
GAZİ ÜNİVERSİTESİ
Gazi
1926

DERİN ÖĞRENME TABANLI UÇTAN UCA TÜRKÇE KONUŞMA SENTEZLEME SİSTEMİ

Mustafa Sami CÜCEN

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TEMMUZ 2023

ETİK BEYAN

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

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

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

Mustafa Sami CÜCEN

12/07/2023

DERİN ÖĞRENME TABANLI UÇTAN UCA TÜRKÇE KONUŞMA SENTEZLEME SİSTEMİ

(Yüksek Lisans Tezi)

Mustafa Sami CÜCEN

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Temmuz 2023

ÖZET

Konuşma sentezleme sistemi, insan benzeri doğal konuşmaları üretmek için geliştirilen bir yapay zeka teknolojisidir. Bu sistem, metin girdilerini alır ve bunları gerçekçi ve akıcı sesli çıktılara dönüştürmek için derin öğrenme algoritmalarını kullanır. Bu çalışmada öncelikle derin öğrenme modelini eğitmek için, Türkçe doğal konuşma örnekleri üzerinde kapsamlı bir veri toplama süreci gerçekleştirilmiştir. Bu veriler, bir genç erkek konuşmacı tarafından İstanbul Türkçesi olarak kaydedilen konuşma örneklerini içermektedir. Bu veriler yaklaşık 13 saat uzunluğundadır. Daha sonra veri kümesi kullanılarak GlowTTS mimarisi ile her biri 261 adım olan 500 çevrimlik model ile eğitilerek derin öğrenme tabanlı bir Türkçe konuşma sentezleme sistemi geliştirilmiştir. Geliştirilen konuşma sentezleme sisteminin performansı farklı ölçütlerle değerlendirilmiştir. Ortalama görüş puanı deneyi, spektrogramların değerlendirilmesi, çapraz korelasyon ve sanal konuşma kalitesi nesnel dinleyici (SKKND) testleri kullanılarak sistemin başarısı analiz edilmiştir. Elde edilen sonuçlara göre, sistemin OGP 2,79, çapraz korelasyon değeri 51,09 ve SKKND puanı 2,32 olarak belirlenmiştir. OGP, kullanıcıların konuşma kalitesini değerlendirmesiyle ortaya çıkan bir ölçüt olarak sistemin tatmin edici bir performans sergilediğini göstermektedir. Çapraz korelasyon değeri ise orijinal ses ve sentezlenen ses arasındaki benzerliğin ortalama olduğunu göstermektedir. SKKND puanı ise konuşmanın algısal kalitesini değerlendiren bir ölçüt olarak sistem tarafından üretilen konuşmanın tatmin edici olduğunu göstermektedir. Bu çalışma, Türkçe konuşma sentezleme sistemlerinin performansını değerlendirmek için nesnel ölçütlerin kullanılabileceğini göstermektedir. Sonuçlar, gelecekteki çalışmalarda sistem iyileştirmelerine ve kullanıcı deneyimini artırmaya yönelik önemli bilgiler sağlamaktadır. Bu sistemin daha önceki Türkçe konuşma sentezleme çalışmalarında karşılaşılan doğallık ve anlaşılırlık sorunlarına çözüm getirmesi amaçlanmıştır.

Bilim Kodu	:	92432
Anahtar Kelimeler	:	Derin Öğrenme, Konuşma Sentezleme, Doğal Dil İşleme, Yapay Zeka, Konuşma Veri Kümesi, GlowTTS, Mel Spektrogram, Konuşma Sentezleme Değerlendirmesi, Çapraz Korelasyon
Sayfa Adedi	:	94
Danışman	:	Doç. Dr. Hüseyin POLAT

DEEP LEARNING BASED END TO END TURKISH SPEECH SYNTHESIS SYSTEM

(M. Sc. Thesis)

Mustafa Sami CÜCEN

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

July 2023

ABSTRACT

The speech synthesis system is an artificial intelligence technology developed to generate human-like natural conversations. This system takes text inputs and utilizes deep learning algorithms to convert them into realistic and fluent vocal outputs. In this study, an extensive data collection process was conducted initially to train the deep learning model using Turkish natural speech samples. The dataset comprises speech examples recorded in Istanbul Turkish by a young male speaker, totaling approximately 13 hours. Subsequently, a Turkish speech synthesis system based on deep learning was developed by training 500 iterations of the GlowTTS architecture, each consisting of 261 steps, using the collected dataset. The performance of the developed speech synthesis system was evaluated using different criteria. Experimentation involved Mean Opinion Score (MOS), spectrogram evaluation, cross-correlation, and the virtual speech quality objective listener (ViSQOL) test to analyze the system's success. The results obtained from these evaluations revealed an MOS of 2.79, a cross-correlation value of 51.09, and an ViSQOL score of 2.32 for the system. The MOS indicates that the system achieved satisfactory performance based on user evaluations of speech quality. The cross-correlation value suggests an average similarity between the original and synthesized speech. Furthermore, the ViSQOL score demonstrates that the system-generated speech was perceived as satisfying when assessing its perceptual quality. This study demonstrates the use of objective criteria in evaluating the performance of Turkish speech synthesis systems. The findings provide crucial information for future research, aimed at system improvements and enhancing user experience. The objective of this system is to address the naturalness and intelligibility issues encountered in previous Turkish speech synthesis studies. This master's thesis aims to present the results of a research conducted on a deep learning-based end-to-end Turkish speech synthesis system.

Science Code	:	92432
Key Words	:	Deep Learning, Speech Synthesis, Natural Language Processing, Artificial Intelligence, Speech Dataset, GlowTTS, Mel Spectrogram, Speech Synthesis Evaluation, Cross-Correlation
Page Number	:	94
Supervisor	:	Assoc. Prof. Dr. Hüseyin POLAT

TEŞEKKÜR

Çalışmalarım boyunca yardımı ve katkılarıyla beni yönlendiren ve bu süreçte yardımını hiç esirgemeyen saygıdeğer Doç. Dr. Hüseyin POLAT hocama ve yine kıymetli tecrübelerinden faydalandığım Bilgisayar Mühendisliği Bölümü'ndeki kıymetli tüm hocalarıma teşekkürü bir borç bilirim. Bu çalışmanın gerçekleştirilmesi sürecinde gerek veri kümesinin temini gerekse geliştirilmesi sürecinde desteklerini üzerimden eksik etmeyen Adıyaman Üniversitesi Bilgisayar Mühendisliği bölüm başkanı Dr. Saadin OYUCU'ya verdiği desteklerden dolayı teşekkür ederim.

İÇİNDEKİLER

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

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

Simgeler	Açıklamalar
F	Sinyal frekansı
f_{ste}	Süre Tahmin Edici
$f_{kodlayıcı}$	Metin kodlayıcı fonksiyonu
Fmel	Mel frekansı
Hz	Hertz
$L_{sınıflandırıcı}$	Lineer sınıflandırıcı
n	Örnek sayısı
N	Pencere boyu
OKKH	Ortalama Karesel Kayıp Hatası (Mean Squared Error Loss)
$P_{X(x\|c)}$	Mel spektrogramların koşullu dağılımını dönüştürme fonksiyonu
$P_{Z(z\|c)}$	Mel spektrogramların metni dönüştürme fonksiyonu
T_{mel}	Giriş mel spektrogramların uzunluğu
$T_{metin\ l}$	Giriş metinlerinin uzunluğu
Gd	Gradyanların durdurulması
x(n)	Zaman düzlemi giriş işareti
X[k]	Ayrık Fourier Katsayıları

Kısaltmalar	Açıklamalar
AFD	Ayrık Fourier Dönüşümü (Discreate Fourier Transform)
BOGP	Bozulma Ortalama Görüş Puanı (Degradation Mean Opinion Score)
DDİ	Doğal Dil İşleme (Natural Language Processing)
DİA	Derin İnanç Ağları (Deep Belief Networks)
DİYUKSB	Derin İki Yönlü Uzun Kısa Süreli Bellek (Deep Bidirectional Long Short-Term Memory)
DKYA	Derin Karışım Yoğunluğu Ağları (Deep Mixture Density Networks)
EÖE	Eşzamanlı Örtüşme Ekleme (Synchronized Overlap Add)
ESA	Evrişimli Sinir Ağı (Convolutional Neural Network)
GD	Gradyanların Durdurulması (Stop Gradient)
HFD	Hızlı Fourier Dönüşümü (Fast Fourier Transform)
KBM	Kısıtlayıcı Boltzmann Makineleri (Restrictive Boltzmann Machines)
KKAD	Konuşma Kalitesinin Algısal Değerlendirmesi (Perceptual Evaluation of Speech Quality)
KOGP	Karşılaştırmalı Ortalama Görüş Puanı (Comparative Mean Opinion Score)
KOKH	Kök Ortalama Kare Hatası (Root Mean Square Error)
KS	Konuşma Sentezleme (Text to Speech)
MHA	Monotonik Hizalama Araması (Monotonic Alignment Search)
MSB	Mel Sepstral Bozukluğu (Mel Cepstral Distortion)
MSKU	Microsoft Sistem Konuşma Uygulaması (Microsoft System Speech Application)
NBİÖ	Neurogram Benzerlik İndeksi Ölçümü (Neurogram Similarity Index Measure)
OGP	Ortalama Görüş Puanı (Mean Opinion Score)
OGP-DKH	Ortalama Görüş Puanı- Dinleme Kalitesi Hedefi (Mean Opinion Score - Listening Quality Objective)
OKT	Otomatik Konuşma Tanıma (Automatic Speech Recognition)
PEÖE	Perde Eşzamanlı Örtüşme Ekleme (Pitch Synchronized Overlap Add)
SAA	Ses Aktivite Algılama (Voice Activity Detection)

Kısaltmalar	Açıklamalar
SKKND	Sanal Konuşma Kalitesi Nesnel Dinleyici (Virtual Speech Quality Objective Listener)
SMM	Saklı Markov Modeli (Hidden Markov Model)
SSİ	Sayısal Sinyal İşlemcisi (Digital Signal Processing)
TOYA	Türkçe Otomatik Yazım Algoritması (Turkish Automatic Spelling Algorithm)
YSA	Yapay Sinir Ağları (Artificial Neural Networks)
ZA-PEÖE	Zaman Alanında Perde Eşzamanlı Örtüşme Ekleme (Time Domain Pitch Synchronized Overlap Add)

1. GİRİŞ

İnsanlık geçmişten günümüze geliştirdiği araçlar ile etkileşim içinde bulunmaktadır. Günümüz dünyasında insanlar ve insanlar adına birçok eylemi gerçekleştiren birçok makine ile etkileşim içinde bulunmaktadır. İnsan-makine etkileşimi için günümüzde en çok kullanılan aracılar bilgisayarlardır. Bu kapsamda son zamanlarda yaşanan teknolojik gelişmeler sonucunda, bilgisayarların basit insan kullanımından, insanlar adına eylemler gerçekleştiren otonom araçlarla vasıtasıyla insan ilişkilerinin kurulmasına doğru giderek kaymaktadır [1]. Bu çerçevede de insanlar ile bilgisayar arasındaki etkileşimi en üst düzeye çıkarmak için Doğal Dil İşleme (DDİ: Natural Language Processing) üzerine çalışmalar yapılmaktadır. Doğal dil işleme kullanılarak etkileşimi etkili hale getirmek için temelde iki farklı teknolojinin geliştirilmesi gerekir. Bunlardan ilki, konuşmanın metne aktarılıp anlamlandırıldığı Otomatik Konuşma Tanıma (OKT: Automatic Speech Recognition) sistemleridir. İkinci teknoloji ise yazılı metinlerin sese dönüştürüldüğü Konuşma Sentezleme (KS: Text to Speech) sistemleridir. Bu teknolojiler üzerine yapılan çalışmalarda derin öğrenme, makine öğrenmesi, istatiksel analiz ve kural tabanlı yaklaşımlar hibrit biçimde kullanılmaktadır.

Konuşma ve doğal dil işleme alanındaki son araştırmalar, makinelerin bir insan gibi doğal dili kullanmasına imkân sağlamaktadır. İnsan dilini işleme yeteneğine sahip hesaplamalı modeller geliştirilmesi, insanların doğal dili nasıl edindiği, depoladığı ve işlediği bilgisinin makinelere öğretilmesidir [2]. Konuşma sentezleme sistemleri, doğal dil metinlerini girdi olarak alarak gerçekçi, doğal ve insan benzeri konuşma sinyalleri üretmeyi amaçlayan sistemlerdir. Konuşma sentezleme sistemleri akustik, dilbilim, sinyal işleme, istatistik ve yapay zekâ gibi birçok farklı disiplinin birlikte kullanılması ile geliştirilmektedir [3]. Konuşma sentezi üzerine yapılan araştırmalarda insan sesinin doğallığına yakın konuşma sentezinin üretilmesi hedeflenmektedir.

Konuşma sentezleme sistemleri, özellikle günümüz bilişim teknolojilerini benimseyen insanlara sosyal yaşamlarında farkındalık oluşturabilecek birçok yeni uygulama alanı sunmaktadır.

Görme veya konuşma engeli olan insanlar için, sesli yanıt veya uyarı sistemleri konuşma

sentezleme sistemlerinin uygulama alanlarından biridir. Bu kapsamda cep telefonları, bilgisayarlar gibi cihazların ekran içeriğini analiz eden ve seslendiren yazılım uygulaması olan ekran okuyucular ile karşılaşılmaktadır.

Eğitim alanı konuşma sentezleme sisteminin uygulama alanlarından bir başkasıdır. Örneğin; insanların ana dil veya yabancı dil kullanımında kelimelerin ve cümlelerin doğru telaffuzunu güçlendirmek için konuşulan dile vurgu yaparak okuryazarlık becerilerinin, yazmanın ve okumanın eğitimini kolaylaştırmak amaçlanmıştır. Çeşitli sesli kitap uygulamaları ile örnek uygulama olarak karşılaşılabilmektedir.

Konuşma sentezleme sistemlerinin bir başka uygulaması insan kullanımı için etkileşimli bilgi işlem sistemlerinin tasarımı, değerlendirilmesi ve uygulanması ile ilgili bir disiplin olan insan-bilgisayar etkileşimli sistemlerdir. Konuşma sentezleme sistemleri aracılığıyla insan-bilgisayar etkileşimli sistemler kabul edilebilir ve etkileşimli olarak kullanıcı dostu hale getirilmeye çalışılmıştır.

