, belli olasılıklarla mümkündür, bu değerlere geçiş olasılıkları denir. Örneğin: xq ile başlayan bir kelimenin olasılığı neredeyse sıfırdır. Bir durumun kendi kendine geçiş yapması, ses kendini tekrarlayacak olursa mümkündür. Markov Modelleri gürültülü ortamlarda oldukça iyi bir performans sergilemektedir çünkü her bir ses birimi ayrıca ele alınmaktadır. Gürültü nedeniyle herhangi bir ses birimi kaybolsa bile, modelin kayıp ses birimini,

ses birimlerinin birbirine geçiş olasılıklarına bakarak tahmin etmesi mümkündür (Deng vd., 2013).

3.10.3 Sinir Ağları (NN)

Sinir ağlarının Markov modelleri ile birçok benzerliği vardır. Her ikisi de grafik olarak temsil edilen istatistiksel modellerdir. Markov modelleri durum geçişleri için olasılıkları kullanırken, sinir ağları bağlantı sağlamlığını ve fonksiyonları kullanır. Önemli bir fark, Markov zincirleri birbirine seri bağlı iken sinir ağlarının temelde paralel bağlı olmasıdır. Konuşma içindeki frekanslar paralel olarak ortaya çıkarken, hece dizileri ve kelimeler esasen seridir. Bu, her iki tekniğin de aslında farklı bir bağlamda çok güçlü olduğu anlamına gelir (Gevaert vd., 2010).

Sinir ağları uygulamasının başlıca zorluğu, bağlantının ağırlıklarını uygun şekilde belirlemektir; Markov modeli içinse zorluk geçiş ve gözlem olasılıklarını bulmaktır. Birçok konuşma tanıma sisteminde, her iki teknik birlikte kullanılmaktadır ve aralarında simbiyotik bir ilişki kurulmaktadır. Sinir ağları, yüksek seviyede paralel ses girdisinden fonem olasılıklarını öğrenmede çok iyi performans gösterirken; Markov modelleri, sinir ağlarının sağladığı fonem gözlem olasılıklarından en olası fonem veya kelimesini dizisini üretmede faydalanabilir. Bu, dilin anlaşılmasında benimsenen hybrid yaklaşımın özüdür (Ray vd., 2017).

3.10.4 Destek Vektör Makinesi (SVM)

“Destek Vektör Makinesi” (Support Vector Machine-SVM), hem sınıflandırma hem de yordama zorlukları için kullanılabilen denetimli bir makine öğrenme algoritmasıdır. Ancak, çoğunlukla sınıflandırma problemlerinde kullanılır. Bu algoritmada, her bir veri öğesi *n*-boyutlu uzayda bir nokta olarak temsil edilir, burada *n* sayısı çıkarımı yapılan öznitelik sayısıdır; ve her bir koordinatın değerini, karşılık gelen özniteliğin o verideki varlığı belirler. Ardından, iki sınıfı en net şekilde ayıran hiper-düzlem bulunarak sınıflandırma yapılır (Jain vd., 2020)(Koolagudi ve Rao, 2012).

3.10.5 Rasgele Orman

Rasgele orman (veya rasgele ormanlar) bir sınıflandırıcıdır; eğitim sırasında birçok karar ağacı oluşturulur ve en çok tekrarlanan karar ağacı, yani o veri kümesinin modu, sınıf olarak belirlenir (Chen vd., 2020).

Rasgele Orman (RF), ortak bir dağılımı olan, rastgele dağıtılmış veri kümeleri için regresyon ağaçlarından oluşur. Algoritma önce eğitim kümesinden rastgele örnekleme yaparak bir örneklem oluşturur; daha sonra her veri örneklemi için ağaç tabanlı bir sıralama oluşturur; ve son olarak bu ağaçlardan oluşan bir orman ortaya koyar (Yu vd., 2018).

4. SİMÜLASYON SONUÇLARI

4.1 Veri Kümesi

Büyük verinin fazlasıyla yaygınlaştığı günümüzde, Google'ın TensorFlow ve AIY ekiplerinin hazırladığı serbestçe kullanılabilen komut konuşma veri kümesi bu alana yoğunlaşan araştırmacılar için güzel bir fırsattır. Bu veri kümesi, *Git (Go)*, *Evet (Yes)* veya *Dur (Stop)* gibi komutlardan oluşan yaklaşık 65 000 adet birer saniyelik ses dosyası içerir. Veri kümesi hakkında daha fazla bilgi (https://en.wikipedia.org/wiki/TensorFlow) adresinde bulunabilir. Kullandığımız dosyanın toplam büyüklüğü 1.4GB'dır.

