T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ARAÇLAR İÇİN TÜRKÇE AKILLI ASİSTAN TASARIMI

Engin ERGÜN

YÜKSEK LİSANS TEZİ

Elektronik ve Haberleşme Mühendisliği Anabilim Dalı

Elektronik Programı

Danışman

Prof. Dr. Tülay YILDIRIM

Haziran, 2022

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ARAÇLAR İÇİN TÜRKÇE AKILLI ASİSTAN TASARIMI

Engin ERGÜN tarafından hazırlanan tez çalışması 21.06.2022 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Elektronik ve Haberleşme Mühendisliği Anabilim Dalı, Elektronik Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Prof. Dr. Tülay YILDIRIM
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Prof. Dr. Tülay YILDIRIM, Danışman
Yıldız Teknik Üniversitesi _____

Doç. Dr. Umut Engin AYTEN, Üye
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Nurullah ÇALIK, Üye
İstanbul Medeniyet Üniversitesi _____

Danışmanım Prof. Dr. Tülay YILDIRIM sorumluluğunda tarafımca hazırlanan Araçlar İçin Türkçe Akıllı Asistan Tasarımı başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Engin ERGÜN

Aileme…

TEŞEKKÜR

Tez çalışmam süresince yardımlarını ve desteğini benden esirgemeyen değerli akademik danışmanım Prof. Dr. Tülay YILDIRIM'a, konuşma sentezleme modülü için veri seti oluşturulmasında bana destek olan dostum Abdulkadir AKDENİZ'e ve kendilerinden çok şey öğrendiğim kıymetli dostlarım Zeynep Gülhan USLU, Oğuz YILMAZ, Osman Nebi AKDENİZ ve Ahmet ÖNER'e sonsuz teşekkürü bir borç bilirim.

Ayrıca, hayatım boyunca beni her zaman destekleyen, haklarını asla ödeyemeyeceğim aileme teşekkürlerimi sunarım.

Engin ERGÜN

İÇİNDEKİLER

KISALTMA LİSTESİ vii
ŞEKİL LİSTESİ viii
TABLO LİSTESİ x
ÖZET xi
ABSTRACT xiii
1.1 Literatür Özeti ... 1
1.2 Tezin Amacı ... 2
1.3 Hipotez ... 3
2.1 Giriş ... 5
2.2 İnsanlardaki Konuşma ve İşitme Sistemleri ... 5
2.2.1 Konuşma Sinyalinin Üretimi ... 6
2.2.2 Sesin Algılanması ... 11
2.3 Fonetik ... 12
2.4 Konuşma Sinyallerinin Yorumlanması ... 14
3.1 Giriş ... 16
3.2 Problem Tanımı ... 17
3.3 Öznitelik Çıkarma ... 18
3.3.1 MFCC ve Dinamik Özellikler ... 18
3.3.2 Log Mel Spektrum ... 25
3.4 Konuşma Tanıma Mimarileri ... 25
3.4.1 WFST Tabanlı Konuşma Tanıma Mimarisi ... 25
3.4.2 Uçtan Uca Konuşma Tanıma Mimarisi ... 33
3.5 Konuşma Tanıma Modellerinin Performanslarının Değerlendirilmesi ... 38
4.1 Giriş ... 39
4.2 Konuşma Sentezleme Sistemlerindeki Zorluklar ... 40
4.2.1 Konuşma İşaretlerinin Doğrusal Olmaması Problemi ... 40
4.2.2 Veri Oluşturma Problemi ... 41
4.2.3 Birden-Çoğa Eşleme Problemi ... 42
4.2.4 Gerçek Zamanda Çalışma Problemi ... 42
4.3 Sentezleyici Model Mimarisi ... 43
4.3.1 Karakter Vektörü ... 44

4.3.2 Konumsal Kodlama 44
4.3.3 Kodlayıcı 45
4.3.4 Varyans Tahminleyici 45
4.3.6 Mel-Spektrogram Kod Çözücü 46
4.4 Ses Kodlayıcı Model Mimarisi 47
4.5 Konuşma Sentezleme Sistemlerinin Başarımlarının ve Performansının Değerlendirilmesi 51
4.5.1 Konuşma Doğallığının Değerlendirilmesi 52
4.5.2 Sistemin Hızının Değerlendirilmesi 52
5.1 Giriş 53
5.2 Dilin Temsil Edilmesi 54
5.2.1 Kelime Vektörleri 54
5.2.2 BERT 56
5.3 İsimlendirilmiş Varlık Tanıma 63
5.4 Niyet Sınıflandırma 64
5.4.1 DIET 65
6.1 Sistem Mimarisi 67
6.2 Kullanıcı Arayüzü 68
6.3 Konuşma Yazılandırma Servisi 69
6.4 Konuşma Sentezleme Servisi 70
6.5 Doğal Dil İşleme ve Diyalog Yönetimi Servisi 70
7.1 Otomatik Konuşma Tanıma 72
7.1.1 Yöntem 72
7.1.2 Deneyler 73
7.2 Konuşma Sentezleme 75
7.2.1 Yöntem 75
7.2.2 Deneyler 76
7.3 Doğal Dil İşleme ve Diyalog Yönetimi 77
7.3.1 Yöntem 77
KAYNAKÇA **83**
TEZDEN ÜRETİLMİŞ YAYINLAR **89**

KISALTMA LİSTESİ

AFD	Ayrık Fourier Dönüşümü
BERT	Bidirectional Encoder Representation from Transformers
CTC	Connectionast Temporal Classification
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
EM	Expectation Maximization
GAN	Generative Adversarial Network
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
KHO	Kelime Hata Oranı
LSTM	Long Short Term Memory
MAE	Mean Absolute Error
MFCC	Mel Frequency Cepstral Coefficient
MSE	Mean Squred Error
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
STFT	Short Time Fourier Transform
TDNN	Time Delayed Neural Network

ŞEKİL LİSTESİ

Şekil 2.1 Ses üretiminden sorumlu olan organların yanal kesit görüntüsü [12]...6
Şekil 2.2 Ötümlü seslerde ses tellerinin açılıp kapanma döngüsü [13]. 7
Şekil 2.3 Sesli harfler (a), patlamalı (b) ve sürtünmeli (c) sessizler için ses yolunun durumu [12]. 8
Şekil 2.4 Ses sinyalinin oluşturulması adımları [15]. 10
Şekil 2.5 İnsanlarda işitme sisteminin yapısı [16] 11
Şekil 2.6 Fonetik olarak hizalanmış bir ses kaydı. 13
Şekil 2.7 Konuşma metni ile hizalanmış bir ses kaydının dalga formu ve spektrogram grafikleri gösterimi. 14
Şekil 3.1 MFCC özellikleri çıkarma adımları [22]. 19
Şekil 3.2 Analog dijital dönüşüm. 19
Şekil 3.3 Ön vurgulama işlemi: (a) ön vurgulamadan önce, (b) ön vurgulamadan sonra [23] 20
Şekil 3.4 Pencereleme işlemi [24]. 21
Şekil 3.5 Sinyallerin kesiminde farklı pencereleme fonksiyonlarının kullanımı [24]. 21
Şekil 3.6 *Hamming* pencere fonksiyonu ile kesilmiş /i/ fonuna ait (a) zaman ve (b) spektrum grafikleri [24]. 22
Şekil 3.7 Hertz ölçeğinin mel ölçeği üzerindeki karşılığı. 23
Şekil 3.8 Mel ölçekli güç spektrumunun elde edilmesi [22]. 24
Şekil 3.9 WFST tabanlı konuşma tanıma mimarisi [26]. 26
Şekil 3.10 HCLG giriş ve çıkışları [28]. 26
Şekil 3.11 H, L ve G FST [29] 27
Şekil 3.12 Zaman gecikmeli sinir ağı mimarisi [35]. 29
Şekil 3.13 Örneklemeli zaman gecikmeli sinir ağı mimarisi [32]. 30
Şekil 3.14 Konuşma tanıma sistemleri için örnek bir kafes yapısı [36].....33
Şekil 3.15 Uçtan uca konuşma tanıma adımları 34
Şekil 3.16 CTC yöntemi ile hizalama işlemi. 35
Şekil 3.17 QuartzNet mimarisi [40].....36
Şekil 3.18 Derinlikli evrişim katmanı örneği [41]. 37
Şekil 3.19 Noktasal evrişim katmanı örneği [41]. 37
Şekil 4.1 Konuşma sentezleme sistemi. 40
Şekil 4.2 Fastspeech2 mimarisi [48], [49]. 44

Şekil 4.3 Multi-band Melgan mimarisi [59]. 49
Şekil 5.1 CBOW ve Skip-gram modeller için ağ mimarisi [62]. 55
Şekil 5.2 Kral, kraliçe, kadın, erkek kelime vektörleri üzerinde yapılan vektörel işlemler [63]. 55
Şekil 5.3 Kelime parçacığı yöntemi kullanılarak elde edilen kelime altı parçacıklar.....57
Şekil 5.4 BERT modeli giriş işlemleri [61]. 59
Şekil 5.5 BERT maskelenmiş dil modeli bileşeni. 60
Şekil 5.6 BERT ön eğitim işlemi [61] (düzenlendi).....61
Şekil 5.7 BERT sıradaki cümle tahmini bileşeni. 62
Şekil 5.8 BERT ince-ayar konfigürasyonları [61]. 62
Şekil 5.9 BERT isimlendirilmiş varlık tanıma konfigürasyonu. 64
Şekil 5.10 DIET mimarisi [68]. 65
Şekil 6.1 Tasarlanan sanal asistanın mimarisi. 67
Şekil 6.2 Uygulama arayüzü. 68
Şekil 6.3 Doğal dil işleme ve diyalog yönetimi servisleri için örnek girdi ve çıktılar. 71
Şekil 7.1 QuartzNet modeli canlı tanıma örneği. 75

TABLO LİSTESİ

Tablo 2.1 Ünlülerin temel ve formant frekansları ortalaması [14]. 9
Tablo 2.2 Seslilerin fonetik özellikleri.....12
Tablo 2.3 Sessizlerin fonetik özellikleri.....12
Tablo 3.1 Okunuş sözlüğü örneği. 32
Tablo 4.1 Konuşma sentezleme modellerinin karşılaştırılması[51]. 43
Tablo 4.2 MelGAN ve Multi-band MelGAN performans karşılaştırması [59]. 48
Tablo 7.1 Konuşma tanıma modellerinin karşılaştırılması.....74
Tablo 7.2 Konuşma sentezleme sistemi için gerçek zamanda çalışma testi.....76
Tablo 7.3 Niyet sınıflandırmasında kullanılabilecek örnekler.....77
Tablo 7.4 Niyet sınıflandırması için oluşturulan veri seti ve örnekleri.....79

ÖZET

Araçlar için Türkçe Akıllı Asistan Tasarımı

Engin ERGÜN

Elektronik ve Haberleşme Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Prof. Dr. Tülay YILDIRIM

Teknolojideki hızlı ilerleme ile beraber makinelerin kontrol yöntemlerinde de birtakım gelişmeler meydana gelmektedir. Günümüzde sıklıkla tercih ettiğimiz grafiksel arayüzlere alternatif olarak geliştirilen sesli kontrol sistemleri, özellikle mobil cihazlar ve giyilebilir teknolojilerde kullanım kolaylığı sağlamaktadır. Genellikle sanal veya akıllı asistan olarak bilinen bu teknolojiler, mobil cihazlar dışında akıllı ev ve araç kontrol sistemlerinde de kullanılabilmektedir. Akıllı ev sistemlerinde özellikle nesnelerin interneti teknolojileri ile ev içerisindeki konforu, araç sistemlerinde ise eller serbest şekilde rota ayarlama, araç içi sıcaklığını değiştirme, multimedya kontrolü gibi seçeneklerle sürüş rahatlık ve güvenliğini arttırmaktadır.

İnsan-makine etkileşiminin en önemli örneklerinden olan sanal asistanlar; konuşma tanıma, doğal dil işleme ve konuşma sentezleme bileşenlerinden oluşmaktadır. Bu bileşenleri kullanan sanal asistanlar; konuşmayı yazılandırmakta, değerlendirmekte ve sesli geri dönüş yapabilmektedirler. Hâliyle; dinleyen, anlayan ve konuşan bir insanı andırdıkları için sanal asistan olarak anılmaktadırlar.

Bu tez çalışması kapsamında, sanal asistan uygulamalarında kullanılabilecek otomatik konuşma tanıma, konuşma sentezleme, niyet sınıflandırma ve isimlendirilmiş varlık tanıma problemleri incelenmiş, araç ortamı içerisinde kullanılabilecek bir sanal asistan uygulaması için literatürdeki uygun yaklaşımlar belirlenmiştir. Ardından belirlenen bu yöntemler ile Türkçe nöral modeller oluşturulmuştur. Modeller, mikro servis mimarisi prensipleri göz önünde bulundurularak birbirinden bağımsız şekilde çalışabilecek servislerin oluşturulmasında kullanılmıştır.

Son olarak bütün servislerin birbiri ile uyumlu bir şekilde çalışması sağlanarak; sürücülerin yol hakimiyetini kaybetmeden hava durumunu ve güncel haberleri sorgulayabileceği, rota oluşturma işlemlerini yapabileceği bir sanal asistan uygulaması hayata geçirilmiştir.

Anahtar Kelimeler: Sanal asistan, konuşma tanıma, doğal dil işleme, konuşma sentezleme

ABSTRACT

Turkish Smart Assistant Design for Vehicles

Engin ERGÜN

Department of Electronic and Communication Engineering

Master of Science Thesis

Supervisor: Prof. Dr. Tülay YILDIRIM

Along with the rapid progress in technology, there are also some developments in the control methods of the machines. Voice control systems, which are developed as an alternative to the graphical interfaces that we often prefer today, provide ease of use, especially in mobile devices and wearable technologies. These technologies, generally known as virtual or smart assistants, are also used in smart home and vehicle control systems apart from mobile devices. In smart home systems, it increases the comfort in the home especially with internet of things applications. In vehicle systems, it increases driving comfort and safety with options such as hands-free route adjustment, changing the interior temperature and multimedia control.

Virtual assistants, one of the most important examples of human-machine interaction, consist of speech recognition, natural language processing and speech synthesis components. Virtual assistants that uses these components can convert speech into text, understand the context of the conversation and make voice feedback. In this case, they are known as a virtual assistant. Because, they resemble a person who can listen, understand and speak.

Within the scope of this thesis study, the problems of automatic speech recognition, speech synthesis, intent classification and named entity recognition problems that can be used in virtual assistant applications have been examined and appropriate approaches in the literature have been determined for a virtual assistant application that can be used in the vehicle environment. After that, Turkish neural models trained with using determined methods. The models have been used to create services that can work independently by taking into consideration the principles of microservice architecture.

Finally, by ensuring that all services work in harmony with each other; a virtual assistant application has been implemented where drivers can inquire about weather conditions and recent news, and create routes without losing their command of the road.

Keywords: Virtual assistant, speech-to-text, natural language processing, text-to-speech

1
GİRİŞ

1.1 Literatür Özeti

Alan Turing, 1950 yılında makinelerin düşünüp düşünemeyeceğini, eğer düşünebilen bir makine olsaydı, bu makineye sorulan soruların karşısında alınacak cevapların gerçek bir insan cevabı kadar gerçekçi olup olmayacağını sorgulayan bir makale yayınlamıştır [1]. Bu makale, insanlar gibi doğal dili anlayabilen ve konuşulan bağlama uygun cevaplar üretebilen makinelerin üretilmesi düşüncesinin önünü açmıştır. Günümüzde karşılaştığımız sanal asistanlar, sohbet botları ve diğer yapay konuşma ajanları bu düşüncenin eserleri olarak karşımıza çıkmaktadır.

Bu makaleden sonra, insanlarla yazılı sohbet kanalı vasıtasıyla etkileşime girebilen ELIZA [2] isimli bilgisayar programı, öncü sohbet botu çalışmalarından olmuştur. Kural tabanlı olarak çalışan ELIZA, bir psikoterapisti simüle etmektedir. Kullanıcının girdiği metni sorgulayıcı bir tavırla tekrar kullanıcıya yönlendirmektedir.

Sohbet botu kategorisinde çıkan bir diğer program ise, psikiyatrist Kenneth Colby tarafından geliştirilen PARRY'dir. PARRY paranoyak şizofreni hastalarını simüle etmektedir. 25 psikiyatrist ile mülakata giren PARRY'ye, 23 psikiyatrist tarafından paranoyak, 2 psikiyatrist tarafından ise konuşma bozukluklarından dolayı beyin hasarı tanısı konulmuştur [3].

1988 yılında geliştirilen Jabberwacky isimli program, daha önceki çalışmalardan farklı olarak, bu alanda yapay zekâ tabanlı yaklaşımların kullanıldığı ilk çalışma olmuştur. Bu program, doğal insan sohbetini; ilginç, eğlenceli ve esprili bir şekilde simüle etmeyi amaçlamıştır [4].

90'lı yılların başlarında internetin yaygınlaşması ile birlikte sohbet botu teknolojileri de artık internetin birer parçası olmaya başlamıştır. Bu bağlamda yapılan ilk çalışma ALICE'dir. ELIZA programından esinlenerek üretilen ALICE,

40000'e yakın kategori hakkında bilgi sahibiyken bu sayı ELIZA için yaklaşık olarak 200'dür. Gözetimli öğrenme ve desen eşleştirme yöntemlerini kullanan ALICE, türünün en güçlü programlarından birisi sayılmaktadır. Turing Testini geçememiş olsa bile, konuşan robotlara verilen Loebner Ödülü'nü 2000, 2001 ve 2004 yıllarında toplamda üç kez kazanmıştır [5].

2000'li yılların başlarında ilk defa, doğal dil yoluyla günlük veri tabanı sorgularının yapılabilmesine olanak sağlayan SmarterChild isimli proje duyurulmuştur. Bu projeyle; film saatleri, maç sonuçları, hisse senedi fiyatları, haberler ve hava durumu hakkında veri tabanlarından bilgi alınabilmiştir [6]. Bu özelliği sayesinde, günümüzde sıklıkla kullanılan sanal asistan kavramının öncü çalışmalarından olmuştur.

2010'lu yıllarda akıllı telefonların yaygınlaşması ile beraber Apple, bir atılım yaparak sesli kişisel asistan uygulaması olan Siri'yi [7] tanıtmıştır. Siri diğer sohbet botu uygulamalarından farklı olarak, metin yerine konuşma sinyallerini giriş olarak kabul etmekte ve yine sesli bir şekilde çıkış verebilmektedir. Telefon üzerinden pek çok işlemin hızlı bir şekilde yapılmasını sağlayan Siri, aynı zamanda Apple Watch gibi giyilebilir teknolojilerin de kullanımını oldukça kolaylaştırmıştır.

2011 yılında Amerika Birleşik Devletleri'nin ulusal bir kanalında yayınlanan Jeopardy isimli bir televizyon yarışmasında, IBM'in geliştirmiş olduğu Watson isimli yapay zekâ tabanlı açık alan soru cevaplama sistemi, tüm zamanların en iyi iki yarışmacısını mağlup ederek birinci olmuştur [8].

Bu çalışmaların ardından; Google Assistant [9], Microsoft Cortana [10], Amazon Alexa [11] gibi oldukça gelişmiş sanal asistan teknolojileri gün geçtikçe gelişerek karşımıza çıkmaktadır.

1.2 Tezin Amacı

İnsan-makine etkileşiminin en önemli örneklerinden olan sanal asistanlar; günlük hayatta birtakım işlemlerin hızlı, kolay ve eller serbest biçimde yapılmasına imkân sağlayan uygulamalardır. Temel bir sanal asistan; konuşma tanıma, doğal dil işleme ve konuşma sentezleme bileşenlerinden oluşmaktadır. Bu bileşenleri kullanan sanal asistanlar; konuşmayı yazılandırmakta, değerlendirmekte ve sesli

geri dönüş yapabilmektedirler. Hâliyle; dinleyen, anlayan ve konuşan bir insanı andırdıkları için sanal asistan olarak anılmaktadırlar.

Bu tez kapsamında, bir sanal asistan uygulaması hayata geçirebilmek için gerekli olan bileşenler analiz edilmiş, güncel literatür taraması yapılarak tez içeriğinde aktarılmıştır. Elde edilen bilgiler kullanılarak, Türkçe dili için uygun modeller oluşturulmuş, araç ortamı için özelleştirilmiş bir sanal asistan uygulaması hayata geçirilmiştir. Bu uygulama ile birlikte sürücünün yol ile göz teması kesilmeden birtakım sorgulamaların yapılması sağlanarak araç sürüş güvenliği ve konforunun arttırılması hedeflenmiştir.

1.3 Hipotez

Bu tez çalışması kapsamında; konuşma tanıma, konuşma sentezleme, niyet sınıflandırma ve isimlendirilmiş varlık tanıma problemleri incelenmiş, literatürdeki güncel yöntemler taranmıştır.

Konuşma tanıma sistemlerinin oluşturulmasında sıklıkla kullanılan ağırlıklı sonlu durum dönüştürücü tabanlı mimari ile uçtan uca konuşma tanıma mimarilerinden olan QuartzNet için Türkçe modeller oluşturularak karşılaştırılmıştır.

İncelenen literatür çalışmalarında, konuşma sentezleme modelleri eğitebilmek için yüksek miktarda ve yüksek kalitede veri ihtiyacının olduğu gözlemlenmiştir. Türkçe konuşma sentezleme modeli eğitebilmek için açık kaynak olarak paylaşılan bir veri kümesi bulunmamaktadır. Buna çözüm olarak, Türkçe konuşma sentezleme modeli eğitebilmek için İngilizce veri kümesinde bulunan yazılar Türkçe okunuşları ile değiştirilmiştir. Elde edilen veri kümesi ile bir konuşma sentezleme modeli eğitilmiştir. Model ile elde edilen çıktıların anlaşılabilirliğinin düşük olması nedeniyle küçük bir konuşma sentezleme verisi oluşturulmuş, önerilen yöntem ile elde edilen model yeni veri kümesi ile eğitime devam ettirilerek nihai sentezleyici model oluşturulmuştur.

Niyet sınıflandırması yapabilmek için araç ortamına özel küçük bir veri kümesi oluşturularak Türkçe model eğitilmiştir.

Yapılan çalışmalar sonucunda üretilen bütün modellerin uyumlu bir şekilde çalışması sağlanarak, araç ortamı için özelleştirilmiş bir akıllı asistan uygulaması hayata geçirilmiştir.

2
KONUŞMA DİLİNİN YAPISI VE ANALİZİ

2.1 Giriş

Bilişim sistemleri dünyasında "dil" kavramı iki farklı anlamda kullanılabilmektedir. Aslında akla gelen iki anlamda da dil, iletişimde kullanılan bir araç olarak karşımıza çıkmaktadır. Bu kavram, insanların kendi aralarındaki iletişimde kullanılıyorsa doğal dil veya konuşma dili, insanların bilgisayar veya makine sistemleri ile iletişim kurmasını sağlıyorsa makine dili olarak adlandırılmaktadır. Doğal dillere Türkçe, İngilizce, Almanca, Fransızca gibi diller, makine dillerine ise Assembly, C, Java, Python vb. örnekler verilebilmektedir.

Makineler ile doğal dillerin anlamlandırılabilmesi için yapılan çalışmalara doğal dil işleme denilmektedir. Doğal dil işleme alanında hem metin hem de konuşma sinyalleri üzerine çalışmalar yapılmaktadır. Sanal asistan sistemleri doğal dil işleme yöntemlerinin en çok kullanıldığı alanlardan bir tanesidir. Bu sistemler ile metin ve ses sinyallerinin işlenmesi sonucunda kullanıcıların istekleri yerine getirilebilmektedir. Bunun için öncelikle konuşma sinyalinin metne dönüştürülmesi, metindeki niyetin saptanması, varsa çeşitli nesnelerin belirlenmesi gerekmektedir. Ardından saptanan niyete göre bir aksiyon alınmalı ve kullanıcıya istenilen işlemin sonucu metin olarak hazırlanmalıdır. Hazırlanan metin seslendirilerek kullanıcıya geri dönüş yapılmalıdır. Tüm bu işlemler aslında insanlardaki birtakım fonksiyonlara karşılık gelmektedir. Özellikle konuşma sinyallerinin yazılandırılması ve metinlerin seslendirilmesi doğrudan insanlardaki işitme ve konuşma sistemleri ile ilişkilidir ve bu sistemlerin tasarlanabilmesi için insanlarda bu olayların nasıl gerçekleştiğini gözlemlemek faydalı olmaktadır.

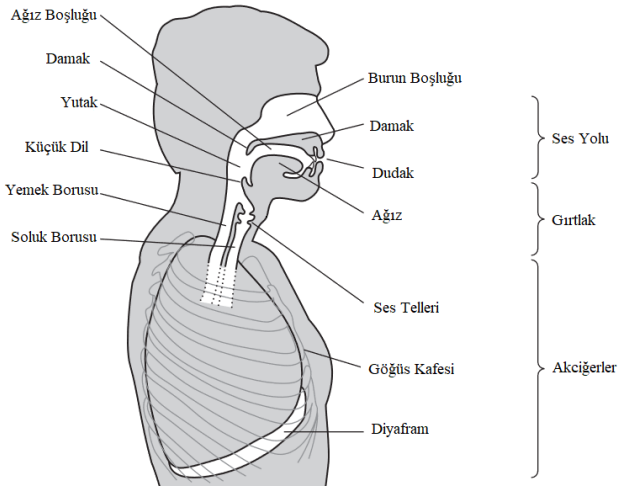
2.2 İnsanlardaki Konuşma ve İşitme Sistemleri

Pek çok algoritma, doğadaki olayların gözlemlenmesi ve sonrasında makinelerin de anlayabileceği şekilde matematiksel olarak ifade edilmesi ile oluşturulmaktadır. İnsanların konuşma ve işitme sistemlerinde yer alan bazı

anatomik unsurlar, makinelere konuşma yazılandırma ve konuşma sentezleme yeteneklerinin kazandırılması için oldukça önemli bilgiler sunmaktadır. Bu bölümde, insan bedeninde sesin üretilmesinden ve algılanmasından sorumlu olan sistemlerin yapısı incelenerek bilgisayar sistemleri için nasıl modeller oluşturulabileceği verilecektir.

2.2.1 Konuşma Sinyalinin Üretimi

Konuşma sinyalleri, bir konuşmacının ses yolunda şekillendirilen hava basıncı dalgaları ile üretilmektedir. Sesin üretilmesinde rol oynayan organlar, üç grup içerisinde yer almaktadır. Bunlar; konuşmak için ihtiyaç duyulan havanın kaynağını oluşturan akciğerler, hava akışını modüle eden gırtlak, ağız-burun organ sistemlerini de içerisinde bulunduran ve sesin çeşitli filtrelerden geçirilerek değiştirilmesini sağlayan ses yoludur. Ses üretiminden sorumlu olan bu organların yanal kesit görüntüsü Şekil 2.1'de gösterilmektedir.



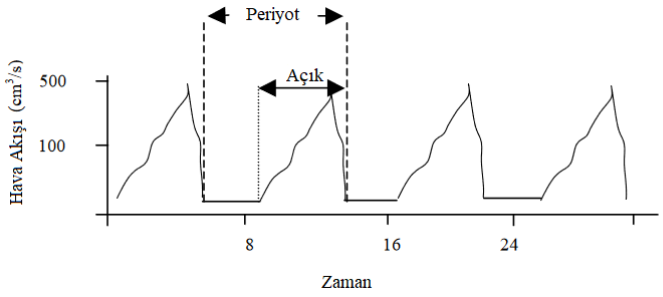
Şekil 2.1 Ses üretiminden sorumlu olan organların yanal kesit görüntüsü [12].

Akciğerlerin bir görevi solunumu sağlamak iken diğer bir görevi ise konuşmak için ihtiyaç duyulan havanın kaynağını oluşturmaktır. Konuşma esnasında, göğüs kafesinin etrafındaki kasların kontrolü ile belirli miktarlarda havanın serbest bırakılması sağlanmakta ve elde edilen hava basıncı filtrelenerek sesler oluşturulmaktadır.

Gırtlak bölümü, özellikle kıkırdak, kas ve bağ dokulardan oluşan bir bölümdür. Konuşma üretimi açısından incelendiği zaman, ses tellerinin kontrolünden sorumlu olan organ sistemlerini içermektedir. Ses telleri, üzeri mukoza ile kaplı,

karşılıklı şekilde konumlanmış iki et parçacığından oluşmaktadır. Bunların uzunlukları erkeklerde yaklaşık olarak 15 mm iken kadınlarda 13 mm'dir. Ses tellerinin durumu yapılan aktiviteye göre üç farklı şekilde sınıflandırılmaktadır. Bunlar; nefes alma, ötümlü ve ötümsüz seslerin telaffuz anları şeklindedir [12].

Nefes alma durumunda ses telleri açık pozisyondadır. Akciğerlerden çıkan hava, rahat bir şekilde boşaltılabilmektedir. Ötümlü seslerin telaffuzu esnasında ise akciğerden çıkan hava sütunu, öncelikle ses tellerinin alt kısmının açılmasını sağlamakta, ardından ses tellerinin üst kısmına doğru ilerlemekte ve tellerin üst kısmı açmaktadır. Hızlı bir şekilde hareket eden hava sütununun arkasında oluşan düşük basınç; ses tellerinin önce tabanının, ardından üst kısmının kapanmasına neden olan bir "Bernoulli Etkisi" oluşturmaktadır. Ses tellerini ardında bırakan hava sütunu, ses yolu rezonatörleri tarafından güçlendirilip değiştirilmekte ve son olarak ses üretilmektedir. Ses sütunları geldikçe döngü devam etmektedir. Ötümlü seslerdeki, ses tellerinin açılıp kapanma döngüsü temel frekans olarak adlandırılmaktadır. Şekil 2.2'de akciğerlerden ses tellerine gelen hava sütunlarının grafiği gösterilmektedir. Genliğin yüksek olduğu bölümlerde ses telleri açık konuma geçmektedir.