Görsel ve işitsel bilgilerin işlenmesini, depolanmasını, iletilmesini ve sunulmasını ve sistemler ile kullanıcıları arasında doğal arayüzlerin kurulmasını içeren telekomünikasyon ve multimedya uygulamaları da konuşma sentezleme sistemlerinin kullanıldığı alanlardandır. Konuşma sentezleme sistemleri en yeni uygulamaları, mobil cihazlar veya internet arama motorları için insanların kendi ana dillerini kullanarak makinelerle etkileşime girebileceği farklı faydalı uygulamalar sağlamak için kullanılmaktadır.

Konuşma sentezleme sistemlerinin çeşitli oyun ve oyuncaklarda da kullanılabilmektedir. Bu sistem, çocukların öğrenmesi için farklı uygulamalar tasarlanarak farklı oyunlara veya oyuncaklara dahil edilebilmektedir. Bu kapsamda akıllı telefonlar gibi cihazlarda çalışacak modern 3D uygulamalardan konuşan oyuncaklara kadar geliştirilmiş teknolojik ürünler mevcuttur.

Türkçe için günümüzde Türkçe için Windows, Android ve iOS gibi işletim sistemlerinin ekran okuma için çeşitli özellikler mevcuttur. Google ve Yandex gibi çeviri ve seslendirme özelliği olan firmaların uygulamaları Türkçe konuşma sentezleme sistemlerinin örneğidir. Firmaların arama merkezleri ile yapılan görüşmelerde karşılaşılan ilk sesler genellikle konuşma sentezleme sistemlerin bir ürünü olarak kullanıcıların karşısına çıkmaktadır. Yine

kitap, dergi gibi metinlerin elektronik ortamda seslendirmeleri genellikle konuşma sentezleme sistemlerinin uygulamasıdır. Bunların yanında Amazon, Google, Microsoft, Mozilla gibi büyük firmaların yanında çeşitli uygulama geliştiriciler web veya mobil uygulaması olarak kullanıcılarına Türkçe konuşma sentezleme deneyimi sunmaktadır.

Yapılan çalışmalar sonucunda insan sesinin doğallığına yakın sentezlenmiş konuşmanın elde edilmesi, konuşma sentezleme sistemi uygulamalarının insanlar tarafından kullanımının artıracağı tahmin edilmektedir. İnsan makine etkileşimi, nesnelerin interneti, çağrı merkezleri gibi alanlarda sıklıkla kullanılacağı ön görülmektedir.

Bu çalışmada, derin öğrenme tabanlı bir Türkçe konuşma sentezleme (Deep Learning-Based Synthesis) sistemi geliştirilmiştir. Derin öğrenme tabanlı sentezleme yönteminde veri kümesi olarak derlem tipi hazırlanmış veri kümeleri kullanılmaktadır. Türkçe derlem tipi hazırlanan veri kümeleri için aşılması gereken çeşitli zorluklar mevcuttur. Türkçe, Ural-Altay dilleri grubuna bağlı sondan eklemeli yapısından dolayı modellenmesi oldukça zor olan bir dildir. Ayrıca Türkçe'de bir kelimenin birkaç farklı telaffuzlarının olabilmesi durumunda, konuşma sentezleme sisteminin hangi telaffuzun doğru olduğuna karar vermesi konusunda da bir zorluk yaşanmaktadır. Türkçe için diğer bir zorluk, metinlerin içerdiği çok sayıda sesteş kelime, sayı ve kısaltmadan dolayı metin normalleştirmedir. Bununla beraber vurgu ve tonlamalarda ve ses yumuşaması, ses sertleşmesi, ses düşmesi, ses eklemesi ve ulama gibi ses değişimlerinde zorluklar yaşanmaktadır. Türkçe için bir başka zorluk, bürünsel (prosodic) analiz sırasında kullanılacak dönüşüm yöntemidir. Tüm bu zorlukların yanında Türkçe veri kümelerinin kolay ulaşılabilir olmaması Türkçe konuşma sentezleme sistemlerinin geliştirilmesini zorlaştırmaktadır. Türkçe için bir dizi konuşma sentezleme çalışması mevcut olsa da Türkçe için konuşma sentezleme sistemindeki eğilimler hakkında genel bir bilgi kaynağının olmaması, bu konudaki araştırmalara yeni başlayanlar için umut kırıcı olmaktadır.

Biçimlendirici (Formant Synthesis), söyleyiş (Articulatory Synthesis) ve birleştirmeli (Concatenation Synthesis) konuşma sentezleme yöntemleri göz önünde bulundurularak yapılan değerlendirmeler de Türkçe dili için birleştirmeli (eklemeli) konuşma sentezleme sistemlerinin anlaşılırlık için iyi sonuçlar verdiği görülmüştür. Ancak aynı yöntemin kullanıldığı konuşma sentezleme sistemlerinde ses birimleri arasındaki geçişlerde ve konuşma doğallığında sorunlar yaşanmaktadır. Bundan sebepten bu çalışmada, doğallık

sorunun önüne geçmek için başka dillerde başarılı sonuç alınan derin öğrenme tabanlı konuşma sentezleme yöntemi tercih edilmiştir.

Bu çalışmada, derin öğrenme tabanlı bir Türkçe konuşma sentezleme sistemi geliştirilmiştir. Bu sistemin daha önceki Türkçe konuşma sentezleme çalışmalarında karşılaşılan doğallık ve anlaşılırlık sorunlarına çözüm getirmesi amaçlanmıştır.

Bu tez çalışması beş bölümden oluşmaktadır. Bu kapsamda ilk bölümde konuşma sentezleme sistemlerinin tanımı, uygulama alanları, yöntemleri ve yapılan çalışmanın amacı verilmiştir, ikinci bölümde konuşma sentezleme sistemlerinin tarihsel süreci incelenmiş ve Türkçe için yapılmış çalışmaların geniş bir literatür taraması verilmiştir. Üçüncü bölümde çalışmada kullanılan materyaller ve metotlar detaylı olarak sunulmuştur. Dördüncü bölümde yapılan çalışmanın deneysel süreci ve bu süreç sonunda yapılan öznel ve nesnel testlerden elde edilen sonuçlardan bahsedilmiştir. Beşinci bölümde ise genel olarak bu tez çalışmasından elde edilen sonuçlar yorumlanmış ve çeşitli öneriler sunulmuştur.

2. LİTERATÜR TARAMASI

Konuşma sentezinin ilk uygulamalarında insan sesinin taklit edilmesi amacıyla bazı ilkel makine ve modeller geliştirilmiştir [4]. 1791'de Wolfgang von Kempelen sadece harflerin değil tam sözcüklerin de üretilebilir olduğunu göstermiştir. Kempelen bir akustik konuşma makinesi geliştirmiştir [5]. Bilim insanları 1930'lu yıllara kadar Kempelen'in geliştirdiği makine üzerinde çalışmalar yapmıştır. 1930'lu yıllarda Bell Laboratuvarlarında konuşmayı otomatik olarak temel tonlarına ve titreşimlerine göre analiz edebilen ses kodlayıcı geliştirilmiştir [4]. Geliştirilen bu kodlayıcı sistem üzerinde ses sentezleme çalışmaları yapan Homer Dudley, Voder isimli ilk elektronik konuşma sentezleme makinesini geliştirmiştir [6]. Voder makinesi insan müdahalesi olmadan konuşma yapma yeteneğine sahip ilk makine olarak bilinmektedir. Mekanik makinelerden elektronik sistemlere geçişten sonra Umeda ve arkadaşları tarafından 1968'de genel İngilizce metin okumayı gerçekleştiren ilk sistem geliştirilmiştir [7].

1970'lerin sonları 1980'lerin başlarında birçok konuşma sentezleme sistemi ticari amaçla üretilmiştir. Yazılımın yanında donanım çözümleri de içeren farklı konuşma sentezleme sistemleri bilgisayarlarda kullanılmaya başlanmıştır. DECtalk, Whistler, MBROLA gibi başarıları dikkate alınabilecek birçok konuşma sentezleme sistemi farklı diller için geliştirilmiştir.

Türkçe konuşma sentezleme sistemi üzerine gerçekleştirilen çalışmalar ise 1990'lı yıllarda başlamıştır [8, 9]. Türkçe konuşma sentezleme sistemleri ile ilgili olarak gerçekleştirilen çalışmalar daha çok veri kümesi hazırlama, dil modeli ve akustik model geliştirme üzerine yoğunlaşmıştır [8, 10].

Çizelge 2.1.Türkçe konuşma sentezleme sistemi çalışmaları

Yıl	Çalışmanın Adı	Yazar Adı	Kurum	Çalışma Türü
1993	A Speech Synthesis System for Turkish Language Based on the Concatenation of Phonemes Taken From a Speaker	İlhan Yaşar Özüm	ODTÜ	Yüksek Lisans Tezi

Çizelge 2.1. (devamı) Türkçe konuşma sentezleme sistemi çalışmaları

Yıl	Çalışmanın Adı	Yazar Adı	Kurum	Çalışma Türü
1994	PC Based Speech Synthesis for Turkish	Kamil Güven	Çukurova Üniversitesi	Yüksek Lisans Tezi
1994	Karma Söz Üretme Yöntemi ile Türkçe Yazılı Metinden Söze Geçme	Murat Servet Erer	İTÜ	Yüksek Lisans Tezi
1998	Text to Speech Synthesizer in Turkish Using Non Parametric Techniques	Kerem Ayhan	ODTÜ	Yüksek Lisans Tezi
1999	Signal Processing Aspect of Test to Speech Synthezier in Turkish	Özgül Salor	ODTÜ	Yüksek Lisans Tezi
2000	Reading Aid for Visually Impaired (a Turkish Text-to-Speech System Development)	Barış Bozkurt	Boğaziçi Üniversitesi	Yüksek Lisans Tezi
2001	Concatenative Speech Synthesis Based on a Sinusoidal Model	Çağla Ömür	ODTÜ	Yüksek Lisans Tezi
2001	Fundamental Frequency Contour Synthesis for Turkish Text to Speech	Erkan Abdullahbeşe	Boğaziçi Üniversitesi	Yüksek Lisans Tezi
2001	An Implementation and Evaluation of Two-Diphone Based Synthesizers for Turkish	Barış Bozkurt Thierry Dutoit	4th ISCA Tutorial and Research Workshop on Speech Synthesis, İskoçya	Konferans Bildirisi
2002	Türkçe Metinden Konuşma Sentezleme	Şifa Özen Özdemir	Hacettepe Üniversitesi	Yüksek Lisans
2002	Turkish Text to Speech System	Barış Eker	Bilkent Üniversitesi	Yüksek Lisans
2002	Duration Analysis and Modeling for Turkish Text-to-Speech Synthesis	Ömer Şayli	Boğaziçi Üniversitesi	Yüksek Lisans Tezi

Çizelge 2.1. (devamı) Türkçe konuşma sentezleme sistemi çalışmaları

Yıl	Çalışmanın Adı	Yazar Adı	Kurum	Çalışma Türü
2002	Automatic Modelling of Turkish Prosody	Banu Oskay	ODTÜ	Yüksek Lisans Tezi
2002	Duration Properties of the Turkish Phonemes	Ömer Şayli Levent M. Arslan A. Sumru Özsoy	11. Uluslararası Türk Dilbilim Konferans	Konferans Bildirisi
2003	A Prosodic Turkish Text-to-Speech Synthesizer	Esra Vural	Sabancı Üniversitesi	Yüksek Lisans Tezi
2003	New Methods for Voice Conversion	Oytun Türk	Boğaziçi Üniversitesi	Yüksek Lisans Tezi
2003	Implementation and Evaluation of a Text-to-Speech Synthesis System for Turkish	Özgül Salor Bryan Pellom Mübeccel Demirekler	EUROSPEECH	Konferans Bildirisi
2004	A Corpus Based Concatenative Speech Synthesis System for Turkish	Haşim Sak	Boğaziçi Üniversitesi	Yüksek Lisans Tezi
2004	A Single Chip Solution for Text-to-Speech Synthesis	Ozan Aktan	Boğaziçi Üniversitesi	Yüksek Lisans Tezi
2004	Duration of Turkish Vowels Revisited	Ebru Arısoy vd.	12th International Conference on Turkish Linguistics,	Konferans
2005	Örnek Bir Dizi Cümle İçin Türkçe Metinden Konuşma Sentezleyici	Asude Karlı	Ankara Üniversitesi	Yüksek Lisans Tezi
2005	Voice Transformation and Development of Related Speech Analysis Tools for Turkish	Özgül Salor	ODTÜ	Doktora Tezi

Çizelge 2.1. (devamı) Türkçe konuşma sentezleme sistemi çalışmaları

Yıl	Çalışmanın Adı	Yazar Adı	Kurum	Çalışma Türü
2005	Modeling Phoneme Durations and Fundamental Frequency Contours in Turkish Speech	Özlem Öztürk	ODTÜ	Doktora Tezi
2005	A Single Chip Solution for Text-to-speech Synthesis	Ozan Aktan I. Faik Başkaya Günhan Dündar	European Conference on Circuit Theory and Design	Konferans Bildirisi
2006	A Corpus-Based Concatenative Speech Synthesis System for Turkish	Haşim Sak Tunga Güngör Yaşar Safkan	Turkish Journal of Electrical Engineering and Computer Sciences	Makale
2007	Görme Engelliler İçin Basılı Doküman Yorumlama ve Seslendirme Sisteminin Gerçekleştirilmesi	Emre Uzun	Karadeniz Teknik Üniversitesi	Yüksek Lisans Tezi
2007	Taşınabilir Cihazlar İçin Türkçe Metinden Konuşma Sentezleme Sistemi	İlker Ünaldı	Hacettepe Üniversitesi	Yüksek Lisans Tezi
2007	Cross Lingual Voice Conversion	Oytun Türk	Boğaziçi Üniversitesi	Doktora Tezi
2008	Türkçe Metinler için Hece Tabanlı Konuşma Sentezleme Sistemi	Rıfat Aşlıyan Korhan Günel	Akademi Bilişim	Konferans
2009	Makine Öğrenme Algoritmalarıyla Türkçe Metin Seslendirme Sistemi Yazılımı	Zeliha Görmez	Fatih Üniversitesi	Yüksek Lisans Tezi
2009	Türkçe Metin Seslendirme	Kenan Güldalı	İTÜ	Yüksek Lisans Tezi
2009	Türkçe Metinden Konuşma Sentezleme Uygulamaları İçin Bir Veri Sözlük Seti ve Yazılım Çerçevesi	Asım Egemen Yılmaz	Gazi Üniversitesi Mühendislik Fakültesi Dergisi	Makale