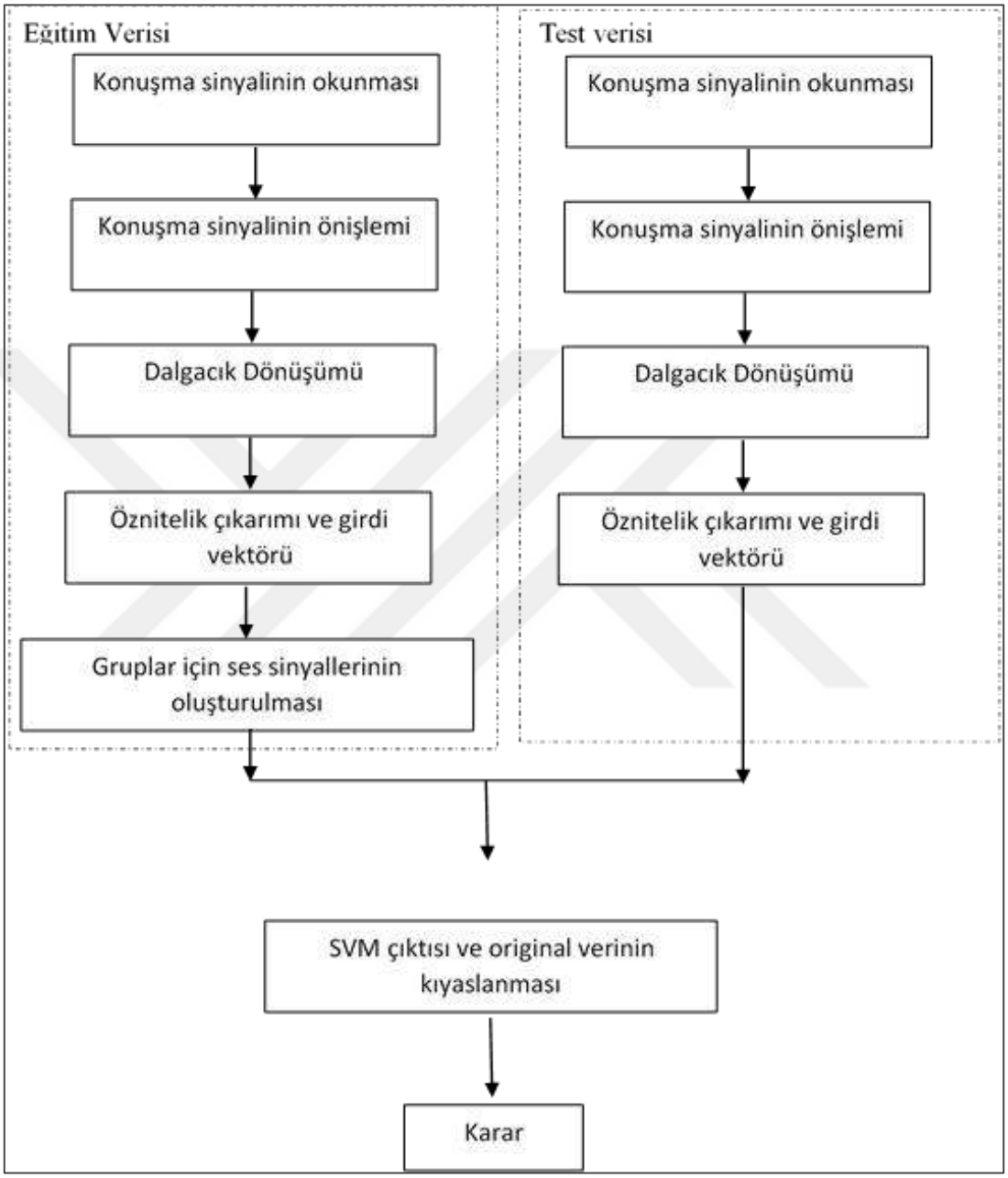
4.1.1 Sürekli

Bu tezde TensorFlow araç kutusunu kullandık. TensorFlow, veri akışı yönelimli programlama için bir çerçeve sunmaktadır. Python programlarında kullanılır ve Python ve C++ koduyla uyumludur. TensorFlow, makine öğrenmesi alanında oldukça popülerdir. Bilimsel araştırmalarda ve ürün geliştirme amaçlı olarak, şu anda konuşma tanıma, Gmail, Google Fotoğraflar ve Google arama gibi ticari Google ürünlerinde farklı ekipler tarafından sunulmaktadır (Eitel vd., 2015). Google Haritalar'da, Sokak Görünümünde çekilen fotoğraflar TensorFlow'a dayalı bir AI kullanılarak analiz edilir (Papernot vd., 2016). Bu ürünlerin birçoğu daha eski bir yazılım olan DistBelief'i kullanıyordu. TensorFlow ilk olarak Google Brain ekibi tarafından Google'ın dahili ihtiyaçları için geliştirilmiş ve daha sonra Apache 2.0 açık kaynak lisansı altında piyasaya sürülmüştür (Mojumder vd., 2018)(Nair ve Hinton, 2010).

TensorFlow, matematiksel işlemleri grafik biçiminde gösterir. Grafik, TensorFlow tarafından gerçekleştirilecek tüm işlemlerin sıralı dizisini temsil eder.

4.2 Deneysel Sonuçlar

Konuşma tanıma için önerilen yöntemin akış şeması şekil 4.1'de gösterilmektedir.

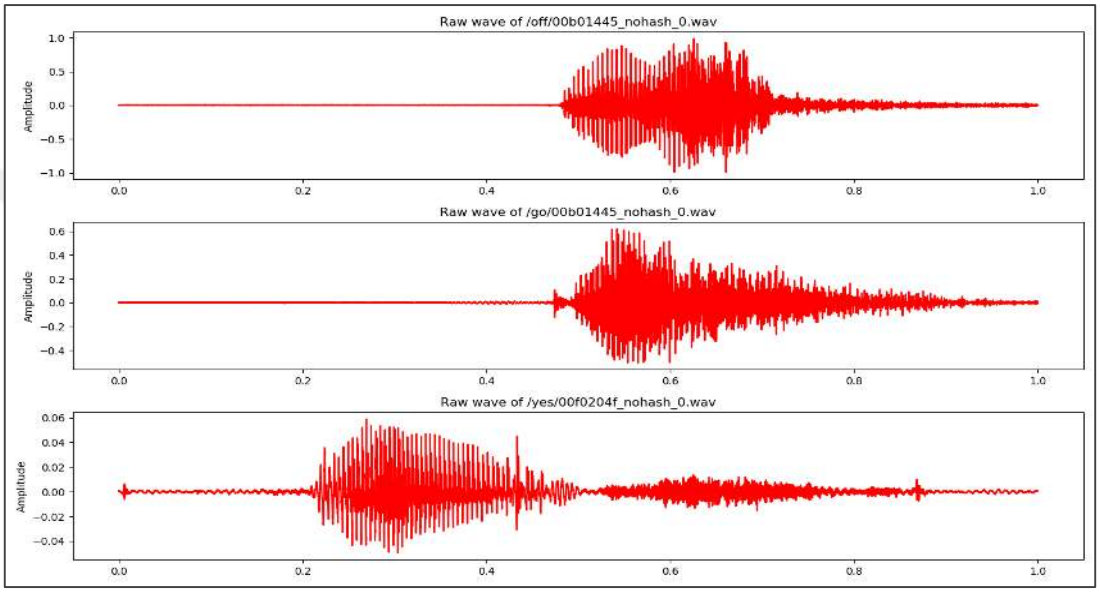


Şekil 4.1 Önerilen yöntemin akış şeması

Bir boyutlu vektörlerin görselleştirilmesi kolaydır, ancak konuşma tanımada ham genlik verileri ile çalışılır. En yaygın yaklaşımlar ses dosyalarını spektrogramlara veya MFCC'ye (Mel-Frequency Cepstral Katsayıları) dönüştürmektir. Bu tezde