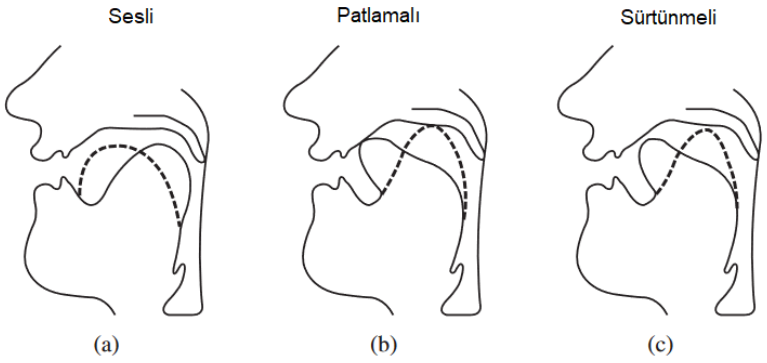
Şekil 2.2 Ötümlü seslerde ses tellerinin açılıp kapanma döngüsü [13].

Ses tellerinin son durumu ise ötümsüz seslerin telaffuzu esnasındaki halidir. Bu durumda ses telleri, nefes alma haline benzemektedir. Fakat, nefes alma durumuna göre birbirlerine daha yakın ve daha gergin durumdadır. Bu esnada da ses telleri üzerindeki hava akışı ses tellerini titreştirmemektedir.

Ses yüksekliği ve sesin perdesi, sesin algılanmasını değiştiren iki önemli özelliktir. Ses yüksekliğinin artışı, ses telleri arasından geçen havanın artışı ile mümkün olabilmektedir. Bu esnada ses telleri daha geniş bir şekilde açılmakta ve daha uzun

bir sürede titreşim döngüsünü tamamlamaktadır. Böylelikle ses tellerinden geçen havanın basıncı artmakta ve üretilen sesin genliği yükselmektedir. Ses tellerindeki titreşimin artması ise perde frekansını yükseltmekte ve daha tiz bir ses oluşturulmasını sağlamaktadır. Sesin perdesi veya temel frekans olarak adlandırılan bu özellik, erkekler için ortalama 125 Hz iken kadınlar için 210 Hz'dir. Bu durumdan dolayı kadınların sesleri erkeklere göre daha tiz şekilde tonlanmaktadır.

Ses yolu; gırtlaktan başlayıp, ağız ve burun boşluklarını da içerisinde bulunduran, dudak ve burun ucunda sonlanan bölümdür. Ses yolu, yetişkin insanlarda yaklaşık olarak 17.5 cm uzunluğundadır. Fakat ağız, dil, diş ve çene hareketleri ile pek çok uzunluk ve değişik kesitlerin oluşturulmasına imkân sağlayan bir mekanizmaya sahiptir. Böylelikle ses tellerinin çıkışında üretilen vızıltı benzeri sesler, burada şekillendirilerek konuşmada kullanılan ses birimleri oluşturmaktadır. Ses yolunun rezonans frekansları, ses bilimi bağlamında, formant frekansları veya basitçe formantlar olarak adlandırılmaktadır ve formantlar farklı ses yolu konfigürasyonları ile değişmektedir [12]. Şekil 2.3'te dilin her bir konumu farklı rezonans boşluklarına karşılık gelmektedir ve farklı seslerin üretilmesine neden olmaktadır.



Şekil 2.3 Sesli harfler (a), patlamalı (b) ve sürtünmeli (c) sessizler için ses yolunun durumu [12].

Konuşma tanıma sistemlerinde, ötümlü seslerin belirlenmesinde kullanılan akustik ölçütlerin başında formant frekansları gelmektedir. Bu bağlamda, Tablo 2.1'de uluslararası fonetik alfabede bulunan sesli fonlar için belirli formant frekansları gösterilmektedir. Tablo; erkek, kadın ve çocuk olmak üzere 76

katılımcıdan alınan ses örnekleri ile oluşturulmuştur. Belirli fonlar için ortalama formant frekansları ölçülmüş ve raporlandırılmıştır.

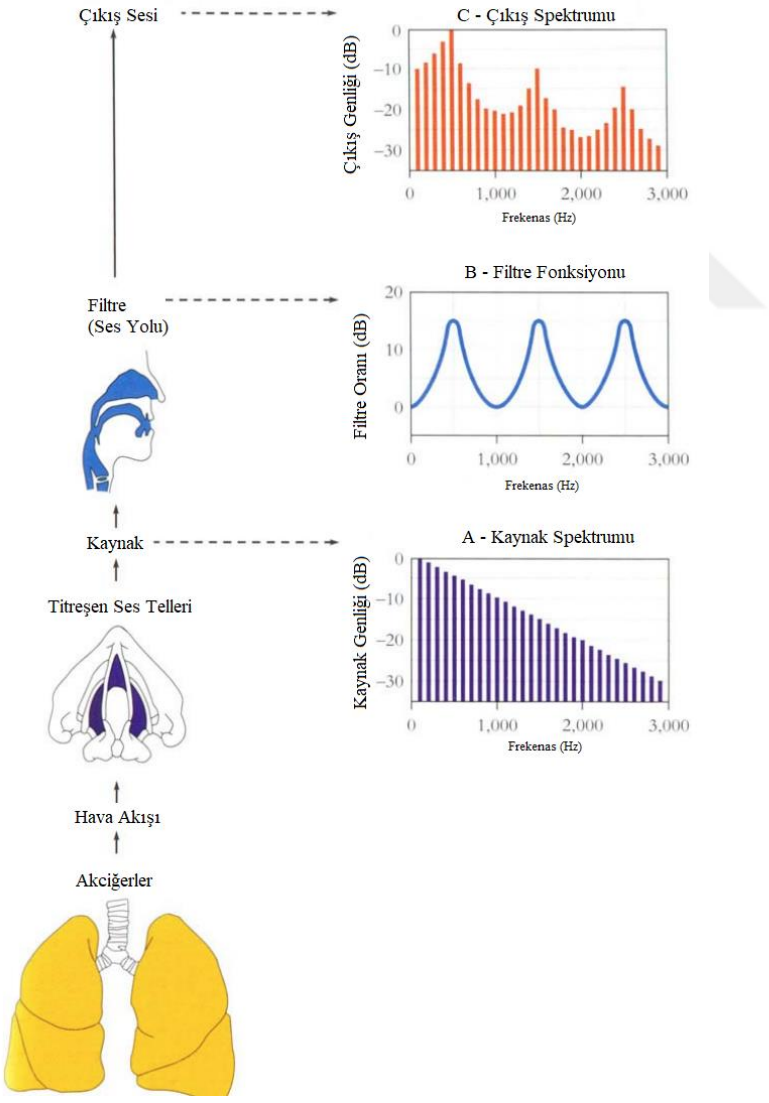
Tablo 2.1 Ünlülerin temel ve formant frekansları ortalaması [14].

		i	ɪ	ɛ	æ	a	ɔ	ʊ	u	ʌ	ɝ
Temel Frekans	E	136	135	130	127	124	129	137	141	130	133
	K	235	232	223	210	212	216	232	231	221	218
	Ç	272	269	260	251	256	263	276	274	261	261
Formant Frekansları		i	ɪ	ɛ	æ	a	ɔ	ʊ	u	ʌ	ɝ
F_1	E	270	390	530	660	730	570	440	300	640	490
	K	310	430	610	860	850	590	470	370	760	500
	Ç	370	530	690	1010	1030	680	560	430	850	560
F_2	E	2290	1990	1840	1720	1090	840	1020	870	1190	1350
	K	2790	2480	2330	2050	1220	920	1160	950	1400	1640
	Ç	3200	2730	2610	2320	1370	1060	1410	1170	1590	1820
F_3	E	3010	2550	2480	2410	2410	2410	2240	2240	2390	1690
	K	3310	3070	2990	2850	2810	2710	2680	2670	2780	1960
	Ç	3730	3600	3570	3320	3170	3180	3310	3260	3360	2160

Tablodaki E, K, Ç hücreleri sırayla erkek, kadın ve çocukları temsil etmektedir. Ötümlü seslerin temel frekansları arasında oldukça küçük farklar olduğu gözlemlenebilmektedir. Bu değerler tek başına seslileri birbirinden ayırabilmek için yeterli görünmemektedir. Fakat diğer formant frekansları da incelendiği zaman, ötümlü seslerin ayırt edilmesini sağlayacak farklar ortaya çıkmaktadır. Bu

yüzden ötümlü seslerin ayırt edilmesi işleminde bir, iki ve üçüncü formant frekanslarına da başvurulmaktadır.

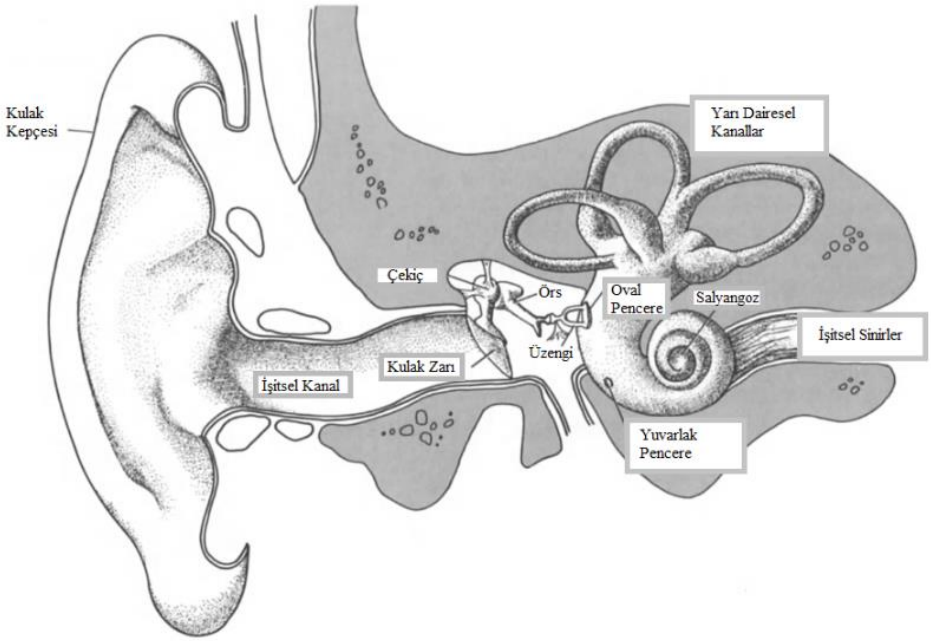
Son olarak Şekil 2.4 ses sinyalinin üretilmesi esnasındaki tüm adımları özetlenmektedir. Kısaca; akciğerlerden çıkan hava, ötümlü sesler üretilirken ses tellerini titreştirmekte ve kaynak spektrumunu oluşturmaktadır. Kaynak spektrumunun genliği; ses tellerinin çıkışında en yüksek, dudaklarda ise en düşük olmaktadır. Ses yolu, kontrollü olarak değiştirilebilen kesit alanları ile kaynak spektrumunu şekillendirmekte ve çıkış spektrumunu oluşturmaktadır. Böylelikle, fonların üretilmesini sağlamaktadır. Çıkış spektrumundaki tepe noktaları formant frekanslarına karşılık gelmektedir. Bu sayede, sinyal analizi ile fonlar birbirinden ayrılabilmektedir.



Şekil 2.4 Ses sinyalinin oluşturulması adımları [15].

2.2.2 Sesin Algılanması

Konuşma işleme sistemlerinin çalışma prensiplerinin anlaşılabilmesi için sesin üretilmesi gibi nasıl algılandığının anlaşılması da oldukça büyük öneme sahiptir. Bu durumda işitme sisteminin fizyolojik olarak incelenmesi faydalı olacaktır.



Şekil 2.5 İnsanlarda işitme sisteminin yapısı [16]

Şekil 2.5'te insan kulak yapısına ait model gösterilmektedir. Kulak kepçesi, yayılan ses dalgalarını toplayarak işitsel kanal vasıtası ile kulak zarına iletmektedir. Kulak zarına ulaşan dalgalar burada titreşime yol açmaktadır. Çekiç, örs ve üzengi kemikleri de bu salınım ile titreşmektedir. Gelen ses dalgası ile hareketlenen üzengi kemiği, kulak salyangozuna küçük vuruşlar yaparak içi sıvı ile dolu olan bu ortamda dalgalanmaya neden olmaktadır. Salyangoz, içerisinde uçları duyma sinirlerine bağlı tüycük demetlerine sahiptir. Bu tüycükler, dalgalanma ile beraber hareket etmektedir ve bu esnada elektriksel sinyaller oluşturularak sinir sistemi ile duyma işlevinin yerine getirilmesi sağlanmaktadır. Dış kulağa yakın olan bölümdeki tüycükler sert yapılıyken, salyangoz yapının sonlarına doğru olan bölümdeki tüycükler daha yumuşak yapılıdır. Bu durumda salyangozun dış tarafı yüksek frekanslı seslere duyarlı iken iç tarafı daha düşük frekanslı seslere duyarlıdır. Kulak içerisindeki bu yapı, arka arakaya dizilmiş ve sadece belirli frekanslara duyarlı bir dizi bant geçiren filtreyi oluşturmaktadır.

2.3 Fonetik

İnsanlar tarafından anlamlı şekilde oluşturulan ses dalgaları, insanların birbirleri ile sözel iletişim kurabilmesini sağlamaktadır. Bahsi geçen ses dalgalarının en küçük formuna fon denilmektedir. Fonların belirli kurallara göre dizilmesi ile kelimeler oluşturulmaktadır. Fon sayısı dillerin yapılarına göre farklılık gösterebilmektedir. Öyle ki, İngilizcede yaklaşık 44 fon bulunurken Türkçe fonetik bir dil olduğu için alfabedeki karakter sayısı aynı zamanda fon sayısıdır.

Fonlar pek çok dilde sesliler (ünlüler) ve sessizler (ünsüzler) olmak üzere ikiye ayrılmaktadır. Sesliler, ses yolunda herhangi bir engel olmaksızın telaffuz edilebilen ses birimleridir. Seslilerin tanımlanmasında çenenin açısı, dudakların biçimi ve dilin devinimi olmak üzere üç farklı özellik kullanılmaktadır bu özellikler Tablo 2.2'de gösterilmektedir.

Tablo 2.2 Seslilerin fonetik özellikleri

Çene Açısının Durumu		Dudakların Biçimi		Dilin Devinimi	
Geniş	Dar	Düz	Yuvarlak	Arka dil	Ön dil
/a/, /e/, /o/, /ö/	/ı/, /i/, /u/, /ü/	/ı/, /i/, /a/, /e/	/o/, /ö/ /u/, /ü/	/ı/, /a/, /o/, /u/	/i/, /e/, /ö/, /ü/

Sessizler ise ses yolunun bir kısmının veya tamamının kapanması ile şekillendirilen seslerdir. Bu sesler üretildikleri yerlere, çıkış biçimine, ses tellerinin titreşim durumuna göre sınıflandırılabilmektedir, ilgili sınıflandırma Tablo 2.3'te gösterilmektedir.

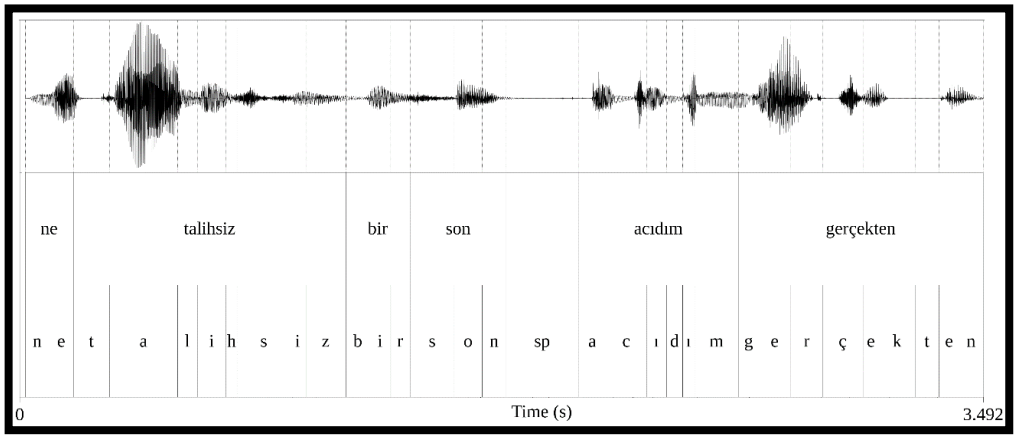
Tablo 2.3 Sessizlerin fonetik özellikleri

Çıkış Biçimleri				
Patlamalı	Geniz	Çarpmalı	Yan Daralma	Sürtünmeli
/b/, /d/, /g/, /p/, /t/, /k/	/m/, /n/	/r/	/l/	/c/, /ç/, /f/, /h/, /s/, /ş/, /v/, /y/, /z/

Tablo 2.3 Sessizlerin fonetik özellikleri (devamı)

Çıkış Yerleri							
Çift – dudak	Dudak – diş	Dilucu – dişardı	Dilucu – dişeti	Dil – öndamak	Dilucu – öndamak	Dil – artdamak	Gırtlak
/b/, /p/, /m/	/f/, /v/	/d/, /t/	/n/, /r/, /s/, /z/	/c/, /ç/, /j/, /ş/, /y/	/l/	/k/, /g/	/h/
Ses Tellerinin Titreşimi							
Ötümlü				Ötümsüz			
/b/, /c/, /d/, /g/, /j/, /l/, /m/, /n/, /r/, /v/, /y/, /z/				/ç/, /f/, /h/, /k/, /p/, /s/, /ş/, /t/			

Ötümlü seslerin (sesli harfler de dahil) telaffuzu esnasında ses telleri titreşirken, ötümsüz seslerde durum tam tersidir. Bu durumda ötümlü ve ötümsüz seslere ait sinyaller incelendiği zaman birbirinden kolaylıkla ayırt edilebilmektedir. Şekil 2.6'da bir konuşma kaydı gösterilmektedir. Bu görselde, ses sinyali ve fonemler birbirleri ile hizalanmıştır. Ötümlü seslere denk gelen konuşma sinyalinin enerjisinin, ötümsüz seslerden daha fazla olduğu görülmektedir.



Şekil 2.6 Fonetik olarak hizalanmış bir ses kaydı.

Akciğerlerden çıkan havanın ses yolu boyunca filtrelenmesi ile farklı fonların üretilmesi sağlanmakta ve bunların anlamlı şekilde birleşimi sayesinde iletişim

kurulabilmektedir. Makineler ile konuşma yazılandırma işleminde de yukarıda bahsi geçen ses özellikleri, fonların birbirinden ayrıştırılmasını sağlamaktadır.

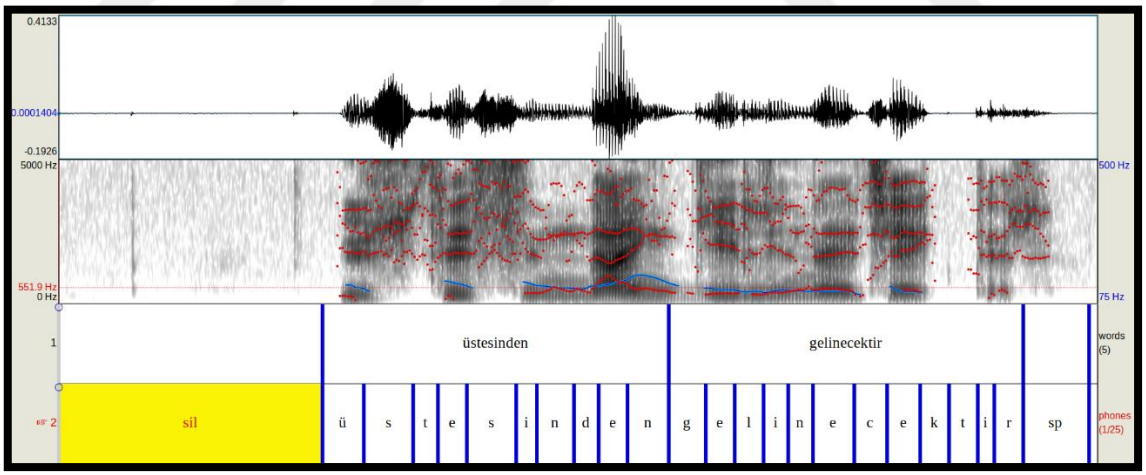
2.4 Konuşma Sinyallerinin Yorumlanması

Konuşma sinyalleri, zaman ve frekans alanlarındaki özelliklerden yararlanarak yorumlanabilmektedir.

Doğrudan zaman alanındaki bir konuşma sinyalinden aşağıdaki bilgiler edinilebilmektedir.

- Sesin enerjisi
- Sesin temel frekansı
- Konuşma ve sessizlik anları
- Sürtünmeli sessizler

Frekans alanından ise özellikle spektrogram grafikleri kullanılarak formant bilgileri gözlemlenebilmektedir. Spektrogram grafiklerinde x-ekseni zaman, y-ekseni frekans, renk koyuluğu ise genlik bilgisini gösteren diğer bir boyuttur. Şekil 2.7'de Praat [17] programı kullanılarak oluşturulmuş grafikler gösterilmektedir.



Şekil 2.7 Konuşma metni ile hizalanmış bir ses kaydının dalga formu ve spektrogram grafikleri gösterimi.

Zaman alanındaki konuşma sinyalinden doğrudan sessizlikleri gözlemleyebilmek mümkündür. Bu sinyale bakılarak doğrudan "üstesinden gelinecektir" cümlesindeki "gelinecek" kelimesi ve "tir" eki arasında bir miktar duraksama yapıldığı yorumu yapılabilir.

Spektrogram incelendiğinde ise sesli harflerde özellikle bazı bölümlerin koyu renkle, bazı bölgelerin ise açık renkle gösterilmesi dikkat çekmektedir. Çizimin daha anlaşılır olabilmesi için koyu renkli olan alaların ortalama frekansları kırmızı noktalarla gösterilmektedir. Koyu renkli olan bölümler aşağıdan yukarıya doğru F1, F2 ve F3 formantları şeklinde ilerlemektedir. /e/ fonemine karşılık gelen F1 formantlarının yaklaşık olarak 500-550Hz dolaylarında olduğu görülmektedir. Türkçedeki /e/ fonemi, uluslararası fonetik alfabede /ɛ/ fonemine karşılık gelmektedir. Tablo 1'de de gösterildiği üzere /ɛ/ fonemine ait F1 frekansı ortalama olarak 530Hz'e karşılık gelmektedir.

Konuşma sinyalleri üzerinde yapabildiğimiz bu yorumların, makineler aracılığı ile yapılması sağlanarak ses kayıtlarının otomatik olarak yazılandırılması gerçekleştirilebilmektedir. Üçüncü bölümde makineler ile otomatik konuşma tanıma işleminin nasıl yapıldığından bahsedilecektir.

3
OTOMATİK KONUŞMA TANIMA

3.1 Giriş

Otomatik konuşma tanıma, insanlar arasında konuşulan dilin makineler aracılığı ile yazıya dönüştürülmesi işlemidir.

Konuşmayı tanıyabilen ilk makine, 1920'li yıllarda ticarileştirilerek satışa sunulan Radio Rex isimli oyuncak olmuştur. Rex, 500Hz'lik enerji seviyesinde tetiklenin bir yayın serbest bırakılması ile kulübesinden çıkan oyuncak bir köpektir. 500Hz, sistemi tetikleyen "Rex" komutunun ilk formantı olan /eh/ ses birimine karşılık gelen frekans seviyesidir [18]. Fakat bu oyuncağı sadece yetişkin erkekler "Rex" komutu ile tetikleyebilmektedir. Bunun nedeni erkek, kadın ve çocukların ses teli ve ses yolundaki farklılıktan kaynaklanmaktadır. Yetişkin erkek sesleri, kadınların ve çocukların seslerine göre daha pes yani daha düşük frekanslıdır. Bu nedenle yetişkin erkeklerdeki yaklaşık olarak 500Hz'lik frekans seviyesine karşılık gelen /eh/ fonu, kadın ve çocuklarda farklı fonlara karşılık gelebilmektedir. Bu oyuncak, ses birimlere ait farklı formantların olduğu, farklı yaş grupları ve cinsiyete göre bu formantların değişim göstermesinin makinelerin de algısını değiştireceğini göstermiştir. Böylelikle, ses tanımak üzere geliştirilecek olan projelerde, bu bilginin de göz önünde bulundurulması gerektiği anlaşılmıştır.

1962 yılında IBM, ses ile kontrol edilebilen bir çeşit hesap makinesi olan Shoebox isimli projesini duyurmuştur. Shoebox, rakamları ve birkaç aritmetik operasyonu, üzerindeki mikrofon sayesinde algılamakta ve daha sonra ses sinyallerinin farklı özelliklerinden yararlanarak bu girdileri sınıflandırmaktadır [19].

1976 yılında DARPA tarafından fonlandırılan bir proje yarışmasının kazananı olan Harpy, yaklaşık olarak üç yaşındaki bir çocuğun kelime dağarcığına sahiptir. Binden fazla kelimeyi kabul edilebilir doğruluk oranı ile tanıyabilmiştir. Bu proje, diğer konuşma tanıma sistemlerinden farklı olarak dil modellerinin kullanıldığı ilk sistem olmuştur [20].

Kısa süre sonra, 1980'li yıllarda Saklı Markov Modelleri (Hidden Markov Model - HMM) [21] ile istatistiksel konuşma tanıma sistemleri geliştirilmeye başlanmıştır. Bununla beraber 10 binden fazla kelime başarı ile tanınmıştır.

2010 sonrasında ise Derin Sinir Ağı (Deep Neural Network - DNN) yapısı ile 1 milyon üzerinde kelime tanıma başarımına ulaşılmıştır. Günümüzde en başarılı otomatik konuşma tanıma sistemleri genellikle derin öğrenme yöntemleri ile oluşturulmaktadır.

3.2 Problem Tanımı

Otomatik konuşma tanıma, bir dizi akustik özelliği bir dizi kelimeye dönüştürme işlemidir. Güncel yöntemlerde; gözetimli makine öğrenmesi ile sistemin girişine verilen akustik özellik vektörlerinin hangi fonlara veya fon gruplarına ait olduğunun sınıflandırılması sağlanmaktadır. Bu ifade matematiksel şekilde Denklem 3.1'deki gibi ifade edilebilmektedir.

$$W^* = argmax\, P(W|X) \tag{3.1}$$

Denklemlerde W^* en olası kelime dizisini, W arama uzayı içindeki kelimeleri, X ise akustik özellik vektörünü temsil etmektedir. Denklem 3.2'ye, Denklem 3.3'te gösterildiği gibi Bayes Teoremi uygulanırsa en olası kelime dizisi W^*'ı akustik model ve dil modeline bağlı şekilde göstermek mümkün hale gelmektedir.

$$P(W|X) = \frac{P(X|W)P(W)}{p(X)} \propto p(X|W)P(W) \tag{3.2}$$

$$W^* = argmax\, p(X|W)P(W) \tag{3.3}$$

Denklem 3.3'te, $p(X|W)$ akustik modeli, $P(W)$ ise dil modelini temsil etmektedir.

Başarılı bir akustik model geliştirebilmek için genellikle ses sinyallerinin, frekans alanındaki özelliklerinden de yararlanarak birtakım vektörler çıkarılmaktadır. Bu işleme öznitelik çıkarma denilmektedir. Çıkarılan öznitelikler, akustik modellerin oluşturulmasında kullanılmaktadır. Akustik modeller; gözlemlenen özellikler ile fonları, dil modelleri ise kelimelerin birbiri ile ilişkisini olasılıklandırmaktadır. Bu

iki olasılığın çarpımı en muhtemel kelime dizisinin bulunmasında kullanılmaktadır.

3.3 Öznitelik Çıkarma

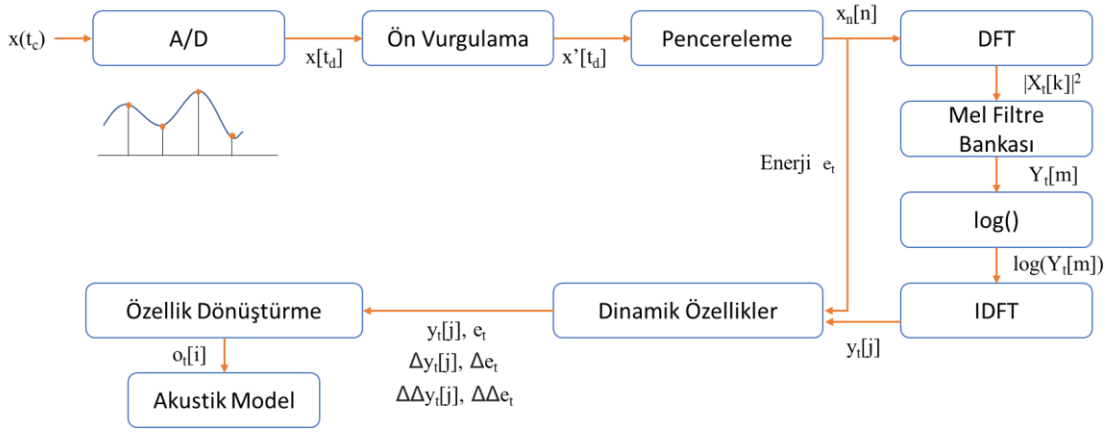
Konuşma tanıma sistemlerinde, konuşmacı ve ortam farklılıkları önemlidir. Konuşmanın akustik olarak modellenmesi, konuşma dalga formundan elde edilen özelliklere karşı, fonemlerin elde edilmesini sağlayacak istatistiksel bilgilerin modellenmesi işlemidir.

Akustik modellerin konuşmacı ve ortam farklılıklarına karşı gürbüz olabilmesi için genellikle farklı ortamlarda ve farklı kişilerce seslendirilmiş çok büyük veri setlerinin (+500 saat) kullanılması tercih edilmektedir.

Konuşma kayıtları, kayıt ortamına bağlı olarak, konuşma dışında bir takım gürültü ve yankıları da içerebilmektedir. Konuşma tanıma modellerinin eğitimi esnasında gürültüden ziyade daha çekirdek bilgi taşıyan bölümlere odaklanmak gerekmektedir. Bu durumda, insanların konuşma ve duyma frekans aralıklarının dikkate alınması önemli bir durumdur. Bunun için doğrudan konuşma sinyalleri ile model eğitmek yerine birtakım öznitelik çıkarma işlemlerinin yapılması genellikle tercih edilmektedir. Bu bağlamda, insan kulağının duyma fizyolojisinden esinlenerek tasarlanan ve en sık kullanılan öznitelik çıkarma yöntemleri Mel-Frekansı Kepstrum Katsayıları (Mel Frequency Cepstral Coefficients - MFCC) ve log mel-spektrumdur.