Çizelge 2.1. (devamı) Türkçe konuşma sentezleme sistemi çalışmaları

Yıl	Çalışmanın Adı	Yazar Adı	Kurum	Çalışma Türü
2010	Türkçe Metinden Konuşma Sentezlemede Yaşanan Sıkıntılar ve Çözüm Yöntemleri	Ş. Murat Canal Sefer Kurnaz Asım Egemen Yılmaz	Havacılık ve Uzay Teknolojileri Dergi	Makale
2010	Türkçe Metin Seslendirme	Tuncay Şentürk	İTÜ	Yüksek Lisans Tezi
2010	Türkçe Metin Seslendirme İçin Doğal Konuşma Sentezleme	Cavit Erdemir	İstanbul Üniversitesi	Yüksek Lisans Tezi
2010	Metinden Konuşma Sentezleme	İbrahim Baran Uslu	TMMOB Elektrik Mühendisleri Odası	Makale
2010	Tek Ses Sentezleme Tümleşik Devresi ile Türkçe Metinden Sentetik Konuşma Üretme	Hakan Tora Çağlar Cengizler	ELECO	Konferans Bildirisi
2010	Türkçe Metinden Konuşma Sentezleme	Yücel Bicil	Sakarya Üniversitesi	Yüksek Lisans Tezi
2011	Beyin Bilgisayar Arayüzleri İçin Türkçe Metinden Konuşma Sentezleme Sistemi	İlhami Sel Davut Hanbay Murat Karabatak	Fırat Üniversitesi Elektrik-Elektronik Bilgisayar Sempozyumu	Konferans Bildirisi
2011	Görme Engelliler İçin Bilgisayar Kullanımının Etkinleştirilmesi, Erişilebilirlik ve Bir Türkçe Hece Tabanlı Konuşma Sentezleme Sisteminin Geliştirilmesi	Güray Arık	Gazi Üniversitesi	Yüksek Lisans Tezi
2012	A small footprint hybrid statistical/unit selection text-to-speech synthesis system for agglutinative languages	Ekrem Güner Cenk Demiroğlu	ICASSP 2012	Konferans Bildirisi

Çizelge 2.1. (devamı) Türkçe konuşma sentezleme sistemi çalışmaları

Yıl	Çalışmanın Adı	Yazar Adı	Kurum	Çalışma Türü
2012	Konuşma İşleme ve Türkçenin Dilbilimsel Özelliklerini Kullanarak Metinden Doğal Konuşma Sentezleme	İbrahim Baran Uslu	Ankara Üniversitesi	Doktora Tezi
2013	Türkçe Metinler İçin Hece Tabanlı Metinden Konuşma Sentezleme Sistemi	İlhami Sel	Fırat Üniversitesi	Yüksek Lisans Tezi
2013	Morfolojik Olarak Zengin Diller İçin Melez İstatistiksel/Birim Seçmeli Metinden Konuşma Sentezleme Sistemi	Ekrem Güner	Özyeğin Üniversitesi	Yüksek Lisans Tezi
2015	Implementation of Turkish text to Speech Synthesis with RC8660 Voice Synthesizer	Timur Karamehmet	Atılım Üniversitesi	Yüksek Lisans Tezi
2016	Türkçe Metin Seslendirme	Tuncay Şentürk Eşref Adalı	Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi	Makale
2016	Text to Speech Synthesis for Turkish Using a DSP Board	Uğur Ayaz	Dokuz Eylül Üniversitesi	Yüksek Lisans Tezi
2017	Eklemeli Sentezleyici Yöntemi Kullanılarak Türkçe Metinden Konuşma Sentezleme	Fatih Uysal	Gazi Üniversitesi	Yüksek Lisans Tezi
2017	Accompaniment Robot with Turkish Speech Synthesis and LIP Synchronization	İbrahim Özcan	Boğaziçi Üniversitesi	Yüksek Lisans Tezi
2017	Implementation of Turkish Text to Speech Synthesis on a Voice Synthesizer Card with Prosodic Features	Hakan Tora İ. Baran Uslu Timur Karamehmet	Anadolu Üniversitesi Bilim ve Teknoloji Dergisi	Makale

Çizelge 2.1. (devamı) Türkçe konuşma sentezleme sistemi çalışmaları

Yıl	Çalışmanın Adı	Yazar Adı	Kurum	Çalışma Türü
2019	Türkçe Metinden Konuşma Sentezlemeye Yönelik Yapılan Çalışmaların İncelenmesi	Gamze Yılmaz	Ahi Evran Üniversitesi	Yüksek Lisans Tezi
2019	Turkish Text to Speech Using Children's Voices Syllables	Yoldaş Erdoğan	Çukurova Üniversitesi	Yüksek Lisans Tezi
2019	Improving Low Resource Turkish Speech Recognition with Data Augmentation and TTS	Hülya Yalçın Ramazan Gökay	International Multi-Conference on Systems, Signals& Devices	Konferans Bildirisi

Özüm (1993), Türkçe için ilk konuşma sentezleme sistemi çalışmasını "A Speech Synthesis System for Turkish Language Based on the Concatenation of Phonemes Taken From a Speaker" isimli yüksek lisans tezi ile yapmıştır. Güven (1994), "PC based speech synthesis for Turkish" isimli yüksek lisans tezinde Türkçe 'deki sesli harflerin seri biçimlendirici konuşma sentezleme yöntemi ile sentezlenmesi sağlanmıştır. Erer (1994)," Karma söz üretme yöntemi ile Türkçe yazılı metinden söze geçme" isimli bir yüksek tezi hazırlamıştır. Ayhan (1998), "Text to Speech Synthesizer in Turkish Using Non Parametric Techniques" isimli yüksek lisans tezinde ZA-PEÖE (Time Domain Pitch Synchronized Overlap Add) algoritması kullanarak Türkçe için referans çalışmalardan birini yapmıştır. Salor (1999), "Signal Processing Aspects of Text to Speech Synthesizer in Turkish" isimli yüksek lisans tezi yapmıştır. Bozkurt (2000), "Reading aid for visually impaired (a Turkish text-to-speech system development)" isimli yüksek lisans tezinde birleştirmeli konuşma sentezleme yöntemini kullanmıştır. Çeşitli kelimelerden oluşturduğu bir veri kümesi kullanmıştır. Birleştirme tekniği olarak ZA-PEÖE kullanmıştır.

Eker (2002), "Türkçe Metin Seslendirme" isimli yüksek lisans tezi çalışmasında; Türkçe bir metni girdi olarak alıp bu metne karşılık gelen bir konuşma sinyali üretmiştir. Bu çalışma MATLAB ortamında gerçekleştirilmiştir ve doğallıktan çok anlaşılır bir çıktı üretmek amaçlanmıştır. Konuşma sentezleme sistemi yöntemlerinden birleştirmeli konuşma

sentezleme yöntemi kullanılmıştır. Veri kümesini oluşturan ses parçacıkları difonlardan oluşturulmuştur. Ancak, veri kümesinin yetersiz olduğu belirtilmiştir. Ses parçalarını birleştirmek için sayısal bir sinyal işleme tekniği olan PEÖE (Pitch Synchronized Overlap Add) algoritmasından faydalanılmıştır. Bu çalışmanın çok üst düzey bir konuşma sentezleme uygulaması olmadığı ve bu çalışma vasıtası ile metni fonetik forma dönüştüren bir ön işleme modülü ile diğer dilleri içinde konuşma sentezleme sisteminin oluşturulabileceği belirtilmiştir.

Salor, Pellom ve Demirekler (2003), "Implementation and Evaluation of a Text-to-Speech Synthesis System for Turkish" isimli çalışmalarında Türkçe için difon tabanlı bir konuşma sentezleme sistemi sunmuştur. Bu sistem için difon tabanlı bir veri kümesi geliştirilmiş ve birim seçmeye dayalı konuşma sentezleme uygulaması olan Festival konuşma sentezleme sistemi kullanılmıştır. Çalışma sonucunda uygulama, 20 dinleyici ile yapılan anlaşılırlık testinde %86,5 oranında anlaşılır bulunmuştur.

Aktan, Başkaya ve Dündar (2005), "A single chip solution for text-to-speech synthesis" isimli çalışmasında konuşma sentezleme sistemleri için bir donanımsal çözüm sunmuşlardır. Oluşturulan bütünleşmiş devre ASCII formatında gelen harfleri konuşmaya dönüştürmektedir. MATLAB ortamında simüle edilen çalışmada birleştirmeli konuşma sentezleme yöntemi kullanılmıştır. Fonemlerden oluşan bir veri kümesi kullanılmıştır.

Sak, Güngör ve Safkan (2006), "A Corpus-Based Concatenative Speech Synthesis System for Turkish" isimli makalede Türkçe için birleştirmeli konuşma sentezleme yönteminin alt başlıklarından olan birim seçmeye dayalı konuşma sentezlemenin bir örneğini sunmuşlardır. Bu çalışmada, fonemlerden oluşan bir veri kümesi kullanılmıştır. Yapılan çalışma sonucu elde edilen uygulama, Ortalama Görüş Puanı (OGP) öznel değerlendirme yöntemi ile 5 üzerinden 4,2 puan alarak test edilmiştir.

Ünaldı (2007), görme engelli insanların kısa mesaj servislerinden (SMS) faydalanmasını sağlamak amacıyla, taşınabilir cihazlarda çalışabilen bir konuşma sentezleme sistemi geliştirilmeyi amaçladığı "Taşınabilir Cihazlar İçin Türkçe Metinden Konuşma Sentezleme Sistemi" isimli yüksek lisans tezi çalışmasını hazırlamıştır. Uygulama Java ortamında gerçekleştirilmiştir. Konuşma sentezleme yöntemlerinden birleştirmeli konuşma sentezleme yöntemi kullanılmıştır. Veri kümesini oluşturan ses parçacıkları difonlardan

oluşturulmuştur. Veri kümesi bilgisayar kapasitesinden dolayı küçük boyutta kullanılmıştır. Ses parçalarını birleştirmek için Zaman Ekseninde Perde Eşzamanlı Örtüştürerek Ekleme olarak çevrilen konuşma sinyallerinin temel frekansını ve süresini değiştirebilen sayısal bir sinyal işleme tekniği olan ZA-PEÖE algoritmasından faydalanılmıştır.

Uzun (2007), görme engelli bireyler için eğitim ve günlük yaşamlarında karşılaştıkları kitap, gazete, belge, ders notu gibi yazılı dokümanları, yardımcı okuyucuya, sesli kitaba veya kabartma baskıya ihtiyaç duymadan, kendileri için seslendirebilecek bir bilgisayar destekli yardımcı eğitim sistemi geliştirmek amacıyla "Görme Engelliler İçin Basılı Doküman Yorumlama ve Seslendirme Sisteminin Gerçekleştirilmesi" isimli bir yüksek lisans çalışması yapmıştır. Çalışmada öncelikle metin, grafik, tablo gibi içerikleri Yapay Sinir Ağı (YSA) kullanılarak karakter tanıma yapılmış ve cümlelerde %98'lik bir başarı elde edildiği belirtilmiştir. Metinlerin seslendirilmesinde birleştirmeli konuşma sentezleme yöntemi kullanılmıştır. Veri kümesini oluşturan ses parçacıkları hecelerden meydana gelmektedir. Uygulama C++ Builder ortamında geliştirilmiştir. Yapılan değerlendirmelerde hecelerin birleştirilmesinde kopukluk olduğu belirtilmiştir.

Aşlıyan ve Günel (2008), "Türkçe Metinler için Hece Tabanlı Konuşma Sentezleme Sistemi" isimli bir konferans için çalışma yapmıştır. Bu çalışmada yine birleştirmeli konuşma sentezleme yöntemi kullanılmış ve ZA-PEÖE algoritmasından faydalanılmıştır. Veri kümesini oluşturan ses parçacıkları hecelerden meydana gelmiştir. Buna ek olarak derlem üzerinden elde edilen konuşmalar TOYA (Turkish Automatic Spelling Algorithm) adlı bir algoritma sayesinde hecelere ayrılmış ve veri kümesine eklenmiştir. Yapılan çalışma sonucu elde edilen uygulama, Bozulma Ortalama Görüş Puanı (BOGP) öznel değerlendirme yöntemi ile 5 üzerinden 3,85 puan alarak test edilmiştir.

Güldalı (2009), "Türkçe Metin Seslendirme" isimli bir yüksek lisans tezi hazırlamıştır. Çalışma C programlama dili kullanılarak gerçekleştirilmiş olup arayüz için QT kütüphanesinden faydalanılmıştır. Çalışmada birleştirmeli konuşma sentezleme yöntemi kullanılmıştır. Kullanılan veri kümesi hecelerden oluşmaktadır. Çalışmanın değerlendirilmesinde ortalama kelime anlaşılırlığı ve kelime içi tonlama bakımında başarı gösterilse de cümle bütününde tonlama başarısı ve ses kalitesinde diğer iki başlık seviyesinde başarı gösterilememiştir.

Yılmaz (2009), Türkçe konuşma sentezleme uygulamalarına yönelik olarak geliştirilen yazılım altyapısı ve beraberindeki veri kümesini tanıtmak amacıyla "Türkçe Metinden Konuşma Sentezleme Uygulamaları İçin Bir Veri Sözlük Seti ve Yazılım Çerçevesi" isimli bir çalışma yapmıştır.

Şentürk (2010), geçmişte yapılan bazı çalışmalara ulama, hece geçişleri v.b. gibi dilbilimsel çalışmalar ekleyerek "Türkçe Metin Seslendirme" isimli yüksek lisans tezi hazırlamıştır. MBROLA aracı kullanılarak hem difonlardan oluşan hem de hecelerin bulunduğu bir veri kümeleri oluşturulmuştur. Eclipse ortamında Java programlama dili ile geliştirilen çalışmada birleştirmeli konuşma sentezleme yöntemi kullanılmıştır. Bu çalışma kapsamında 3 farklı uygulama elde edilmiştir; difon tabanlı birleştirmeli yöntem, hece tabanlı birleştirmeli yöntem ve vurgulu birleştirmeli yöntem. Bu uygulamalar anlaşılırlık üzerinden değerlendirilmiş ve difon tabanlı birleştirmeli yöntem %91,5, hece tabanlı birleştirmeli yöntem %96,16 ve vurgulu birleştirmeli yöntem %98,13'lük değer almıştır.

Erdemir (2010), "Türkçe Metin Seslendirme İçin Doğal Konuşma Sentezleme" isimli bir yüksek lisans tezi çalışması yapmıştır. Çalışma MATLAB ortamında C++ ve Java programlama dili kullanılarak gerçekleştirilmiştir. Çalışmada birleştirmeli konuşma sentezleme yöntemi kullanılmıştır. Kullanılan veri kümesi Türkçedeki ikili hecelerden yani difonlardan oluşmaktadır. Çalışmanın değerlendirilmesinde ortalama kelime anlaşılırlığı ve kelime içi tonlama bakımında başarı gösterilse de cümle bütününde tonlama başarısı ve ses kalitesinde diğer iki başlık seviyesinde başarı gösterilememiştir. Bu çalışma anlaşılırlık ve doğallık yönünden ayrı ayrı olarak 5 üzerinden dinleyiciler üzerinden test edilerek değerlendirilmiştir. Bu kapsamda anlaşılırlık yönünden ortalama 4,56'lık ve doğallık yönünden 3,73'lük değer elde edilmiştir.