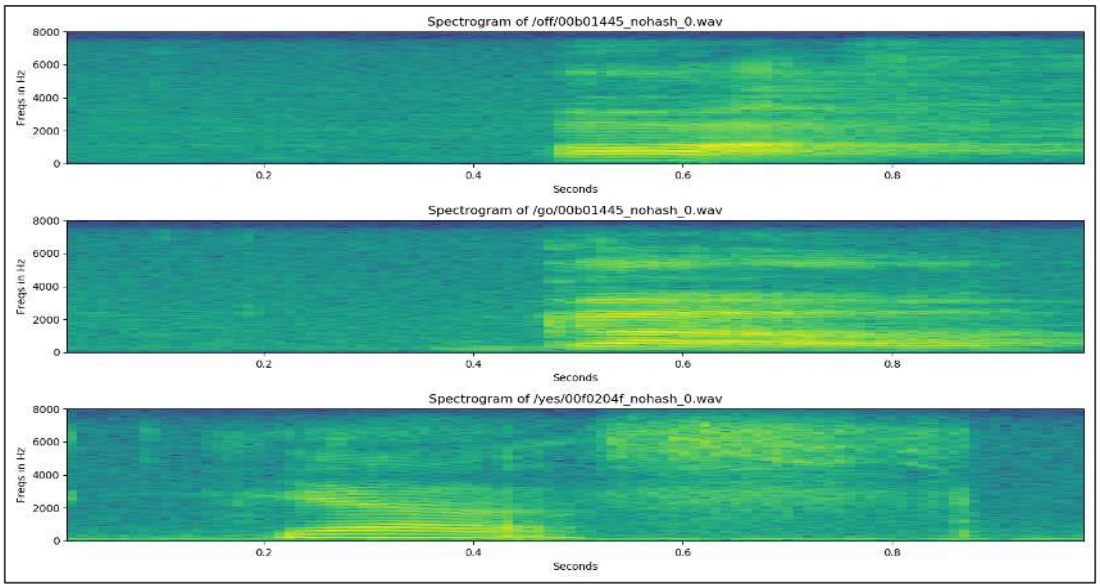
spektrogram yöntemi uygulanacaktır (daha net söylemek gerekirse log-spektrogram, çünkü görselleştirilmesi daha iyidir).

Spektrogram, ses dosyasının frekans bölgesinde temsilidir (ham verilerde var olan zaman bölgesinden farklı olarak). Ham verileri spektrogramlara dönüştürmek için Kısa Zamanlı Fourier Dönüşümü (STFT) uygulanır. Örnek ses sinyalleri şekil 4.2'de gösterilmektedir.



Şekil 4.2 Ses Sinyalleri

Log-spektrogramların grafikleri şekil 4.3'te gösterilmiştir.



Şekil 4.3 Seslerin Spektrogramı

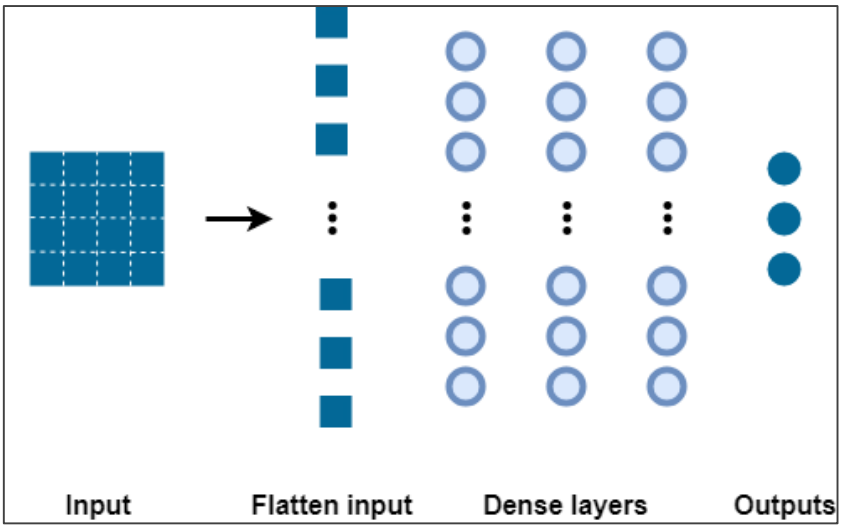
4.3 Off (Kapalı), Go (git) ve Yes (Evet) Komutlarının Log-Spektrogramları

Nyquist teoremine göre örnekleme hızı 16 000 ise, frekanslar [0, 8 000] aralığındadır. Spektrogramın frekans çözünürlüğü yaklaşık 8 000/98 ≈ 80 Hz iken, insan kulağı 3,6 Hz çözünürlükte çalışır ve spektrogramdan çok daha hassastır. Bununla birlikte, amaçlarımız doğrultusunda, mevcut frekans çözünürlüğü yeterli olacaktır.

Önerilen yöntemin adımları:

1) Eğitim/test ayrıştırması da dahil olmak üzere seçilen kelimeler için verinin hazırlanması
2) Ham verilerin log-spektrogramlara dönüştürülmesi
3) Makine öğrenmesi modelinin eğitilmesi
4) Test verileri üzerinde doğruluk kontrolü

Veri kümesi eğitim ve test alt kümeleri olmak üzere iki gruba ayrılır. Sinir ağlarının erken durması durumu için bir doğrulama verisine ihtiyaç olacaktır, bu nedenle eğitim verilerinin bir kez daha bölünmesi gerekir. CNN Mimarisi şekil 4.4'de gösterilmiştir.



Şekil 4.4 CNN Mimarisi

4.4 Sinir Ağları Mimarisi

Mimari, Keras Model nesnesini döndüren derin fonksiyonda uygulanır. Son katman, sigmoid aktivasyon fonksiyonu ile sınıflandırma olasılıklarını çıktılamak için kullanılır.