3.3.1 MFCC ve Dinamik Özellikler

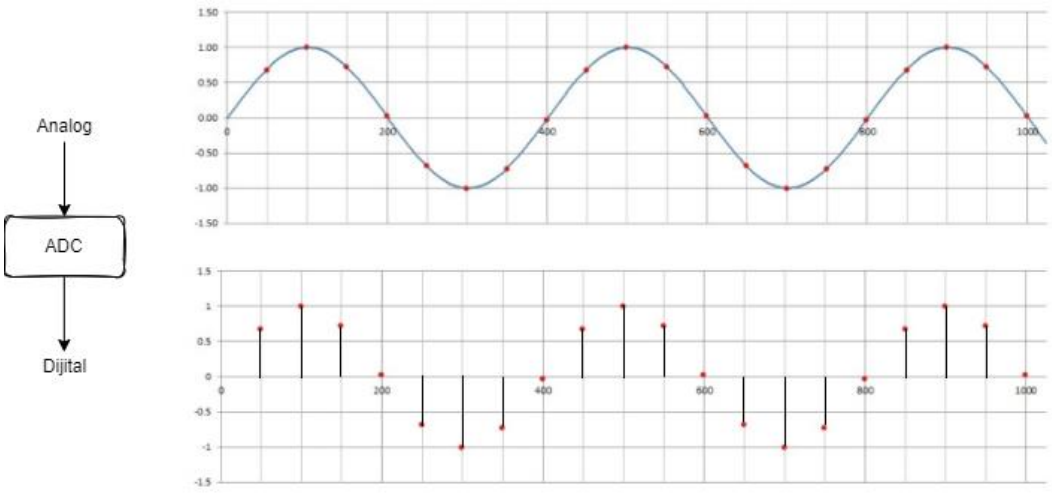
MFCC, insan kulağının duyma fizyolojisini temel alarak tasarlanmış bir özellik çıkarma yöntemidir. Şekil 3.1'de MFCC özellikleri çıkarılırken izlenilen adımlar gösterilmektedir.



Şekil 3.1 MFCC özellikleri çıkarma adımları [22].

Analog-Dijital Dönüşümü

Mikrofonun çıkışından analog olarak elde edilen işaretin, makine tarafından işlenebilmesi için dijitalleştirilmesi gerekmektedir. Konuşma sinyallerindeki bilginin büyük bölümü, 8 kHz altındaki frekans bantlarında yer almaktadır. Örnekleme esnasında örtüşme problemleri ile karşılaşılmaması için Nyquist Teoremi dikkate alınarak örnekleme yapılmaktadır. 8 kHz bir sinyal için 16 kHz örnekleme frekansının tercih edilmesi uygun olmaktadır. Eğer konuşma tanıma sistemi telefon konuşmalarının yazılandırılabilmesi için geliştirilecekse 8 kHz örnekleme frekansı kullanılmaktadır. Bunun nedeni telefon görüşmelerinin bant genişliğinin maksimum 4 kHz olmasından kaynaklanmaktadır. Bu durumda, konuşma tanıma sistemlerinde genellikle 8 kHz veya 16 kHz örnekleme frekansları kullanılmaktadır.



Şekil 3.2 Analog dijital dönüşüm.

Ön Vurgulama

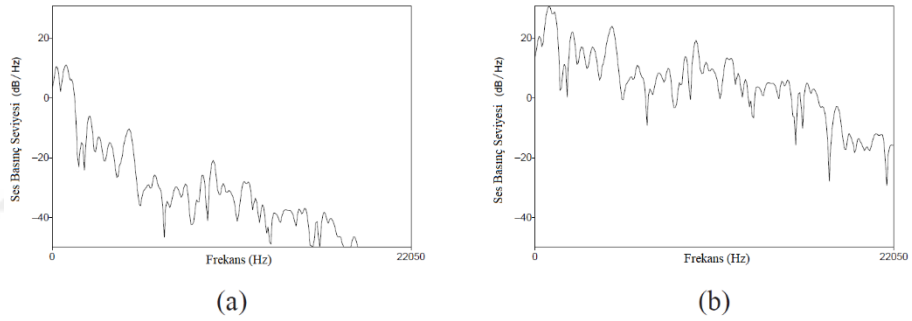
Ön vurgulama, yüksek frekanslardaki enerji miktarını arttırmaktadır. Yüksek frekanslardaki enerjiyi artırmak, bu frekans seviyelerindeki formantlara ait bilgilere ulaşmayı kolaylaştırmaktadır. Böylelikle ön vurgulama işlemi, fon algılama doğruluğunu arttırmaktadır.

Ön vurgulama, yüksek frekansları güçlendirmek için Denklem 3.4'teki gibi bir filtre kullanmaktadır.

$$x'[t_d] = x[t_d] - \alpha x[t_d - 1] \qquad (3.4)$$

$$0.95 < \alpha < 0.99$$

Şekil 3.3, ön vurgulama işlemi öncesinde ve sonrasında işaretin değişimini göstermektedir.



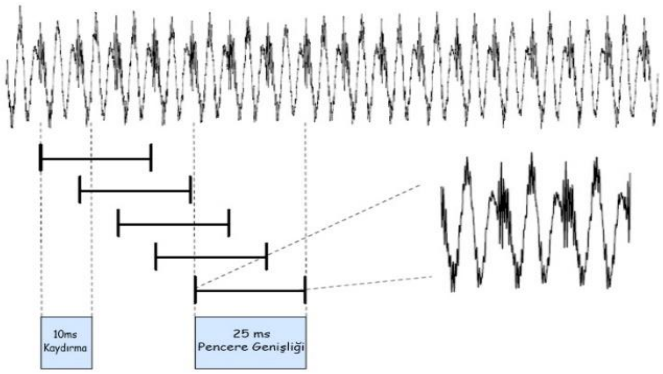
Şekil 3.3 Ön vurgulama işlemi: (a) ön vurgulamadan önce, (b) ön vurgulamadan sonra [23]

Pencereleme

Pencereleme işlemi, sinyali 25ms genişliğindeki bölütlere ayırmak için kullanılmaktadır. Pencere genişliği, bir saniye içerisinde 4 fona sahip 3 kelime söylenildiği ve her fonun 3 alt durumunun olduğu göz önüne alınarak hesaplanmıştır. Bu durumda bir saniyede 36 durum oluşmaktadır ve her durum yaklaşık 28ms sürmektedir. Bu nedenle pencere genişliğinin 25ms olarak alınması uygun bir yaklaşım olmaktadır.

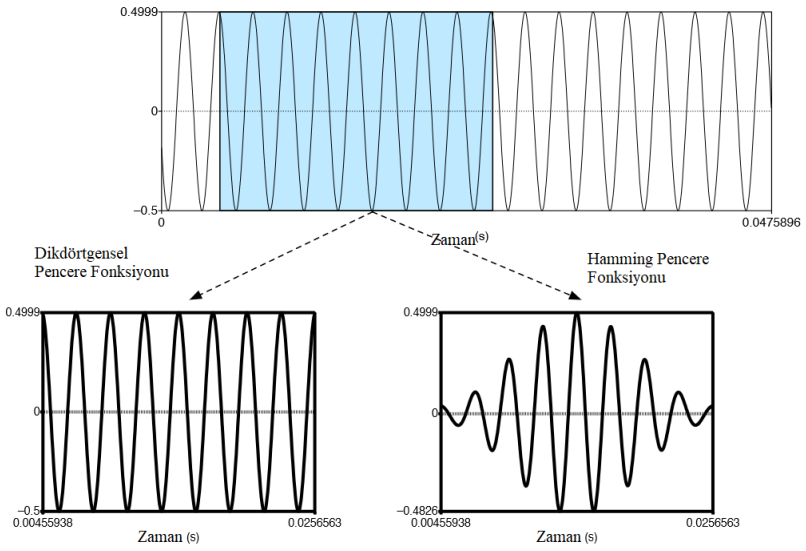
Konuşma sürekli olan bir eylem olduğu için ses çıkışını şekillendiren fonların üretilmesini sağlayan ağız eklemlerinin hareketi, fonlar arası geçişi etkilemektedir. Yani bir fon kendinden önce ve sonra gelen fonlara göre değişiklik göstermektedir.

Bu değişimi yakalayabilmek için elde edilen pencere Şekil 3.4'te görüldüğü gibi 10ms aralıklarla kaydırılmaktadır.



Şekil 3.4 Pencereleme işlemi [24].

Sinyali dikdörtgensel pencere kullanarak kesmek, kesilen yerlerde meydana gelen ani genlik değişimlerinden dolayı gürültüye sebep olmaktadır. Bu durumun önüne geçmek için sinyal, *Hamming* veya *Hanning* gibi pencere fonksiyonları kullanılarak kesilmektedir. Dikdörtgensel ve *Hamming* pencereleri ile kesilen sinyallerin sonucu Şekil 3.5'te gösterilmektedir.



Şekil 3.5 Sinyallerin kesiminde farklı pencereleme fonksiyonlarının kullanımı [24].

$$y[n] = w[n]s[n] \tag{3.5}$$

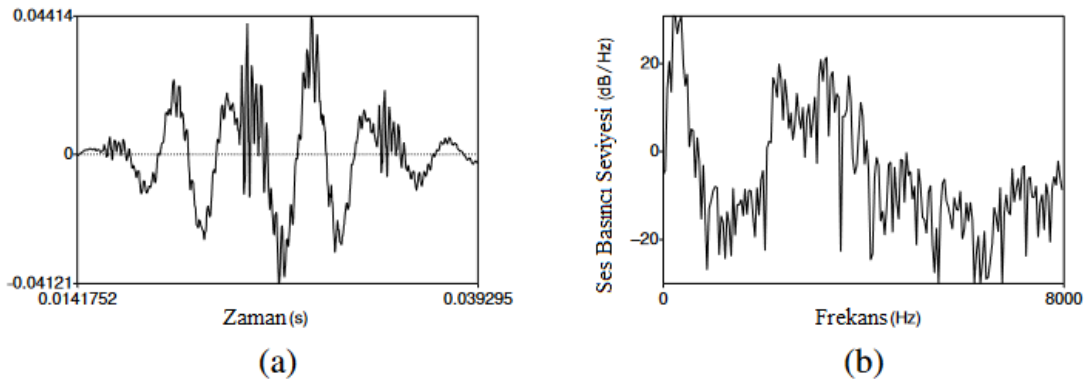
$$w[n] = (1 - a) - \text{acos}\left(\frac{2\pi n}{L-1}\right) \qquad (3.6)$$

$$Hamming(a = 0.46164), \qquad Hanning(a = 0.5)$$

Denklem 3.5'te $s[n]$ kesilecek olan sinyali, $w[n]$ pencere fonksiyonunu, $y[n]$ kesilmiş sinyali ve L ise pencere genişliğini ifade etmektedir.

Ayrık Fourier Dönüşümü (AFD)

Ayrık zamanlı sinyalin, ayrık frekans bantlarındaki spektral bilgisini çıkartılabilmesi için Ayrık Fourier Dönüşümü yapılmaktadır. Böylelikle farklı frekans bantlarının ne kadar enerjiye sahip olduğu anlaşılabilmektedir.



Şekil 3.6 *Hamming* pencere fonksiyonu ile kesilmiş /i/ fonuna ait (a) zaman ve (b) spektrum grafikleri [24].

Şekil 3.6'da soldaki grafik /i/ fonemine ait bir sinyalin *Hamming* pencere fonksiyonu ile kesilmiş halini yansıtmaktadır. Sağdaki grafik ise zaman alanındaki sinyalin Fourier Dönüşümü alınarak elde edilmiştir.

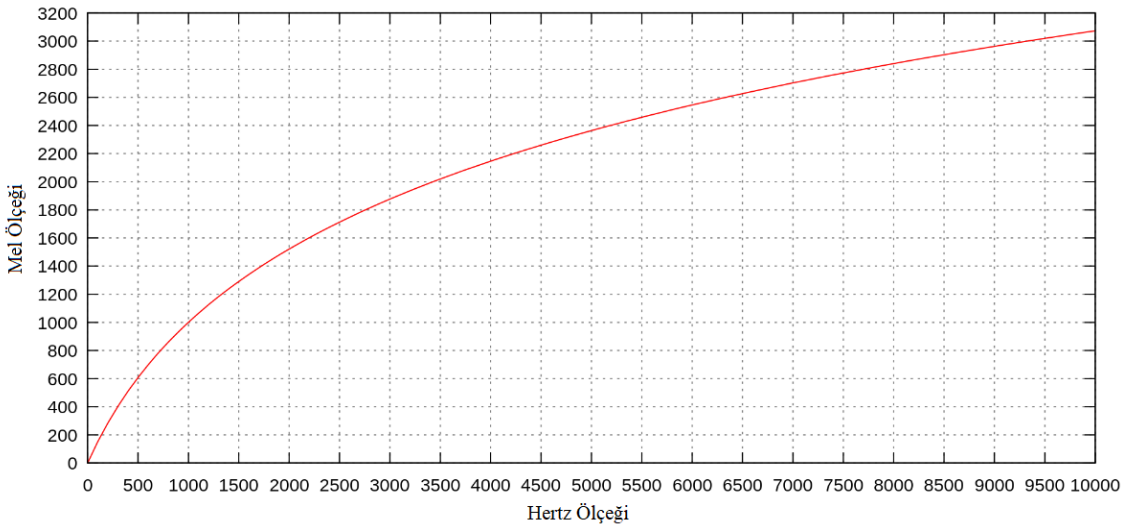
Denklem 3.7'de Ayrık Fourier Dönüşümü gösterilmektedir. AFD hesaplanırken karmaşıklığının az olması nedeni ile Hızlı Fourier Dönüşümü algoritması tercih edilmektedir.

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn} \qquad (3.7)$$

Mel Filtre Bankası ve Log

Bir ses kaynağı tarafından üretilen ses ile insanların algıladıkları sinyallerinin frekansları aynı değildir. İnsanlar tarafından algılanan ses yüksekliği frekansa göre değişim göstermektedir. Ayrıca, algılanan frekansın çözünürlüğü frekans arttıkça azalmaktadır. Bu yüzden insanlar, yüksek frekanslara karşı daha az hassastır. 0 ila 1kHz arası ses sinyallerinin insanlar tarafından algılanması neredeyse doğrusalken, 1kHz'den sonra frekans artışı ile logaritmik olarak azalmaktadır [25]. İnsanlar tarafından algılanan ses ile kaynaktan çıkan sesin arasındaki ilişki Şekil 3.7'deki gibi Mel ölçeği ile gösterilmektedir. Bu ilişki Denklem 3.8 ile hesaplanabilmektedir.

$$M(f) = 1127\ln(1 + f/700) \tag{3.8}$$



Şekil 3.7 Hertz ölçeğinin mel ölçeği üzerindeki karşılığı.

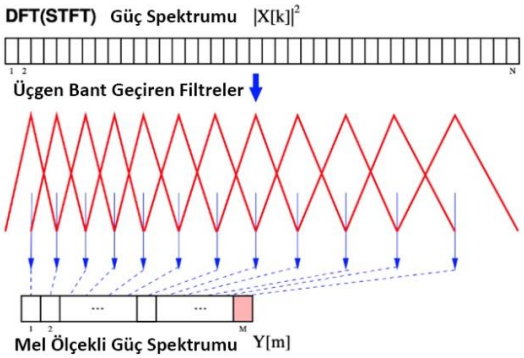
Bölüm 2'de anlatıldığı gibi mel filtre bankası insanların ses algılama yöntemini taklit etmeye çalışmaktadır. Kulakta bulunan salyangozun en dış kısmı yüksek frekanslara duyarlı iken iç kısım daha düşük frekanslı sinyallere duyarlıdır. Salyangoz içerisindeki tüycükler ile biyolojik olarak bant filtreleme işlemi yapılmaktadır.

MFCC ile öznitelik çıkarımı yapılırken, insan kulağı yapısındaki gibi bant geçiren filtreler kullanılmaktadır. AFD aşamasından sonra sinyalin gücü hesaplanmaktadır. Buna, AFD güç spektrumu denilmektedir. Ardından elde edilen

güce üçgen bant geçiren filtreler uygulanarak Mel-ölçekli güç spektrumu hesaplanmaktadır.

$$Y_t[m] = \sum_{k=1}^{N} W_m[k]|X_t[k]|^2 \tag{3.9}$$

Denklem 3.9'da k güç spektrumu numarasını, m ise mel filtre numarasını göstermektedir. Filtreler uygulandıktan sonra Şekil 3.8'de gösterilen özellik vektörü elde edilmektedir.



Şekil 3.8 Mel ölçekli güç spektrumunun elde edilmesi [22].

Kepstrum

Kepstrum analizi, ses tellerinden kaynaklı titreşim ile ses yolunda uygulanan filtrelerin oluşturduğu sinyallerin ayrıştırılması için kullanılmaktadır. Ses tellerinden kaynaklanan titreşim, fonların ayırt edilebilmesi için önemli bir bilgi taşımamaktadır [22]. Bu sinyal ayrıştırılarak diğer formant frekansları daha anlaşılır hale getirilebilmektedir.

Dinamik Özellikler

Dinamik özellikler, konuşmacılara ait konuşma tarz ve temposunun tanınmasında büyük bir rol oynamaktadır. Bunun için önceki durumlarda elde edilen özelliklerin zamana göre türevleri bu değişimleri göstermektedir. Enerji ile birlikte elde edilen toplam 13 özelliğin zamana göre birinci ve ikinci türevleri alınarak dinamik özellikler çıkarılmaktadır. Çerçeveler arasındaki geçişlerde meydana gelen değişimlerin ölçülmesi için bu özelliklerin çıkarılması önemlidir. Dinamik özelliklerle beraber her çerçeve için 39 adet özellik elde edilmektedir. Zamana göre türev alma işlemi Denklem 3.10'da gösterilmektedir.

$$d(t) = \frac{c(t+1) - c(t-1)}{2} \tag{3.10}$$

3.3.2 Log Mel Spektrum

Log Mel spektrum özellikleri MFCC ile oldukça benzerlik göstermektedir. Log Mel spectrum özelliklerinin çıkarılması için Başlık 3.3.1 altında anlatıldığı gibi;

- ✓ Analog Dijital Dönüşüm
- ✓ Pencereleme
- ✓ Ayrık Fourier Dönüşümü
- ✓ Mel Filtre Bankası ve Log

işlemleri takip edilmektedir.

3.4 Konuşma Tanıma Mimarileri

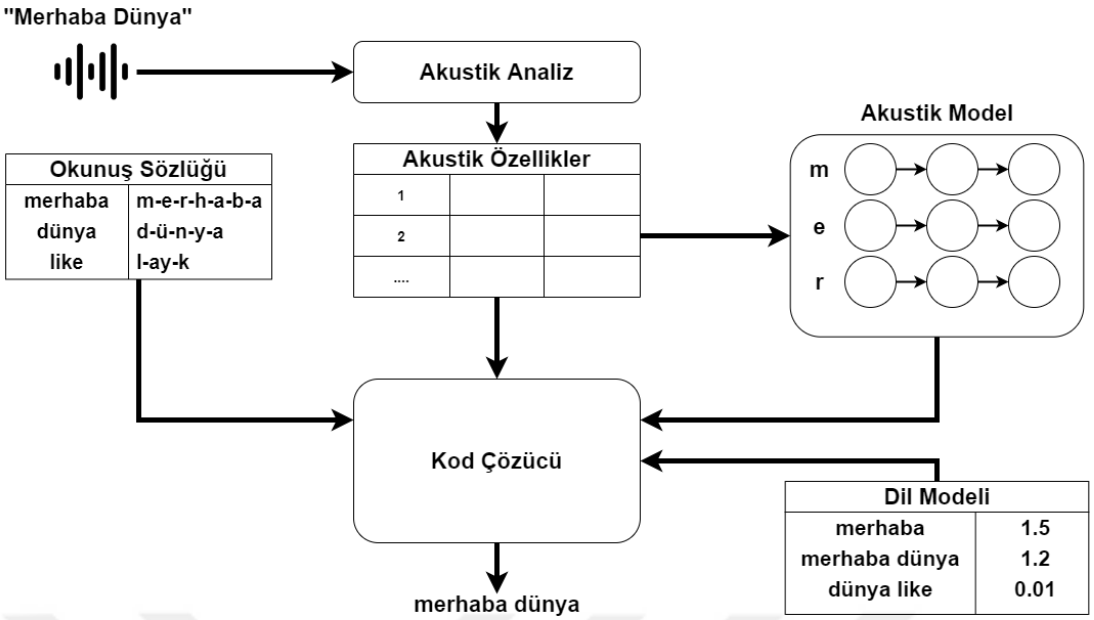
Güncel konuşma tanıma mimarilerinde ağırlıklı olarak iki yaklaşım kullanılmaktadır. İlk yaklaşım; akustik model, dil modeli ve okunuş sözlüğü modellerinin ağırlıklı sonlu durum dönüştürücü (Weighted Finite State Transducer - WFST) üzerine kurgulanmasına dayanırken, ikinci çözüm ise uçtan-uca eğitilmiş bir akustik model ile bütün problemi ele almayı hedeflemektedir.

Her iki yöntemde de genellikle doğrudan zaman alanındaki sinyalleri kullanmak yerine, konuşma için daha çekirdek bilgi taşıyan özniteliklerin çıkarılması ile akustik modellerin eğitimi sağlanmaktadır.

3.4.1 WFST Tabanlı Konuşma Tanıma Mimarisi

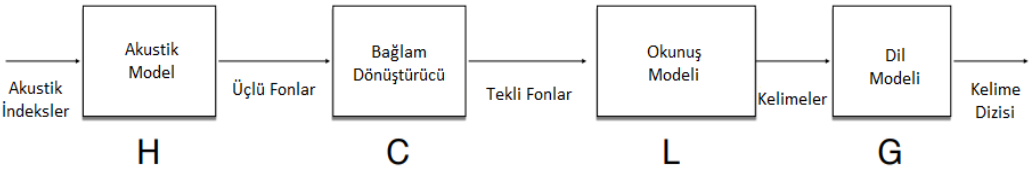
İstatistiksel konuşma tanıma mimarilerinden olan bu yaklaşım; Şekil 3.9'da gösterildiği gibi akustik model, dil modeli ve okunuş sözlüğü kullanılarak sistemin girişindeki akustik özellik vektörlerine karşı en olası kelime sekansını üretmektedir.

Bahsedilen üç model arasındaki ilişkilerin ağırlıklı sonlu durum dönüştürücüleri üzerinde modellenmesi ile oldukça büyük bir arama grafı oluşturulmaktadır. Elde edilen graf kullanılarak, modelin girişine verilecek akustik özellikler ile çıkışta en olası kelime dizisi üretilmeye çalışılmaktadır.



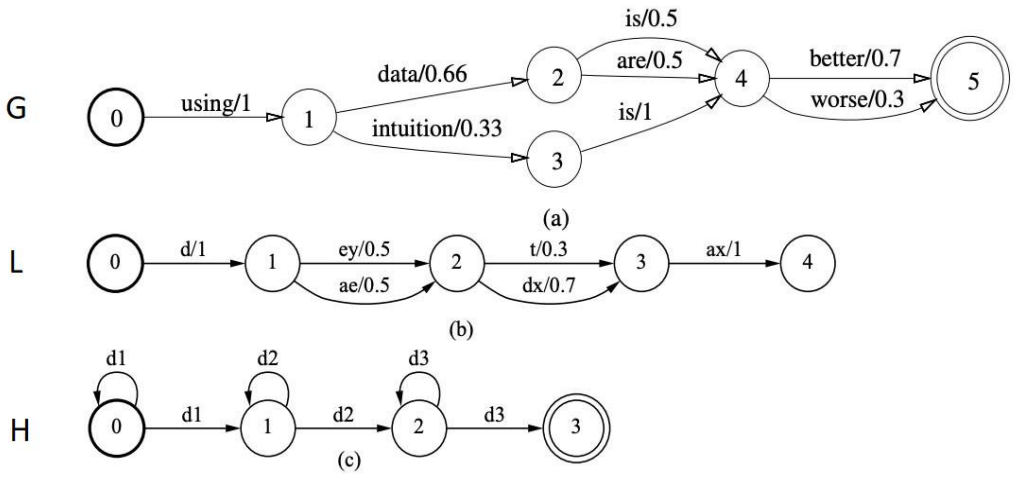
Şekil 3.9 WFST tabanlı konuşma tanıma mimarisi [26].

Konuşma işleme sistemlerinin tasarlanmasında sıklıkla kullanılan Kaldi [27] aracı, HCLG ismi verilen bir sonlu durum dönüştürücü kullanarak, dilbilgisi açısından doğru çıktı üretebilen bir otomatik konuşma tanıma sisteminin oluşturulmasını sağlamaktadır. HCLG olarak isimlendirilmiş graftaki her harf Şekil 3.10'da gösterildiği gibi ayrı bir sonlu durum dönüştürücüye karşılık gelmektedir. Bütün dönüştürücülerin birleştirilmesi ile HCLG grafı oluşturulmaktadır.



Şekil 3.10 HCLG giriş ve çıkışları [28].

HCLG grafı Şekil 3.11'de de gösterildiği gibi fonlar ve kelimeler arasındaki durumların geçiş olasılıklarını tutmaktadır.



Şekil 3.11 H, L ve G FST [29]

H: Bağlam bağımlı fonların temsil edilmesi için kullanılmaktadır. Her ses kendinden önceki ve sonraki ses göre farklılık gösterilebilmektedir. Burada fonlar üçlü gruplar halinde temsil edilmektedir.

C: Üçlü fonların tekli fonlar haline dönüştürülmesinde kullanılmaktadır.

L: Kelimeleri oluşturan fonlar arasındaki geçişleri temsil etmektedir.

G: Dil modelinden gelen olasılıklar kullanılarak, kelimeler arası geçişler temsil edilmektedir.

3.4.1.1 Akustik Model

Akustik modeller, akustik öznitelik girişleri ile fonemler veya benzeri dil öğeleri için olasılıksal bir dağılım oluşturmaktadır. Bu modeller eğitilirken veri seti içerisindeki ses kayıtlarına ait öznitelikler ve yazı çiftleri arasında istatistiksel bağlantı kurulmaktadır.

Başlarda sıklıkla Gauss Karışım Modelleri (Gaussian Mixture Model - GMM) ve Saklı Markov Modelleri (Hidden Markov Model - HMM) ile oluşturulan akustik modeller, ardından derin sinir ağları (Deep Neural Network - DNN) ile daha başarılı şekilde modellenmiştir [30]. Sonrasında ise akustik model oluşturmak için yinelemeli sinir ağları (Recurrent Neural Network - RNN) kullanılmaya başlanmıştır. Yinelemeli sinir ağları, konuşma gibi zamana bağlı dizilerden oluşan girişleri daha güçlü bir şekilde temsil edebilmektedir. Bu bağlamda, akustik model oluşturmada yinelemeli sinir ağlarından olan LSTM'lerin (Long Short Term

Memory) kullanılması DNN'lere kıyasla daha iyi sonuçların alınmasını sağlamıştır [31]. Ancak, yinelemeli sinir ağlarındaki yineleme bağlantıları, parallelleştirmedeki güçlük nedeniyle eğitimi yavaş kılmaktadır. Yinelemeli sinir ağları yerine, uzun süreli hafıza bağımlılıklarını ileri beslemeli şekilde temsil edebilen zaman gecikmeli sinir ağlarının akustik modelleme için kullanılması, eğitim süresini hızlandırmakta ve yinelemeli sinir ağları ile kıyaslanabilir modellerin oluşturulmasını sağlamaktadır [32].

Uslu'ya ait tez çalışmasında da GMM, DNN ve TDNN (Time Delay Neural Network) tabanı akustik model yaklaşımları incelenmiş olup, TDNN modelinin daha başarılı olduğu gösterilmiştir [33]. Bu çalışmada da TDNN tabanlı akustik model kullanılarak oluşturulan bir otomatik konuşma tanıma sistemi ile uçtan uca eğitilmiş bir konuşma tanıma modeli karşılaştırılacaktır.

GMM Tabanlı Akustik Model

Gauss karışım modeli (GMM), gözetimsiz öğrenme algoritmalarındandır ve ağırlıklı Gauss dağılımlarının birleştirilmesi sonucundan oluşturulmaktadır. Özellikle, doğrusal olmayan veri dağılımlarının modellenmesi için kullanılabilmektedir. Tam bir Gauss karışım modeli; ortalama vektörleri, kovaryans matrisleri ve karışım ağırlıkları ile parametrelendirilmektedir [34]. GMM parametreleri, eğitim verileri üzerinde Beklenti Maksimizasyon (Expectation Maximization - EM) algoritması kullanılması sonucunda hesaplanabilmektedir.

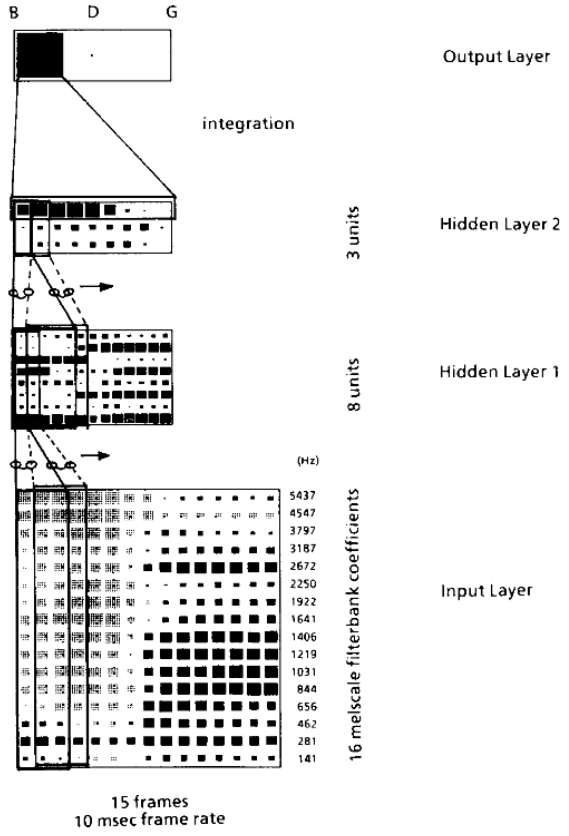
GMM eğitiminin hızlı olması nedeniyle günümüzde de konuşma tanıma sistemlerinin tasarlanmasında sıklıkla kullanılmaktadır. Akustik özellikler ile fonlar arasındaki benzerliklerin modellenmesini sağlayan GMM, ses ve metin verilerinin hizalanması için kullanılabilmektedir.