Uslu (2010), "Metinden Konuşma Sentezleme" isimli makalesinde konuşma sentezleme sistemlerinin yapısını incelemiştir. Bu kapsamda öncelikle konuşma sentezleme sisteminin geleneksel mimarisi hakkında bilgi vermiştir. Daha sonra konuşma sentezleme sistemlerinin genel yöntemleri hakkında bilgi verilmiştir. Makalenin ilerleyen bölümlerinde ise elde edilen konuşmanın değerlendirilmesinde anlaşılırlığının ve doğallığının önemli bir faktör olduğu belirtilmiştir. Son olarak gelecekte yapılacak olan çalışmalarda özellikle bu konulara dikkat edilmesi gerektiği vurgulanmıştır.

Tora ve Cengizler (2010), “Tek Ses Sentezleme Tümleşik Devresi ile Türkçe Metinden Sentetik Konuşma Üretme” isimli bir bildirisinde gerçekleştirdiği bir konuşma sentezleme uygulamasını açıklamıştır. Uygulamada oluşturdukları arayüz, İngilizce için hazırlanmış SpeakJet isimli tümleşik devreye uygulanmıştır. SpeakJet tümleşik devresindeki İngilizce alafon tablosu uygun Türkçe fonemlerle değiştirilmiştir. Yapılan değerlendirmelerde seslendirmelerin başarı ile gerçekleştirildiği ancak anlaşılabilirlik ve doğallık konusunda istenen sonucun tam olarak elde edilemediği belirtilmiştir. Bu çalışmanın tümleşik devreler üzerinden yapılacak konuşma sentezlemeleri için temel bir çalışma olacağı belirtilmiştir.

Bicil (2010), “Türkçe Metinden Konuşma Sentezleme” isimli bir yüksek lisans tezi hazırlamıştır. Birleştirmeli konuşma sentezleme yöntemi kullanılarak geliştirilen sistemde difonların (ikili seslerin) ve trifonların (üçlü seslerin) kaydedilmesiyle hazırlanmış bir veri kümesi kullanılmıştır. Ses parçalarının birleştirilmesi için ZA-PEÖE algoritmasından faydalanılmıştır. Çalışma sonucu elde edilen uygulama, OGP öznel değerlendirme yöntemi ile 5 üzerinden 3,87 puan alarak test edilmiştir.

Sel, Hanbay ve Karabatak (2011), “Beyin Bilgisayar Arayüzleri İçin Türkçe Metinden Konuşma Sentezleme Sistemi” isimli bir konferans bildirilerinde, Microsoft .NET 2.0 ve Microsoft Visual C# ortamlarında birleştirmeli konuşma sentezleme sistemini kullanmışlardır.

Arık (2011), “Görme Engelliler İçin Bilgisayar Kullanımının Etkinleştirilmesi, Erişilebilirlik ve Bir Türkçe Hece Tabanlı Konuşma Sentezleme Sisteminin Geliştirilmesi” isimli yüksek lisans çalışması hazırlamıştır. Visual Studio 2008 geliştirme ortamında C#, Microsoft .NET programlama dilini kullanarak birleştirmeli konuşma sentezleme yöntemine dayalı bir konuşma sentezleme sistemi geliştirmeye çalışmıştır. Çalışmada, bir ve iki harften oluşan hecelerin olduğu bir veri kümesi kullanılmıştır. 10 katılımcıya 10 tane cümlenin dinletilmesiyle yapılan değerlendirme sonucunda 3,7’lik OGP elde edilmiştir.

Güner ve Demiroğlu (2012), “A small footprint hybrid statistical/unit selection text-to-speech synthesis system for agglutinative languages” isimli bildirilerinde istatiksel parametrik konuşma sentezleme yöntemlerinden olan Saklı Markov Modelini (SMM: Hidden Markov Model) kullanarak hibrit bir konuşma sentezleme sistemi geliştirdiklerini ifade etmişlerdir. Aynı zamanda geliştirdikleri birleştirmeli konuşma sentezleme sistemi ve

hibrit konuşma sentezleme sistemi A/B testine tabi tutulmuştur. Teste katılan dinleyicilerin %53,9'u hibrit sistemi tercih etmiştir.

Uslu (2012), "Konuşma İşleme ve Türkçenin Dilbilimsel Özelliklerini Kullanarak Metinden Doğal Konuşma Sentezleme" isimli doktora tezi çalışmasında sentezlenmiş konuşmaya doğal konuşmadaki özelliklerin kazandırılmasını amaçlamıştır. Birleştirmeli konuşma sentezleme yöntemi kullanılan çalışmada veri kümesi difonlardan meydana gelmektedir. Çalışmada vurgu ve tonlamalar üzerine ezgi modelleri geliştirilmiş ve modele eklenmiştir. Çalışma öznel değerlendirme yöntemlerinden Karşılaştırmalı Ortalama Görüş Puanı (KOGP) ile değerlendirilmiş ve geliştirilen ezgi model, mevcut modelin başarı ortalamasını 5 üzerinden 1,45 değerinde artırmıştır.

Sel (2013), "Türkçe Metinler İçin Hece Tabanlı Metinden Konuşma Sentezleme Sistemi" isimli bir yüksek lisans çalışmasında, MATLAB ve C# programlama dilini kullanarak 2 farklı uygulama geliştirmiştir. MATLAB ile geliştirilen sistemde birleştirmeli konuşma sentezleme sistemi tercih edilmiş ve hecelerden oluşan bir veri kümesi kullanılmıştır. Hecelerden oluşan ses parçalarını birleştirmek için PEÖE algoritmasından faydalanılmıştır. C# ortamında hazırlanan diğer sistemde ise Microsoft .NET kütüphanelerinden MSKU (Microsoft System Speech Application) kullanılmıştır. Ses veri kümesi olarak MBROLA projesi [11] ile hazırlanan difon veri kümesi kullanılmıştır.

Güner (2013), "Morfolojik Olarak Zengin Diller İçin Melez İstatistiksel/Birim Seçmeli Metinden Konuşma Sentezleme Sistemi" isimli bir çalışma hazırlamıştır. Bu çalışmada, birleştirmeli konuşma sentezleme yöntemi örneklerinden olan birim seçmeli metinden konuşma sentezleme ve istatiksel parametrik konuşma sentezleme (Statistical Parametric Synthesis) yöntemi örneklerinden SMM tabanlı metinden konuşma sentezleme yöntemi karşılaştırılmıştır. Birim seçmeli metinden konuşma sentezlemede, ses parçalarının birleştirilmesi sonucu ortaya çıkan doğallık probleminin SMM tabanlı sistemde yaşanmadığı belirtilmiştir. Birim seçmeli metinden konuşma sentezlemede daha büyük veri kümesi kullanıldığı için de ses kalitesinin daha iyi olduğu belirtilmiştir. Bu yüzden morfolojik olarak zengin diller için bu iki sistemden de özellikler alan bir melez metinden konuşma sentezleme sistemi önerilmiştir.

Karamehmet (2015), "RC8660 Ses Sentezleyici ile Türkçe Metinden Konuşma Sentezleme"

isimli yüksek lisans tezinde, RC8660 gömülü sisteminin Türkçe'ye uyarlanması için çalışmıştır. RC8660 kartının özelliklerinin kullanıldığı çalışmada İngilizce fonemlerle yüklü gelen karta Türkçe fonemler tanımlanarak yeni bir sözlük oluşturulmuştur. RC Studio ortamında hız, ifade, perde, biçim frekansı, ton, gecikme ve telaffuz gibi özellikler değiştirilerek bu ayarların konuşma üzerindeki etkileri de test edilmiştir. 20 katılımcının yaptığı test sonuçlarında ortalama 3,8'lik OGP elde edilmiştir.

Ayaz (2016), "Text to Speech Synthesis for Turkish using a DSP Board" isimli yüksek lisans tezinde, konuşma sentezleme işlemini sayısal sinyal işleme kartı üzerinde gerçekleştirmiştir. Bilgisayardan girilen Türkçe metin, hece ve kelimelerinin arasına çeşitli ifadeler eklenmesiyle dilbilimsel olarak analiz edilmekte ve daha sonra sayısal sinyal işleme kartına gönderilmektedir. Sayısal sinyal işleme kartı her bir sese karşılık gelen sayısal ifadeleri belleğinde kalıcı olarak saklamaktadır. Eşleşen bu sayısal ifadeler tekrar ses sinyaline dönüştürülerek ses kodlayıcı-kod çözücü işlemci vasıtasıyla dışarı aktarılmaktadır. Bu çalışma sonucunda yapılan öznel değerlendirmede anlaşılırlık ve doğallık açısından ortalama bir sonuç alındığı belirtilmiştir.

Uysal (2017), "Eklemeli Sentezleyici Yöntemi Kullanılarak Türkçe Metinden Konuşma Sentezleme" isimli yüksek lisans tezinde, ses parçaları olarak hecelerin kullanıldığı birleştirmeli konuşma sentezleme yöntemini kullanmıştır. Bu çalışmayı diğer çalışmalardan ayıran özelliği dinamik bir veri kümesinin kullanılmasıdır.

Özcan (2017), "Accompaniment Robot with Turkish Speech Synthesis and LIP Synchronization" isimli yüksek lisans tezinde, robot-insan ilişkisi kapsamında, robotun ses üretimi, konuşma esnasındaki dudak pozisyonlarının konuşma ile uyumlu hareketi ve robotun söylenen cümleleri ilişki kurarak anlayabilmesi için bir çalışma yapmıştır. Robotun ses üretimi için biçimlendirici konuşma sentezleme yöntemi kullanılmıştır. Yeterince eğitim verisi olmadığı için bu yöntemin tercih edildiği belirtilmiştir.

Tora, Uslu ve Karamehmet (2017), "Implementation of Turkish Text to Speech Synthesis on a Voice Synthesizer Card with Prosodic Features" isimli çalışmalarında, birleştirmeli konuşma sentezleme yöntemi kullanılan RC8660 gömülü sisteminin Türkçe uygulaması kullanılmıştır. Çalışmada Türkçe ses birimleri kullanılmamıştır. İngilizce fonemler, Türkçe fonemlerle eşleştirilmiştir. Buna ek olarak bürünsel özellikler RC8660 gömülü sistemine

uyarlanmıştır. Bu çalışma sonucunda yapılan öznel testlerde ortalama 3,29'luk OGP elde edilmiştir.

Erdoğan (2019), "Turkish Text to Speech Using Children's Voices Syllables" isimli yüksek lisans tez çalışmasında, çocuk seslerinden oluşan bir ses veri kümesi tasarlamış ve sentezlenecek sesin çocuk sesi olmasını hedeflemiştir. Çalışmada, konuşma sentezleme yöntemlerinden birleştirmeli konuşma sentezleme yöntemi kullanılmıştır. Veri kümesini oluşturan ses parçacıkları difonlardan oluşturulmuştur. Ses parçalarını birleştirmek için sayısal bir sinyal işleme tekniği olan EÖE (Synchronized Overlap Add) algoritmasından faydalanılmıştır. Çalışma sonucunda yapılan değerlendirmelerde10 dinleyicinin katıldığı öznel testte 100 üzerinden ortalama 70 puan elde edilmiştir.

Gökay ve Yalçın (2019), "Veri Geliştirme ve Metin Okuma ile Düşük Kaynaklı Türkçe Konuşma Tanıma İyileştirme" isimli çalışmalarında, konuşma tanıma araştırmacılarının karşılaştığı en büyük sorunlardan biri olan veri eksikliğine odaklanmıştır. Veri büyütme ve konuşma sentezlemenin konuşma tanıma üzerindeki etkilerini görmek için çok sınırlı eğitim verileriyle bazı deneyler gerçekleştirmiştir. Veri artırmanın yanı sıra, iki farklı konuşma sentezleme yöntemi kullanılarak sentezlenmiş konuşma üretilmiştir. İlk konuşma sentezi yaklaşımında, konuşma sentezleyici olarak Google Translate Text to Speech (gTTS) kullanılmış, ikinci konuşma sentezi yaklaşımında ise uçtan uca bir Türkçe konuşma sentezleme sistemi geliştirilmiştir. Son olarak, tüm bu alternatif yöntemlerin düşük kaynaklı diller için konuşma tanıma üzerindeki etkileri incelenmiştir. Sonuçlar, bazı veri artırma veya konuşma sentezi tekniklerinin, düşük kaynaklı diller için konuşma tanımayı iyileştirmede iyi performans sağladığını göstermiştir.

3. MATERYAL VE METOTLAR

Konuşma sentezleme sistemlerinin geliştirilmesi için birleştirmeli sentezleme, biçimlendirici (formant) sentezleme, söyleyiş sentezleme, istatiksel parametrik sentezleme ve son yıllarda kullanılmaya başlayan derin öğrenme tabanlı konuşma sentezleme gibi farklı yöntemler kullanılmaktadır.

Birleştirmeli konuşma sentezleme sistemleri, temel olarak konuşma bilgisinden uygun birimlerin seçilmesini, seçilen birimleri ekleyen algoritmaları ve birleştirme sınırlarını yumuşatmak için sinyal işleme çalışmalarını içermektedir [12]. Türkçe gibi sondan birleştirmeli bir yapıya sahip diller için konuşma sentezleme sistemi olarak en uygun sistemin birleştirmeli konuşma sentezleme sistemi olduğu belirtilmektedir [13]. Bu teknik için kullanılan veri kümeleri konuşma birimlerinin örneklerinden oluşmaktadır. Bu birimler sözcükler, heceler, yarı heceler, ses birimleri, çift sesler veya üçlü sesler olabilmektedir. Birim uzunluğu, sentezlenen konuşmanın doğallık ve anlaşılırlığını doğrudan etkilemektedir.

Biçimlendirici konuşma sentezleme sistemleri, ses yolu aktarım fonksiyonunun, biçimlendirici frekansları ve biçimlendirici genlikleri benzetilerek üretilebilmesi üzerine gerçekleştirilen çalışmalardır [14].

Söyleyiş konuşma sentezleme sistemleri, insanın söyleyiş davranışının doğrudan modellenmesiyle konuşmayı üretmektedir. Bu nedenle prensipte yüksek kaliteli konuşma sentezi üretmek için en başarılı yöntem olarak kabul edilmektedir. Ancak pratikte uygulanması en zor yöntemlerden biridir [15].