İlk tam bağlantılı katmandaki 512 nöronla kullanılan 177×98 matris ile temsil edilen bir konuşma sinyali girdisi, 177*98*512 = 8 881 152 farklı sayıdaki ağırlığın optimize edilmesini gerektirir. Takip eden katmanların eklenmesiyle bu muazzam sayı daha da artar. Bu durumun hızlı bir şekilde aşırı uymaya yol açabileceği açıktır. Bu sorunun çözümü, bir girdinin boyutlarını akıllıca azaltmaktır. Evrişimli Sinir Ağları (CNN), girdinin bir konuşma sinyali olduğu durumlarda açıkça kullanılır; biz de bunları bir sonraki bölümde uygulayacağız.

4.5 Mimari: Çok Katmanlı Derin Evrişimli Sinir Ağı

Evrişimli Sinir Ağları, büyük ölçekli konuşma tanımada kullanılan en popüler derin öğrenme mimarisidir. Bu araştırma tezi, ikili sınıfta konuşma tespiti için uygulanan çok katmanlı bir CNN ile ilgilidir. Derin Evrişimli Sinir Ağları (DCNN), evrişimsel bir katmandan yoğun/tam bağlantılı bir katmana kadar çok katmanlı ağlar üzerinde gerçekleştirilen çeşitli işlemlere sahiptir. Bu çalışmada uygulanan DCNN, 2 evrişimli tabaka, 1 birleştirme tabakası, 1 evrişim tabakası, 1 birleştirme tabakası, 1 yassıltma tabakası, 2 tam bağlantılı tabakadan oluşmaktadır.

4.5.1 Evrişim

İlk evrişimsel katman, 1X64 konuşma sinyali girdisini 1X5 boyutunda 20 çekirdek ile filtreler. (4.1) denkleminde verilen Evrişim işlemi, tüm konuşma sinyali girdisinde gerçekleştirilir ve 20 farklı çekirdek uygulanarak 20 öznitelik haritası elde edilir. Öznitelik haritası, önceki katmana uygulanan tek bir filtrenin çıktısıdır. Tek bir filtre, önceki katmanın tamamı üzerinden, her seferinde bir piksel hareket ettirilerek geçirilir. Filtre, önceki katmanın girdisi ile birlikte evrişim işlemine tabi tutulur.

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau).g(t-\tau)d\tau \quad (4.1)$$

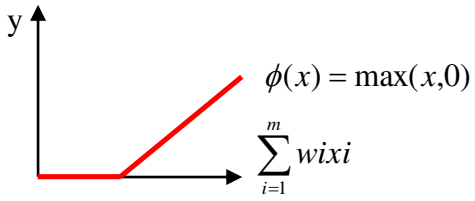
Evrişimli katmanlarda birden çok öznitelik haritası oluşturmak için birden çok filtre uygulanabilir. İkinci evrişimsel katmanda, 30 öznitelik haritası elde etmek için birinci evrişimsel katmanın konuşma sinyali çıktısı üzerinde 1X5 boyutunda 30 çekirdekle evrişim gerçekleştirilir. Üçüncü evrişim tabakası, ilk birleştirme tabakasından sonra devreye girer. Üçüncü evrişimsel katmanda, 50 öznitelik haritası oluşturmak için 1X5 boyutunda 50 çekirdek kullanılır. DCNN'deki bulanıklık, kabartma, kenar, düzleştirme, keskinleştirme gibi çekirdekler modelin kendisi tarafından seçilir. Baskın bir şekilde kullanılan çekirdek kenar filtresidir. Evrişimi gerçekleştirirken, DCNN pikseller arasındaki uzamsal ilişkileri korur. Küçük öznitelikler ortadan kaldırılmayıp korunur. Evrişimden sonraki bir konuşma sinyali çıktı boyutu (O), şu formülle hesaplanır:

$$O = ((W-K+2P)/S) +1 \quad (4.2)$$

W bir konuşma sinyalinin girdisinin yüksekliği/genişliği, K Çekirdek boyutu, P Dolgulama ve S Adım'dır. Örneğin, bir konuşma sinyali girdisinin yüksekliği (W) 64, çekirdek (K) 3X3, Adım (S) 1 ise ve dolgulama (P) sıfırsa, konuşma sinyali çıktısının boyutu (O) 62X62 olur .

4.5.2 ReLU-Rektifiye Doğrusal Birim

Evrişimsel katmanın çıktısı olarak elde edilen öznitelik haritasında negatif değerler bulunur. Tüm negatif değerleri kaldırmak ve yalnızca pozitif değeri elde tutmak için öznitelik haritaları üzerinde ReLU adlı bir aktivasyon fonksiyonu uygulanır. Bu aynı zamanda modelin doğrusal olmama özelliğini de arttırır. Rektifiye Fonksiyonu şekil 4.5'de gösterilmiştir.



Şekil 4.5 Rektifiye Fonksiyonu

ReLU fonksiyonu, doğrusal olmamayı artırmak için ve aynı zamanda konuşma sinyalindeki hassas özellikleri korumak için tüm evrişimsel katmanlara uygulanır. (Scherer vd., 2010)'de ReLU, performansını artırmak için kısıtlı Boltzmann Makinesine uygulanmıştır. Ancak model verimliliğini artırmak için ReLU'nun DCNN'lerde de uygulanması mümkündür.