TDNN Tabanlı Akustik Model

Akustik modelleme için geliştirilecek olan yapay sinir ağı mimarilerinde; zamana bağlı olayların modellenebilmesi, lineer olmayan karar yüzeylerinin belirlenebilmesi ve yeterli miktarda ağırlık kullanılarak mümkün olduğunca küçük bir model oluşturabilmek önemlidir. Bu noktada, bu özelliklere uygun olan zaman gecikmeli sinir ağı mimarisi (TDNN) [35] önerilmiştir. Zaman gecikmeli sinir ağı

mimarisi, belirli bir miktar zaman gecikmesi ile girişteki öznitelikleri bir sonraki katmana taşımaktadır. Böylelikle, o anki gözlem ile geçmiş ve gelecekteki gözlemler arasında ilişki kurulabilmektedir. Bu özelliği ile TDNN mimarisi, yapısal olarak evrişimli sinir ağlarına benzemektedir.

Şekil 3.12'de TDNN mimarisi gösterilmektedir. TDNN mimarisi, zamansal bağımlılıkları modelleyebilmek için her katmanda belirli bir çerçeve genişliğindeki özellikleri kabul ederek bir sonraki katman için bir vektör oluşturmaktadır. Giriş özellikleri birinci gizli katmana aktarılırken üç, birinci gizli katmandan ikinci gizli katma geçişte ise beş genişliğinde çerçeve kullanılmıştır. Katman derinliği arttıkça kısa zamanlı gözlemler ile daha geniş zamanlı gözlemler oluşturulabilmektedir.



Şekil 3.12 Zaman gecikmeli sinir ağı mimarisi [35].

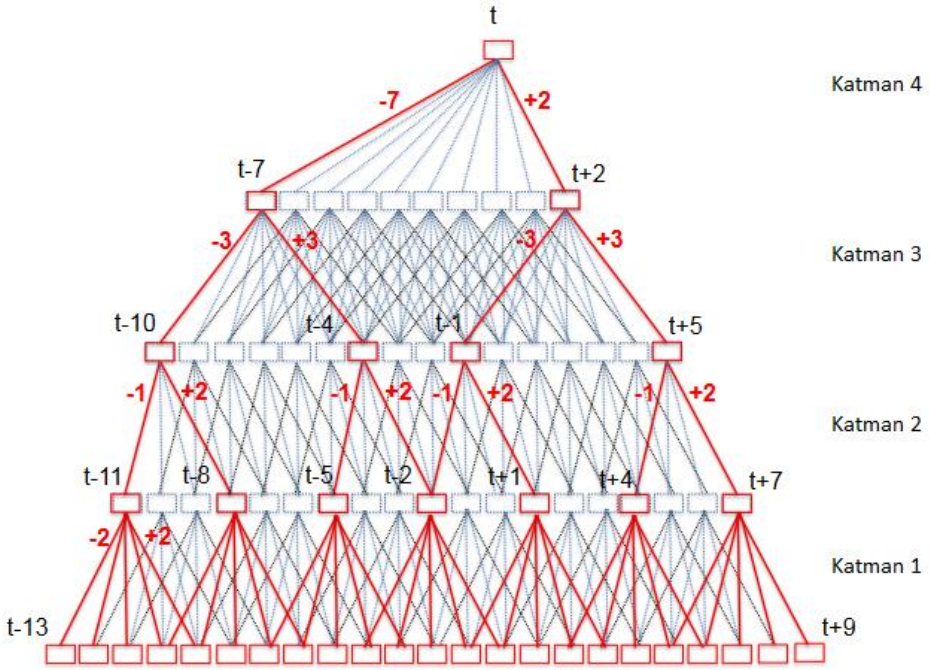
TDNN mimarisinde, giriş katmanından çıkış katmanına doğru ilerlenirken daha geniş ölçekte zaman gecikmesinin kullanılması, kısa süreli zamansal gözlemler ile daha uzun süreli zamansal gözlemlerin elde edilmesini sağlamaktadır. Yani, çıkış katmanına yakın olan katmanlar, daha geniş ölçekte zamansal gözlemleri içermektedir.

Bu ağ mimarisinin zaman bağımlılıklarını ileri beslemeli bir şekilde modelleyebilmesi, eğitim süresini hızlandırdığı için konuşma tanıma gibi büyük miktarda veri ile eğitim yapılan uygulamalarında sıklıkla tercih edilmektedir.

Akustik model oluşturmada sıklıkla tercih edilen zaman gecikmeli sinir ağları, girişte akustik özellikleri kabul ederek çıkışta fon veya fon gruplarına ait olasılıksal bir dağılım sunmaktadır. Gradyan inişi algoritmasından yararlanarak ortalama kare hatasının geri yayılımı ile TDNN'lerde öğrenme gerçekleştirilebilmektedir.

Temel bir TDNN mimarisinde, komşu zaman adımlarında hesaplanan aktivasyonların girdi bağlamları arasında büyük örtüşmeler vardır [32]. Bu durumda, komşu düğümlerin tamamını kullanmak yerine bir bölümünün kullanılması, parametre sayısının azaltılarak modelin küçülmesini ve eğitimin hızlanmasını sağlamaktadır. Bunun için örneklemeli zaman gecikmeli sinir ağı mimarisi önerilmiştir [32].

Şekil 3.13'te örneklemeli zaman gecikmeli sinir ağı mimarisi gösterilmektedir. Kırmızı renk ile belirtilen ağırlıklar alt örneklemeleri temsil etmektedir.



Şekil 3.13 Örneklemeli zaman gecikmeli sinir ağı mimarisi [32].

3.4.1.2 Dil Modeli

Dil modeli, kelimelerin yan yana gelme olasılıklarını hesaplayarak gelecekteki kelimenin tahmin edilmesinde kullanılabilmektedir. Günlük hayatta sıklıkla kullanılan mesajlaşma uygulamalarında, yazılan kelimelerden sonra bazı kelimelerin önerilmesi dil modelleri sayesinde yapılmaktadır. Buna ek olarak, yazım araçlarında, yazım ve dil bilgisi hatalarının düzeltilmesinde de kullanılmaktadır.

Dil modelleri; konuşma tanıma sistemlerinde akustik olarak birbirine benzeyen kelimelerin, cümle bağlamına bakılarak doğru tahmin edilmesinde ve kelimelerin yanlış telaffuz edilse dahi doğru olarak yazılandırılmasında önemli bir rol oynamaktadır. Bu duruma fonetik olmayan bir dil olan İngilizce için örnek vermek daha anlaşılır olacaktır. Gece anlamına gelen *night* kelimesi ile şövalye anlamına gelen *knight* kelimelerinin akustik özellikleri neredeyse aynıdır. Konuşma tanıma sistemlerinde, dil modelinin kullanılması konuşma bağlamına göre doğru kelimenin bulunmasına yardım etmektedir.

Dil modeli üretiminde çeşitli yaklaşımlar olmasına karşın, konuşma tanıma sistemlerinde n-gram dil modelleri ağırlıklı olarak kullanılmaktadır. N-gram n tane kelimeden oluşan bir diziyi temsil etmektedir.

Denklem 3.11 ve 3.12'de 3-gram dil modeli için hesaplama örneği gösterilmektedir.

$$P(w_i|w_{i-2}, w_{i-1}) = \frac{c(w_{i-2}, w_{i-1}, w_i)}{c(w_{i-2}, w_{i-1})} \tag{3.11}$$

Denklem 3.11'de w ile gösterilen semboller kelimeleri temsil etmekte, w_i kelimesinden önce gelen iki kelime sırayla w_{i-1} ve w_{i-2} olarak gösterilmektedir. Derlem içerisinde w_{i-2}, w_{i-1}, w_i kelimelerinin yan yana geldiği durumların toplam sayısının, sadece w_{i-2}, w_{i-1} kelimelerinin yan yana gelme durumlarının toplam sayısına oranı, w_i kelimesinden önce w_{i-2} ve w_{i-1} kelimelerinin gelme olasılığını göstermektedir.

$$P(w_1, \dots, w_L) = \prod_{i=1}^{L+1} P(w_i | w_{i-2}, w_{i-1}) \quad (3.12)$$

Denklem 3.11'deki olasılığın, Denklem 3.12'de olduğu gibi bütün kelime dizisi üzerinde hesaplanması ile dil modeli oluşturulabilmektedir.

Dil modelleri oluşturulurken olasılıksal dağılımların konuşma dilini iyi temsil edebilmesi için milyonlarca cümleden oluşan büyük derlemlerin kullanılması faydalı olmaktadır.

3.4.1.3 Okunuş Sözlüğü

Kelimelerin okunuşlarının temsil edildiği bir sözlüktür. Türkçe, fonetik yapısından dolayı yazıldığı gibi okunmaktadır. Bu nedenle sözlük içerisindeki kelime ve okunuş karşılıkları genellikle aynıdır. Fakat günlük konuşma dilinde kelimeler, okunduğundan farklı şekillerde telaffuz edilebilmektedir. Bu telaffuz farklılıklarına karşı doğru yazılandırma işleminin yapılabilmesi için diğer okunuş varyantlarının da Tablo 3.1'de gösterildiği gibi okunuş sözlüğüne eklenmesi faydalı olacaktır.

Tablo 3.1 Okunuş sözlüğü örneği.

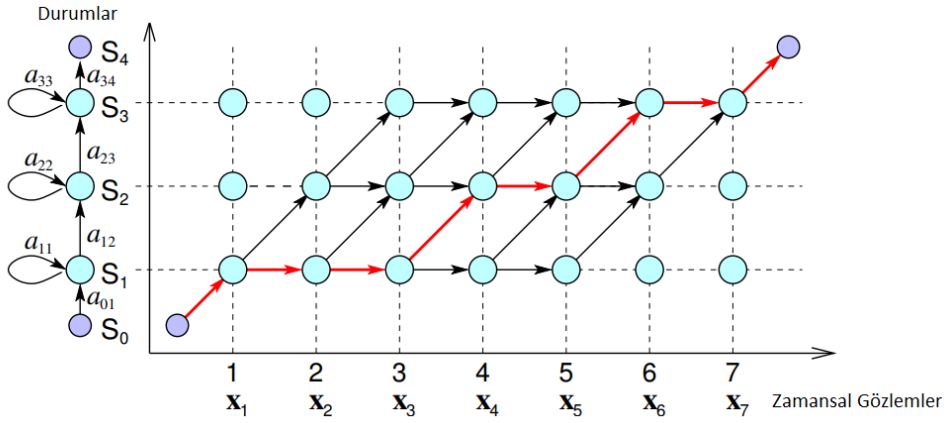
Kelime	**Okunuşu**
geliyorum	g e l i y o r u m
geliyorum	g e l i y o m
gideceğim	g i d e c e ğ i m
gideceğim	g i t c e m

3.4.1.4 Kod Çözücü

Kod çözme aşaması, Denklem 3.3'te gösterildiği gibi akustik gözlemlere karşı en olası kelime sekansının bulunması işlemidir. Akustik model, okunuş sözlüğü ve dil

modelleri ile oluşturulmuş graf içerisinden akustik gözlemlere göre en olası kelime sekansının bulunması için genellikle Viterbi Algoritması kullanılmaktadır.

Şekil 3.14'te olasılığı en yüksek kelime sekansının bulunmasında kullanılacak kafes yapısı gösterilmektedir. Şekildeki kafes yapısı, ağırlıklı sonlu durum dönüştürücü modeller ile tasarlanmış bir konuşma tanıma sistemini temsil etmektedir. Viterbi algoritması, zamansal gözlemler ile bu kafes üzerindeki olasılığı en yüksek olan yolu arayarak en muhtemel kelime dizisini bulmaktadır.



Şekil 3.14 Konuşma tanıma sistemleri için örnek bir kafes yapısı [36].

t anındaki Viterbi olasılığı Denklem 3.13'te gösterilmektedir. Hesaplamada; $v_{t-1}(i)$ ile bir önceki adımdaki Viterbi olasılığı, a_{ij} ile i durumundan j durumuna geçiş olasılığı ve $b_j(o_t)$ ile mevcut j durumundaki o_t gözlem olasılığı gösterilmektedir. Bütün olasılıkların çarpımının maksimum değeri yeni Viterbi yolunun bulunmasını sağlamaktadır.

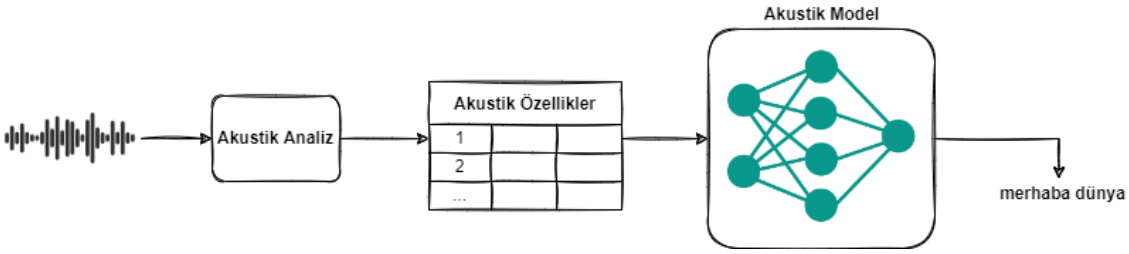
$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i) a_{ij} b_j(o_t) \tag{3.13}$$

3.4.2 Uçtan Uca Konuşma Tanıma Mimarisi

Uçtan uca konuşma tanıma sistemleri, okunuş sözlüğü ve dil modeline ihtiyaç duyulmadan, başarılı akustik modeller oluşturularak konuşma tanıma problemini

tek adımda çözmeyi hedeflemektedir. Bu sayede, modellerin eğitim ve çıkarım süreçleri oldukça kolaylaşmaktadır.

Şekil 3.15'te uçtan uca konuşma tanıma sistemlerinin oluşturulmasında genel olarak izlenilen yol gösterilmektedir.



Şekil 3.15 Uçtan uca konuşma tanıma adımları

Uçtan uca konuşma tanıma sistemlerinde de akustik analiz ile elde edilecek MFCC veya Mel spektrogram gibi akustik özniteliklerin kullanımı sıklıkla tercih edilmektedir.

Bu mimarilerde bulunan akustik model bölümü genellikle yapay sinir ağı yapılarından oluşmaktadır. Yapay sinir ağları, giriş özelliklerinin daha iyi bir biçimde temsil edilmesi sağlamakta ve çıkışta tahmin edilecek sınıflar için olasılıksal bir dağılım sunmaktadır. Uçtan uca konuşma tanıma sistemlerinin güncel literatür çalışmalarında, akustik model bölümünde sıklıkla Bağlantıcı Zamansal Sınıflandırma (Cconnectionist Temporal Classification - CTC) [37] yöntemin kullanıldığı görülmektedir [38]–[40].

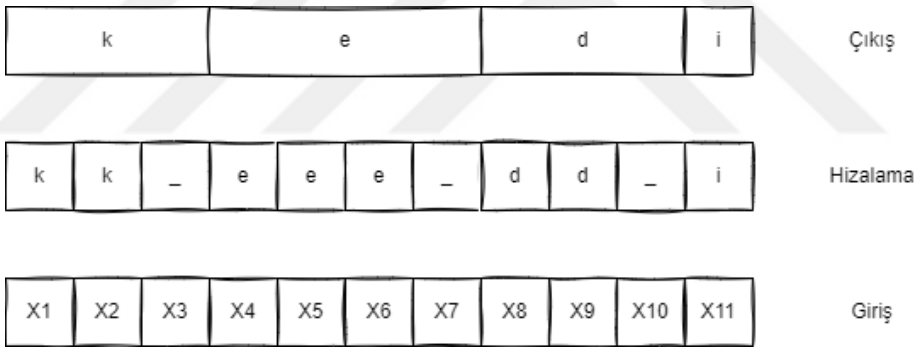
3.4.2.1 Bağlantıcı Zamansal Sınıflandırma (Connectionist Temporal Classification - CTC)

Konuşma tanıma modellerinin eğitimi için ses verileri ve bu verilere ait metinlere ihtiyaç duyulmaktadır. Veri seti içerisindeki bir konuşma parçacığının metin olarak içeriği bilinmesine karşın, ses ile metin arasında fon seviyesinde bir hizalama bulunmamaktadır. Eğer bu hizalama bilgisine sahip olunsaydı, giriş özellikleri ile çıkış sembollerini eşleyebilmek için pek çok farklı yöntemin kullanılması mümkün olurdu. İlk etapta bu işlemin insan gücü ile etiketlenerek yapılması durumu düşünülebilse de büyük veri setleri için bu durum oldukça maliyetli olacaktır. Bu durumda, uçtan uca konuşma tanıma sistemlerinde sıklıkla kullanılan Bağlantıcı Zamansal Sınıflandırma (CTC) tekniği, giriş özellikleri ve

çıkış arasında Şekil 3.16'daki gibi hizalama yaparak bu zorluğun önüne geçebilmektedir. Bu teknik, yinelemeli sinir ağlarından faydalanarak dizi verilerinin etiketlenmesini sağlamaktadır. Giriş dizisine bağlı olarak, tüm etiket dizisi için çıkışta olasılıksal bir dağılım üretilmekte ve doğru etiket değerleri ile bu olasılık dağılım maksimize edilmeye çalışılmaktadır. Benzer şekilde amaç fonksiyonu, Denklem 3.14'te de olduğu gibi negatif benzerlik olasılığının minimize edilmesi ile de kurgulanabilmektedir [37].

$$O^{ML}(S, \mathcal{N}_w) = -\sum_{(\mathbf{x},\mathbf{z})\in S} \ln(p(\mathbf{z} \mid \mathbf{x})) \tag{3.14}$$

Amaç fonksiyonu türevlenebilir olduğu için doğrudan hatanın geri yayılımı ile eğitim gerçekleştirilebilmektedir.



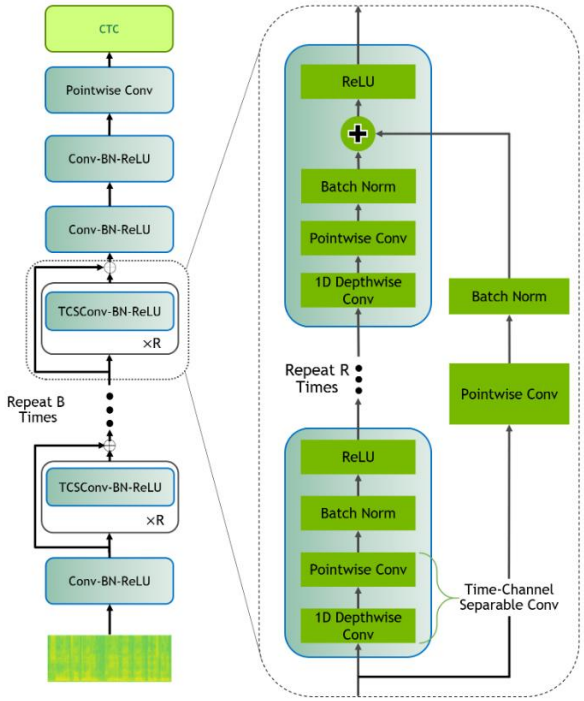
Şekil 3.16 CTC yöntemi ile hizalama işlemi.

Şekil 3.16'da gösterilen örnekte giriş vektörünün 11, çıkış dizisinin ise 4 uzunluğunda olduğu görülebilmektedir. CTC yöntemi, giriş ve çıkış özelliklerini birbirine eşleyebilmek için her girişe karşı bir tahmin yaparak hizalama işlemi gerçekleştirmektedir. Bu noktada tahmin edilecek semboller; alfabedeki karakterler ("[a-z]"), boşluk (" ") ve CTC için özel bir karakter olan bir ayraç ("_") sembollerinin oluşturduğu küme içerisinden seçilmektedir. Hizalamanın yapılmasından sonra ayraç karakteri görülene kadar tahmin edilen semboller birleştirilmektedir ve çıkış dizisi üretilmektedir.

3.4.2.2 QuartzNet

Uçtan uca konuşma tanıma sistemlerindeki MFCC veya Mel spektrogram benzeri girdi temsillerinin, derin ve geniş kapasiteli ağlar ile daha iyi bir biçimde temsil edilmesi ve ardından CTC yöntemi ile sınıflandırma yapılması sonucunda, ağırlıklı sonlu durum dönüştürücü gibi geleneksel konuşma tanıma mimarilerini geride bırakabilecek başarımlara ulaşılmıştır [39], [40].

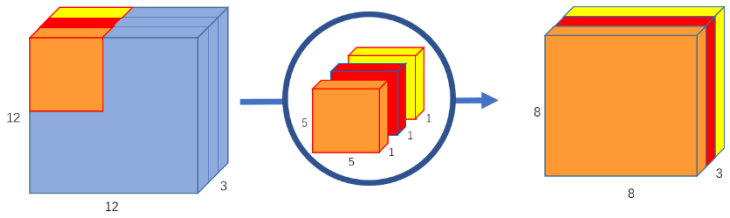
QuartzNet mimarisi de benzer bir yöntem kullanarak girdi temsillerini kabul etmekte; derinlemesine ayrılabilir evrişim katmanı, toplu normalleştirme ve ReLU katmanlarına sahip bloklar ile daha yoğun gösterimlerin elde edilmesini sağlamaktadır [40]. Oluşturulan mimarinin CTC kaybı ile eğitilmesi sonucunda oldukça başarılı modeller oluşturulabilmektedir. Şekil 3.17'de QuartzNet mimarisi gösterilmektedir.



Şekil 3.17 QuartzNet mimarisi [40].

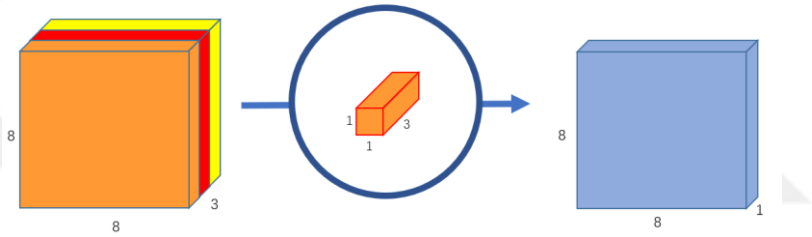
QuartzNet mimarisi, derinlemesine ayrılabilir evrişim modülünden oluşan blokları kullanmaktadır. Bu modülün kullanılmasının sebebi, sıradan evrişimli ağlara göre hesaplama maliyetinin daha düşük olmasıdır.

Derinlemesine ayrılabilir evrişim modüllerinin her biri derinlikli ve noktasal evrişim katmanlarından oluşmaktadır. Derinlikli evrişim katmanında, Şekil 3.18'e benzer şekilde her kanal için ayrı bir filtre uygulanmaktadır.



Şekil 3.18 Derinlikli evrişim katmanı örneği [41].

Noktasal evrişim katmanında ise 1x1 boyutunda bir filtre kullanılarak Şekil 3.19'da gösterildiği gibi derinlikli evrişim katmanında oluşan çıktıların doğrusal kombinasyonunun oluşturulması sağlanmaktadır [40].



Şekil 3.19 Noktasal evrişim katmanı örneği [41].

QuartzNet mimarisinin tasarımında, derinlemesine ayrılabilir evrişim katmanları özelleştirilerek tek boyutlu zaman-kanal ayrılabilir evrişim katmanı olarak kullanılmıştır. Tek boyutlu zaman-kanal ayrılabilir evrişimler; her kanalda ayrı ayrı ancak K zaman dilimlerinde çalışan, K çekirdek uzunluğuna sahip tek boyutlu derinlemesine evrişim katmanına ve her zaman diliminde bağımsız olarak ancak tüm kanallarda çalışan noktasal evrişim katmanına sahiptir [40].

Tekrarlı ve benzer evrişim işlemleri sonunda CTC bloğu kullanılarak evrişim katmanlarından elde edilen özelliklerin sınıflandırılması sağlanmaktadır.

3.5 Konuşma Tanıma Modellerinin Performanslarının Değerlendirilmesi

Konuşma tanıma sistemlerinin değerlendirilmesinde kelime hata oranı (KHO) metriği kullanılmaktadır. Otomatik olarak yazılandırma işlemi yapılırken üç farklı hata durumu ile karşılaşılmaktadır. Bunlar; kelime ekleme, silme ve yer değiştirme hatalarıdır. Kelime hata oranının hesaplamasında kullanılan yöntem Denklem 3.15'te gösterilmektedir.

$$KHO = 100x \frac{Ekleme + Silme + Yer\ Değiştirme}{Toplam\ Kelime\ Sayısı} \tag{3.15}$$

4
KONUŞMA SENTEZLEME

4.1 Giriş

İnsan-makine etkileşimi için oldukça önemli olan konuşma sentezleme, metinlerin yapay olarak seslendirilmesini sağlayan uygulamalardır. Bu çalışmalar, görme engelliler için geliştirilebilecek uygulamalarda, interaktif sesli yanıt sistemlerinde ve özellikle de sanal asistan sistemlerinde sıklıkla kullanılmaktadır.

Derin öğrenme yöntemlerinden önceki konuşma sentezleme çalışmalarında, fonların kural tabanlı [42] veya istatistiksel [43] olarak birbirine eklenmesi ile konuşma sentezleme işlemi yapılabilmekteydi. Bu durumda üretilen sesler, insan sesi doğallığından oldukça uzak olmaktaydı.

2016 yılında WaveNet [44] çalışması ile konuşma sentezleme sistemlerinin derin öğrenme yaklaşımları kullanılarak başarılı bir şekilde modellenmesi sağlandı. Konuşma sentezleme sistemlerinde derin öğrenme çağının başlaması ile beraber, yeterli miktarda veri sağlandığı takdirde, neredeyse insan sesi ile ayırt edilemeyecek doğallıkta konuşma sentezleme çıktıları alınmaya başlandı. Bu bağlamda tasarlanacak konuşma sentezleme mimarilerinde, üretilen seslerin insan sesi doğallığında olması ve bu seslerin gerçek zamanda üretilmesi temel tasarım kriterleri olarak sayılmaya başlanmıştır.

Bu noktada, konuşmanın doğallığının değerlendirilmesi, ortalama görüş puan (Mean Opinion Score - MOS) olarak adlandırılan ve konuşma sentezleme modeli tarafından üretilen seslerin bir heyet tarafından dinlenerek doğallığının puanlandırılması yöntemi ile gerçekleştirilmektedir.

Sistemin gerçek zamanda çalışabilirlik durumu ise gerçek zaman faktörü (Real Time Factor - RTF) hesabı ile yapılmaktadır. Gerçek zaman faktörü, üretilen ses sinyalinin birim süresi için sistem tarafından ne kadar zaman harcandığının hesaplanması işlemidir.

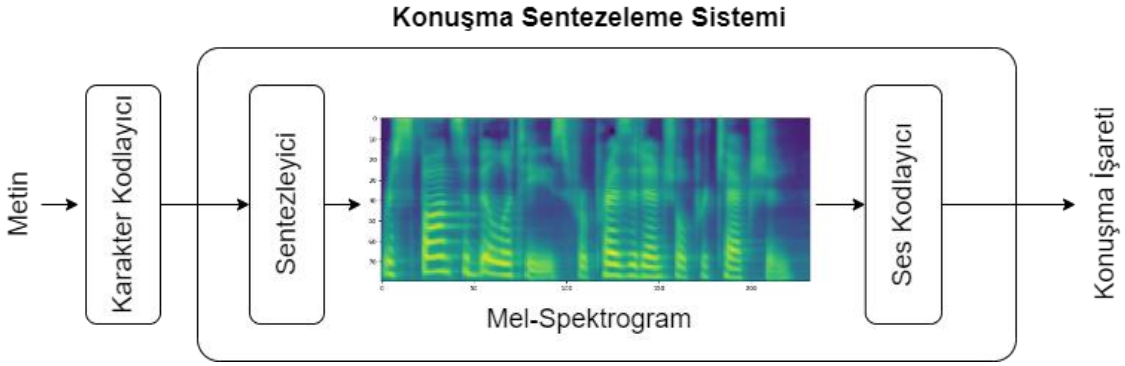
4.2 Konuşma Sentezleme Sistemlerindeki Zorluklar

Konuşma sentezleme sistemlerinin tasarlanması adımlarında karşılaşılan temel problemler vardır. Bu problemler ve ilgili probleme ait çözümler aşağıdaki başlıklarda incelenecektir.

4.2.1 Konuşma İşaretlerinin Doğrusal Olmaması Problemi

Konuş sentezleme çalışmalarında metin gibi oldukça sıkıştırılmış, ilkel bir veri tipinden, konuşma işareti gibi kompleks ve doğrusal olmayan bir dizi elde edilmeye çalışılmaktadır [45]. Bu problemin çözülmesi için genellikle iki farklı modelin bir arada kullanılması önerilmektedir. Sentezleyici model olarak isimlendirilen ilk model, metin girişine karşı ses işaretlerinin orta seviyeli gösterimleri olan mel-spektrogramları üretmek için kullanılmaktadır. Mel-spektrogramlar ses işaretine nazaran daha az kompleks, fakat daha çok kayıplıdırlar. Fakat derin öğrenme modellerinde eğitimin daha kolay şekilde yakınsamasını sağlamaktadırlar.

Ses kodlayıcı olarak adlandırılan diğer model ise mel-spektrogramları kullanarak zaman alanındaki ses işaretlerin üretilmesini sağlamaktadır. Bu yöntem izlenerek tasarlanan konuşma sentezleme sisteminin genel yapısı Şekil 4.1'de verilmiştir.