İstatistiksel parametrik konuşma sentezleme sistemlerinde konuşma sentezi için model tabanlı bir yaklaşım izlenmektedir. Birleştirmeli sistemlerin aksine model tabanlı bir yaklaşımda, birimleri depolamak yerine her bir birime karşılık gelen modeller bir havuzda saklanmaktadır. Model tabanlı yaklaşımda ölçeklendirilen konuşma örneklerinden elde edilen parametreler için gerekli modeller istatistiksel yöntemlerle hazırlanmaktadır [16].

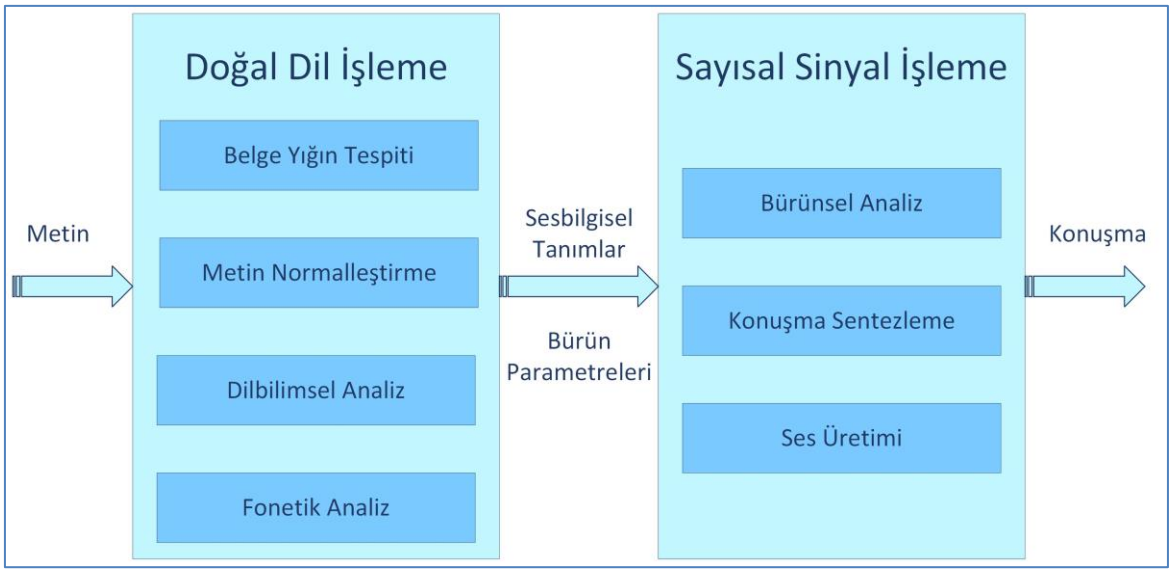
Doğal özelliklerin öğrenilmesinde etkili olduğu kanıtlanmış derin sinir ağları ile dil

özelliklerinden elde edilen akustik özellikler doğrudan haritalanmaktadır [3]. Böylelikle dil özellikleri ile akustik özellikler arasında doğrudan bir bağ kurulmaktadır. Derin öğrenme temelli konuşma sentezleme sistemleri karmaşık modeller içermekte ve bu modellerin eğitilmesi sırasında dil özelliklerinin daha iyi öğrenilebilmesi için büyük veri kümesine ihtiyaç duyulmaktadır. Derin öğrenme tabanlı konuşma sentezleme yöntemlerinin geliştirilmesinde metin ve ses dosyalarının birebir eşleştirilmesi ile hazırlanan derlem tipi veri kümeleri kullanılmaktadır. Veri kümesinin büyüklüğü, konuşma sürelerinin uzunluğu, konuşmalarda geçen kelime ve benzersiz kelime sayısı geliştirilen sistemin başarısını doğrudan etkilemektedir.

Doğallık ve anlaşılırlık üzerine gerçekleştirilen değerlendirmelerde birleştirmeli konuşma sentezleme sistemlerinin anlaşılırlık için daha iyi sonuçlar verdiği belirtilmiştir. Ancak doğallık değerlendirmesinde ses birimleri arasındaki geçişlerde konuşma doğallığında sorunlar yaşanmaktadır. Doğallık sorunun önüne geçmek için son yıllarda uygulanmaya başlanan derin öğrenme tabanlı konuşma sentezleme yöntemleri önerilmiştir.

3.1. Geleneksel Konuşma Sentezleme Sisteminin Mimarisi

Konuşma sentezleme sistemleri, farklı diller için farklı gereksinimlerle karşılaşabilir, ancak temel olarak iki ana bileşenden oluşurlar: Doğal Dil İşleme (DDİ: Natural Language Processing) ve Sayısal Sinyal İşleme (SSİ: Digital Signal Processing). Doğal dil işleme bölümü, bilgisayar sistemlerinin bir doğal dili çözümleme, anlama, yorumlama ve üretme yeteneğine sahip olmasını sağlamakla ilgilenmektedir [17]. Bu bölümde, metnin fonetik bir uyarlamasını üreten dilsel biçimcilik, çıkarım motorları ve mantıksal çıkarımlar gibi unsurlar bulunmaktadır. Ayrıca, ses bilgisel tanımlamalar ve bürün parametreleri (vurgu, ezgi, durak, ulama, ton, uzunluk gibi konuşma diline özgü ögeler) de doğal dil işleme bölümünde yer almaktadır. Diğer bir bileşen olan sayısal sinyal işleme bölümü, sembolik bilgileri konuşmaya dönüştürmekle ilgilenmektedir. Bu bölümde matematiksel modeller, algoritmalar ve hesaplamalar kullanılmaktadır [18]. Geleneksel bir konuşma sentezleme sisteminin mimarisi Şekil 3.1'de gösterilmiştir.



Şekil 3.1. Geleneksel konuşma sentezleme sistemlerinin mimarisi

Şekil 3.1'de yer alan doğal dil işleme bölümü, metni konuşmaya çevirmek için ön işlemler yapmaktadır. Bu bölümün temel amacı, sentezlenecek konuşmanın oluşturulmasında kullanılacak ses bilgilerini elde etmektedir. Metinde kısaltmalar ('Dr.', 'Sn.', 'Doç.' gibi) bulunuyorsa, bunlar açık yazılışlarına çevrilmelidir. Ayrıca, sayılar (tarihler, 1. gibi kısaltmalar vb.) metne dönüştürülmelidir. Sinyal işleme modülü ise ses bilgisel ve bürünsel bilgileri kullanarak konuşmayı oluşturmaktadır. Bu modül, elde edilen metinsel bilgileri matematiksel modeller, algoritmalar ve hesaplamalar aracılığıyla ses bilgisine dönüştürmektedir. Bu şekilde, konuşma sentezlenmekte ve istenen dilde doğal bir ses elde edilmektedir [19].

3.1.1. Doğal dil işleme

Ana işlevi bir doğal dili çözümleme, anlama, yorumlama ve üretme olan bilgisayar sistemlerinin tasarımı olarak tanımlanan doğal dil işleme bölümü Şekil 3.1'de Belge yapısı tespiti, metin normalleştirme, dilbilimsel analizler ve fonetik analizler bölümlerinden oluşmaktadır. Bu bileşenler genel olarak girilen metnin fonetik bir uyarlamasını üretirken, içerlerinde dilsel biçimcilik, çıkarım motorları ve mantıksal çıkarımlar yer almaktadır.

Belge yığın tespiti (document structure detection (recognition)), doğal dil işleme alanında kullanılan bir yöntemdir. Bu yöntem, bir belgenin içerisindeki farklı yapısal bileşenleri tanımak ve belgeyi bölümlere ayırmak için kullanılmaktadır [20].

Belgeler genellikle başlık, paragraf, liste, tablo, başlık altı, görsel, dipnot gibi farklı yapısal bileşenleri içermektedir. Belge yığın tespiti, bu bileşenleri belge içinde tanımlamak için dilbilimsel ve yapısal analiz yöntemlerini kullanılmaktadır. Konuşma sentezleme sistemlerinin doğal dil işleme aşamasında kullanılan bir yöntem olarak belge yığın tespiti, metinlerin yapısal düzenini tespit etmek ve belge içeriğini daha iyi anlamak için önemli bir araçtır.

Konuşma sentezleme sistemi, metinleri konuşma sesine dönüştürmek için kullanılır ve belgenin yapısını doğru bir şekilde anlamak, daha doğal ve akıcı bir konuşma sentezi elde etmek için gereklidir. Belgenin yapısal bileşenlerini tespit etmek, konuşma sentezleme sistemlerinin vurgu, tonlama ve akıcılık gibi özellikleri daha doğru bir şekilde uygulamasına yardımcı olmaktadır. Özetle bu işlem metin analizi olarak yorumlanabilmektedir. Metin analizi bölümü, giriş metnini analiz eden ve yönetilebilir kelime listesi halinde organize eden ön işleme bölümüdür. Başka bir ifade ile metin analiz etme bölümünde rakamlar, kısaltmalar gibi ifadeleri gerektiğinde tam metne dönüştürülmektedir.

Bir başka doğal dil aşaması metin normalleştirmedir. Metin normalleştirme, genellikle metin bir şekilde işlenmeden önce gerçekleştirilmektedir. Metin normalleştirme bölümü, metnin telaffuz edilebilir bir forma dönüştürülmesidir. Bu sürecin temel amacı, sözcükler arasındaki noktalama işaretlerini ve duraklamaları belirlemektir. Genellikle metin normalleştirme işlemi, tüm harfleri küçük veya büyük harfe dönüştürmek, noktalama işaretlerini, aksan işaretlerini, gereksiz kelime veya çok yaygın kelimeleri ve diğer aksanları metinden çıkarmak için yapılmaktadır [21]. Metin normalleştirmede yapılan bazı çıkarma işlemeleri İngilizce ‘ye benzer dillerde, aynı kelime farklı bağlamda farklı şekilde telaffuz edilebilirken, Türkçe’de karakterler yazıldığı gibi okunduğu ve yalnızca bir telaffuzları vardır. Bu nedenle, bağlama göre telaffuz değişimi Türkçe bir zorluk değildir. Ancak Türkçe‘de diğer dillerde olamayan çeşitli karakterlerin ASCII karşılığı olamaması metin normalleştirmede en önemli sorunlardan biridir. Türkçe için sayılar, telaffuz edilirken bir tekrar dahilinde artarak devam ettiği için bir dizi ortak telaffuz kuralı tasarlanabilmektedir. Türkçe’deki kısaltmalar üzerine hem eğitim hem de eğitim sonrasında giriş metni üzerinde çalışabilecek ek bir liste hazırlanmalıdır. Bu listede yer alacak kısaltmalar için Türk Dil Kurumu’nun belirlediği kısaltmalardan yararlanılabilinir.

Doğal dil işlemedeki üçüncü bölüm dilbilimsel analizlerdir. Doğal dil işleme, metin

verilerinin dilbilgisel yapısını anlamak ve çözümlemek için kullanılan yöntemlerin bir bütünüdür. Bu süreçte, dilbilimsel analizler metinlerin cümle yapısını, sözcük anlamını, bağlamsal ilişkileri ve dilbilgisel kuralları anlamak için kullanılan tekniklerdir. Bu analizler, metinlerin dilbilgisel ve anlamsal öğelerini incelemek ve anlamak amacıyla uygulanmaktadır [22].

Dilbilimsel analizlerin bir örneği, biçimbilimsel (morfolojik) analizdir. Bu analiz, bir kelimenin yapısal bileşenlerini inceleyerek kökünü, eklerini ve çekim/çoğul yapısını anlamayı hedefler. Örneğin, "baharlar" kelimesini morfolojik olarak analiz ederken, "baharlar" kelimesinin kökü olan "bahar" ve çoğul ekini temsil eden "-lar" bileşenleri ayırt edilebilmektedir.

Sözdizimsel (sentaks) analizi ise cümle yapısını ve dilbilgisel kuralları inceler. Bu analiz, cümlelerin dilbilgisi yapısına uygunluğunu belirlemek için kullanılır. Sözdizimsel analizi sırasında, cümlenin özne, yüklem, nesne ve zarf tümceleri gibi yapısal unsurları ve bu unsurlar arasındaki ilişkileri çözümlemeye çalışılmaktadır.

Anlamsal (semantik) analiz, metindeki sözcüklerin ve ifadelerin anlamını çıkarmayı hedeflemektedir. Bu analiz, kavramlar arasındaki ilişkileri, eş anlamlılık, zıtlık, benzerlik gibi anlamsal özellikleri çözümlemek için kullanılmaktadır. Böylece, metindeki anlam karmaşıklığı anlaşılır hale gelmektedir ve metnin daha iyi anlaşılması sağlanmaktadır.

Edimsel (pragmatik) analiz ise ifadelerin gerçek dünya bağlamında nasıl işlev gördüğünü ve anlamlandırıldığını incelemektedir. Bu analiz, dilin kullanım amacını, ima edilen anlamları ve konuşma niyetlerini anlamak için kullanılmaktadır. Örneğin, bir şakanın veya ironinin dilin örtük anlamlarını anlamak edimsel analizin bir parçasıdır.

Bu dilbilimsel analizler, metinlerin içerdiği dilbilgisel ve anlamsal unsurları çözümlemek ve anlamak için kullanılmaktadır. Doğal dil işleme uygulamalarında, bu analizler metinlerin işlenmesi, anlamlandırılması, etiketlenmesi ve daha genel olarak doğal dilin anlaşılması ve üretilmesi süreçlerinde önemli bir role sahiptir.

Fonetik analiz konuşma sentezleme sistemlerini oluşturan iki bölümden biri olan doğal dil işleme bölümü altında gerçekleşmektedir. Metin ifadelerinden konuşma üretmek için, bir

dizi grafikten foneme dönüştürme kurallarına ihtiyaç vardır. Fonetik analiz, ortografik sembollerin fonolojik birimlere dönüştürülmesini içermektedir [23]. Giriş metnindeki kelimelerin her birinin fonetik gösterimi ilgili dillerin telaffuz edilebilir biçimlerine dayalı olarak elde edilmektedir. Metinden foneme dönüştürme için uygun yöntemin seçilmesi ve farklı dillerde uygun dönüştürme kurallarının hazırlanması konuşma sentezleme sistemlerinde önemli bir konudur. Ortografik sembollerin uygun fonolojik birimlere dönüştürülmesi, sözlük tabanlı [24] veya kural tabanlı dönüşüm [25] kullanılarak yapılabilmektedir.