4.5.3 Ortalama Birleştirme/Alt Örnekleme

Konuşma sinyallerinin öznitelik haritası tıkalı veya döndürülmüş olmak üzere farklı yönlerdedir. Birleştirme işlemi parametre sayısını azaltır, böylece aşırı uyumlamayı da azaltmış olur. Birleştirme, ağın işlem süresini de azaltır. Birleştirme, Uzamsal Değişimsizliğe izin verir ve özniteliklerin döndürülmüş, ölçeklendirilmiş veya tıkalı olması bu işlem için farketmez. Öznitelikler korunur ve bu birleştirme katmanında veri kaybı olmaz. Evrişimli ağlarda farklı birleştirme işlemlerinin değerlendirilmesi (Min vd., 2017)'da ele alınmıştır. Lenet (Min vd., 2017)'de, maksimum birleştirme/aşağı örnekleme (max pooling/down sampling) kullanılmıştır, bu işlemde belirli bir komşuluktaki maksimum değer bulunur ve diğer bütün değerler bu maksimum değerle değiştirilir. Burada tartışılan çalışmada ortalama birleştirme/alt örnekleme kullanılmıştır, bunda ise, pikselin değeri çevresinin ağırlıklı ortalamasıyla değiştirilir. Az sayıda model hem maksimum hem de ortalama birleştirme tekniklerini aynı anda kullanır. İkinci ve üçüncü evrişimsel katmanlardan sonra iki birleştirme katmanı sırasıyla eklenir. Daha düşük boyutta bir öznitelik haritası oluşturmak için ikinci ve üçüncü evrişimsel katmanların çıktısına 2X2'lik ortalama birleştirme filtresi uygulanır. Birleştirme yaptıktan sonra bir konuşma sinyalinin çıktı boyutu (O) formül (4.3)'te şu şekilde verilir:

$$O = ((W-K)/S) +1 \tag{4.3}$$

Burada W, evrişimsel katmandan gelen bir konuşma sinyalinin yüksekliği/genişliği, K, çekirdek boyutu ve S, adım'dır. Örneğin, evrişimsel katmandan konuşma sinyali yüksekliği (W) 62, çekirdek (K) 3X3 ise ve birleştirme adımı (S) 2 olursa, konuşma sinyali çıktı boyutu (O) 31X31 olur.

4.5.4 Yassıltma

Birleştirme katmanının çıktısı indirgenmiş veya küçültülmüş bir öznitelik haritasıdır ve bunun tek bir vektörler kümesine dönüştürülmesi gerekir. Yassıltma bu işlemi yerine getirmeye yardımcı olur. Yassıltma, dönüştürülen öznitelik haritasını sinir ağına yönlendirir.

4.5.5 Tam / Yoğun Bağlantı

Tam Bağlantı, nöronların bir önceki katmandaki tüm aktivasyonlarla tam bağlantılarının olduğu sinir ağı katmanıdır. Aktivasyonları bir matris çarpımı ve ardından bir sapma dengelemesidir. Bu modelde iki tam bağlantı katmanı vardır. Model, kayıp fonksiyonunu denklem (4)'te verildiği gibi hesaplar ve eğitim konuşma sinyallerini kullanarak modeli eğitmek için ağırlıkları ayarlamak üzere geriye doğru yayılım yapar. Daha sonra test konuşma sinyalleri, kayıp/maliyet fonksiyonunu bulmak için modele beslenir ve çok katmanlı mimari için doğruluk hesaplanır. Tam bağlantının çıkış katmanı, nesneyi iki sınıfta algılamak için sigmoid etkinleştirme fonksiyonunu uygular. Çok sınıflı nesne algılama ve sınıflandırma durumunda, bir softmax aktivasyon fonksiyonu uygulanabilir.

$$\boldsymbol{H(p,q) = -\sum_x p(x).log\, q\,(x)} \tag{4.4}$$

4.6 Evrişimli Sinir Ağları (CNN)

CNN, sinir ağının başlangıcında konuşma sinyalininin boyutunu küçültmek ve bir önişleme tabi tutmak için ek katmanlar kullanır. CNN'in temel mimarisi şunları içerir:

4.6.1 Evrişimli Katman

Giriş sinyalini filtrelemek ve bazı ek konuşma sinyali özniteliklerini çıkarmak için evrişimsel operatör kullanır.

4.6.2 Aktivasyon Fonksiyonu

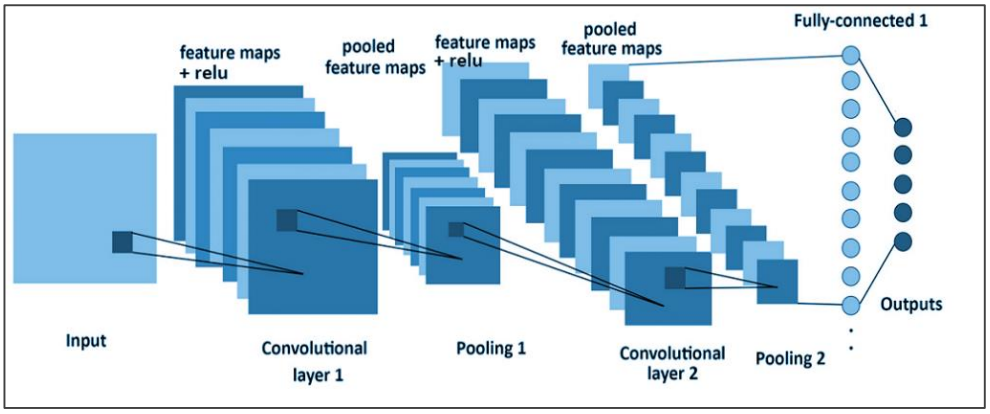
Evrişimsel katmanın çıktılarına rektifiye fonksiyonu gibi doğrusal olmayan fonksiyonlar uygular.

4.6.3 Birleştirme Katmanı

max()veya sum() gibi fonksiyonlar uygulayarak alt örnekleme işlemi gerçekleştirir.

4.6.4 Tam-Bağlantılı Katman

Önceki katmandaki her nöron, bir sonraki katmandaki her bir nöronla bağlantılıdır ve bu katmanların sonuncusu sinir ağının çıktılarını üretir. Önerilen Evrişimli Sinir Ağı mimarisi şekil 4.6'da gösterilmektedir.