Şekil 4.1 Konuşma sentezleme sistemi.

Bilgisayarlar bilindiği üzere sayılar üzerinde işlem yapabilen araçlardır. Bu yüzden, konuşma sentezleme sistemlerinin girişlerindeki metinlerin sayısal olarak ifade edilmesi gerekmektedir. Şekil 4.1'de de gösterildiği üzere, sistemin girişine verilecek olan metin, öncelikle karakter veya fonem kodlayıcı ile sayı dizisine

dönüştürülmektedir. Bu işlem, basitçe metin içerisindeki her bir karakterin, alfabedeki sıra numarası ile temsil edilmesi şeklinde olabilmektedir.

Şekil 4.1'de görsellenen sistem sonucunda elde edilecek vektör boyutları aşağıdaki gibi olacaktır.

- karakterKodlayıcı(metin) →
  kodlanmışKarakterler = [boyut=(karakterSayısı)]
- sentezleyici(kodlanmışKarakterler) →
  melSpektrogram = [boyut=(melFiltreSayısı, süre)]
- sesKodlayıcı(melSpektrogram) →
  konuşmaİşareti = [boyut =(örneklemeFrekansı*süre)]

"Merhaba dünya" kelime dizisine karşılık gelecek bir konuşma kaydının 1 saniye, örnekleme frekansının ise 16 kHz olacağı varsayılsın. Konuşma sentezleme sisteminde yalnızca bir model kullanılmış olsa, sistemin girişine verilecek olan 13 uzunluğundaki bir vektör ile 16000 uzunluğundaki bir vektör elde edilmeye çalışılacaktır. Bu durum, eğitilecek olan modelin yakınsamasını oldukça zorlaştıracaktır. Sentezleyici ve ses kodlayıcı modellerin bir arada kullanılması öğrenme problemini kolaylaştırmaktadır.

4.2.2 Veri Oluşturma Problemi

İnsan konuşması doğallığına yakın konuşma sentezleme sistemlerinin geliştirilebilmesi için profesyonel şekilde oluşturulmuş verilere ihtiyaç duyulmaktadır. Veri seti temel olarak bir konuşma kaydı ve konuşma kaydına ait metinden oluşmaktadır. Konuşmalar mümkün olduğunca temiz bir şekilde kayıt edilmelidir. Aksi takdirde, geliştirilecek konuşma sentezleme sisteminin doğallığı oldukça düşecektir. Bunun için kayıtların stüdyo ortamında yapılması faydalı olacaktır. Bunun dışında, sesini profesyonel bir şekilde kontrol edebilen konuşma sanatçıları tarafından ses kayıtlarının verilmesi birden-çoğa eşleme probleminin olumsuz etkilerinin azaltılmasına yardımcı olacaktır. Tüm bu durumlar incelendiği zaman, veri seti oluşturulması zaman ve finansal açıdan oldukça maliyetli olabilmektedir.

4.2.3 Birden-Çoğa Eşleme Problemi

Konuşma sentezleme modellerinin eğitilebilmesi için metin ve bu metinlere karşılık gelen konuşma kayıtlarına ihtiyaç duyulmaktadır. Bu veriler içerisinde sıklıkla tekrar eden kelime ve kelime grupları ile karşılaşmak çok muhtemeldir. Bu durum, birden-çoğa eşleme problemi olarak adlandırılmaktadır ve konuşma sentezleme sistemleri ile elde edilecek konuşma sinyallerinin doğallığını olumsuz etkilemektedir.

Bu duruma çözüm olarak özbağlanımlı ağların [44], [46], [47] kullanımı önerilebilir fakat, bu ağların eğitim ve çıkarım süreleri özbağlanımlı olmayan ağlara göre daha uzun olmaktadır.

Özbağlanımlı olmayan ağlarda ise giriş ve çıkış arasındaki bilgi boşluğunu azaltacak, sesin perde ve enerji bilgilerinin doğrudan sentezleyici ağda kullanılması, birden-çoğa eşleme probleminin olumsuz etkilerini azaltmaktadır [48].

4.2.4 Gerçek Zamanda Çalışma Problemi

Konuşma sentezleme uygulamalarında özbağlanımlı ağlar ile üretilen sesler, insan sesi doğallığına oldukça yakındır. Fakat özbağlanımlı ağların doğası gereği her “t-1” anda üretilen çıkış “t” anında giriş olarak kullanılmaktadır. Bu durum, özbağlanımlı ağların paralel olarak çalışmasını kısıtlamaktadır. Dolayısıyla eğitim ve sentezleme anında bu ağların yavaş çalışmasına neden olmaktadır. Sanal asistanlar gibi gerçek zamanda çalışması beklenen sistemlerde genellikle özbağlanımlı olmayan ağların tercih edilmesi daha uygun olmaktadır. FastSpeech [49], FastSpeech2 [48], Flow-TTS [50] gibi özbağlanımlı olmayan ağlar, özbağlanımlı ağlarla karşılaştırılabilir düzeyde insan sesine yakın çıktı verebilmektedir. Tablo 4.1’de konuşma sentezleme modellerinin MOS ve RTF değerlerinin karşılaştırılması gösterilmektedir.

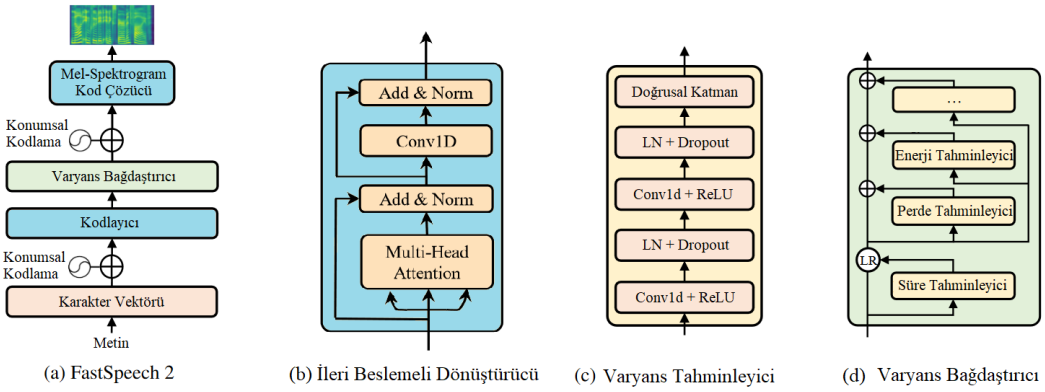
Tablo 4.1 Konuşma sentezleme modellerinin karşılaştırılması[51].

Model	**MOS**	**RTF**
Test Verisi	4.56 ± 0,09	-
HifiGAN(melSpektrogram(Test Verisi))	4.47 ± 0,10	-
Tacotron2	4.03 ± 0,12	$1{,}35 \times 10^{-1}$
FastSpeech2	3.83 ± 0,14	**$4{,}21 \times 10^{-3}$**
Glow-TTS	3.62 ± 0,13	$9{,}39 \times 10^{-3}$
BVAE-TTS	3.16 ± 0,13	$4{,}21 \times 10^{-3}$
VAENAR-TTS	**4.15 ± 0,12**	$7{,}45 \times 10^{-3}$

MOS değerinin yüksek olması, sistemin daha insan konuşması doğallığında ses üretilmesi anlamına gelirken, RTF değerinin düşük olması sistemin çok daha hızlı çalıştığının göstergesidir.

4.3 Sentezleyici Model Mimarisi

Sanal asistan uygulamalarında, diyalog sistemi tarafından üretilecek metin çıktısının, sesli olarak kullanıcıya döndürülmesi işleminin gerçek zamanda yapılması gerekmektedir. Tablo 4.1'de gösterildiği üzere FastSpeech2 [48] mimarisinin MOS değerinin kabul edilebilir seviyede olması ve RTF faktörünün düşük olması, bu çalışmada FastSpeech2 mimarisinin tercih edilmesinde önemli rol oynamıştır.



Şekil 4.2 Fastspeech2 mimarisi [48], [49].

Şekil 4.2 (a)'da FastSpeech2 mimarisi uçtan uca gösterilmekte; (b), (c) ve (d) de ise FastSpeech2 mimarisi içerisinde kullanılan diğer bloklar verilmektedir.

Metin girişinden mel-spektrogramın elde edilmesine kadar olan süreç aşağıdaki alt başlıklarda incelenecektir.

4.3.1 Karakter Vektörü

Türkçe için metinler birer karakter dizisi olarak ifade edilebilmektedir. İngilizce gibi fonetik olmayan dillerde, karakter dizisi yerine fonemlerin kullanılması tercih edilebilmektedir. Her iki durumda da metinlerin ilk olarak vektörel olarak ifade edilmesi gerekmektedir. Bunun için her bir fonemin alfabedeki veya fonetik sözlük içerisindeki fonem sıra numaraları kullanılabilmektedir.

4.3.2 Konumsal Kodlama

FastSpeech2 mimarisi yinelemeli veya evrişimli olmadığından, karakterlerin sırasının model içerisinde kullanılabilmesi için karakter vektörüne konumsal kodlama uygulanarak, konum bilgisinin karakter vektörüne enjekte edilmesi gerekmektedir.

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}) \tag{4.1}$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}}) \tag{4.2}$$

Denklem 4.1 ve 4.2 kullanılarak, karakter vektörüne konum enjekte etme işlemi yapılabilmektedir. Burada pos konumu, i boyutu ve d_{model} ise vektör boyutunu

temsil etmektedir. Böylelikle, konumsal kodlama ile karakter vektörünün her boyutu bir sinüzoide karşılık gelmektedir [52].

4.3.3 Kodlayıcı

Kodlayıcı, girişteki fonem dizisini, gizli fonem dizisi vektörüne dönüştürmektedir. Nadet özdeş katman yığınından oluşan kodlayıcı, Şekil 4.2 (b)'de gösterildiği gibi, her katmanında çok kafalı kendi kendine dikkat mekanizması ve ileri beslemeli yapılardan oluşmaktadır [52]. Her iki yapının da çıkışında katman normalizasyonu yapılmaktadır [53]. Kodlayıcının çıkışında, girişte konumsal olarak kodlandırılmış vektör elemanlarının, birbirleri ile ilişkilendirilmiş ağırlıkları elde edilmektedir.

4.3.4 Varyans Tahminleyici

Şekil 4.2 (c)'de gösterildiği üzere, varyans tahminleyici bir dizi evrişimsel sinir ağı ile beraber doğrusal katmanlardan oluşmaktadır. Bu blokta, girişteki metne bağlı olarak; sesin uzunluğu, perde frekansı ve enerjisi tahmin edilmeye çalışılmaktadır. Bu üç tahminleme için de aynı ağ yapısı kullanılmaktadır.

4.3.4.1 Süre Tahminleyici

Süre tahminleyici, girişinde fonemlere ait gizli katman dizilerini alarak, çıkışında her bir fonem için süre tahmin etmektedir. Eğitim aşamasında önce veri hazırlanırken, veri setinde bulunan her bir konuşma parçacığı (yazı ve ses çifti) için hizalama işlemi yapılarak, fonemler için süre bilgisinin çıkarılması gerekmektedir. Şekil 2.6'da bu hizalama işlemi gösterilmiş olup, fonemlere ait sınırlar çizdirilmiştir. Süre tahminleyicinin eğitiminde Denklem 4.3'te gösterilen ortalama kare hata (Mean Squared Error -MSE) fonksiyonu kullanılarak optimizasyon yapılmaktadır [48].

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - x_i)^2 \tag{4.3}$$

4.3.4.2 Perde Frekansı Tahminleyici

Perde frekansı, konuşma içerisindeki duyguların taşınmasında önemli rol oynamaktadır. Ren ve diğerleri [48], perde frekansı konturlarındaki çeşitliliklerin

daha iyi tahmin edilebilmesi için CWT yöntemini kullanarak perde frekansı serisinin perde spektrogramına ayrıştırılmasını ve bunun perde frekansı tahminleyicinin eğitilmesinde kullanılacak MSE fonksiyonunda hedef olarak kullanmayı önermişlerdir.

4.3.4.3 Enerji Tahminleyici

Enerji doğrudan elde edilecek sesteki genliği kontrol etmektedir. Enerji tahminleyicide hedef enerji olarak, her bir STFT çerçevesinin genliğinin L2 normu kullanılmakta ve enerji tahminleyici, MSE fonksiyonu ile optimize edilmektedir [48].

4.3.5 Varyans Bağdaştırıcı

Varyans bağdaştırıcı, sentezleyicinin girişi ve çıkışı arasındaki bilgi boşluğunun azaltılmasında varyans tahminleyicilerin sağladığı bilgileri kullanmaktadır. Kısaca sentezleyici, sadece metin gibi düşük boyutlu bir vektörle değil, aynı zamanda giriş metni üzerinden tahmin edilecek süre, perde frekansı ve enerji gibi vektörlerle de beslenmektedir. Bu çözüm sayesinde özbağlanımlı olmayan bu ağ mimarisi, özbağlanımlı ağlarla kıyaslanabilir doğallıkta ses çıktısı verebilmektedir.

4.3.6 Mel-Spektrogram Kod Çözücü

Mel-spektrogram kod çözücü, ileri beslemeli dönüştürücü blok ve devamında doğrusal katmandan oluşmaktadır. Varyans bağdaştırıcıdan alınan çıkışların tekrar konumsal kodlama fonksiyonundan geçirilmesi sonucunda elde edilen çıktıları kabul etmektedir. Çıkışta ise mel-spektrogram üretmektedir. Denklem 4.4'te gösterilen ortalama mutlak hata (Mean Absolute Error - MAE) fonksiyonu kullanarak modelin optimizasyonu sağlanmaktadır.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n} \tag{4.4}$$

4.4 Ses Kodlayıcı Model Mimarisi

Sentezleyici modelden elde edilen mel-spektrogramların zaman alanındaki ses işaretlerine dönüştürülerek dinlenebilir hale getirilmesi için ses kodlayıcılardan faydalanılmaktadır. Ses kodlayıcılar da sentezleyici modellerde olduğu gibi özbağlanımlı [44], [54], [55] ve özbağlanımsız olacak şekilde tasarlanabilmektedir. Özbağlanımlı olmayan modeller, yüksek oranda paralelleştirilebildiği için gerçek zamanda çalışacak sistemlerde daha çok tercih edilmektedir. Ses kodlayıcılar son zamanlarda sıklıkla çekişmeli üretici ağlarla (Generative Adversarial Networks - GAN) [56] tasarlanmaktadır.

Çekişmeli Üretici Ağlar, üretken ve ayrıştırıcı iki ağın birbiri ile çekişmesi prensibine dayanan bir öğrenme yöntemi kullanmaktadır. Üretici ağ, gerçeğe yakın örnekler üretmeye çalışırken, ayrıştırıcı ağ ise bu örneklerin gerçek mi yoksa sahte mi olduğunu anlamaya çalışmaktadır. Bu iki ağın çekişmesi sürecinde ayrıştırıcı ağ, gerçek ve sahte örneklerin ayrımını daha iyi yapmakta, üretici ağ ise daha gerçekçi örnekler üretmektedir.

GAN'lardaki öğrenme prensibi, özellikle bilgisayarlı görü alanında başarılı çalışmaların yapılmasını sağlamıştır. Resim üretimi konusundaki başarıları bilinen GAN'lar, resimden-resime veya videodan-videoya dönüşüm amacı ile de kullanılmaktadır. Bilgisayarlı görüdeki başarılarının ardından GAN'lar, benzer yaklaşımlarla konuşma sentezleme alanında da çalışma konusu olmuştur. Öyle ki, bu alanda da oldukça başarılı sonuçlar elde edilmiştir.

Neekhara ve diğerleri [57] GAN'ları kullanarak, Mel-spektrogramlarını genlik spektrogramlarına eşlemeyi, ardında da faz tahmini yaparak ham ses dalga formunu elde etmeyi önermişlerdir. Daha sonra Kumar ve diğerlerinin çalışmalarıyla MelGAN [58] mimarisi önerilmiş, mel-spektrogramlar kullanılarak doğrudan ham ses dalga formları elde edilmiştir. MelGAN ile GPU (GTX 1080Ti) üzerinde gerçek zamanın 100 katı hızında, CPU üzerinde ise 2 katı hızda sentezleme performansı yakalanmıştır [58]. Ardından Multi-band MelGAN [59] mimarisinde, MelGAN'ın zayıf yönlerine odaklanılmış, MOS ve RTF skorları daha da iyileştirilmiştir. Tablo 4.2'de MelGAN ve Multi-band MelGAN ağlarının MOS ve RTF karşılaştırmaları gösterilmiştir.

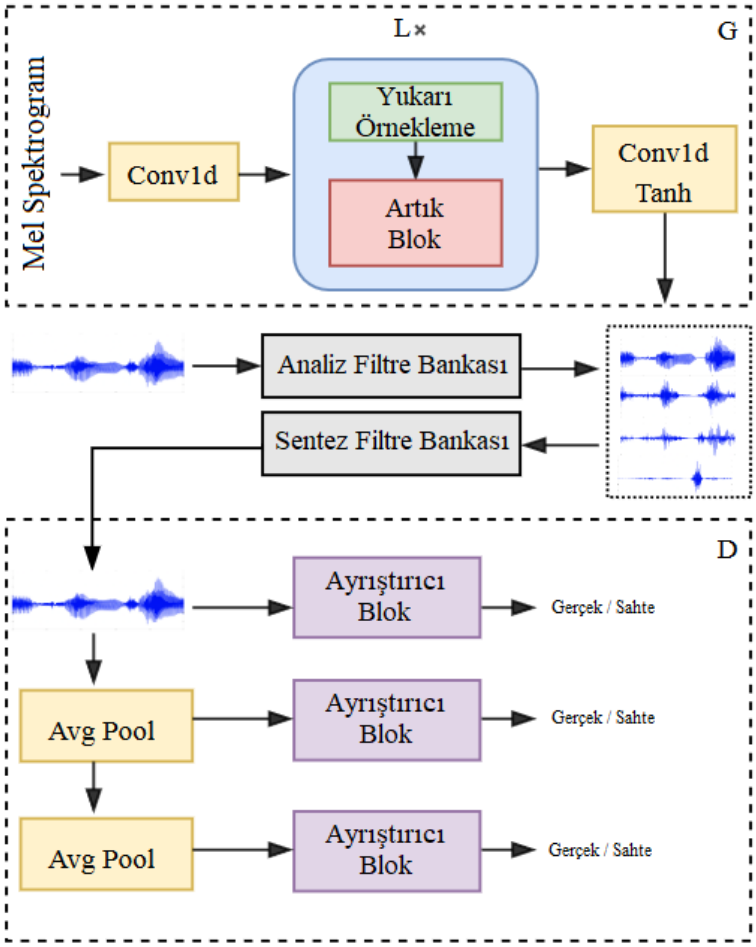
Tablo 4.2 MelGAN ve Multi-band MelGAN performans karşılaştırması [59].

Sentezleyici Model	**Ses Kodlayıcı Model**	**RTF**	**MOS**
Tacotron2	MelGAN	0,2	3,87
	Full-band MelGAN	0,22	4,18
	Multi-band MelGAN	**0,03**	**4,22**
Referans			4,58

MelGAN mimarisinin başarımını arttırmak için Multi-band MelGAN aşağıdaki yöntemleri kullanmaktadır [59].

- Üretken ağdaki algılayıcı alan sayısı arttırılmıştır.
- Gerçek ve sahte seslerin ayrımını daha iyi yapabilen bir kayıp fonksiyonu kullanılmıştır.
- Üretken ağın girişine verilen mel-spektrogramlar ile bir yerine dört alt bant için üretim yapılmış, üretilen her sinyal ayrıştırıcı ağa verilmeden önce tek bir sinyal haline getirilmiştir.

Temel MelGAN mimarisinin üretici ağı, mel dizisi ile dalga form frekanslarının eşleşebilmesi için yukarı örneklemenin yapılmasını sağlayacak transpoze edilmiş evrişimli ağ bloğu yığını kullanmaktadır. Şekil 4.3'te gösterildiği gibi, yukarı örnekleme bloğunu, artık blok takip etmektedir. Bu blok algılayıcı alan sayısını arttırabilmek için genişletilmiş evrişimli ağdan oluşmaktadır.



Şekil 4.3 Multi-band Melgan mimarisi [59].

Ayrıştırıcı ağda, birden fazla ayrıştırıcı bloğun kullanılması üretilen sesteki metalikliği azaltmak için önemli bir görev yapmaktadır. Buradaki her bir ayrıştırıcı blok, sesin farklı frekans aralıklarındaki özellikleri öğrenmek niyetindedir. Temel MelGAN'ın kullanmış olduğu amaç fonksiyonları Denklem 4.5 ve 4.6'da gösterilmektedir.

$$\min_{D_k} \mathbb{E}_x[(D_k(x)-1)^2] + \mathbb{E}_{s,z}[D_k(G(s,z))^2] \tag{4.5}$$

$$\min_{G} \mathbb{E}_{s,z}\left[\sum_{k=1}^{K} D_k(G(s,z)-1)^2\right] \tag{4.6}$$

Denklemlerde; D_k, k numaralı ayrıştırıcıyı, x ham dalga formunu, s giriş mel-spektrogramını, z ise Gauss gürültü vektörünü temsil etmektedir.

Multi-band MelGAN mimarisi, MelGAN mimarisindeki özellik eşleştirme kayıp fonksiyonu yerine çok çözünürlüklü STFT kayıp fonksiyonunu önermektedir. Bu güncelleme ile gerçek-sahte örnekler arasındaki ayrım daha iyi yapılmakta ve aynı zamanda eğitim hızlanmaktadır [59].

Tek çözünürlüklü STFT kayıp fonksiyonu ile hedef dalga formu ve üretici ağ tarafından tahmin edilen dalga formu arasındaki L_{sc} (Denklem 4.7) ve L_{mag} (Denklem 4.8) kayıpları minimize edilmektedir.

$$L_{sc}(x,\tilde{x}) = \frac{\| \, |STFT(x)| - |STFT(\tilde{x})| \, \|_F}{\| \, |STFT(x)| \, \|_F} \tag{4.7}$$

$$L_{mag}(x,\tilde{x}) = \frac{1}{N} \| \log|STFT(x)| - \log|STFT(\tilde{x})| \, \|_1 \tag{4.8}$$

$\| \, . \, \|_F$ ve $\| \, . \, \|_1$ sırayla Frobenius ve L_1 normları göstermektedir. $|STFT(\,.\,)|$, STFT ile hesaplanan büyüklüğü ve N ise büyüklük içerisindeki eleman sayısını temsil etmektedir.

Çok çözünürlüklü STFT amaç fonksiyonu, Denklem 4.9'da gösterildiği gibi, farklı analiz parametreleri (FFT boyutu, pencere boyutu, atlama boyutu) ile hesaplanmış M adet tek çözünürlüklü STFT kayıplarının ortalaması ile bulunmaktadır.

$$L_{\text{mr\_stft}}(G) = \mathbb{E}_{x,\tilde{x}} \left[\frac{1}{M} \sum_{m=1}^{M} \left(L_{sc}^{m}(x,\tilde{x}) + L_{mag}^{m}(x,\tilde{x}) \right) \right] \tag{4.9}$$

Üretici ağın eğitimi için Denklem 4.10'da gösterilen amaç fonksiyonu kullanılmaktadır.

$$\min_{G} \mathbb{E}_{s,z} \left[\lambda \sum_{k=1}^{K} (D_k(G(s,z)) - 1)^2 \right] + \mathbb{E}_s \left[L_{mr\_stft}(G) \right] \tag{4.10}$$

Tam bant ölçeği için de kayıp fonksiyonu tekrar düzenlendiğinde Denklem 4.11 elde edilmektedir.

$$L_{mr\_stft}(G) = \frac{1}{2}\left(L^{full}_{fmr\_stft}(G) + L^{sub}_{smr\_stft}(G)\right) \quad (4.11)$$

Denklem 4.11'de $L^{full}_{fmr\_stft}$, $L^{sub}_{smr\_stft}$ sırayla tam bant ve alt bantlardaki çok çözünürlüklü STFT kayıplarına karşılık gelmektedir.

Özetle; Multi-band MelGAN, mel-spektrogram girdileri ile paralel olarak alt bant sinyallerinin üretilmesini sağlayan üretici ağa sahiptir. Çok çözünürlüklü STFT kayıp fonksiyonunun hesaplanabilmesi için analiz filtreleri kullanılmaktadır. Ardından alt bant sinyalleri ile tam bant sinyalinin oluşturulması için sentez filtresi kullanılmakta ve tam bant sinyali için çok çözünürlüklü STFT kaybı hesaplanmaktadır.

Eğitim;

1. Üretici ağ ve ayrıştırıcı ağların parametrelerinin başlatılması,
2. Üretici ağın Denklem 4.11 kullanılarak yakınsayana kadar eğitilmesi,
3. Ayrıştırıcı ağın Denklem 4.5 kullanılarak eğitilmesi,
4. Üretici ağın Denklem 4.10 kullanılarak eğitilmesi,
5. Üretici ve ayrıştırıcı ağ yakınsayana kadar 3. ve 4. adımların tekrar edilmesi.

adımlarını takip etmektedir [59].

Eğitim tamamlandıktan sonra mel-spektrogram girişleri ile zaman alanındaki sinyallerin üretilmesi için sadece üretici ağ kullanılmaktadır.

4.5 Konuşma Sentezleme Sistemlerinin Başarımlarının ve Performansının Değerlendirilmesi

Konuşma sentezleme sistemlerinin birbirleri ile karşılaştırılabilmesi için iki önemli değerlendirme kullanılmaktadır. Alt başlıklarda da inceleneceği üzere bunlar; sistemin verdiği çıktının insan konuşması doğallığının saptanması ve gerçek zamanlı sistemlerde çalışmaya ne kadar uygun olduğunun belirlenmesine yöneliktir.

4.5.1 Konuşma Doğallığının Değerlendirilmesi

Konuşma sentezleme sistemlerinde, üretilen çıktıların insan sesine yakınlığının ölçülebilmesi için MOS metriği kullanılmaktadır. MOS, insan yargısına dayalı bir metriktir. Genellikle bağımsız dinleyici bir gruba, bir takım ses kayıtlarının dinletilmesi ve bu kayıtların insan sesine yakınlığının 1 ila 5 (0,5 puan aralıklar ile) arasında oylanması istenmektedir. 5 en yüksek skor olup, ses kaydının insan sesi ile ayırt edilemez olduğunu göstermektedir.

Değerlendirme heyeti genellikle;

1. Sistemin eğitildiği rerferans ses kayıtlarını,
2. Sistemin eğitildiği ses kayıtlarının mel-spektrogramlarının elde edilmesi ve ardından ses kodlayıcıdan geçirilerek elde edilen ses kayıtlarını,
3. Konuşma sentezleme sistemi tarafından elde edilen ses kayıtlarını,
4. Karşılaştırılacak konuşma sentezleme sistemi tarafından elde edilen ses kayıtlarını,

puanlandırmaktadır. Bu puanlama yapılırken, ses kayıtlarının elde edildiği yöntem değerlendirme heyetinden gizli tutulmaktadır.

Bu değerlendirme yöntemi, insan yargısına ve yoruma dayalı olduğu için sistemlerin birbiri ile karşılaştırılabilmesini zorlu kılmaktadır.

4.5.2 Sistemin Hızının Değerlendirilmesi

Konuşma sentezleme sistemlerinde sistemin gerçek zamanda çalışıp çalışmadığının hesaplanabilmesi için gerçek zaman faktörü hesaplama işlemi yapılmaktadır. Gerçek zaman faktörü, 1 saniye uzunluğundaki bir ses dosyasının üretilmesinin ne kadar zaman aldığının hesaplanması işlemidir. Eğer sistem 1 saniyelik bir ses işaretini, 1 saniye içerisinde üretebiliyorsa gerçek zamanda çalışmaktadır.

5
DOĞAL DİL İŞLEME

5.1 Giriş

Akıllı asistan veya arkasında yatan sohbet botu teknolojisi, yapılandırılmamış metinlerin doğal dil işleme yöntemleri ile anlamlı hale getirilmesi ve ardından gerekli sorgulamalar doğrultusunda, kullanıcılara uygun cevapların döndürülmesi işlemlerini temel alan uygulamalardır. Bu sistemler, kullanım amacına göre açık ve kapalı alan olmak üzere ikiye ayrılmaktadır.

Açık alan sohbet botları gelişmiş sohbet kabiliyetine sahiptir. Bu tür botlar özellikle karşılıklı konuşmalar için tasarlanmıştır ve kullanıcı girdilerine karşı her daim mantıklı cevap üretmeleri beklenmektedir. Bu tür botlar; şov, psikoterapi, öneri sistemleri ve eğitimde kullanılabilmektedir. Google'ın geliştirdiği Meena [60] isimli sohbet botu bu sistemlerin en iyi örneklerindendir.

Öte yandan, kapalı alan veya alana özel sohbet botları, sadece belirli komut kümelerine odaklanmakta ve kısıtlı şekilde cevap üretebilmektedirler. Örneğin, bir yemek şirketi için tasarlanmış sohbet botu sadece sipariş almak ve siparişin takibini yapmaktan sorumludur. Kısaca sadece kendi alanı hakkında bilgi sahibidir. Bu tür botlar genellikle belirli birtakım işlerin yapılmasından sorumludurlar. Genellikle sanal asistanlar bu yapıda kurgulanmaktadır. Hava durumu, adres veya güzergâh sorgulama, not alma, alarm kurma gibi işlemlerin yapılmasında görev alabilmektedirler.

Sanal asistan veya benzeri sohbet ajanlarının, kullanıcı girişlerini anlamlandırabilmesi için iki konsept oldukça önem arz etmektedir. Bunlar niyet sınıflandırma ve isimlendirilmiş varlık tanımadır. Fakat her şeyden önce, doğal dil işleme çalışması yapabilmek için kelimeleri veya kelime altı parçacıkları, makinelerin anlayacağı şekilde sayılarla ifade etmek gerekmektedir.