Sözlük tabanlı dönüşüm yaklaşımı, bir dilin tüm kelimelerinin ve doğru telaffuzlarının saklandığı bir sözlük kullanılmaktadır. Bu yöntem, hızlı bir şekilde doğru sonuçlar üretebilir, ancak belirli bir kelime sözlüğünde olmayan kelimelerde tamamen başarısız olabilmektedir. Örneğin, "fani" ve "fanila" kelimeleri benzer olsa da "fani" kelimesindeki "a" harfi uzun olarak telaffuz edilmektedir. Sözlük boyutu büyüdükçe bellek gereksinimi de artmaktadır.

Diğer yandan, kural tabanlı yaklaşım, telaffuz kurallarını kelimenin yazımına dayanarak uygular ve doğru telaffuzları belirlemektedir. Bu yöntemde herhangi bir sözlük saklanmasına ihtiyaç duyulmaz ve herhangi bir girdi üzerinde çalışabilmektedir. Ancak, düzensiz yazımlar veya telaffuzlar ortaya çıktığında, kuralların karmaşıklığı önemli ölçüde artmaktadır. Bu nedenle, her iki yöntemde de metinden foneme (ses birimi) dönüşüm için daha yüksek kalitede sonuçlar elde etmek en büyük hedeftir.

Bu amaçla, daha fazla çalışma ve araştırma yapılması gerekmektedir. Geliştirilecek bir sistem, hem geniş bir sözlük tabanını kullanarak hem de kural tabanlı yöntemleri uygulayarak daha doğru ve kapsamlı sonuçlar elde etmeye odaklanmalıdır.

3.1.2. Sayısal sinyal işleme

Girilen metnin fonetik bir uyarlamasını üreten doğal dil işleme bölümüne ek olarak ses bilgisel tanımlamalara ve bürün parametrelerinin eklenmesiyle sembolik bilgiler elde edilmektedir. Elde edilen bu sembolik bilgileri konuşmaya dönüştüren sayısal sinyal işleme bölümü içerisinde genel bir yapı matematiksel modeller, algoritmalar ve hesaplamalar yer almaktadır [18]. Bu matematiksel modeller, algoritmalar ve hesaplamalar, sayısal sinyal işlemenin bir parçası bürünsel analiz, konuşma sentezleme ve ses üretimi içerisinde

gerçekleşmektedir.

Bürünsel analiz, daha doğal bir şekilde elde edilmiş konuşma sentezleme sonucu için insan konuşmasının farklı bürünsel özelliklerinin bir analizini içermektedir [26]. Bürünsel özellikler, ses perdesi, süre ve zaman içindeki vurgu ve tonlamaları içerebilir ve cinsiyet, yaş, duygular gibi özellikler üzerinde iyi bir kontrol ile çok doğal sesli konuşma modellenebilmektedir [27]. Ses özelliklerinin uygun şekilde tanımlanması ve uygulanması, son derece doğal konuşma bölümlerinin üretilmesine yardımcı olabilmektedir. Bununla birlikte, konuşmanın duygusal içeriklerini belirlemek, zor ve devam eden bir süreç olan doğal konuşmayı doğru bir şekilde oluşturmak için farklı konuşma parametrelerinin tanımlanmasını ve farklı sinyal işleme yöntemlerinin kullanılmasını gerektirmektedir [28]. Dildeki farklı biçimleri kapsayan ve dile özgü bürün modellemeyi kapsayan genel bir akustik veri tabanı oluşturmanın zor olduğu tahmin edilmektedir.

Konuşma sentezi, içerisinde matematiksel modeller, algoritmalar ve hesaplamalar yer alan ve aldığı sembolik bilgileri konuşmaya dönüştüren sayısal sinyal işleme bölümünde gerçekleşmektedir. Sentezlenmiş konuşma, yapılan fonetik ve bürünsel analizler sonucu elde edilen sembolik bilgilerin çeşitli konuşma sentezleme yöntemlerinden birinin kullanılmasıyla bu aşamada üretilmektedir.

Konuşma sentezleme sistemlerinin geliştirilmesi için literatürde farklı yöntemler kullanılmaktadır. Bu yöntemler arasında birleştirmeli sentezleme, biçimlendirici sentezleme, söyleyiş sentezleme, istatiksel parametrik sentezleme ve derin öğrenme tabanlı konuşma sentezleme yöntemleri bulunmaktadır.

Birleştirmeli konuşma sentezleme sistemleri, konuşma bilgisinden uygun birimlerin seçilmesini, birimleri birleştirmek için algoritmaları ve birleştirme sınırlarını yumuşatmak için sinyal işleme tekniklerini içermektedir. Bu yöntemde, birleştirmek üzere seçilen birimlerin eklenmesiyle konuşma sentezi oluşturulmaktadır [12].

Biçimlendirici konuşma sentezleme sistemleri, ses yolundaki aktarım fonksiyonunun, biçimlendirici frekansları ve biçimlendirici genlikleri taklit etmek için yapılan çalışmalardır. Bu yöntemde, ses yolundaki organların davranışları modellemek suretiyle konuşma sentezi üretilmektedir [14].

Söyleyiş konuşma sentezleme sistemleri, insanın söyleyiş davranışını doğrudan modellemektedir ve bu yöntem en yüksek kalitede konuşma sentezi üretmek için en başarılı yöntem olarak kabul edilmektedir. Ancak, pratikte uygulanması en zor yöntemlerden biridir [15].

İstatistiksel parametrik konuşma sentezleme sistemleri ise model tabanlı bir yaklaşımı benimsemektedir. Bu yöntemde, her bir birim için modeller bir havuzda saklanır ve konuşma sentezi için ölçeklendirilmiş konuşma örneklerinden elde edilen parametreler kullanılarak bu modeller istatistiksel yöntemlerle oluşturulmaktadır [16].

Türkçe gibi sondan birleştirmeli bir yapıya sahip diller için birleştirmeli konuşma sentezleme sisteminin en uygun yöntem olduğu belirtilmektedir [13]. Bu yöntemde kullanılan veri setleri, konuşma birimlerinin örneklerinden oluşmaktadır. Birim uzunluğu, sentezlenen konuşmanın doğallığını ve anlaşılırlığını etkileyen bir faktördür.

Yapılan değerlendirmeler doğrultusunda, birleştirmeli konuşma sentezleme sistemlerinin anlaşılırlık açısından daha iyi sonuçlar verdiği belirtilmektedir. Ancak doğallık değerlendirmelerinde, ses birimleri arasındaki geçişlerde hala sorunlar yaşanmaktadır. Bu nedenle, son yıllarda derin öğrenme tabanlı konuşma sentezleme yöntemleri önerilmekte ve doğallık sorunlarının üstesinden gelmeye çalışılmaktadır.

Derin öğrenme tabanlı konuşma sentezleme yöntemleri, dil özelliklerinden elde edilen akustik özellikleri doğrudan haritalamak için etkili olduğu kanıtlanmış derin sinir ağlarından faydalanmaktadır. Bu yöntemde dil özellikleri ile akustik özellikler arasında doğrudan bir bağlantı kurulmaktadır [3]. Bu sayede, konuşmanın doğal özellikleri daha iyi öğrenilebilir hale gelmektedir.

Derin öğrenme tabanlı konuşma sentezleme sistemleri karmaşık modeller içermektedir ve bu modellerin geliştirilmesi için genellikle güçlü donanım altyapısına ihtiyaç duyulmaktadır. Ayrıca, dil özelliklerinin daha iyi öğrenilebilmesi için geniş bir eğitim veri kümesine ihtiyaç vardır. Bu nedenle, metin ve ses dosyalarının birebir eşleştirildiği derlem tipi veri kümeleri derin öğrenme tabanlı konuşma sentezleme yöntemlerinin geliştirilmesinde kullanılmaktadır.

Derin öğrenme tabanlı sistemlerin başarısını doğrudan etkileyen faktörler arasında veri kümesinin büyüklüğü, konuşma süreleri, kullanılan kelimelerin sayısı ve benzersiz kelimelerin sayısı bulunmaktadır. Daha özet bir ifade ile, derin öğrenme tabanlı konuşma sentezleme yöntemleri, doğallık sorunlarını aşmaya yönelik çabalarda önemli bir adımdır. Ancak bu yöntemlerin karmaşıklığı, donanım gereksinimleri ve geniş veri kümelerine ihtiyaç duymaları, geliştirilmeleri sürecinde dikkate alınması gereken faktörlerdir.

Ses üretimi aşaması, konuşma sentezi mimarisinde sentezlenen metnin ses haline getirildiği aşamadır. Bu aşama, metin temsilinin sayısal bir forma dönüştürülmesini ve sonrasında bu sayısal verilerin hoparlör veya kulaklık gibi ses çıkış aygıtlarına yönlendirilerek ses sinyallerinin üretilmesini içermektedir. Ses üretimi sürecinde, sentezlenen metnin doğru bir şekilde ses haline getirilmesi için diğer ses işleme teknikleri de kullanılabilmektedir. Örneğin, filtreleme yöntemleri kullanılarak istenmeyen gürültülerin azaltılması veya sesin daha net bir şekilde duyulmasını sağlayan özellikler kullanılabilmektedir. Ses üretimi bölümü son çalışmalarda ses kodlayıcı [29] bölümüne de denk gelmektedir.

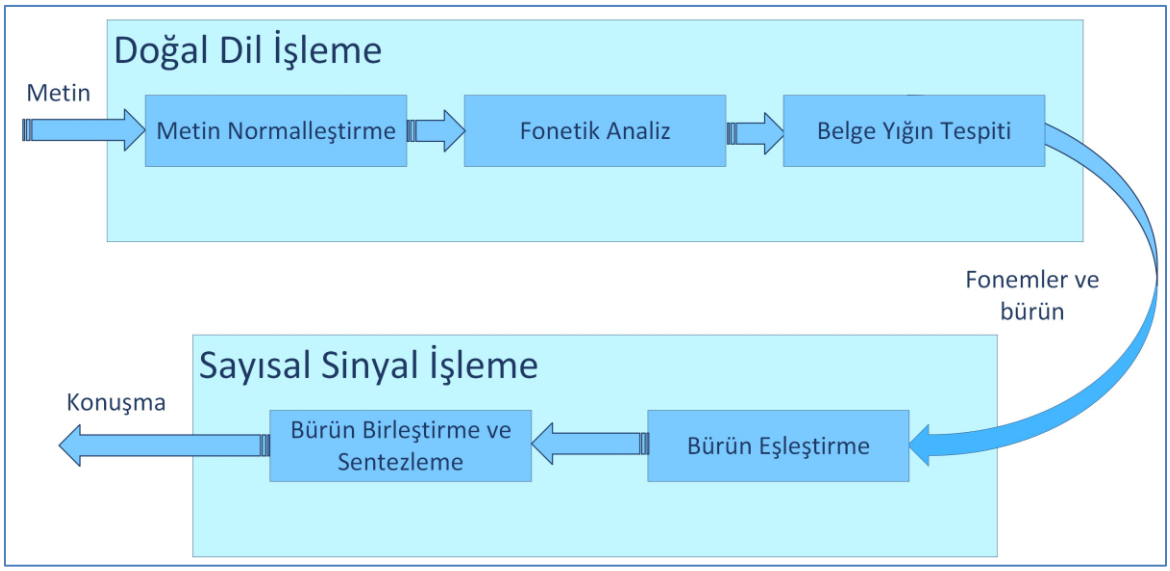
3.2. Konuşma Sentezleme Sistemi Yöntemleri

Literatürdeki çalışmalar incelendiğinde birleştirmeli konuşma sentezleme, biçimlendirici konuşma sentezleme, söyleyiş konuşma sentezleme, istatiksel parametrik konuşma sentezleme ve son yıllarda sıklıkla kullanılmaya başlayan derin öğrenme tabanlı konuşma sentezleme yöntemlerine rastlanılmıştır. Bölüm 3.1'de verilen geleneksel mimari belirtilen konuşma sentezleme sistemi yöntemleri için de geçerlidir.

3.2.1. Birleştirmeli konuşma sentezleme sistemi

Birleştirmeli konuşma sentezleme sistemleri, konuşma bileşenlerinden uygun birimlerin seçilmesini, bu birimleri birleştiren algoritmaları ve birleştirme sınırlarını yumuşatmak için bazı sinyal işleme yöntemlerini içeren sistemlerdir [12]. Bu yöntemde doğal veya önceden kaydedilmiş konuşma birimleri bir araya getirilerek konuşma üretilmektedir. Bu birimler, sözcükler, heceler, yarı heceler, ses birimleri, çift sesler veya üçlü sesler olabilmektedir. Birim uzunluğu, sentezlenen konuşmanın kalitesini etkilemektedir. Daha uzun birimler kullanıldığında doğallık artmakta ve daha az birleştirme noktasına ihtiyaç duyulmaktadır. Bununla birlikte, daha uzun birimler için daha fazla donanım altyapısı ve bellek

gerekmektedir. Ayrıca, depolanan birimlerin sayısı artmaktadır. Daha kısa birimler kullanıldığında ise daha az donanım altyapısı ve bellek gerekmektedir. Ancak, örnek toplama ve etiketleme teknikleri daha karmaşık hale gelmektedir [15]. Şekil 3.2'de geleneksel bir birleştirmeli konuşma sentezleme sisteminin mimarisi görülmektedir.



Şekil 3.2. Birleştirmeli konuşma sentezleme sistemi geleneksel mimarisi

Birleştirmeli konuşma sentezleme sistemlerinin genel çalışma yapısının geleneksel konuşma sentezleme sistemleri mimarisinden çok büyük bir farkı yoktur. Bu sistemler genel olarak dil verisi hazırlığı ile başlamaktadır. Metin normalizasyonu, giriş metni düzenleme ve standartlaştırma işlemidir. Bu adımda, metindeki özel karakterler, noktalama işaretleri, sayılar, kısaltmalar vb. gibi metin içindeki farklılıklar düzeltilmektedir. Ayrıca, metindeki büyük-küçük harf ayrımı da düzeltilmektedir.

Daha sonra özellik çıkarımı yapılmaktadır. Fonetik analiz adımı, metindeki kelimelerin fonetik açıklamalarını elde etmek için kullanılmaktadır. Bu adımda, her kelimenin fonetik gösterimi belirlenmektedir. Fonemler, ses birimlerini temsil eden seslerin en küçük birimleridir. Her kelimenin fonem dizisi belirlendikten sonra, ilgili ses birimlerinin dil verisindeki kaynakları bulunmaktadır.

Bir sonraki aşamada bürün üretimi yapılmaktadır. Bürün üretimi adımı, metindeki vurgu desenlerinin belirlenmesini sağlamaktadır. Bu adımda, kelimelerin vurgulu veya vurgusuz olma durumları belirlenmektedir. Bunun yanı sıra, metindeki cümlelerin tonlaması, vurgu

yerleşimi ve süreleri de belirlenmektedir. Bu adım, sentezlenen konuşmanın doğal ve akıcı bir şekilde duyulmasını sağlamaktadır.