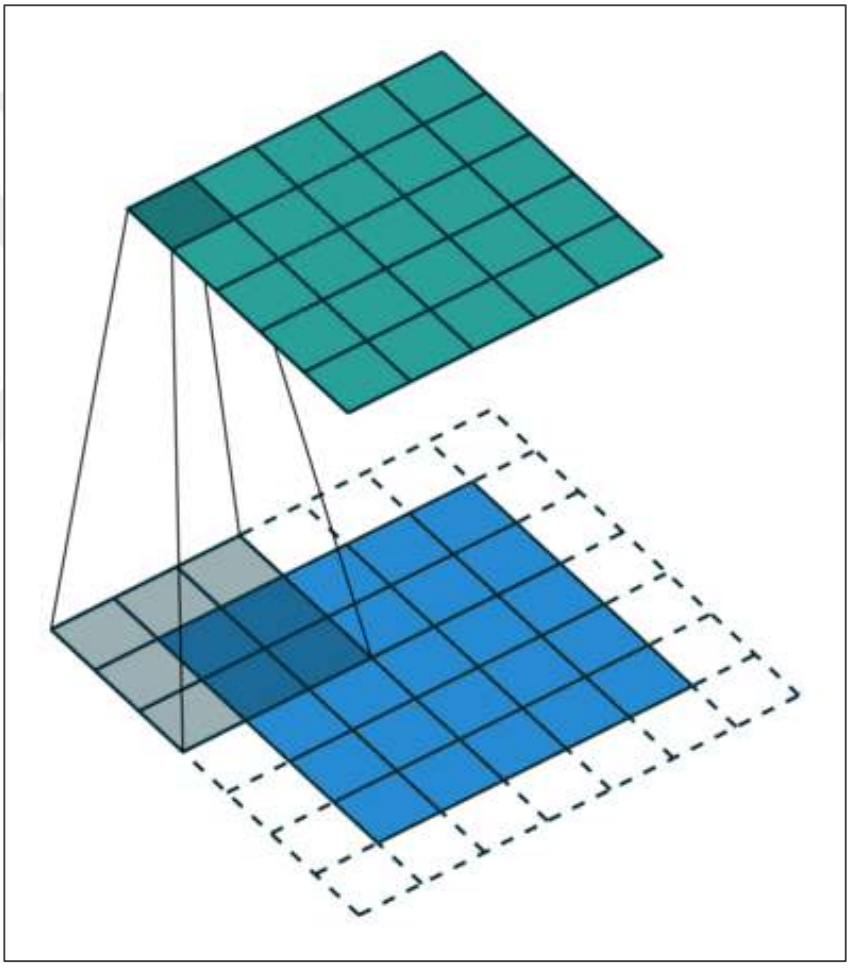


Şekil 4.6 Evrişimli Sinir Ağı Mimarisi

Başka bir deyişle, evrişimsel katman ve birleştirme katmanı konuşma sinyalinin girdisinin yüksek seviye özniteliklerini temsil ederler. Birleştirme katmanı, aşırı uyumlamayı kontrol etmek için konuşma sinyalinin boyutunu azaltır. Ayrıca, evrişimsel ve birleştirme katmanları, geri yayılım algoritması sırasında kullanım için hala geçerlidirler, böylece sinir ağı azalan eğim yaklaşımları kullanılarak eğitilebilir.

CNN’i Keras'ta uygulayacak olursak; deep_cnn fonksiyonu Keras Model nesnesini döndürür, bu da evrişim/birleştirme katmanlarından ve bir yoğun katmandan oluşan CNN mimarisini temsil eder. Ek olarak, kararlılığı artırmak ve fazla uyumlamayı önlemek için batch normalizasyon yöntemi kullanılmaktadır.

Burada üç adet evrişimli katman kullandık. Her durumda katmanlarda 32 filtre vardı, 3×3 pencere, adım = 1 ve aynı dolgulama kullanılmıştır. Bir evrişim tabakası filtresi, şekil 4.7'de gösterilmektedir.

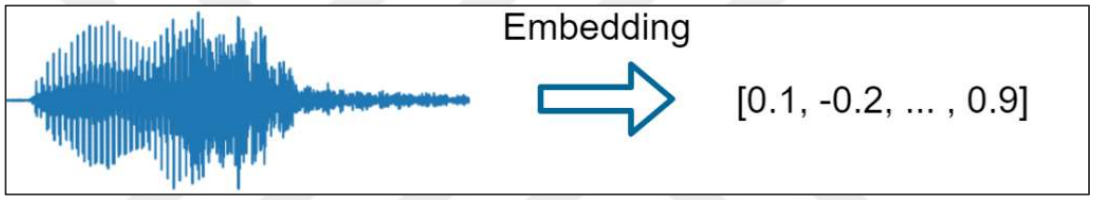


Şekil 4.7 3×3 pencereli, adım = 1 ve aynı dolgulamaya sahip olan evrişim tabakası

Şekil 4.8'de gösterildiği gibi, birleştirme katmanları, max() işlemini kullanarak ve alt örnekleme yaparak girdi boyutunu yarıya indirmektedir.

Şekil 4.8 2×2 pencereli ve adım = 2 olan MaxPooling katmanı

Adımların boyutları ve filtre sayısı hiper-parametrelerdir ve daha da optimize edilebilir. Sinyalin sayılara gömme işlemi şekil 4.9'da gösterilmektedir.



Şekil 4.9 Sinyali sayılara gömme

Programın çıktısı aşağıdaki Tablo 4.10 'da gösterilmiştir.

Tablo 4.1 Programın çıkış verileri

Layer (type)	Output Shape	Param #
inputs (InputLayer)	(None, 177, 98, 1)	0
block1_conv (Conv2D)	(None, 177, 98, 32)	320
block1_pool (MaxPooling2D)	(None, 89, 49, 32)	0
block1_norm (BatchNormalizat	(None, 89, 49, 32)	128
block2_conv (Conv2D)	(None, 89, 49, 32)	9248
block2_pool (MaxPooling2D)	(None, 45, 25, 32)	0
block2_norm (BatchNormalizat	(None, 45, 25, 32)	128
block3_conv (Conv2D)	(None, 45, 25, 32)	9248
block3_pool (MaxPooling2D)	(None, 23, 13, 32)	0
block3_norm (BatchNormalizat	(None, 23, 13, 32)	128

flatten (Flatten)	(None, 9568)	0
dense (Dense)	(None, 64)	612416
dense_norm (BatchNormalizati	(None, 64)	256
dropout (Dropout)	(None, 64)	0
pred (Dense)	(None, 3)	195