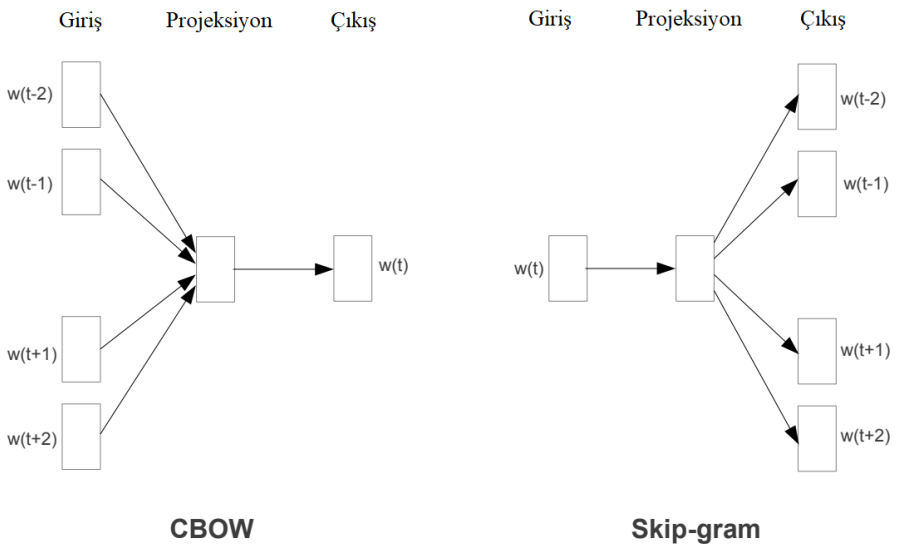
5.2 Dilin Temsil Edilmesi

Doğal dilin en önemli unsurlarından olan metinlerin, bilgisayarlar tarafından işlenebilmesi için öncelikle sayısal ifadelere dönüştürülmeleri gerekmektedir. Bu durumda, kelimelerin doğrudan vektörel olarak ifade edilmesini sağlayacak yaklaşımlar olduğu gibi, son zamanlarda dilin daha kapsamlı şekilde modellenmesini sağlayan BERT [61] benzeri yaklaşımlar da bulunmaktadır. BERT, dili oldukça iyi ifade edebilmektedir ve pek çok doğal dil işleme probleminin çözümü için alt yapı oluşturmaktadır. Fakat BERT konseptine geçilmeden önce kelime vektörlerinin incelenmesi faydalı olacaktır.

5.2.1 Kelime Vektörleri

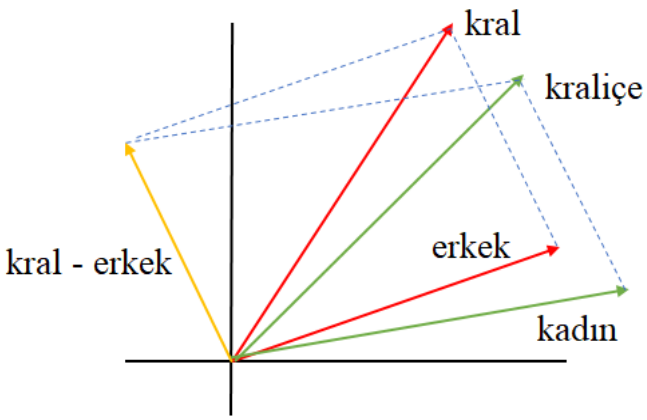
Kelimelerin bilgisayarlar tarafından anlaşılabilecek biçimde sayısal olarak temsil edilmesi ile kelime vektörleri oluşturulmaktadır. Geçmiş yaklaşımlarda kelimeleri temsil edebilmek için *one-hot* vektörlerinden yararlanılmaktaydı. Fakat bu vektörler, kelimeler arasında anlamsal bir ilişki kuramadığı için bazı doğal dil işleme görevleri için yetersiz kalmaktaydı.

Ardından CBOW ve Skip-gram modelleri ile kelimeler, bir kelime uzayında aralarındaki anlamsal ilişki kaybedilmeden vektörize edilebilmiştir [62]. Şekil 5.1'de CBOW ve Skip-gram ağlarının mimarisi gösterilmektedir. CBOW bir kelimeyi komşularına göre tahmin etmeye çalışırken, Skip-gram bir kelimenin komşularını tahmin etmeye çalışmaktadır. Her iki durumda da giriş olarak, bir kelime listesi içerisindeki kelimelerin *one-hot* vektörleri kullanılmaktadır. Çıkış katmanında *softmax* aktivasyon fonksiyonu kullanılmaktadır ve hedef kelime (CBOW) veya kelimeler (Skip-gram) için kelime listesi boyutunda olasılıksal bir dağılım elde edilmektedir. Bu dağılım, kelimelerin vektör karşılıklarını oluşturmaktadır.



Şekil 5.1 CBOW ve Skip-gram modeller için ağ mimarisi [62].

Kelimelerin anlamsal yakınlıklarının kelime uzayında temsil edilebilmesi, kelimeler arasına vektörel hesaplamaların yapılmasına olanak sağlamıştır. Şekil 5.2'de bazı kelimeler ve bu kelimeler arasındaki vektörel uzaklıklar gösterilmektedir.



Şekil 5.2 Kral, kraliçe, kadın, erkek kelime vektörleri üzerinde yapılan vektörel işlemler [63].

Yeterli büyüklükteki bir derlem üzerinde eğitilmiş CBOW ve Skip-Gram modelleri ile kelimeler üzerinde yapılan basit vektörel toplama ve çıkarma işlemleri sonucunda, kelime vektörlerinin, kelimeler arasındaki anlamsal ilişkiyi tutabildiği

gösterilmiştir. Aşağıda kelimeler arasındaki anlamsal ilişkileri gösteren örnekler yer almaktadır.

- $X =$ vektör ("kral") $-$ vektör ("erkek") + vektör ("kadın")
    - $X =$ "kraliçe"

Kral kelimesini temsil eden vektörden, erkek kelimesini temsil eden vektörün çıkarılması sonucunda “kraliyet” kelimesine benzer bir vektör oluşmaktadır. Bu vektöre, kadın kelimesinin vektörünün eklenmesi sonucunda “kraliçe” kelimesinin vektörü elde edilebilmektedir.

- $X =$ vektör ("Almanya") $-$ vektör ("Berlin") + vektör ("Türkiye")
    - $X =$ "Ankara"

Benzer şekilde, Almanya vektöründen Berlin vektörü çıkartıldığında ortaya çıkan yeni vektör, “başkent” bilgisini taşımaktadır. Bu vektöre, Türkiye kelimesinin vektörü eklendiği zaman “Ankara” kelimesine karşılık gelecek vektör elde edilebilmektedir.

5.2.2 BERT

BERT, Google yapay zekâ ekibi tarafından tasarlanmış, metin formundaki doğal dilin bilgisayarlar için anlamlı hale getirilmesini sağlayan bir dil temsilidir. Bunun yapılmasında *Transformer* [52] mimarisinin kodlayıcı bloğunun yığın şeklinde kullanılması önerilmektedir.

BERT, ön eğitim ve ince-ayar evrelerinden oluşmaktadır. Ön eğitim; BERT’in dilin yapısını anlaması ve metin bağlamı üzerine çıkarım yapabilmesini sağlamaktadır. İnce ayar ise ön eğitim aşamasından geçmiş BERT modelinin, etiketlenmiş veriler kullanılarak, belirli bir doğal dil işleme probleminin çözülmesi için tekrar eğitilmesi işlemidir.

BERT, dili temsil ederken maskelenmiş dil modeli ve sıradaki cümle tahmini yöntemlerini kullanarak, dilin vektör uzayında iyi bir biçimde kodlanmasını sağlamaktadır. Böylelikle doğal dil işleme üzerine yapılacak pek çok çalışmada kodlayıcı olarak önceden eğitilmiş bir BERT modelinin, kod çözücü olarak da sadece basit bir doğrusal katmanın kullanılması ile dahi çok başarılı sonuçlar elde edilebilmektedir.

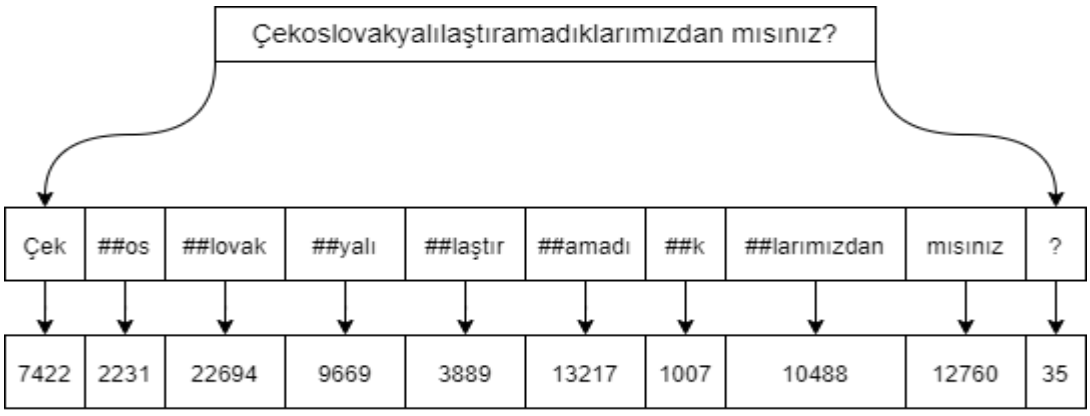
BERT doğrudan kelimeleri kullanmak yerine, kelimelerin simgeleştirilmesi ile elde edilen kelime altı parçacıkları giriş olarak kabul etmektedir. Böylece kelime dağarcığı problemi ile karşılaşılmadan bütün kelimeler ifade edilebilmektedir.

5.2.2.1 Ön Eğitim

BERT'in ön eğitimi; simgeleştirme, maskelenmiş dil modeli ve sıradaki cümle tahmini aşamalarından oluşmaktadır.

Simgeleştirme

CBOW ve Skip-gram yöntemleri ile oluşturulan kelime vektörleri, İngilizce gibi kelime sayısı çok fazla olmayan dillerin temsilinde başarılıdır. Ancak, Japonca ve Korece gibi oldukça fazla karakter envanterine sahip Asya dillerinin, bu modeller kullanılarak temsil edilmesi durumunda, dağarcık dışında kalan çok fazla kelime olabilmektedir. Aynı durum sondan eklemeli bir dil olan Türkçe için de geçerlidir. Bu durumda, kelimeleri olduğu gibi kullanmak yerine kelime altı parçacıkların kullanılması, dil içindeki bütün kelimelerin temsil edilmesini sağlayacaktır. Bu bağlamda, BERT modellerinin de kullandığı kelime parçacığı (Word Piece*)* [64] yöntemi ile oluşturulan kelime altı parçalar, dağarcık dışında herhangi bir kelime kalmadan bütün sözcükleri ifade edilebilmektedir. Şekil 5.3'te "Çekoslovakyalılaştıramadıklarımızdan mısınız?" cümlesinin kelime parçacığı yöntemi kullanılarak elde edilen kelime altı parçacıkları ve bu parçacıkların kelime listesi içerisindeki sıra numaraları gösterilmektedir.



Şekil 5.3 Kelime parçacığı yöntemi kullanılarak elde edilen kelime altı parçacıklar.

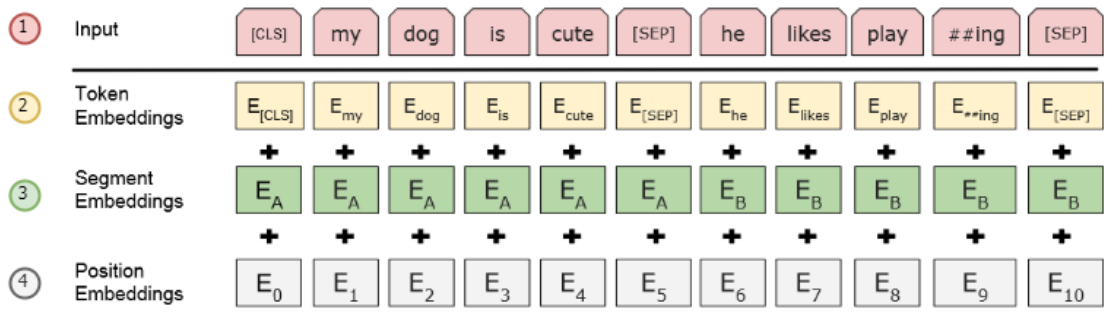
Kelime parçacığı yöntemi, kelime dağarcığında en sık görülen karakter kombinasyonlarının yinelemeli olarak kelime dağarcığına eklenmesi prensibine dayanan bir yöntem kullanmaktadır ve bu yöntem aşağıdaki adımları takip etmektedir [64].

1. Kelime birimi envanteri, temel karakterlerle başlatılır.
2. 1. Maddedeki kelime envanterini kullanarak eğitim verileri üzerinde bir dil modeli oluşturulur.
3. Mevcut kelime envanterinde bulunan iki birim birleştirilerek yeni bir kelime altı birim oluşturulur. Bu yeni birim kelime envanterine eklendikten sonra envanterdeki eleman sayısı 1 artacaktır. Yeni kelime birimi modele eklenirken, eğitim verilerinin olasılığını en fazla arttıracak şekilde tüm olası durumlar arasından seçilir.
4. Önceden tanımlanmış bir kelime birimi sınırına ulaşılana veya olasılık artışı belirli bir eşiğin altına düşene kadar 2'ye gidilir.

Girdi/Çıktı Temsili

BERT modeli için önceden belirlenmiş özel simgeler bulunmaktadır. Bu simgeler ve kullanım amaçları aşağıdaki gibidir.

- [CLS]: Sınıflandırma anlamına gelen bu sembol, modelin girişindeki ilk cümlenin başlangıcına konulmaktadır. Modelin çıkışında, buradan alınacak vektörle duygu analizi benzeri cümle sınıflandırma işlemleri yapılabilmektedir.
- [SEP]: Sıradaki cümle tespiti görevi için önemli olan [SEP] belirteci, cümleleri birbirinden ayırmak için kullanılmaktadır. Eğer modele tek bir cümle verilecekse, ilk cümlenin sonuna konulur ve ikinci cümle verilmez.
- [MASK]: Bu belirteç maskelenmiş dil modeli için önemlidir. Maskeli kelimeyi modele göstermek için kullanılır.



Şekil 5.4 BERT modeli giriş işlemleri [61].

Modelin giriş temsillerinin oluşturulması için Şekil 5.4'te gösterildiği gibi üç adet vektör operasyonu yapılmaktadır.

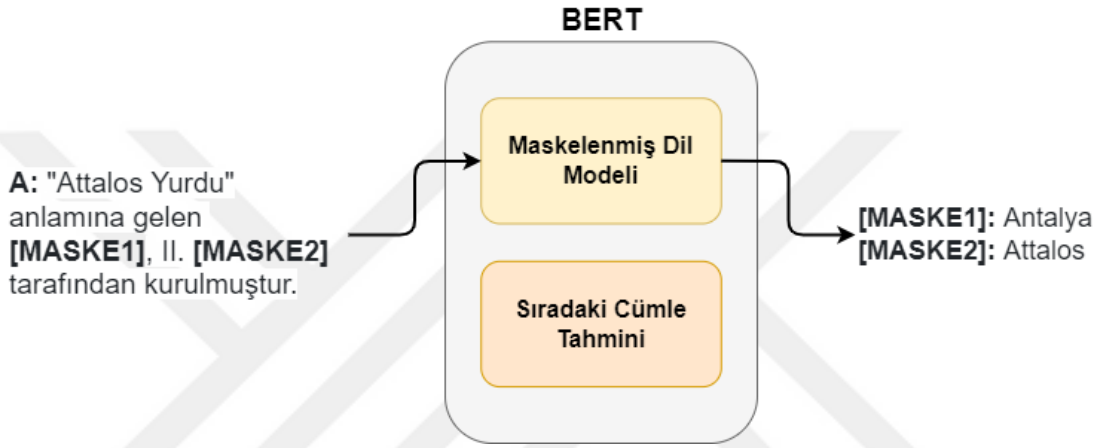
1. Girişler, kelime parçacığı yöntemi ile kelime altı parçacıklar halinde temsil edilmektedir.
2. 30000 (toplam kelime altı parçacık sayısı) x 768 (BERT gizli katman boyutu) boyutlu bir matris oluşturulmaktadır. Bu matrisin ağırlıkları eğitim sırasında öğrenilmektedir.
3. Cümlelerin ayrıştırılabilmesi için A ve B cümlelerinin kelime vektörlerine sırayla gizli katman uzunluğunda 0 ve 1 elemanlarından oluşan vektörler eklenmektedir.
4. Kelime sıralarının vektör içerisinde temsil edilebilmesi için Denklem 4.1 ve 4.2'de olduğu gibi konumsal kodlama uygulanmakta ve elde edilen vektörler önceki adımlardaki vektörlerle toplanmaktadır.

BERT modelinin çıkışında ise her bir kelime altı parçacık için 768 boyutlu bir vektör elde edilmektedir. İnce-ayar aşamasında bu ağırlıklar kullanılarak isimlendirilmiş varlık tanıma ve sözcük türü tespiti gibi doğal dil işleme çalışmaları için modeller eğitilebilmektedir. [CLS] özel belirteci için oluşturulan vektörle, ince-ayar aşamasında cümle sınıflandırma görevleri yapılabilmektedir.

Maskelenmiş Dil Modeli

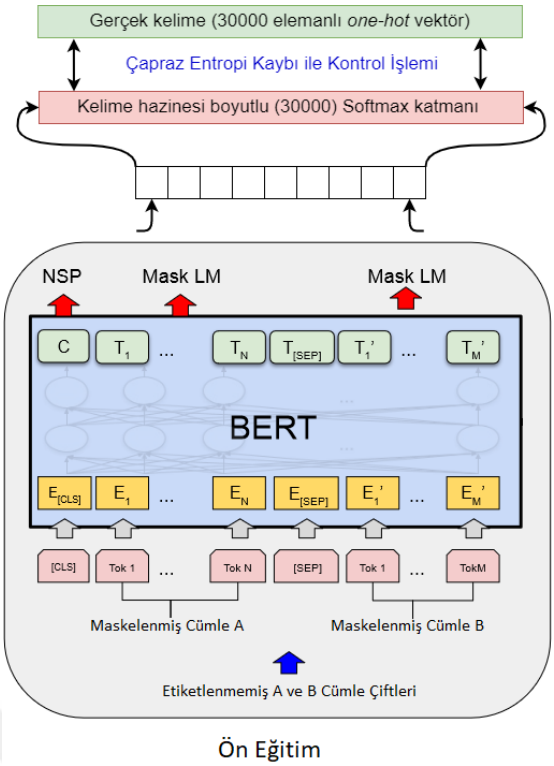
BERT, dili temsil edebilmek için maskelenmiş dil modeli tekniğini önermektedir. Bu teknik, modelin girişine verilen kelime altı parçacıkları rastgele olacak şekilde maskelemekte, ardından maskelenmiş bu parçacığın, kelime dağarcığı içerisindeki sırasını bulmaya çalışmaktadır. Bu durumda, herhangi bir veri etiketleme işlemine

ihtiyaç duyulmadan kendi kendine gözetimli öğrenme yapılmaktadır. Gözetimli öğrenmede kullanılacak sınıflar maskelenmiş kelime altı parçacıklara karşılık gelmektedir. Şekil 5.5'te modelin girişine verilen kelimeler, rastgele olacak şekilde maskelenmiştir. BERT'in maskelenmiş dil modeli bileşeninin amacı, maskelenmiş kelimelerin orijinal kelime listesi içerisindeki sırasının bulunmasını sağlamaktır [61]. Bu bağlamda, maskelenmiş dil modeli ile kelimelerin sadece soldan sağa veya sağdan sola değil, cümle bağlamına göre iki yönlü olarak da temsil edilmesi sağlanmaktadır.



Şekil 5.5 BERT maskelenmiş dil modeli bileşeni.

Maskelenmiş dil modeli eğitimi esnasında girişteki simgeleştirilmiş kelime altı parçacıların rastgele %15'i BERT için özel bir simge olan [MASK] ile değiştirilmektedir. Çıkışta ise [MASK] simgesi yerine orijinal kelime tahmin edilmeye çalışılmaktadır. Fakat bu işlem, kelimelerin doğru şekilde temsil edilebilmesi için yeterli değildir. Çünkü bu durumda, sadece maskelenmiş semboller tahmin edilecektir ve ince-ayar evresinde [MASK] belirteçleri kullanılmamaktadır. Bunun yerine maskelenmek üzere seçilen kelimeler %80 ihtimalle maskelenmekte, %10 ihtimaller rastgele bir kelime ile değiştirilmekte ve %10 ihtimalle olduğu gibi bırakılmaktadır [61]. Bu şekilde modelin çıkışında elde edilecek kelime vektörleri, ince-ayar yapabilmek için daha uygun olmaktadır.



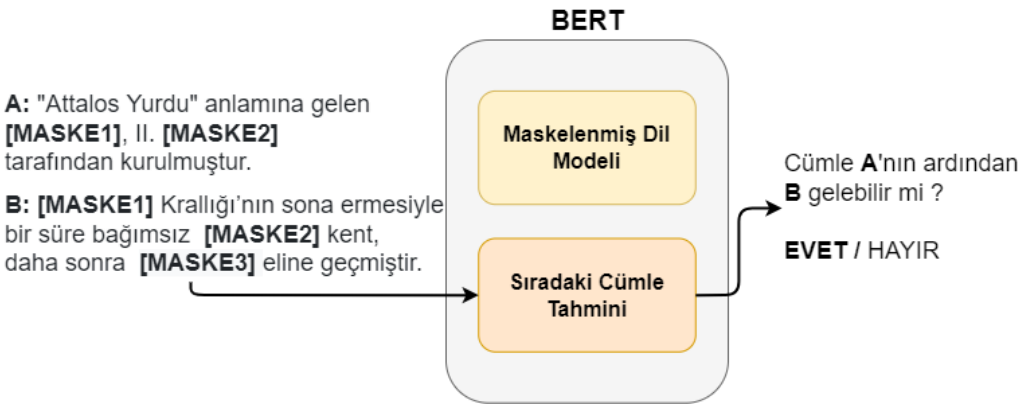
Şekil 5.6 BERT ön eğitim işlemi [61] (düzenlendi).

Şekil 5.6'da maskelenmiş dil modeli eğitimi gösterilmektedir. Bu işlem yapılırken, *softmax* aktivasyon fonksiyonuna sahip son gizli katmandan sağlanan makelenmiş kelime ağırlıkları ile orijinal kelime listesinin arasındaki Denklem 5.1'de gösterilen çapraz entropi kaybı hesaplanmaktadır.

$$L(\Theta) = -\sum_{i=1}^{k} y_i \log(\hat{y}_i) \tag{5.1}$$

Sıradaki Cümle Tahmini

Doğal dil çıkarımı ve soru cevaplama gibi doğal dil işleme çalışmaları, cümleler arasındaki ilişkilerin çıkarılması prensibine dayanmaktadır [61]. Bu yüzden BERT modelinde, cümleler arasındaki ilişkilerin de temsil edilmesini sağlayacak sıradaki cümle tahmini bileşeni eklenmiştir. Bu bileşen, girişte verilen iki cümlenin birbiri ardına gelip gelemeyeceğini tahmin etmeye çalışmaktadır. Şekil 5.7'de sıradaki cümle tahmini bileşeni için bir örnek gösterilmektedir.



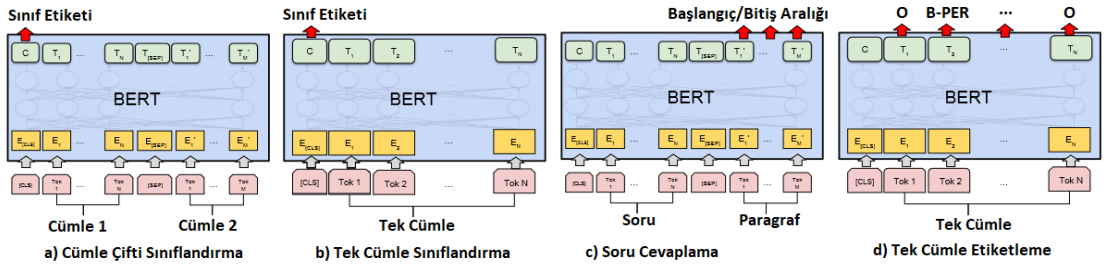
Şekil 5.7 BERT sıradaki cümle tahmini bileşeni.

BERT modelinin eğitimi esnasında, Şekil 5.7'de gösterilen A cümlesinden sonra gelecek B cümlesi, %50 ihtimalle ya sıradaki cümle ya da derlem içerisinden rastgele bir cümle olmaktadır. Böylelikle model, cümleler arasındaki ilişkileri de çıkararak dili daha iyi modelleyebilmektedir.

5.2.2.2 İnce Ayar

Ön eğitim aşamasından geçen BERT modeli, dilin yapısını ve cümleler arasındaki ilişkileri modelleyebilmektedir. Model artık az miktar veri ve kaynakla, kısa süre içerisinde, doğal dil işleme için önemli olan belirli görevleri için eğitilmeye hazırdır. Yapılacak doğal dil işleme görevine göre, modelin uygun çıkışlarına eklenecek basit doğrusal katmanlarla dahi oldukça iyi çıkarımlar yapılabilmektedir.

Şekil 5.8'de BERT modelinin farklı doğal dil işleme işlemleri için nasıl konfigüre edileceği gösterilmektedir.



Şekil 5.8 BERT ince-ayar konfigürasyonları [61].

Aşağıda her bir model yapısının ne tür doğal dil işleme çalışmaları için uygun olacağı sıralanmıştır.

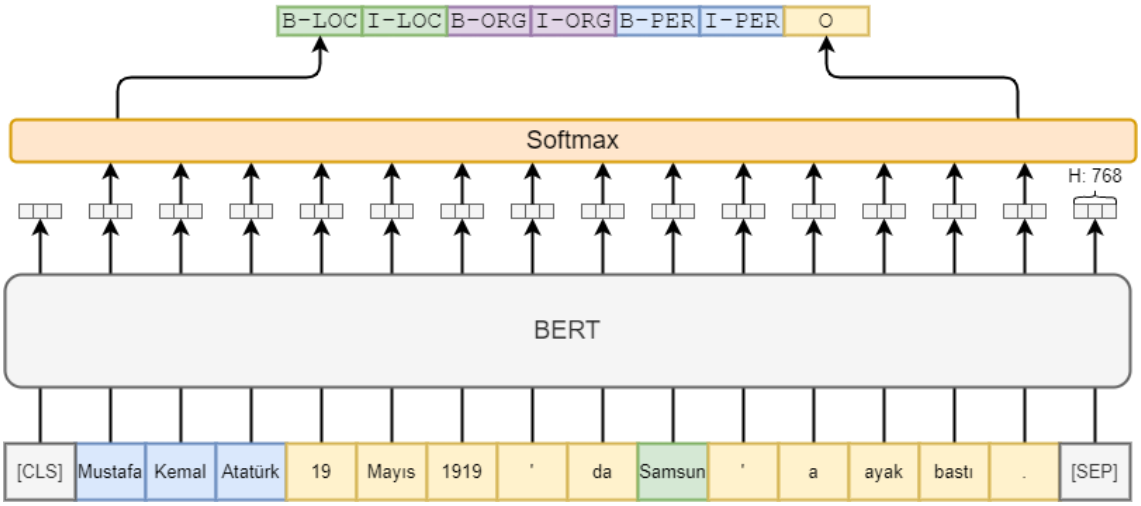
a) **Cümle Çifti Sınıflandırma:** Verilen iki cümle arasındaki ilişkinin saptanmasını gerektiren uygulamalarda kullanılabilmektedir. Örnek olarak; iki cümlenin birbirini takip edip edemeyeceğinin sınıflandırması, iki cümlenin aynı anlamı taşıyıp taşımadığının sınıflandırılması verilebilir.
b) **Tek Cümle Sınıflandırma:** Tek cümlenin sınıflandırılmasını gerektiren uygulamalarda kullanılabilmektedir. Örnek olarak; duygu analizi ve niyet sınıflandırılması verilebilir.
c) **Soru Cevaplama:** Soru ve cevabın aranacağı paragrafın verilmesi ile cevabın paragraf içerisinde nerede olacağının belirlenmesi uygulamalarında kullanılabilmektedir.
d) **Tek Cümle Etiketleme:** Cümle içerisindeki kelime parçacıklarına karşı üretilecek vektörlerin sınıflandırılması için kullanılabilmektedir. Örnek olarak; isimlendirilmiş varlık tanıma, sözcük türlerinin belirlenmesi verilebilir.

5.3 İsimlendirilmiş Varlık Tanıma

Sohbet ajanları için oldukça önemli olan isimlendirilmiş varlık tanıma işlemi, ham bir metin içerisindeki varlıkların çıkarılmasını, böylelikle metinlerin daha anlamlı hale getirilmesini sağlamaktadır. Genellikle; insan, yer, kuruluş isimlerinin tespit edilebilmesi için önemlidir. Bunun dışında tarih, para birimi, mail adresleri, sanat eserleri vb. varlıkların belirlenmesi için de kullanılabilmektedir.

Bir sanal asistandan, A noktasından B noktasına gitmek için bir rota oluşturması istenildiğinde, asistanın, A ve B noktalarının birer konuma karşılık geldiğini bilmesi gerekmektedir. Bu noktada, isimlendirilmiş varlık tanıma modelleri bu işlemin yapılmasını sağlamaktadır.

İsimlendirilmiş varlık tanıma modeli eğitebilmek için Başlık 5.2.2.2'de anlatıldığı gibi BERT kullanılabilmektedir. Bunun için ön eğitim aşamasından geçmiş BERT modelinin sonuna, sınıf sayısı kadar nörona sahip bir doğrusal katman eklenmelidir. BERT modelinin çıkışında her bir kelime için oluşacak ağırlıkların, *softmax* aktivasyon fonksiyonundan geçirilerek, çapraz entropi kaybının minimize edecek şekilde tekrardan eğitilmesi ile bu ağırlıklara karşılık gelen kelime altı parçacıklar sınıflandırılabilmektedir.



Şekil 5.9 BERT isimlendirilmiş varlık tanıma konfigürasyonu.