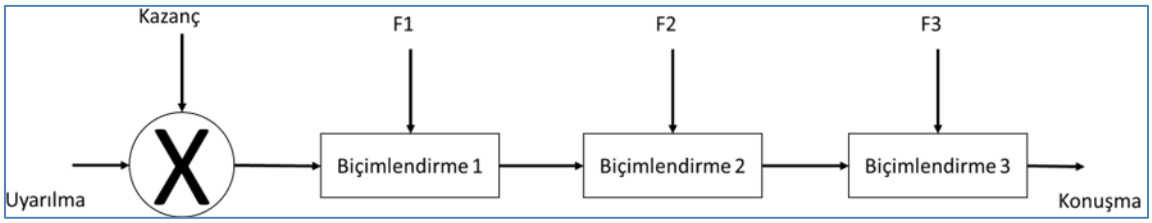
Üretilen bürünlerin daha sonra eşleştirmesi aşamasına geçilmektedir. Bürün eşleştirme adımı, sentezlenen konuşmanın doğru vurgu desenlerini yansıtmasını sağlamaktadır. Bu adımda, metindeki vurgu desenleri ile dil verisindeki kaynakların uyumlu olması için bir eşleştirme işlemi gerçekleştirilmektedir. Bu sayede, sentezlenen konuşma daha doğal ve akıcı bir şekilde duyulmaktadır.

Son olarak ise birim birleştirme ve sentezleme aşamasına geçilmektedir. Birim birleştirme ve sentez adımı, önceden belirlenen fonetik, bürünsel ve sinyal işleme bilgilerine dayanarak sentezlenmiş konuşma üretimini gerçekleştirmektedir. Ses birimleri uygun şekilde birleştirilmekte ve sentezlenmiş konuşma sinyali elde edilmektedir. Bu adımda, sentezlenen konuşma sesi, doğal ve anlaşılır bir şekilde duyulmaktadır.

3.2.2. Biçimlendirici konuşma sentezleme sistemi

Biçimlendirici sentezleme, tarihsel olarak en temel konuşma sentezleme yöntemidir. Bu yöntemde, ses yolunun aktarım fonksiyonu, biçimlendirici frekansları ve biçimlendirici genlikleri benzetilerek modellenmektedir. Biçimlendirici konuşma sentezleme, yapay olarak biçimlendirici özelliklerin yeniden oluşturulmasını içermektedir. Biçimlendirici sentezlemede, ses yolundaki filtreleme etkisi dikkate alınarak konuşmanın oluşturulması hedeflenmektedir. Ses yolunun özellikleri, bir konuşma sentezleyici tarafından modellenmekte ve bu model üzerinden konuşma üretilmektedir [14].

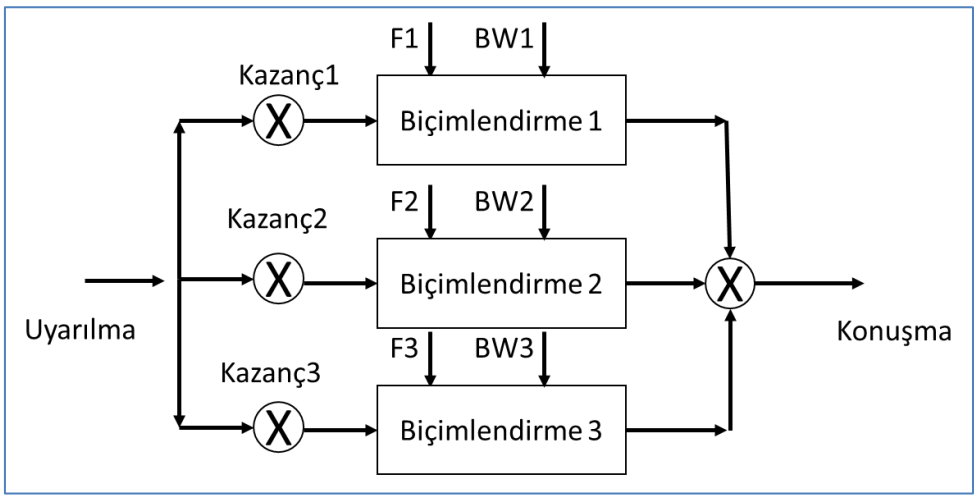
Biçimlendirici sentezleme, seri ve paralel olmak üzere iki temel türe ayrılabilmektedir. Seri biçimlendirici sentezleme, önce bir ses kaynağı üretilir ve ardından bu kaynak sesi, ses yolundaki filtreleme etkisiyle biçimlendirerek konuşma üretilmektedir. Paralel biçimlendirici sentezlemede ise ses kaynağı ve ses yolunun filtreleme etkisi eşzamanlı olarak modellenmektedir ve konuşma sentezlenmektedir.



Şekil 3.3. Seri biçimlendirici konuşma sentezleme sistemi

Şekil 3.3'te gösterilen seri biçimlendirici sentezleme sistemleri, sadece biçimlendirici frekanslarına ihtiyaç duyan bir denetim bilgisi kullanılmaktadır. Ünlülerin sentezlenmesi için, biçimlendirici frekanslarındaki genlik değerlerinin ayrı ayrı kontrol edilmesine gerek yoktur. Seri biçimlendirici sentezleme, sürtünücü ve patlamalı ünsüzleri sentezlemede paralel yapıya göre daha az başarılı olduğu belirtilmektedir. Ancak paralel yapıya kıyasla daha az denetim bilgisine sahip olduğu için daha kolay bir şekilde gerçekleştirilebilmektedir.

Bu sistemde, biçimlendirici frekansları, konuşmanın ağız boşluğunda oluşan rezonanslar ve belli bir ünlünün nasıl şekillendirildiğiyle ilgili bilgileri ifade edilmektedir. Bu frekans bilgileri, konuşma sesinin doğru bir şekilde üretilmesi için kullanılmaktadır. Seri biçimlendirici sentezleme, ünlü seslerin sentezlenmesinde kullanılan bir yöntemdir. Ancak sürtünücü ve patlamalı ünsüzlerin sentezlenmesi konusunda seri biçimlendirici sentezleme yöntemi bazen yetersiz kalabilmektedir. Bu tür ünsüzlerin doğru bir şekilde sentezlenmesi için daha fazla denetim bilgisine ve detaylı modellemeye ihtiyaç duyulabilmektedir [19].



Şekil 3.4. Paralel biçimlendirici konuşma sentezleme sistemi

Şekil 3.4'te gösterilen paralel biçimlendirici sentezleme yöntemi, her bir biçimlendirici frekansı için ayrı frekans, genlik ve bant genişliği parametrelerine ihtiyaç duymaktadır. Bu nedenle, daha karmaşık bir yapıya sahiptir. Bu yöntem, geniz (nasal) ünsüzlerin ve sürtünücü ünsüzlerin sentezinde daha iyi sonuçlar verdiği gözlemlenmiştir. Ancak ünlülerin sentezini gerçekleştirmede başarılı olamamaktadır.

Paralel biçimlendirici sentezleme yöntemi, konuşmayı üretirken konuşma organlarının fiziksel özelliklerini taklit etmeye çalışmaktadır. Bu yöntemde, konuşma organlarındaki titreşimlerin frekans, genlik ve bant genişliği gibi parametrelerine odaklanmaktadır. Bu şekilde, farklı ünsüzlerin doğru şekilde üretilebilmesi amaçlanmaktadır. Ancak ünlülerin sentezlenmesi daha zor bir süreçtir. Ünlülerin üretimi, konuşma organlarının açıklığı ve pozisyonu gibi birçok faktöre bağlıdır. Bu nedenle, ünlülerin sentezlenmesi paralel biçimlendirici sentezleme yöntemiyle tam olarak başarılamamıştır [19].

3.2.3. Söyleyiş konuşma sentezleme sistemi

Söyleyiş sentezleme, insanın söyleyiş davranışının doğrudan modellenmesi yoluyla konuşma üretme yöntemidir. Bu yaklaşım, insan ses üretim mekanizmasında nasıl oluşturulduğunu en hassas şekilde modelleyerek yüksek kaliteli konuşma üretmeyi hedeflemektedir. Bu nedenle, söyleyiş sentezleyicileri prensipte en tatmin edici ve teorik olarak en doğru yaklaşıma sahip olabilmektedir. Ancak, ses yolundaki organların davranışlarını modellemek zorlu bir görevdir. Ses telleri, ağız ve geniz boşlukları, dil, dudaklar vb. gibi organların davranışlarını tam olarak modellemek ve bu organların görüntülerini elde etmek zorlu bir süreçtir. Bu nedenle, söyleyiş sentezlemenin pratikte uygulanması diğer yöntemlere göre daha zorlu bir yöntemdir.

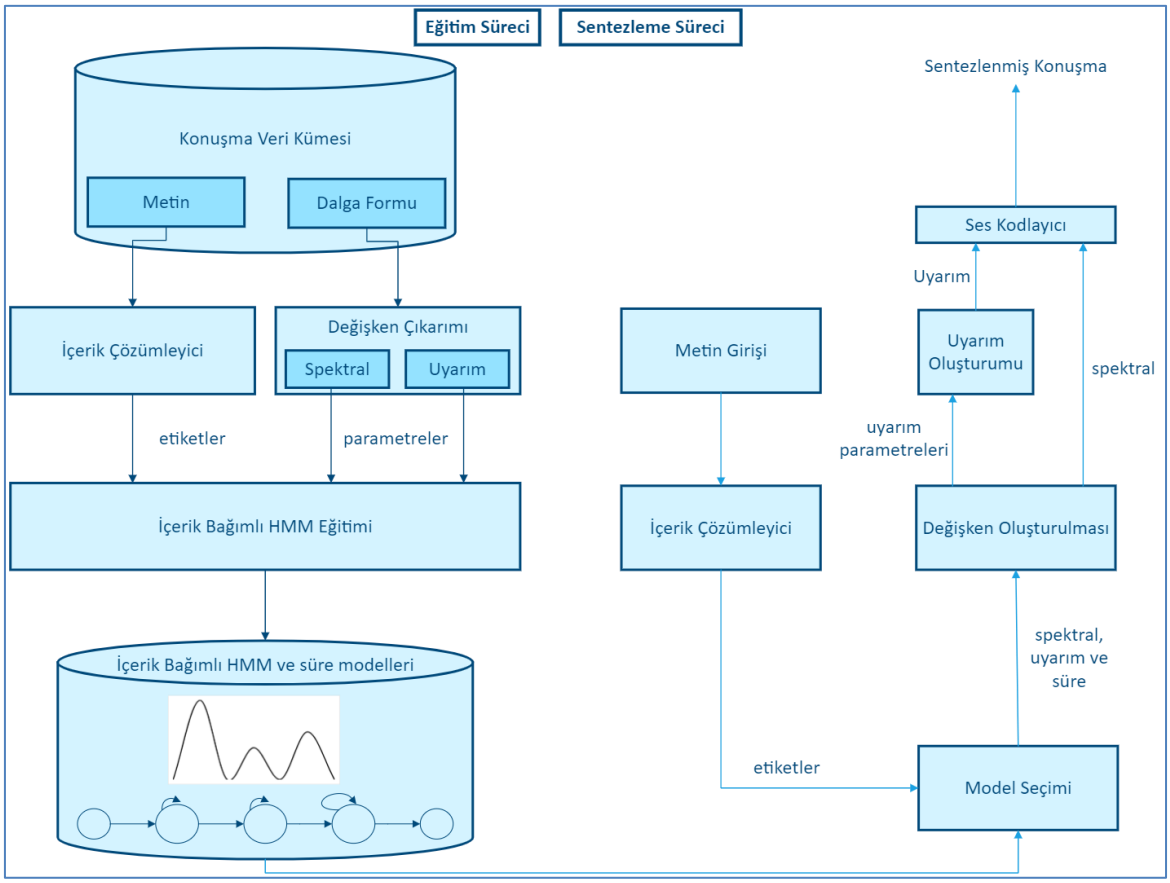
Ses yolundaki organların davranışlarını modellemek için kullanılan yöntem, çok sayıda parametreyi içerir ve bu modelin elde ettiği konuşma kalitesinin biçimlendirici sentezleyicilere göre daha iyi sonuçlar vermesi beklenmektedir. Özellikle konuşmanın geçiş bölümlerinde daha doğal sonuçlar elde etmek mümkün olabilmektedir. Ancak, bu yöntemde bir kalite veya özellik artarken başka bir kalite veya özellik azalabilmektedir. Bu tür bir takas durumu, diğer yönteme göre daha sık ortaya çıkabilmektedir. Sonuç olarak, söyleyiş sentezleme yöntemi konuşma üretiminde yüksek kaliteli sonuçlar elde etme potansiyeline sahip olsa da pratikte uygulanması oldukça zor ve karmaşık bir yöntemdir [15, 19].

3.2.4. İstatiksel parametrik konuşma sentezleme sistemi

İstatistiksel parametrik konuşma sentezleyicileri, konuşma sentezi için model tabanlı bir yaklaşım kullanılmaktadır. Bu yöntemde, birimlerin depolanması yerine, her bir birime karşılık gelen modeller bir havuzda saklanmaktadır. Model tabanlı yaklaşımda, konuşma parametreleri parametrelendirilmektedir ve bu parametreler için istatistiksel yöntemler kullanılarak modeller oluşturulmaktadır. Bu nedenle, bu yaklaşıma istatistiksel parametrik konuşma sentezi denir.

Bu yaklaşımda, konuşma üretimi için önceden elde edilen konuşma örneklerinden istatistiksel bilgiler kullanılmaktadır. Bu bilgiler, konuşma parametrelerinin istatistiksel dağılımını ve ilişkilerini tanımlamak için kullanılmaktadır. Ardından, verilen bir metne dayanarak istatistiksel model kullanılarak konuşma parametreleri sentezlenmektedir. Bu parametreler daha sonra konuşma sentezini oluşturmak için sinyal işleme teknikleriyle kullanılmaktadır.

İstatistiksel parametrik konuşma sentezi, genellikle daha doğal ve akıcı konuşma üretimine olanak sağlamaktadır. Ancak, bu yaklaşımın dezavantajı, büyük miktarda veri ve hesaplama gücü gerektirmesidir. İstatistiksel modellerin oluşturulması ve kullanılması zaman alıcı ve kaynak yoğun olabilmektedir [16].