Total params: 632.067
Trainable params: 631.747
Non-trainable params: 320

5. SONUÇ VE TARTIŞMA

Konuşma tanıma tercihen teknik uygulamalarda, örneğin zaman çizelgesi bilgileri gibi otomatik diyalog sistemlerinde kullanılır. Sadece sınırlı bir kelime haznesinin kullanıldığı yerlerde konuşmacıdan bağımsız konuşma tanıma başarılı bir şekilde uygulanabilir. İngilizce seslendirilen 0'dan 9'a kadar olan rakamları tanıma sistemleri neredeyse %100 tanıma oranına ulaşmaktadır. "Konuşmacıya bağlı" konuşma tanıma kullanılırken de çok yüksek tanıma oranlarına ulaşılabilmektedir. Ancak, yüzde 95'lik doğruluk oranı bile çok düşük olarak algılanabilir, çünkü çok fazla iyileştirmeye ihtiyaç duyulacaktır. "Konuşmacıya bağlı" bir konuşma tanıma sisteminin başarısı için belirleyici faktör, kullanıcı ve sistem arasındaki etkileşimdir; bu ise kullanıcının kişisel tanıma sonucunu doğrudan veya dolaylı olarak etkiler. Uygulamaya kaynak teşkil edecek sözlüğün boyutuna ve esnekliğine ek olarak, akustik kaydın kalitesi de belirleyici bir rol oynar. Doğrudan ağzın önüne yerleştirilen mikrofonlar (örneğin, kulaklıklar veya telefonlar), daha uzakta konumlandırılan oda mikrofonlarına göre önemli ölçüde daha yüksek algılama doğruluğu sağlar. Ancak yine de, pratikte sonucu en fazla etkileyen faktörler, diktedeki kesin telaffuzla beraber uyumlu ve akıcı bir konuşmadır, böylece kelime ilişkileri ve kelime dizisi olasılıkları tanıma sürecine en uygun şekilde dahil edilebilir.

Bu tezde insan konuşma sinyallerinden kelimeleri tanımak için derin öğrenme yöntemi kullanılmıştır. Önerilen yöntemin uygulanması için Python adım adım kullanılır ve sunulur. Çalışmamız neticesinde, konuşma tanıma için CNN'in (evrişimli sinir ağları) oldukça başarılı bir araç olduğu görülmüştür, bu durum daha fazla etiketin söz konusu olması halinde daha sofistike bir mimari inşası yönünde cesaret verici olmuştur. Daha farklı çözümleri denemek için Kaggle'ın TensorFlow Speech Recognition Challenge platform kullanılmıştır.

KAYNAKLAR

Aadit, M. N. A., Kirtania, S. G., & Mahin, M. T. (2016). Pitch and formant estimation of bangla speech signal using autocorrelation, cepstrum and LPC algorithm. In *2016 19th International Conference on Computer and Information Technology (ICCIT)* (pp. 371-376). IEEE.

Amin, T. B., & Mahmood, I. (2008). Speech recognition using dynamic time warping. In *2008 2nd international conference on advances in space technologies* (pp. 74-79). IEEE.

Arısoy, E. (2004). *Turkish dictation system for radiology and broadcast news applications* (Doctoral dissertation, Master's thesis, Bogaziçi University, Istanbul, Turkey).

Arısoy, E., Sak, H., & Saraçlar, M. (2007). Language modeling for automatic Turkish broadcast news transcription. In *Eighth Annual Conference of the International Speech Communication Association*.

Arisoy, E., & Saraçlar, M. (2015). Multi-stream long short-term memory neural network language model. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Arisoy, E., & Saraglar, M. (2018). Turkish broadcast news transcription revisited. In *2018 26th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

Arisoy, E., Can, D., Parlak, S., Sak, H., & Saraçlar, M. (2009). Turkish broadcast news transcription and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(5), 874-883.

Arnold, D. J., Balkan, L., Meijer, S., Humphreys, R. L., & Sadler, L. (1994). Machine Translation: an Introductory Guide. Blackwells-NCC, London. 1994.

Assaleh, K. T., & Mammone, R. J. (1994). New LP-derived features for speaker identification. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 630-638.

Aşlıyan, R. (2008). *Design and implementation of Turkish speech recognition engine* (Doctoral dissertation, DEÜ Fen Bilimleri Enstitüsü).

Avci, E. (2007). An automatic system for Turkish word recognition using discrete wavelet neural network based on adaptive entropy. *Arabian Journal for Science and Engineering*, *32*(2), 239.

Avci, E., Turkoglu, I., & Poyraz, M. (2005). Intelligent target recognition based on wavelet packet neural network. *Expert Systems with Applications*, *29*(1), 175-182.

Can, B., & Artuner, H. (2013). A syllable-based Turkish speech recognition system by using time delay neural networks (TDNNs). In *2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR)* (pp. 219-224). IEEE..

Charpentier, F., Micca, G., Schukat-Talamazzini, E., & Thomas, T. (1995). The recognition component of the SUNDIAL project. In *Speech Recognition and Coding* (pp. 345-348). Springer, Berlin, Heidelberg.

Chen, L., Su, W., Feng, Y., Wu, M., She, J., & Hirota, K. (2020). Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Information Sciences*, *509*, 150-163.

Chi, D., Caron, M., Templeton, R., & Stewart, T. (2004). *U.S. Patent Application No. 10/427,353*.

Danieli, M., & Gerbino, E. (1995). Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI spring symposium on Empirical Methods in Discourse Interpretation and Generation* (pp. 34-39).

Deng, L. (2016). Deep learning: from speech recognition to language and multimodal processing. *APSIPA Transactions on Signal and Information Processing*, *5*.

Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8599-8603). IEEE.

Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., & Burgard, W. (2015). Multimodal deep learning for robust RGB-D object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 681-687). IEEE.

Fadhilah, H., Djamal, E. C., Ilyas, R., & Najmurrokhman, A. (2018). Non-halal ingredients detection of food packaging image using convolutional neural networks. In *2018 International symposium on advanced intelligent informatics (SAIN)* (pp. 131-136). IEEE.