Şekil 5.9'da BERT'in isimlendirilmiş varlık tanıma görevi için ince-ayar konfigürasyonu gösterilmektedir. Kişi, yer ve kuruluş isimlerinin sınıflandırılmasını sağlayacak, Türkçe isimlendirilmiş varlık tanıma modeli eğitebilmek için WikiAnn [65] veri seti kullanılabilmektedir. Alternatif olarak WikiAnn veri seti ile daha önceden eğitilerek yayınlanan modelin [66] kullanılması da mümkündür. WikiAnn veri seti, "B-LOC", "I-LOC", "B-ORG", "I-ORG", "B-PER", "I-PER", "O" olmak üzere 7 adet sınıfa sahiptir. Etiketler içerisinde yer alan "B" ve "I" belirteçleri, gruplardan oluşan kelimelerin sınırlarını belirlemekte fayda sağlamaktadır. "B" başlangıcı, "I" ise bir önceki kelimeye ait etiketin devam ederek bir bütün olduğunu göstermektedir. "LOC","ORG" ve "PER" sırayla; yer, kuruluş ve kişi isimlerinin etiketleridir. "O" ise diğer kelimeleri temsil etmek için kullanılmaktadır.

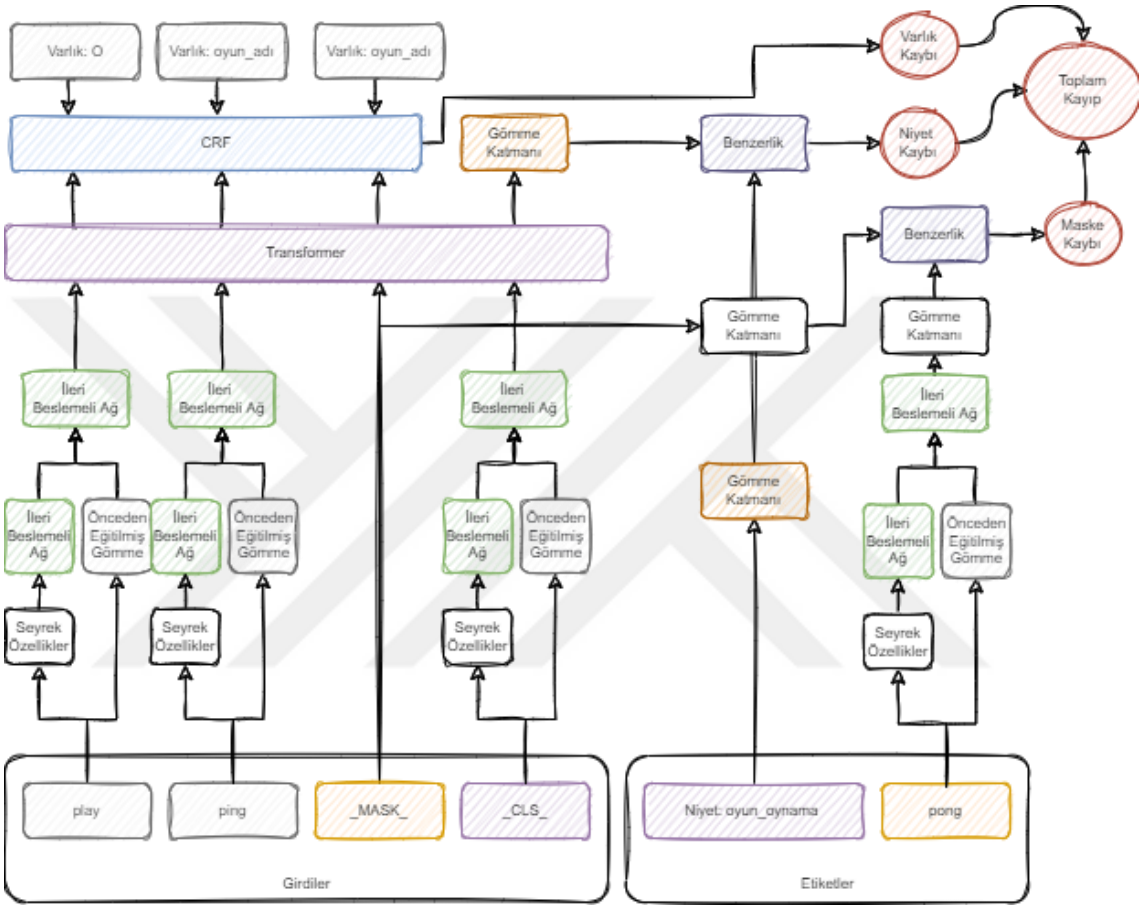
5.4 Niyet Sınıflandırma

Günlük hayatta tek bir istek onlarca farklı şekilde dile getirilebilmektedir. Aslında söylenmek veya yapılmak istenilen şey, tek bir duruma karşılık gelmektedir. Bu gibi bir durum, doğal dil işlemede niyet sınıflandırma problemi olarak bilinmektedir.

Niyet sınıflandırmak için de BERT modelinin kullanılması mümkündür. Fakat RASA [67] ekibi, doğrudan diyalog sistemlerinde kullanılabilecek DIET [68] isminde daha hafif bir mimari önermiştir.

5.4.1 DIET

Bu model hem niyet hem de isimlendirilmiş varlıkların çıkarılmasına imkân sağlayan bir yapıda tasarlanmıştır. Şekil 5.10'da gösterildiği gibi üç kayıp fonksiyonu ile oluşturulan toplam hatanın minimize edilmesi ile modelin öğrenmesi gerçekleşmektedir.



Şekil 5.10 DIET mimarisi [68].

DIET mimarisi de BERT'e benzer şekilde "__MASK__" ve "__CLS__" özel belirteçlerine sahiptir. Bu belirteçler burada da benzer amaçlar için kullanılmaktadır. "__MASK__" belirteci giriş içerisinden rastgele bir kelimeyi maskelemektedir. Ardından maskelenmeyen kelimelerin oluşturduğu ağırlıklardan da faydalanarak maskelenen kelime tahmin edilmeye çalışılmaktadır. Böylece kelimelerin daha iyi bir şekilde temsil edilmesi sağlanmaktadır.

“__CLS__” belirteci ise bütün girişi temsil edilecek bir vektörün oluşturulmasını sağlamaktadır. Böylelikle buradan elde edilecek vektör doğrudan niyet sınıflandırmada kullanılabilmektedir.

Girişlerde kelimelerin ifade edilebilmesi için iki farklı yöntem kullanılabilmektedir. BERT [61], GloVe [69] ve ConveRT [70] gibi önceden eğitilmiş modellerden sağlanacak yoğun gösterimler kullanılabileceği gibi, kelimelerin *one-hot* kodlama ve karakter n-gramları gibi daha seyrek gösterimleri de tercih edilebilmektedir. Her iki yöntemle de elde edilen çıktılar, dönüştürücü bloğa verilmeden önce ileri beslemeli bir ağdan geçirilmektedir.

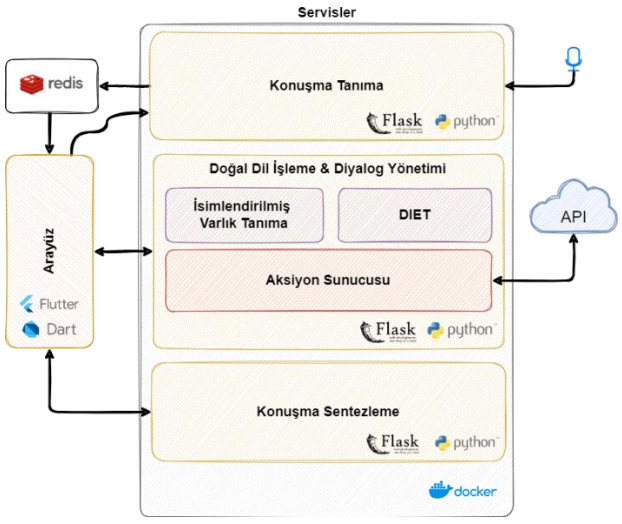
CRF [71] bloğunda isimlendirilmiş varlık tanıma için negatif log olasılığı [72] ile kayıp hesaplanmaktadır. Niyet sınıflandırma ve maskeleme de ise nokta çarpım kayıp fonksiyonu kullanılmaktadır [68]. Bütün kayıpların toplamının minimize edilmesi ile öğrenme sağlanmaktadır.

6
UYGULAMA TASARIMI

6.1 Sistem Mimarisi

Sanal asistan uygulamaları genellikle kullanıcıların yapacağı sorgulamaları ve işlemleri sesli olarak kabul etmektedir. Sesli girişlerin işlenebilmesi için ilk olarak konuşma sinyallerinin metne dönüştürülmesi gerekmekte, ardından doğal dil işleme yöntemleri ile metinlerdeki niyet ve varlıkların çıkarılması sağlanmaktadır. Böylelikle ham metin, makinelerin anlayacağı forma dönüşmekte ve gerekli sorgulama işlemlerinin kolaylıkla yapılması sağlamaktadır. Kullanıcının yapmak istediği işleme göre sanal asistanın vereceği cevaplar, görsel ve işitsel formda olabilmektedir.

Bu tez çalışmasında, araç ortamı için özelleştirilerek tasarlanan sanal asistan uygulamasının mimarisi Şekil 6.1'de gösterilmektedir. Uygulama oluşturulurken mikro servis mimarisi prensipleri temel alınmıştır. Böylelikle uygulama daha yönetilebilir olarak tasarlanmıştır. Servis yönetiminin sağlanabilmesi için Docker [73] teknolojisi kullanılmıştır. Servisler, olağan konfigürasyon ile yerel makinede çalışacak şekilde düzenlenmiştir. Ancak, minimal değişikliklerle uzak sunucularda da çalışabilecek hale getirilebilmektedir.

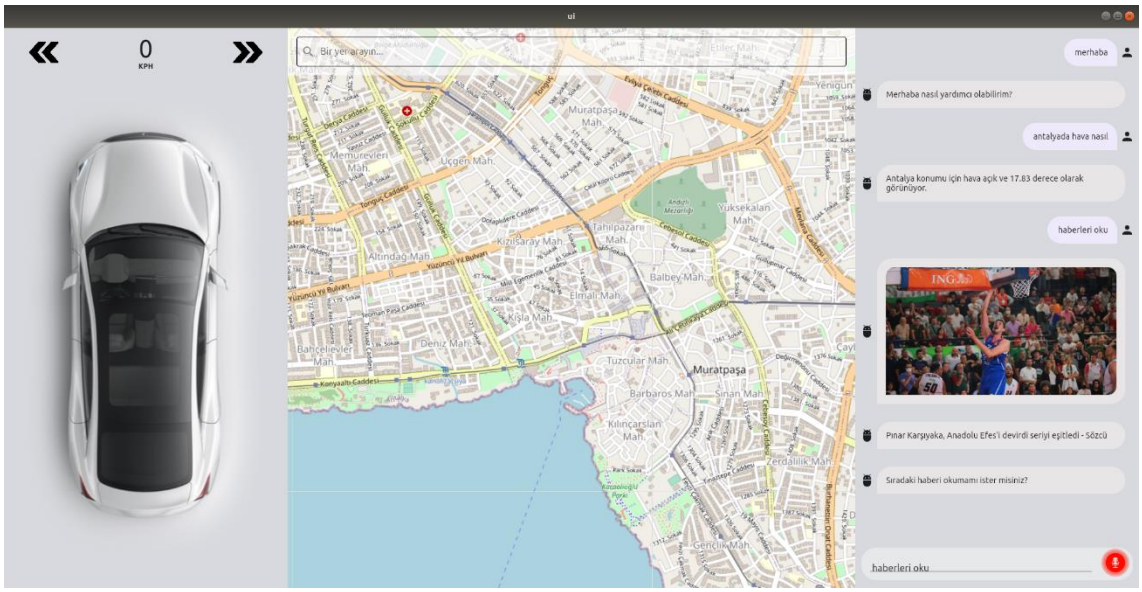


Şekil 6.1 Tasarlanan sanal asistanın mimarisi.

Uygulama, kullanıcının servislerle etkileşimini sağlayacak bir arayüz üzerine inşa edilmiştir. Arayüz, konuşama tanıma servisinin tetiklenmesine imkân sağlayan bir butona sahiptir. Butona basılıyken, konuşma tanıma servisi aktive edilmektedir ve servis, mikrofon üzerinden alınan konuşma sinyallerinin yazıya dönüştürülmesi sağlanmaktadır. Bu esnada mikrofondan alınan veriler, anlık olarak işlenmektedir ve sonuçlar RAM üzerinde değer tutabilen Redis veri tabanına yazılmaktadır. Veri tabanına işlenen her bir değerden sonra, arayüz üzerinde yeni kelime sekansı gösterilmektedir. Kullanıcının yapacağı sesli girişin tamamlanması ile oluşan nihai metin, yapılandırılmak üzere doğal dil işleme servislerine gönderilmektedir. Bu servislerde öncelikle metin içerisindeki özel varlıklar saptanmaya çalışılmakta, ardından metindeki niyet sınıflandırılmaktadır. Niyetin saptanması ile uygun aksiyon tetiklenmekte ve gerekli olan prosedürler işletilerek uygun cevapların üretilmesini sağlamaktadır. Üretilen cevap, metin olarak arayüz üzerinde gösterilmekte ve aynı zamanda konuşma sentezleme servisi kullanılarak sesli şekilde kullanıcıya dönülmektedir.

6.2 Kullanıcı Arayüzü

Kullanıcı arayüzü, Flutter Framework'ü [74] kullanılarak tasarlanmış ve Dart [75] dili ile kodlanmıştır. Flutter, her platform için uygulama geliştirme esnekliği sunmasından dolayı tercih edilmiştir. Tasarlanan arayüz Şekil 6.2'de gösterilmektedir.



Şekil 6.2 Uygulama arayüzü.

Servisler ile etkileşim, sağ tarafta bulunan bölüm üzerinden gerçekleşmektedir. En alt kısımda bulunan mikrofon sembollü buton ile sesli bir şekilde giriş yapılabilmektedir. Tanınan kelimeler butonun sol tarafında bulunan metin kutusu içerisine yazdırılmaktadır. Butonun serbest bırakılması ile birlikte kullanıcı sorgusu, diyalog servisine gönderilerek işlenmektedir. Aksiyon sunucusu, sorguya karşılık gelen uygun cevabı hazırlayarak geri dönüş yapmaktadır. Dönülen cevaplar, metin içerikli olabileceği gibi resim formatlarında da olabilmektedir.

Uygulamanın orta bölümünde bulunan harita üzerinde, sanal asistan kullanılarak güzergâh oluşturma işlemleri yapılabilmektedir.

6.3 Konuşma Yazılandırma Servisi

Konuşma yazılandırma modeli, Bölüm 3'te anlatılan ağırlıklı sonlu durum dönüştürücü prensibine dayalı olarak Kaldi Konuşma Tanıma Aracı [76] kullanılarak eğitilmiştir. Bu model; Python [77] programlama dili ve Flask Framework'ü [78] kullanılarak, konuşma yazılandırması yapabilecek bir mikro servis haline getirilmiştir.

Konuşma yazılandırma servisinin tetiklenerek mikrofondan veri akışının başlatılabilmesi için bir uç nokta (https://localhost:5000/recognize-microphone) oluşturulmuştur. Bu uç noktaya gelecek istek ile mikrofondan veri akışı başlatılmakta, ardından elde edilen veriler, modelin kabul edeceği formata dönüştürülebilmesi için ön işleme tabi tutulmaktadır. Ön işleme esnasında ilk olarak sinyal 16 kHz örnekleme frekansına getirilmekte, ardından MFCC özellikleri çıkarılmaktadır. MFCC özellikleri modele giriş olarak verildikçe, çıkışta kelime sekansı elde edilmektedir. Elde edilen sonuçlar, mikrofondan veri alma işlemi kesilene kadar Redis veri tabanına *Pub/Sub* mekanizması kullanılarak yazılmaktadır.

Pub/Sub mekanizması nesnelerin interneti uygulamalarında sıklıkla kullanılmaktadır. Bu mekanizma temel olarak kanal, yayıncı (publisher) ve abone (subscriber) bileşenlerinden oluşmaktadır. Bir yayıncı tarafından belirli bir kanalda paylaşılan veriler, o kanalı dinleyen abonelere veriler güncellendikçe bildirilmektedir. Bu serviste de konuşma tanıma modelinden çıkan her kelime

dizisi, Redis veri tabanında oluşturulan “Konuşma Tanıma” kanalında yayınlanmaktadır. Arayüz de bu kanalın abonelerindendir. Bu sayede “Konuşma Tanıma” kanalına yazılan her bir veriden arayüzün de haberi olmaktadır ve ekranlarda bu çıktılar gösterilmektedir.

6.4 Konuşma Sentezleme Servisi

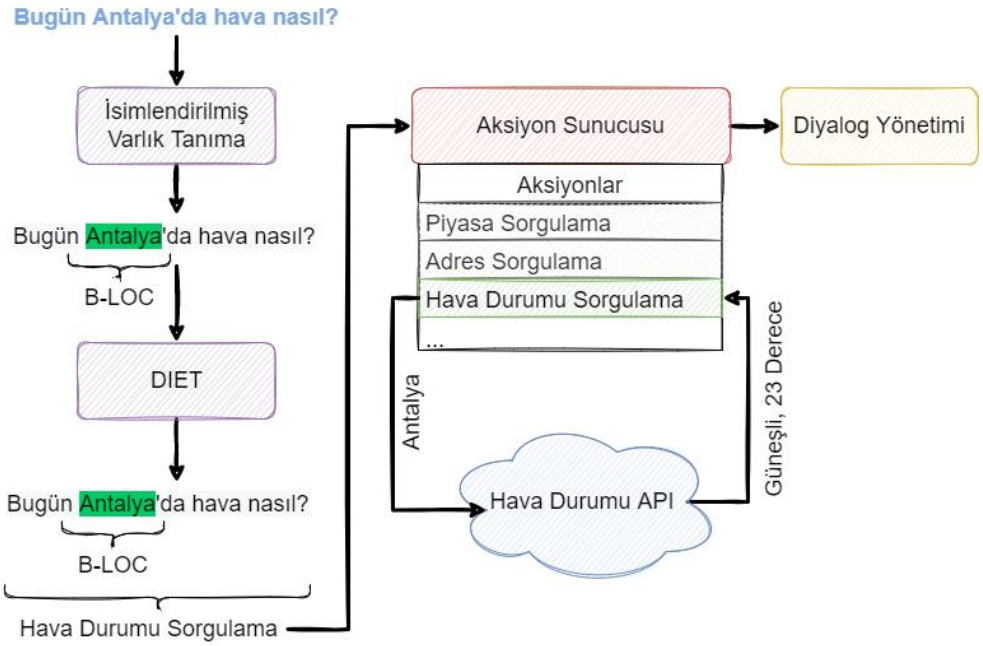
Konuşma sentezleme sistemi için Bölüm 4’te bahsedilen FastSpeech2 ve Multi-band MelGAN mimarileri kullanılmıştır. Bu modeller, Türkçe veriler ile eğitilmiştir. Bu servis de konuşma yazılandırma servisi gibi Python programlama dili ve Flask Framework’ü kullanılarak mikro servis haline getirilmiştir. Diyalog sistemi tarafından arayüze döndürülen metin, ardından seslendirilmek üzere bu servise gönderilmektedir. Servisten dönülen cevap ile elde dilen çıktı, arayüz üzerinden dinlenebilmektedir.

6.5 Doğal Dil İşleme ve Diyalog Yönetimi Servisi

Doğal dil işleme ve diyalog yönetimi servisleri; metinlerin anlamlandırılmasını, uygun cevapların üretilmesini ve sohbetlerin birbiri ile ilişkilendirilmesini sağlamaktadır. Sohbetleri birbirleri ile ilişkilendirilmesi ve konuşma takibinin sağlanabilmesi için açık kaynak kodlu Rasa [79] aracı kullanılmıştır.

Metnin anlamlandırılarak cevap üretilmesi noktasında izlenen prosedür Şekil 6.3’te bir örnek ile gösterilmiştir. Kullanıcı tarafından akıllı asistana, “Bugün Antalya’da hava nasıl?” sorusunun yönlendirildiği varsayılsın. Bu metin ham hali ile makineler için çok fazla anlamlı değildir. Çünkü yapılandırılmamıştır. Metnin yapılandırılması için metindeki önemli varlıkların belirlenmesi ve niyetin tespit edilmesi gerekmektedir. İsimlendirilmiş varlık tanıma modülü, daha önceden belirlenmiş sınıfları tanıma üzere eğitilmiş bir modelden oluşmaktadır. Ham metnin bu modelden geçmesi ile belirlenen sınıflara uygun bir kelimenin olup olmadığı tespit edilmektedir. Bu örnek için, yer ismi bildiren “Antalya” kelimesi tespit edilmiştir. Böylelikle metin, makine için daha anlamlı hale gelmiştir. Ardından kullanıcının niyetinin anlaşılması, çıktı için ne tür bir prosedürün işletilmesi gerektiğinin anlaşılması için önemlidir. Bu bağlamda DIET mimarisi, metin üzerinde diğer isimlendirilmiş varlıkların tanınmasını sağlamakla beraber

metindeki niyeti de tespit etmektedir. Böylelikle nasıl bir aksiyon alınacağı belirlenecektir. Aksiyon sunucusu, belirlenen niyete göre uygun işlemlerin yapılmasını sağlamakla yükümlüdür. Bu örnek için, hava durumunun sorgulanmasını sağlayacaktır. Bu işlemin cevabının üretilebilmesi için uzak sunucuya, belirlenen yer ismi ile bir istek atılacak ve elde edilen cevap kullanılarak bir metin üretilecektir. Üretilen metin arayüz üzerinden kullanıcıya sunulacaktır.



Şekil 6.3 Doğal dil işleme ve diyalog yönetimi servisleri için örnek girdi ve çıktılar.

Başlık 5.4.1'de anlatıldığı üzere DIET mimarisi hem isimlendirilmiş varlıkların hem de metinlerdeki niyetin çıkarılması için kullanılabilmektedir. Fakat harici bir isimlendirilmiş varlık tanıma modelinin kullanılması sistemin doğruluğunu arttıracaktır. Özellikle araçlar için özelleştirilmiş akıllı asistan sistemlerinde, yer isimlerinin yüksek doğruluk oranı ile çıkarılması navigasyon uygulamaları için oldukça önem arz etmektedir. Bu bağlamda yer, kişi ve kurum isimlerinin çıkarılmasında, önceden eğitilmiş BERT [66] modeli kullanılmıştır. Bunun dışındaki varlıkların çıkarılmasında ve niyetin sınıflandırılmasında DIET mimarisi kullanılmıştır.

7
YÖNTEM VE DENEYLER

7.1 Otomatik Konuşma Tanıma

7.1.1 Yöntem

Otomatik konuşma tanıma modellerinin geliştirilmesi için Bölüm 3'te anlatılan ağırlıklı sonlu durum dönüştürücü tabanlı model ve uçtan uca konuşma tanıma mimarilerinden QuartzNet [40] modeli kullanılmıştır. Her iki model de 550 saatlik Boğaziçi verisi [80] ile eğitilmiştir. İki modelin eğitiminin sonucunda elde edilen modeller teste tabi tutularak sanal asistan uygulaması için hangisinin kullanılacağı belirlenmiştir.

7.1.1.1 Ağırlıklı Sonlu Durum Dönüştürücü Tabanlı Model

Ağırlıklı sonlu durum dönüştürücü tabanlı model eğitilirken Kaldi konuşma tanıma aracı kullanılmıştır. Kaldi konuşma tanıma aracı; ağırlıklı olarak C++ dili ile geliştirilmiştir. Kaldi'nin /egs dizini altında, farklı konuşma tanıma problemleri ve farklı veri setleri için oluşturulmuş örnekler yer almaktadır. Bu çalışmada kullanılan model /egs/librispeech adımları takip edilerek oluşturulmuştur.

Adımlar sırayla özetlenecek olursa;

1. Okunuş sözlüğü kullanılarak fonların belirlenmesi ve L.fst'nin oluşturulması,
2. MFCC özniteliklerinin çıkarılması çıkarılması ve normalize edilmesi,
3. Tekli ve üçlü fonlar için GMM modelinin oluşturulması ve hizalama işlemlerinin yapılması,
4. TDNN akustik modelinin eğitilmesi,
5. Dil modelinin oluşturulması (G.fst),
6. HCLG.fst grafının oluşturulması

ile konuşma tanıma sistemi için gerekli modeller hazır hale gelmektedir. Bu adımlardan elde edilen akustik model ve HCLG grafı kod çözme esnasında kullanılmaktadır.

7.1.1.2 Uçtan Uca Konuşma Tanıma Modeli

Uçtan uca konuşma tanıma sistemi gerçeklemek için QuartzNet modeli kullanılmıştır. QuartzNet modeli, çıkış katmanında alfabedeki karakter sayısı ve bir boşluk karakteri ile toplamda 30 nörona sahiptir. Modelin eğitimi, girişteki log mel-spektrogram özellikleri ile çıkıştaki karakterler arasındaki CTC kaybının minimize edilmesi ile gerçekleştirilmiştir.

Buna ek olarak, bu akustik modelden elde edilecek sonuçlar 3-gramlık bir dil modeli kullanılarak yeniden skorlandırılmıştır. Yeniden skorlama işlemi Denklem 7.1'de gösterilmektedir.

$$skor_{son} = skor_{akustikModel} + \alpha * skor_{dilModeli} + \beta * Uzunluk_{dizi} \qquad (7.1)$$

Denklemdeki son skor akustik model, dil modeli ve dizinin uzunluğundan gelecek skorların toplamından oluşmaktadır. α terimi, dil modeline ne kadar önem verilmesi gerektiğini ayarlamayı sağlamaktadır. Büyük α değerleri, dil modeline daha çok önem verilmesi anlamına gelmektedir. β ise bir ceza terimidir. Negatif değerler kısa sonuçların üretilmesini sağlarken pozitif değerler daha uzun sekansların oluşturulmasını sağlamaktadır.

7.1.2 Deneyler

Bir önceki bölümde anlatılan konuşma tanıma modellerinin performanslarının test edilebilmesi için üç adet veri seti kullanılmıştır. Bu setler, Boğaziçi test seti içerisinden oluşturulmuş 1 saat uzunluğundaki test verisi, Youtube'dan elde edilen 1 saat uzunluğundaki test verisi ve CommonVoice [81] Türkçe veri seti içerisindeki test verilerinden oluşmaktadır. Bu test setleri kullanılarak elde edilen model performansları Tablo 7.1'de gösterilmektedir. Modellerin test edilmesinde kelime hata oranı metriği kullanılmıştır. Kelime hata oranını düşük olan modellerin tanıma kabiliyeti daha yüksektir. Bir numaralı deneyde, ağırlıklı sonlu durum dönüştürücü tabanlı ve TDNN akustik modelini kullanan konuşma tanıma sisteminin sonuçları gösterilmektedir. İkinci deneyde ise uçtan uca konuşma

tanıma mimarilerinden olan QuartzNet'in performans değerleri gösterilmektedir. Üçüncü deneyde ise ikici deneyden farklı olarak 3-gramlık dil modeli ile yeniden skorlandırma yapılarak elde edilen başarı gösterilmektedir.

Tablo 7.1 Konuşma tanıma modellerinin karşılaştırılması.

			Kelime Hata Oranı			
Deney No	**Akustik Model**	**Dil Modeli**	**TV Yayını**	**Youtube Yayını**	**Common Voice TR Test Seti**	**Ortalama**
1	TDNN	3-gram	%12,29	**%17,59**	%22,97	%17,62
2	QuartzNet	-	%18,47	%27,21	%29,22	%24,97
3	QuartzNet	3-gram	**%11,95**	%20,41	**%19,66**	**%17,34**

İkinci ve üçüncü deneylerde, en iyi sonuçların aranması noktasında ışın araması yöntemi kullanılmıştır. Işın genişliği olarak 64 seçilmiştir. Buna ek olarak Denklem 7.1'de gösterilen α ve β değerleri 1 olarak belirlenmiştir. Üçüncü deneyde kullanılan dil modelinin, sistemin performansını oldukça arttırdığı gözlemlenmektedir. Birinci ve üçüncü deneylerde kullanılan modellerin performanslarının birbirlerine oldukça yakın olduğu gözlemlenmektedir. Ortalama değerde ise üç numaralı deneyin modelinin daha başarılı olduğu görülmektedir.

Bu deneye ek olarak, her iki model türü de mikrofondan akan veri ile test edilmiştir. Ağırlıklı sonlu durum makinesi yaklaşımı oldukça başarılı sonuçlar üretirken, QuartzNet modeli kelime sınırlarını belirlemede ve doğru tahminler yapmada güçlük çekmiştir. "Ankara'dan İnstabul'a gitmek istiyorum." cümlesi için QuartzNet modeli ile Şekil 7.1'deki çıktı alınmıştır.

```
Listening...
era anistanbula gitmek istiyoru
```

Şekil 7.1 QuartzNet modeli canlı tanıma örneği.

Canlı konuşma yazılandırma sistemleri için ağırlıklı sonlu durum makinesi yaklaşımının daha uygun olduğu saptanmıştır ve sanal asistan uygulaması için bu model tercih edilmiştir.

7.2 Konuşma Sentezleme

7.2.1 Yöntem

Derin öğrenme yöntemleri ile başarılı konuşma sentezleme modelleri geliştirebilmek için stüdyo ortamında kaydedilmiş uzun süreli (20+ saat) ve yüksek kaliteli veriye ihtiyaç duyulmaktadır. Türkçe için açık olarak paylaşılmış böyle bir veri seti bulunmamaktadır. Bu tez çalışması kapsamında hazırladığımız makalede, LJSpeech [82] gibi İngilizce bir veri setinin, Türkçe konuşma sentezleme çalışmaları için kullanılıp kullanılamayacağı tartışılmıştır [83]. Önerilen çalışma ile İngilizcede bulunmayan “ı, ö, ü, ç, ş” karakterlerine karşı başarılı bir şekilde fon üretimi yapılabilmiştir. Ancak, elde edilen model ile üretilen konuşmalar, İngiliz bir turistin Türkçe konuşmasını andırması ve anlaşılabilirliğinin düşük olması sebebiyle doğrudan nihai uygulamada kullanılmamıştır. Bunun yerine, stüdyo ortamında 2.5 saatlik konuşma kaydı alınarak küçük bir veri seti oluşturulmuştur. Ardından önerilen yöntem ile elde edilen modelin eğitimine bu küçük veri seti ile devam edilerek nihai sentezleyici (FastSpeech2) model elde edilmiştir.