Foote, J. (1997). A similarity measure for automatic audio classification. In *Proc. AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*.

Fu, K. S. (1976). Pattern recognition and image processing. *IEEE transactions on*

computers, *100*(12), 1336-1346.

Gevaert, W., Tsenov, G., & Mladenov, V. (2010). Neural networks used for speech recognition. *Journal of Automatic control*, *20*(1), 1-7.

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.

Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, *2015*.

Hu, G. S. (2005). *Introduction to digital signal processing* (pp. 26-31). Tsinghua University Press, Beijing.

Jain, M., Narayan, P., Balaji, A., Bhowmick, & Muthu, R. K. (2020). "Speech emotion recognition using support vector machine," *arXiv Prepr. arXiv2002.07590*, 2020.

Kalia, A., Sharma, S., Pandey, S. K., Jadoun, V. K., & Das, M. (2020). Comparative Analysis of Speaker Recognition System Based on Voice Activity Detection Technique, MFCC and PLP Features. In *Intelligent Computing Techniques for Smart Energy Systems* (pp. 781-787). Springer, Singapore.

Karakaş, M. (2010). *Computer based system control using voice input* (Doctoral dissertation, DEÜ Fen Bilimleri Enstitüsü).

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. *International journal of speech technology*, *15*(2), 99-117.

Lawrence, R. (2008). *Fundamentals of speech recognition*. Pearson Education India.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541-551..

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *Ismir* (270, pp. 1-11).

Loizou, P. C. (2013). *Speech enhancement: theory and practice*. CRC press.

Mantha, V., Duncan, R., Wu, Y., Zhao, J., Ganapathiraju, A., & Picone, J. (1999). Implementation and analysis of speech recognition front-ends. In *Proceedings*

IEEE Southeastcon'99. Technology on the Brink of 2000 (Cat. No. 99CH36300) (pp. 32-35). IEEE.

McCree, A. V. (2005). *U.S. Patent No. 6,889,185*. Washington, DC: U.S. Patent and Trademark Office.

Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in bioinformatics*, *18*(5), 851-869.

Mojumder, S. A., Louis, M. S., Sun, Y., Ziabari, A. K., Abellán, J. L., Kim, J., & Joshi, A. (2018). Profiling dnn workloads on a volta-based dgx-1 system. In *2018 IEEE International Symposium on Workload Characterization (IISWC)* (pp. 122-133). IEEE.

Mueen, A., & Keogh, E. (2016). Extracting optimal performance from dynamic time warping. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2129-2130).

Muhammad, H. Z., Nasrun, M., Setianingsih, C., & Murti, M. A. (2018). Speech recognition for English to Indonesian translator using hidden Markov model. In *2018 International Conference on Signals and Systems (ICSigSys)* (pp. 255-260). IEEE.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 807-814).

Papernot, N., Faghri, F., Carlini, N., Goodfellow, I., Feinman, R., Kurakin, A., & Matyasko, A. (2016). Technical report on the cleverhans v2. 1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*.

Peckham, J. (1991). Speech Understanding and Dialogue over the telephone: an overview of the ESPRIT SUNDIAL project. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

Peker, M., Sen, B., & Delen, D. (2015). Computer-aided diagnosis of Parkinson's disease using complex-valued neural networks and mRMR feature selection algorithm. *Journal of healthcare engineering*, *6*(3), 281-302.

Phadke, S., Limaye, R., Verma, S., & Subramanian, K. (2004). On design and implementation of an embedded automatic speech recognition system. In *17th International Conference on VLSI Design. Proceedings.* (pp. 127-132). IEEE.

Rabiner, L. R., Juang, B. H., Levinson, S. E., & Sondhi, M. M. (1985). Recognition of isolated digits using hidden Markov models with continuous mixture

densities. *AT&T technical journal*, *64*(6), 1211-1234.

Raj, S., & Shaji, A. (2016). Design and implementation of reconfigurable digital filter bank for hearing aid. In *2016 International Conference on Emerging Technological Trends (ICETT)* (pp. 1-6). IEEE.

Ray, S., Bansal, S., Gupta, A., Gupta, D., & Shaikh, F. (2017). Understanding Support Vector Machine algorithm from examples (along with code). *Analytics Vidhya*, *13*, 19.

Saunders, J. (1996). Real-time discrimination of broadcast speech/music. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings* (2, pp. 993-996). IEEE.

Scheirer, E., & Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In *1997 IEEE international conference on acoustics, speech, and signal processing* (2, pp. 1331-1334). IEEE.

Scherer, D., Müller, A., & Behnke, S. (2010). Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks* (pp. 92-101). Springer, Berlin, Heidelberg.

Xu, J., Yu, J., Peng, Y. N., & Xia, X. G. (2011). Radon-Fourier transform for radar target detection, I: generalized Doppler filter bank. *IEEE Transactions on Aerospace and Electronic Systems*, *47*(2), 1186-1202.

Yu, Y., Abadi, M., Barham, P., Brevdo, E., Burrows, M., Davis, A.& Isard, M. (2018). Dynamic control flow in large-scale machine learning. In *Proceedings of the Thirteenth EuroSys Conference* (pp. 1-15).

Zeqiri, B., Lee, N. D., Hodnett, M., & Gélat, P. N. (2003). A novel sensor for monitoring acoustic cavitation. Part II: prototype performance evaluation. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, *50*(10), 1351-1362.

Zhang, T., & Kuo, C. C. J. (1998). Hierarchical system for content-based audio classification and retrieval. In *Multimedia Storage and Archiving Systems III* (3527, pp. 398-409). International Society for Optics and Photonics.

ÖZGEÇMİŞ

Adı Soyadı	: Mustafa Ahmed ELBERRI
Doğum Yeri ve Yılı	: Libya 1987
Medeni Hali	: Evli
Yabancı Dili	:Ingilizce, Arapça
E-posta	: albrry1@gmail.com

Eğitim Durumu

Lise	: Ibn Galboun Lisesi, 2005
Lisans	: Misurata Üniversitesi, 2011

Mesleki Deneyim

İş Yeri	: Libyan Iron And Steal Company, 2012