Ses kodlayıcı modelin (Multi-band MelGAN) eğitilmesinde ise;

1. İngilizce veri seti olan LibriTTS [84] içerisinden rastgele olarak ayrılmış 10 saat veri,
2. METU Turkish Microphone Speech v1.0 [85] veri setinin tamamı,
3. Oluşturduğumuz veri setinin tamamı

kullanılmıştır. Tüm veriler 16 kHz örnekleme frekansına getirilmiştir. Eğitim öncesi ön hazırlık evresinde oluşturulan veri seti içerisindeki örneklerin mel-

spektrogramları çıkarılmıştır. Eğitim ilk olarak, Multi-band MelGAN mimarisindeki üretici ağın STFT kayıp fonksiyonu ile eğitilmesiyle başlatılmıştır. Üretici ağın belirli bir miktar yakınsamaya başlamasıyla, ayrıştırıcı ağ da devreye sokularak eğitime devam edilmiştir. Üretici ağın uygun ses sinyallerini üretmeye başlaması ile birlikte eğitim tamamlanmıştır. Sentezleyici model ile birlikte ses kodlayıcı modelin beraber kullanılması sonucunda yazıdan sese dönüşüm işlemi başarılı bir şekilde yapılabilmiştir.

7.2.2 Deneyler

Konuşma sentezleme modelleri için sentezleme süreleri oldukça önemlidir. Sanal asistan benzeri uygulamalarda kullanıcıya dönülecek ses sinyallerinin hızlı bir biçimde üretilmesi gerekmektedir. Bu bağlamda, Tablo 7.2'de konuşma sentezleme sistemi ile üretilen ses çıktıları için sentezleyici ve ses kodlayıcı modellerin ne kadar sürede üretim yapabildiği gösterilmiş ve gerçek zaman faktörü hesaplanmıştır. Deney esnasında test donanımı olarak 3,5 GHz işlemci frekansına sahip Intel i5 4690K kullanılmıştır.

Tablo 7.2 Konuşma sentezleme sistemi için gerçek zamanda çalışma testi.

Cümle No	Sentezlenen Sesin Süresi	FastSpeech2 Üretim Süresi	Multi-band MelGAN Üretim Süresi	Gerçek Zaman Faktörü
1	1.76 saniye	0.84 saniye	0.30 saniye	0.65
2	3.08 saniye	0.36 saniye	0.17 saniye	0.17
3	3.0 saniye	0.16 saniye	0.17 saniye	0.11
4	3.23 saniye	0.15 saniye	0.17 saniye	0.10
5	5.73 saniye	0.18 saniye	0.19 saniye	0.06
6	7.88 saniye	0.29 saniye	0.41 saniye	0.09
7	3.31 saniye	0.16 saniye	0.17 saniye	0.10

Tablo 7.2 Konuşma sentezleme sistemi için gerçek zamanda çalışma testi (devamı).

8	3.45 saniye	0.16 saniye	0.18 saniye	0.10
9	1.61 saniye	0.12 saniye	0.21 saniye	0.21
10	3.42 saniye	0.24 saniye	0.38 saniye	0.18

7.3 Doğal Dil İşleme ve Diyalog Yönetimi

7.3.1 Yöntem

Günlük yaşamda yapmak istediğimiz eylemleri başkalarına aktarırken yüzlerce farklı şekilde cümle kurabilmemiz mümkündür. Bir sanal asistan ile etkileşim esnasında da bu durumlarla karşılaşılmaktadır. Kullanıcıların farklı şekillerde ifade ettiği durumlardaki niyetlerin saptanması noktasında niyet sınıflandırma yöntemleri kullanılmaktadır.

Bir sanal asistan veya sohbet botu uygulaması geliştirmeden önce ilk olarak tasarlanacak sistemin hangi yetkinliklere sahip olacağının belirlenmesi faydalı olacaktır. Sanal asistanın yetkinliklerine uygun niyet ve niyet örneklerinin oluşturulması niyet sınıflandırıcı modelin eğitimi için gereklidir. Tablo 7.3'te selamlama, hava durumu sorgulama, rota oluşturma ve haber sorgulama yetkinliklerine sahip bir sanal asistan için oluşturulan niyet örnekleri gösterilmektedir.

Tablo 7.3 Niyet sınıflandırmasında kullanılabilecek örnekler.

Örnekler	**Niyet**
Selam Merhabalar Selamlar bot Merhaba sanal asistanım	Selamlama

Tablo 7.3 Niyet sınıflandırmasında kullanılabilecek örnekler (devamı).

Bugün ……'da hava nasıl olacak? Hava ……'da nasıl? …… konumu için hava durumu sorgula. …… konumunda hava nasıl? Meteorolojiden …… için hava durumu sonuçlarını getir.	Hava Durumu Sorgulama
…… konumundan ……'ya gitmek istiyorum. …… ve …… arasındaki güzergahı çiz. ……'dan ……'ya gideceğim yolu gösterir misin?	Rota Oluşturma
Haberleri oku. Güncel haberleri görüntüle. Gazeteleri okur musun? Gazete manşetlerini oku. Güncel haberleri göster.	Haber Sorgulama

Sanal asistan uygulamalarında yapılacak işleme karar verilebilmesi için öncelikle kullanıcının niyetinin anlaşılması gerekmektedir. Böylelikle kullanıcının isteğine karşı uygun bir aksiyonun alınması sağlanabilmektedir.

Sanal asistanlar için diğer bir önemli unsur ise cümlelerdeki isimlendirilmiş varlıkların tespit edilmesidir. Böylelikle tespit edilen varlığa özel sorgulamaların yapılması sağlanabilmektedir. Tablo 7.3'te hava durumu sorgulama ve rota oluşturma işlemlerinin yapılabilmesi için bir şehir veya yerleşim yeri bilgisine ihtiyaç duyulmaktadır. Böylelikle tespit edilen konum için uzak sunuculardan o bölge için gerekli bilgiler elde edilebilmektedir.

Bu çalışmada, niyet sınıflandırma ve isimlendirilmiş varlık tanıma problemlerini tek bir sinir ağı mimarisi ile çözmeyi hedefleyen DIET mimarisi kullanılmıştır.

Tablo 7.4'te DIET mimarisini eğitebilmek için oluşturulan niyetler, niyetlere ait örnekler ve toplam örnek sayısı gösterilmektedir. Toplamda 9 niyet için 67 örnek oluşturulmuştur.

Tablo 7.4 Niyet sınıflandırması için oluşturulan veri seti ve örnekleri.

Niyet	Örnek	Örnek Sayısı
rota_sorgulama	Ankara'dan İstanbul'a rota oluştur.	10
selamlama	Merhaba	9
hatır_sorma	Nasılsın	9
hava_durumu_sorgulama	Ankara'da hava nasıl	9
haritada_gösterme	Haritada Isparta'yı göster.	7
sinyal_verme	Sağa dönmek istiyorum.	7
haberler_sorgulama	Haberleri oku	6
onaylama	Evet	5
reddetme	Hayır	5

Veri seti oluşturulurken her bir örnekteki niyet, isimlendirilmiş varlıklar ve alınacak aksiyon aşağıdaki formatta hazırlanmıştır.

- Intent: rota_sorgulama
- User: Ankara'dan İstanbul'a rota oluştur.
- Entities:
    - LOC: Ankara
    - LOC: İstanbul
- Action: rota_sorgulama_aksiyonu

Kullanıcıdan gelecek "Ankara'dan İstanbul'a rota oluştur." talebi veya benzerleri için iki adet "LOC" etiketine sahip isimlendirilmiş varlık bulunması sonucunda rota sorgulama aksiyonu tetiklenecektir. Bu aksiyon ile uzak bir konum belirleme ve

rota oluşturma işlemleri için hizmet veren bir sunucuya istek atılmaktadır. İlk olarak bu konumlar için coğrafik konum bilgileri elde edilmekte, ardından bu bilgiler kullanılarak rota oluşturma işlemleri için tekrar sunucuya istekte bulunulmaktadır. Bu isteğe karşılık güzergahı oluşturacak konum noktaları elde edilmektedir. Daha sonrasında bu noktalar arayüz üzerinde gösterilebilmektedir.

DIET mimarisine ek olarak, isimlendirilmiş varlık tanıma işlemlerinin daha başarılı bir şekilde yapılabilmesi için çok daha fazla örnek ile eğitilmiş bir BERT modeli [66] de sisteme harici olarak entegre edilmiştir. Böylelikle konum belirleme işlemlerinin daha yüksek doğrulukla yapılması sağlanabilmektedir.

8
SONUÇ VE ÖNERİLER

Bu tez çalışması kapsamında; dinleme, anlama ve konuşma özellikleri ile insanları andıran, günlük yaşamda birtakım işlemlerin eller serbest bir şekilde yapılmasına imkân sağlayan sanal asistan uygulamalarının bileşenlerin analiz edilmiş ve araç ortamlarında kullanılabilecek bir sanal asistan uygulaması hayata geçirilmiştir.

Sanal asistanların konuşma sinyallerinin metne dönüştürülmesini sağlayan otomatik konuşma tanıma bileşeni, insanlardaki dinleme yetisine karşılık gelmektedir. Konuşma sinyallerinin yazıya dökülebilmesi için literatürde farklı yaklaşımlar bulunmaktadır. Bu çalışmada, ağırlıklı sonlu durum dönüştürücüler üzerine kurgulanan konuşma tanıma mimarisi ile uçtan uca konuşma tanıma mimarilerinden olan QuartzNet kullanılarak Türkçe modeller oluşturulmuş ve test edilmiştir. Yapılan deneyler sonucunda QuartzNet modelinin daha düşük kelime hata oranına sahip olduğu ortaya çıksa da mikrofondan akan veri ile yapılan deney sonucunda QuartzNet modelinin kelimeleri başarılı bir şekilde oluşturamadığı gözlemlenmiştir. Kaldi Konuşma Aracı ile üretilen; akustik model, dil modeli ve okunuş sözlüğü modellerinin ağırlıklı sonlu durum dönüştürücüler üzerine kurgulandığı model ise akan veride QuartzNet'e kıyasla çok daha başarılı sonuçların ortaya çıkmasını sağlamıştır. Sonuç olarak, canlı konuşma tanıma sistemleri için ağırlıklı sonlu durum dönüştürücü yaklaşımının daha doğru bir tercih olacağı gözlemlenmiştir.

İnsanlardaki konuşma yetisine karşılık gelen ve metinlerin ses sinyallerine dönüştürülmesini sağlayan konuşma sentezleme bileşenin gerçek zamanda çalışabilmesi sanal asistanlar için oldukça önemlidir. Bu noktada, metin girdileri ile mel-spektrogramların oluşturulmasında FastSpeech2 mimarisi kullanılmıştır. Mel spektrogram girdilerinin ses sinyellerine dönüştürülmesinde ise Multi-band MelGAN mimarisi kullanılmıştır. Bu iki sinir ağının kullanımı ile gerçek zamanda metinlerin seslendirilmesi sağlanmıştır.

Bu çalışma kapsamında metinlerdeki niyet ve varlıkların saptanması noktasında DIET ve BERT mimarilerinden faydalanılmıştır. DIET, isimlendirilmiş varlık tanıma ve niyet sınıflandırma problemlerini tek bir sinir ağı mimarisi ile çözebilmektedir. Tercihen, yapılacak işlemlerin önemine göre DIET modelini destekleyecek doğal dil anlama yöntemlerinin kullanılabilmesi de mümkündür. Örneğin, araç ortamlarında kullanılacak bir akıllı asistan uygulaması için yer ismi bildiren kelimelerin bulunabilmesi oldukça önemlidir. Bunun için çok büyük veri kümeleri ile eğitilmiş harici isimlendirilmiş varlık tanıma modellerinin kullanılması navigasyon benzeri uygulamaların daha etkili bir biçimde kullanılmasını sağlayacaktır. Bu çalışmada da; yer, kişi ve kuruluş isimlerinin bulunmasını sağlayacak harici bir BERT modeli [66] kullanılmıştır.

Oluşturulan bütün modeller, birbirlerinden bağımsız olarak çalışabilecek birer mikro servis olarak tasarlanmıştır. Oluşturulan servislerin birbiri ile uyumlu bir şekilde çalışmasını sağlayacak bir arayüz tasarlanarak bir uygulama ortaya çıkarılmıştır.

Bu çalışma kapsamında araç ortamı için kurgulanan uygulamanın benzer yöntemler ile farklı alanlar için de kurgulanabilmesi mümkündür. Bu noktada niyet sınıflandırma ve isimlendirilmiş varlık tanıma modellerinin uygun veriler ile eğitilmesi gerekmektedir.

Sistem, anlatıldığı hali ile düşük seviyeli donanımlar üzerinde performanslı şekilde çalışmaya uygun değildir. Modellerin düşük seviyeli donanımlarda da çalışabilmesi için derin öğrenme modellerindeki ağırlıkların nicemleme işlemi ile Float32'den Int8'e indirilmesi mümkündür. Böylelikle model boyutu ve bellek kullanımı 4 kat azalarak düşük seviyeli donanımlarda da çalışabilir hale getirilebilmektedir.

KAYNAKÇA

[1] A. M. Turing, “I.—COMPUTING MACHINERY AND INTELLIGENCE”, *Mind*, c. LIX, sy 236, ss. 433-460, Eki. 1950, doi: 10.1093/mind/LIX.236.433.

[2] J. Weizenbaum, “ELIZA—a computer program for the study of natural language communication between man and machine”, *Commun. ACM*, c. 9, sy 1, ss. 36-45, Oca. 1966, doi: 10.1145/365153.365168.

[3] K. M. Colby, A. P. Goldstein, ve L. Krasner, "*Artificial Paranoia: a Computer Simulation of Paranoid Processes."* Saint Louis: Elsevier Science, 2014. Erişim: 30 Nisan 2022. [Çevrimiçi]. Erişim adresi: http://qut.eblib.com.au/patron/FullRecord.aspx?p=1838570

[4] “jabberwacky - About Thoughts - An Artificial Intelligence AI chatbot, chatterbot or chatterbox, learning AI, database, dynamic - models way humans learn - simulate natural human chat - interesting, humorous, entertaining”. http://www.jabberwacky.com/j2about (erişim 01 Mayıs 2022).

[5] R. S. Wallace, “The Anatomy of A.L.I.C.E.”, içinde *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, R. Epstein, G. Roberts, ve G. Beber, Ed. Dordrecht: Springer Netherlands, 2009, ss. 181-210. doi: 10.1007/978-1-4020-6710-5_13.

[6] E. Adamopoulou ve L. Moussiades, “Chatbots: History, technology, and applications”, *Mach. Learn. Appl.*, c. 2, s. 100006, Ara. 2020, doi: 10.1016/j.mlwa.2020.100006.

[7] “Siri”, *Apple (Türkiye)*. https://www.apple.com/tr/siri/ (erişim 01 Mayıs 2022).

[8] D. A. Ferrucci, “Introduction to ‘This is Watson’”, *IBM J. Res. Dev.*, c. 56, sy 3.4, s. 1:1-1:15, May. 2012, doi: 10.1147/JRD.2012.2184356.

[9] “Google Asistan - Kişisel Google’ınız”, *Assistant*. https://assistant.google.com/intl/tr_tr/ (erişim 01 Mayıs 2022).

[10] “Cortana - Your personal productivity assistant”. https://www.microsoft.com/en-us/cortana (erişim 01 Mayıs 2022).

[11] "Amazon Alexa – Learn what Alexa can do | Amazon.com”. https://www.amazon.com/b?ie=UTF8&node=21576558011 (erişim 01 Mayıs 2022).

[12] T. F. Quatieri, "*Discrete-time speech signal processing: principles and practice*." Upper Saddle River, NJ: Prentice Hall, 2002.

[13] X. Huang, A. Acero, ve H.-W. Hon, "*Spoken language processing: a guide to theory, algorithm, and system development*." Upper Saddle River, NJ: Prentice Hall PTR, 2001.

[14] G. E. Peterson ve H. L. Barney, “Control Methods Used in a Study of the Vowels”, s. 11.

[15] Shree, *cs224s*. 2022. Erişim: 02 Nisan 2022. [Çevrimiçi]. Erişim adresi: https://github.com/sknadig/cs224s/blob/7b1dd38225768288085f184fdcc93bafe8512ea0/224s.17.lec2.pdf

[16] P. H. Lindsay ve D. A. Norman, "*Human information processing: An introduction to psychology"*, 2d ed. New York: Academic Press, 1977.

[17] D. W. Paul Boersma, "Praat: doing Phonetics by Computer". https://www.fon.hum.uva.nl/praat/ (erişim 17 Nisan 2022).

[18] J. R. Pierce, "Whither Speech Recognition?", *J. Acoust. Soc. Am.*, c. 46, sy 4B, ss. 1049-1051, Eki. 1969, doi: 10.1121/1.1911801.

[19] M. A. Fox, C. J. Aschkenasi, ve A. Kalyanpur, "Voice recognition is here comma like it or not period", *Indian J. Radiol. Imaging*, c. 23, sy 3, ss. 191-194, Tem. 2013, doi: 10.4103/0971-3026.120252.

[20] B. T. Lowerre, "THE HARPY SPEECH RECOGNITION SYSTEM", s. 19.

[21] L. R. Rabiner ve B. H. Juang, "An Introduction to Hidden Markov Models", s. 12.

[22] H. Shimodaira ve S. Renals, "Speech Signal Analysis", s. 40.

[23] D. Jurafsky ve J. H. Martin, "*Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*", 2nd ed. Upper Saddle River, N.J: Pearson Prentice Hall, 2009.

[24] D. Jurafsky ve J. H. Martin, "*Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*." 2022. [Çevrimiçi]. Erişim adresi: https://web.stanford.edu/~jurafsky/slp3/ed3book_jan122022.pdf

[25] L. Muda, M. Begam, ve I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", c. 2, sy 3, s. 6, 2010.

[26] P. Jyothi, "lecture1.pdf". https://www.cse.iitb.ac.in/~pjyothi/cs753/slides/lecture1.pdf (erişim 29 Mayıs 2022).

[27] D. Povey *vd.*, "The Kaldi Speech Recognition Toolkit", s. 4.

[28] P. Jyothi, "lecture6.pdf". https://www.cse.iitb.ac.in/~pjyothi/cs753/slides/lecture6.pdf (erişim 11 Mayıs 2022).

[29] M. Mohri, F. Pereira, ve M. Riley, "Speech Recognition with Weighted Finite-State Transducers", içinde *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, ve Y. A. Huang, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, ss. 559-584. doi: 10.1007/978-3-540-49127-9_28.

[30] G. Hinton *vd.*, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups", *IEEE Signal*

Process. Mag., c. 29, sy 6, ss. 82-97, Kas. 2012, doi: 10.1109/MSP.2012.2205597.

[31] H. Sak, A. Senior, ve F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition", *ArXiv14021128 Cs Stat*, Şub. 2014, Erişim: 11 Mayıs 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1402.1128

[32] V. Peddinti, D. Povey, ve S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts", içinde *Interspeech 2015*, Eyl. 2015, ss. 3214-3218. doi: 10.21437/Interspeech.2015-647.

[33] Z. G. Uslu, "SESLİ TIBBİ RAPORLARIN YAZIYA DÖNÜŞTÜRÜLMESİ", Yıldız Teknik Üniversitesi, 2019.

[34] M. Vyas, "A Gaussian Mixture Model Based Speech Recognition System Using Matlab", *Signal Image Process. Int. J.*, c. 4, ss. 109-118, Ağu. 2013, doi: 10.5121/sipij.2013.4409.

[35] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, ve K. J. Lang, "Phoneme recognition using time-delay neural networks", *IEEE Trans. Acoust. Speech Signal Process.*, c. 37, sy 3, ss. 328-339, Mar. 1989, doi: 10.1109/29.21701.

[36] H. Shimodaira ve S. Renals, "Hidden Markov Models and Gaussian Mixture Models", s. 81.

[37] A. Graves, S. Fernandez, F. Gomez, ve J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks", s. 8.

[38] Z. Xiao, Z. Ou, W. Chu, ve H. Lin, "Hybrid CTC-Attention based End-to-End Speech Recognition using Subword Units", arXiv, arXiv:1807.04978, Eyl. 2018. doi: 10.48550/arXiv.1807.04978.

[39] J. Li *vd.*, "Jasper: An End-to-End Convolutional Neural Acoustic Model", arXiv, arXiv:1904.03288, Ağu. 2019. doi: 10.48550/arXiv.1904.03288.

[40] S. Kriman *vd.*, "QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions", *ArXiv191010261 Eess*, Eki. 2019, Erişim: 11 Mayıs 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1910.10261

[41] C.-F. Wang, "A Basic Introduction to Separable Convolutions", *Medium*, 14 Ağustos 2018. https://towardsdatascience.com/a-basic-introduction-to-separable-convolutions-b99ec3102728 (erişim 24 Mayıs 2022).

[42] L. Lee, C. Tseng, ve C. Hsieh, "Improved tone concatenation rules in a formant-based Chinese text-to-speech system", *IEEE Trans. Speech Audio Process.*, c. 1, sy 3, ss. 287-294, Tem. 1993, doi: 10.1109/89.232612.

[43] A. Pradhan, A. S. S, A. Prakash, K. Veezhinathan, ve H. Murthy, "A syllable based statistical text to speech system", içinde *21st European Signal Processing Conference (EUSIPCO 2013)*, Eyl. 2013, ss. 1-5.

[44] A. van den Oord *vd.*, “WaveNet: A Generative Model for Raw Audio”, *ArXiv160903499 Cs*, Eyl. 2016, Erişim: 25 Ocak 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1609.03499

[45] C. Jemine, “Master thesis : Automatic Multispeaker Voice Cloning”, s. 38.

[46] Y. Wang *vd.*, “Tacotron: Towards End-to-End Speech Synthesis”, *ArXiv170310135 Cs*, Nis. 2017, Erişim: 25 Ocak 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1703.10135

[47] J. Shen *vd.*, “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions”, *ArXiv171205884 Cs*, Şub. 2018, Erişim: 25 Ocak 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1712.05884

[48] Y. Ren *vd.*, “FastSpeech 2: Fast and High-Quality End-to-End Text to Speech”, *ArXiv200604558 Cs Eess*, Mar. 2021, Erişim: 25 Ocak 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/2006.04558

[49] Y. Ren *vd.*, “FastSpeech: Fast, Robust and Controllable Text to Speech”, *ArXiv190509263 Cs Eess*, Kas. 2019, Erişim: 25 Ocak 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1905.09263

[50] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, ve J. Xiao, “Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow”, içinde *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May. 2020, ss. 7209-7213. doi: 10.1109/ICASSP40776.2020.9054484.

[51] H. Lu *vd.*, “VAENAR-TTS: Variational Auto-Encoder Based Non-AutoRegressive Text-to-Speech Synthesis”, içinde *Interspeech 2021*, Ağu. 2021, ss. 3775-3779. doi: 10.21437/Interspeech.2021-2121.

[52] A. Vaswani *vd.*, “Attention Is All You Need”, *ArXiv170603762 Cs*, Ara. 2017, Erişim: 02 Nisan 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1706.03762

[53] J. L. Ba, J. R. Kiros, ve G. E. Hinton, “Layer Normalization”, *ArXiv160706450 Cs Stat*, Tem. 2016, Erişim: 04 Mayıs 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1607.06450

[54] S. Mehri *vd.*, “SampleRNN: An Unconditional End-to-End Neural Audio Generation Model”, *ArXiv161207837 Cs*, Şub. 2017, Erişim: 07 Mayıs 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1612.07837

[55] N. Kalchbrenner *vd.*, “Efficient Neural Audio Synthesis”, *ArXiv180208435 Cs Eess*, Haz. 2018, Erişim: 25 Ocak 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1802.08435

[56] I. J. Goodfellow *vd.*, “Generative Adversarial Networks”, *ArXiv14062661 Cs Stat*, Haz. 2014, Erişim: 25 Ocak 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1406.2661

[57] P. Neekhara, C. Donahue, M. Puckette, S. Dubnov, ve J. McAuley, “Expediting TTS Synthesis with Adversarial Vocoding”, *ArXiv190407944 Cs Eess*, Tem. 2019, Erişim: 07 Mayıs 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1904.07944

[58] K. Kumar *vd.*, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis", *ArXiv191006711 Cs Eess*, Ara. 2019, Erişim: 25 Ocak 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1910.06711

[59] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, ve L. Xie, "Multi-band MelGAN: Faster Waveform Generation for High-Quality Text-to-Speech", *ArXiv200505106 Cs Eess*, Kas. 2020, Erişim: 25 Ocak 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/2005.05106

[60] D. Adiwardana *vd.*, "Towards a Human-like Open-Domain Chatbot", *ArXiv200109977 Cs Stat*, Şub. 2020, Erişim: 09 Nisan 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/2001.09977

[61] J. Devlin, M.-W. Chang, K. Lee, ve K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *ArXiv181004805 Cs*, May. 2019, Erişim: 02 Nisan 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1810.04805

[62] T. Mikolov, K. Chen, G. Corrado, ve J. Dean, "Efficient Estimation of Word Representations in Vector Space", *ArXiv13013781 Cs*, Eyl. 2013, Erişim: 08 Mayıs 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1301.3781

[63] "Math with Words – Word Embeddings with MATLAB and Text Analytics Toolbox", *Loren on the Art of MATLAB*. https://blogs.mathworks.com/loren/2017/09/21/math-with-words-word-embeddings-with-matlab-and-text-analytics-toolbox/ (erişim 08 Mayıs 2022).

[64] M. Schuster ve K. Nakajima, "Japanese and Korean Voice Search", içinde *International Conference on Acoustics, Speech and Signal Processing*, 2012, ss. 5149-5152.

[65] "wikiann · Datasets at Hugging Face". https://huggingface.co/datasets/wikiann (erişim 10 Mayıs 2022).

[66] "savasy/bert-base-turkish-ner-cased · Hugging Face". https://huggingface.co/savasy/bert-base-turkish-ner-cased (erişim 10 Mayıs 2022).

[67] "Open source conversational AI", *Rasa*, 01 Aralık 2020. https://rasa.com/ (erişim 10 Mayıs 2022).

[68] T. Bunk, D. Varshneya, V. Vlasov, ve A. Nichol, "DIET: Lightweight Language Understanding for Dialogue Systems", *ArXiv200409936 Cs*, May. 2020, Erişim: 02 Nisan 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/2004.09936

[69] J. Pennington, R. Socher, ve C. Manning, "Glove: Global Vectors for Word Representation", içinde *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, ss. 1532-1543. doi: 10.3115/v1/D14-1162.

[70] M. Henderson, I. Casanueva, N. Mrkšić, P.-H. Su, T.-H. Wen, ve I. Vulić, "ConveRT: Efficient and Accurate Conversational Representations from

Transformers”, *ArXiv191103688 Cs*, Nis. 2020, Erişim: 10 Mayıs 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1911.03688

[71] J. Lafferty, A. McCallum, ve F. C. N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, s. 10.

[72] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, ve C. Dyer, “Neural Architectures for Named Entity Recognition”, *ArXiv160301360 Cs*, Nis. 2016, Erişim: 10 Mayıs 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1603.01360

[73] “Home - Docker”, 10 Mayıs 2022. https://www.docker.com/ (erişim 13 Mayıs 2022).

[74] “Flutter - Build apps for any screen”. //flutter.dev/ (erişim 12 Mayıs 2022).

[75] “Dart programming language | Dart”. https://dart.dev/ (erişim 12 Mayıs 2022).

[76] "*Kaldi Speech Recognition Toolkit*." Kaldi, 2022. Erişim: 12 Mayıs 2022. [Çevrimiçi]. Erişim adresi: https://github.com/kaldi-asr/kaldi

[77] “Welcome to Python.org”, *Python.org*. https://www.python.org/ (erişim 12 Mayıs 2022).

[78] “Welcome to Flask — Flask Documentation (2.1.x)”. https://flask.palletsprojects.com/en/2.1.x/ (erişim 12 Mayıs 2022).

[79] T. Bocklisch, J. Faulkner, N. Pawlowski, ve A. Nichol, “Rasa: Open Source Language Understanding and Dialogue Management”, *ArXiv171205181 Cs*, Ara. 2017, Erişim: 02 Nisan 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1712.05181

[80] Saraçlar, Murat, “Turkish Broadcast News Speech and Transcripts”. inguistic Data Consortium, s. 14566923 KB, 16 Mayıs 2012. doi: 10.35111/ZXEV-1K65.

[81] “Mozilla Common Voice”. https://commonvoice.mozilla.org/ (erişim 15 Mayıs 2022).

[82] K. Ito ve L. Johnson, “The LJ Speech Dataset”. 2017. [Çevrimiçi]. Erişim adresi: https://keithito.com/LJ-Speech-Dataset/

[83] E. Ergün ve T. Yıldırım, “Is it possible to train a Turkish text-to-speech model with English data?”, *RASE Recent Adv. Sci. Eng.*, c. 2, sy 1, s. 5, 2022, doi: 10.14744/rase.2022.0001.

[84] H. Zen *vd.*, “LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech”, *ArXiv190402882 Cs Eess*, Nis. 2019, Erişim: 25 Ocak 2022. [Çevrimiçi]. Erişim adresi: http://arxiv.org/abs/1904.02882

[85] Salor, Ozgul, Ciloglu, Tolga, Pellom, Bryan, ve Demirekler, Mubeccel, “Middle East Technical University Turkish Microphone Speech v 1.0”. Linguistic Data Consortium, s. 680960 KB, 18 Mayıs 2006. doi: 10.35111/SK8B-SS58.

TEZDEN ÜRETİLMİŞ YAYINLAR

Makaleler

[1] E. Ergün ve T. Yıldırım, “Is it possible to train a Turkish text-to-speech model with English data?”, *RASE Recent Adv. Sci. Eng.*, c. 2, sy 1, s. 5, 2022, doi: 10.14744/rase.2022.0001.