Şekil 3.5. SMM tabanlı konuşma sentezleme modeli [30]

İstatistiksel konuşma sentezleme alanında yaygın olarak kullanılan temel yöntem, Saklı Markov Modeli (SMM) olarak bilinmektedir. SMM tabanlı konuşma sentezleme sistemlerinde, ses birimleri (fonemler, heceler vb.) modellenirken içeriğe bağımlı SMM'ler kullanılmaktadır. Bu yöntemle gerçekleştirilen konuşma sentezlemesi, eğitim ve sentez olmak üzere iki aşamada gerçekleştirilmektedir.

Eğitim sürecinde, birkaç saatlik ses verisi, çok değişkenli bir Gauss dağılımıyla modellenmektedir. Makine, hangi özniteliklerin her bir ses biriminde kullanılacağını belirlemek için veri setinden özniteliklerin dağılımını öğrenmektedir. Bu eğitim aşamasında, her bir ses birimi için oluşturulan modeller, ilgili öznitelikleri temsil etmektedir.

Sentez aşamasında ise, sentezlenecek metne göre oluşturulan modeller bir araya getirilerek yumuşak geçişli öznitelikler oluşturulmaktadır. Bu öznitelikler daha sonra ses kodlayıcıya verilerek konuşma sinyalleri elde edilmektedir. Yani, sentez aşamasında metindeki her bir bölümün karşılık gelen modelleri kullanılarak konuşma sesleri üretilmektedir.

Bu yöntem, istatistiksel modellerin kullanılmasıyla konuşma sentezlemesi gerçekleştirir ve genellikle iyi sonuçlar vermektedir. Ancak, doğru ve akıcı konuşma üretmek için büyük miktarda eğitim verisi ve karmaşık istatistiksel hesaplamalar gerektirmektedir [30].

3.2.5. Derin öğrenme tabanlı konuşma sentezleme sistemi

Derin öğrenme tabanlı konuşma sentezleme sistemleri, dil özelliklerini akustik özelliklere doğrudan haritalayan derin sinir ağlarının olağanüstü etkinliğini kullanmaktadır. Bu sistemler, verilerin doğal özelliklerini öğrenmek için derin sinir ağlarını kullanmaktadır ve dil özelliklerinden akustik özelliklere doğrudan bir haritalama gerçekleştirmektedir. Konuşma sentezi için derin öğrenme tabanlı birçok model önerilmiştir, bunlar arasında Kısıtlanmış Boltzmann Makineleri (KBM), Çoklu Dağıtık Derin İnanç Ağları (Çoklu Dağıtık DİA), Derin Karışık Yoğunluk Ağları (DKYA), Derin İki Yönlü Uzun-Kısa Vadeli Bellek (DİYUKSB), WaveNet, Tacotron ve DeepVoice3 gibi yöntemler bulunmaktadır.

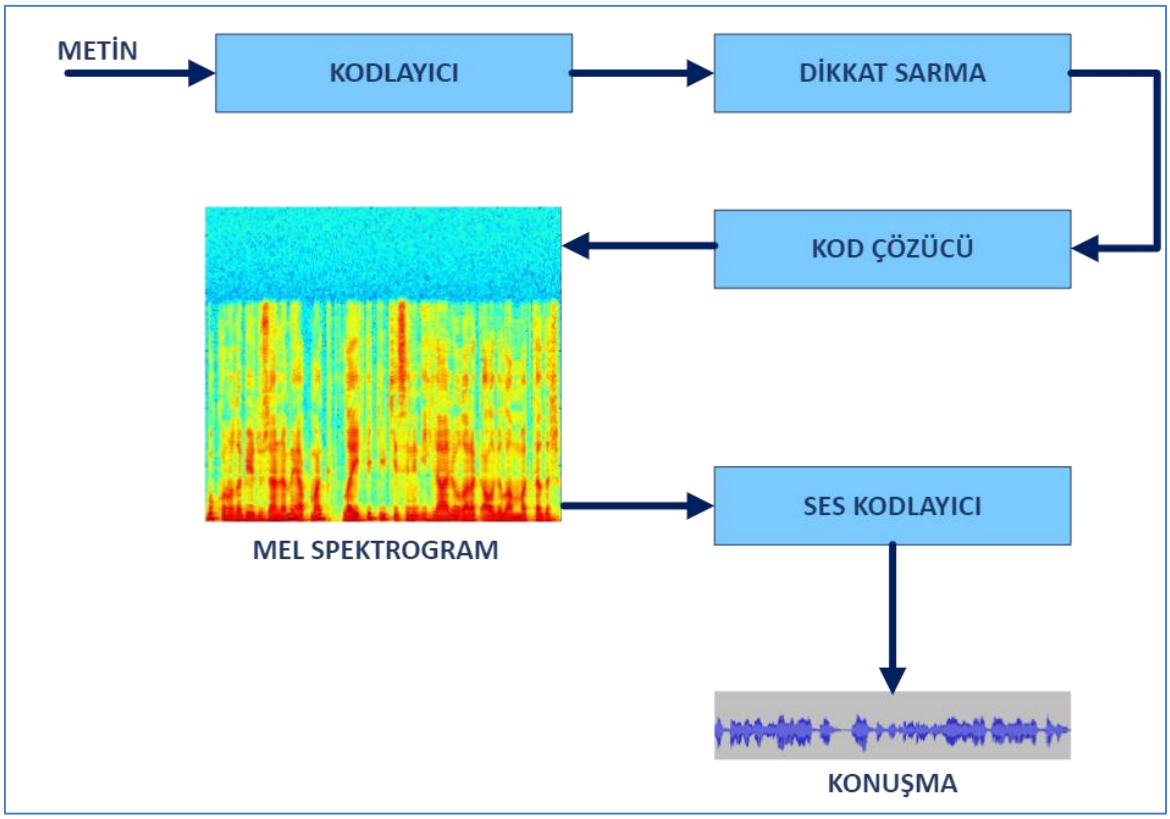
Bu yöntemlerin her birinin çeşitli avantajları ve dezavantajları vardır. KBM'ler, yüksek boyutlu spektral zarfların dağılımlarını daha iyi tanımlayabilme yeteneği nedeniyle sentez sorununu hafifletebilmektedir. Ancak eğitim verilerinin parçalanmasıyla ilgili bir sorunları vardır. Çoklu dağıtık DİA, bağlam bilgisini ve akustik özelliklerin ortak dağılımını modelleyebilme yeteneğiyle KBM'lerden farklıdır ve eğitim verilerinin parçalanma sorunu yaşamamaktadır. Ancak oluşturulan konuşmanın kalitesi düşük olabilmektedir. DKYA, tek kiplik problemini çözebilmekte, ancak sınırlı bağlamlardan yararlanabilmektedir ve her çerçeve bağımsız olarak eşlenmektedir. DİYUKSB, bağlamsal bilgilerden tam olarak yararlanabilmektedir, ancak dalga biçimini sentezlemek için bir ses kodlayıcıya ihtiyaç duymaktadır. WaveNet, işlenmemiş ses dalga formlarını güçlü bir şekilde üretebilmektedir, ancak yavaş çalışmakta ve ön uçtan gelen hatalar doğrudan konuşma sentezini etkileyebilmektedir. Tacotron ise tamamen uçtan uca konuşma sentez modeli olarak yüksek kaliteli konuşma dalga formları üretebilmektedir, ancak modeli eğitmek maliyetli olabilmektedir. Evrişimli Sinir Ağları (ESA) DeepVoice3, paralel hesaplamayı mümkün kılmaktadır ve hızlıdır, ancak konuşma kalitesi açısından düşük sonuçlar üretebilmektedir [3].

Konuşma sentezleme sistemleri arasında birleştirmeli konuşma sentezleyici sistemlerin Türkçe dil yapısına en uygun teknik olduğu görülmektedir. Bu sistemler doğallık ve

anlaşılırlık açısından iyi sonuçlar vermektedir, ancak ses birimleri arasındaki geçişlerde doğallık sorunları yaşanabilmektedir. Ayrıca, bu sistemler yüksek boyutlu veri kümelerine ihtiyaç duymaktadır.

Bu tekniklerin uygulanması sırasında araştırmacılar farklı sorunlarla karşılaşmıştır. Bu sorunların giderilmesi amacıyla farklı teknikler üzerinde çalışmalar yapılmış ve istatiksel parametrik konuşma sentezleme sistemleri geliştirilmiştir. Bu yeni sistemlerde eğitim süresi masraflı olsa da uçtan uca konuşma sentez modeli kullanılarak yüksek kaliteli dalga formları üretme başarıları elde edilmiştir. Türkçe için henüz bu yöntemlerle yapılmış çalışmalar bulunmasa da İngilizce için yapılan çalışmalarda doğallık açısından Ortalama Görüş Puanı (OGP) değerinin 4.526 gibi yüksek bir değere ulaşıldığı görülmüştür. İngilizcede doğal insan konuşması için OGP'nın 4.582 olduğu göz önüne alındığında, başarılı sonuçlar elde edildiği belirtilmektedir [31].

Bu bulgular, konuşma sentezleme sistemlerinin Türkçe diline uygun olarak birleştirmeli konuşma sentezleyici sistemlerin tercih edilebileceğini ve istatiksel parametrik yöntemlerin yüksek kalitede konuşma sentezi sağlama potansiyeline sahip olduğunu göstermektedir. Türkçe üzerinde yapılan ileri çalışmalarla bu yöntemlerin etkinliği ve doğallığı daha da artırılabilmektedir.



Şekil 3.6. Derin öğrenme tabanlı konuşma sentezleme sistemi genel mimarisi

Diğer dillerde yapılan çalışmalar, derin öğrenme tabanlı uçtan uca konuşma sentezleme modelleriyle yüksek kalitede dalga formu üretiminin mümkün olduğunu göstermektedir. Bu yöntem genellikle derin sinir ağları kullanılarak gerçekleştirilmektedir. Şekil 3.6'da derin öğrenme tabanlı uçtan uca konuşma sentezleme modelinin temel bir yapısı görülmektedir. Bu yapıda kodlayıcı (encoder), dikkat sarmalama (attention wrapping), kod çözücü (decoder) ve ses kodlayıcı (vocoder) bulunmaktadır. Bu yapıda temel olarak iki işlem gerçekleştirilmektedir. İlk olarak, girilen metin kodlayıcıdan kod çözücüye kadar mel spektrogramlarına dönüştürülmektedir.

Spektrum, zaman alanından frekans alanına dönüşmüş sinyalleri ifade etmektedir. Zamanla değişen ses sinyallerinin frekansları da zamanla değişmektedir. Bu nedenle, bu sinyallerin spektrumunu temsil etmek için sinyal çerçeveleri üzerinde Ayrık Fourier Dönüşümü (AFD: Discrete Fourier Transform) uygulanarak spektrogram elde edilmektedir. Eşit mesafelerdeki frekanslar, dinleyiciye eşit derecede algılanmaktadır, bu nedenle bir perde birimine "mel skalası" denmektedir. Bu nedenle, mel spektrogramu, spektrogramun mel ölçeğine dönüştürülmüş halidir.

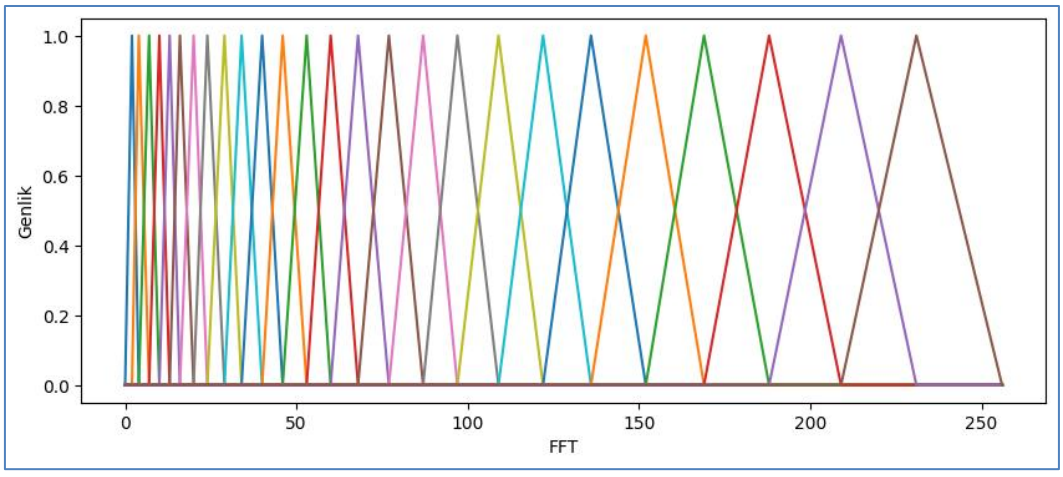
Mel spektrogramları elde etmek için, ses sinyali öncelikle zaman-dalga formuna göre küçük parçalara bölünmektedir. Bu parçalara "çerçeve" (frame) denir ve genellikle 20-40 milisaniye gibi bir süreyi kapsamaktadır. Her çerçeve, bir pencere fonksiyonu ile çarpılmaktadır. Hanning penceresi, en yaygın kullanılan pencere fonksiyonlarından biridir [32]. Hanning penceresi, her çerçeve için bir ağırlıklandırma sağlamaktadır ve spektrum analizinde istenmeyen yan etkileri azaltmaktadır [33].

Daha sonra her pencere için AFD yapılmaktır. Bu işlem yapılırken genellikle hesaplama maliyetini azaltmak için kullanılan AFD uygulaması Hızlı Fourier Dönüşümüdür (HFD: Fast Fourier Transform) [34]. Bir sinyalin AFD'ü Denklem 3.1'de gibi tanımlanmaktadır.

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\left(-jkn\frac{2\pi}{N}\right)}, \quad k=0, 1, \dots, N-1 \tag{3.1}$$

Dönüşüm sonucu elde edilen spektrogramlar mel ölçeğe mel filtre bankaları vasıtası ile dönüştürülmektedir. Bu filtre bankaları üçgen filtrelerden oluşmaktadır ve belirli frekans aralıklarında duyarlıdır. Her bir filtrenin tepkisi, frekans bileşenlerinin Mel ölçeği ile ağırlıklı bir kombinasyonunu temsil etmektedir. Mel ölçeğe dönüşüm için genellikle Denklem 3.2'de görülen denklem kullanılmaktadır. Bu denklem baz alınarak Şekil 3.7'de örnek bir mel filtre bankası gösterilmiştir.

$$f_{mel} = 2595 * \log_{10}\left(1+\frac{f}{700}\right) \tag{3.2}$$



Şekil 3.7. Örnek mel filtre bankası

Şekil 3.6'daki yapıda yapılan ikinci işlemde de ses kodlayıcılar mel spektogramları sese dönüştürmektedir. Bu yapıda kodlayıcı, girilen metin ifadesini algılayıp kendi içinde hecelere bölünmesini sağlamaktadır. Dikkat sarma bölümü bu yapıda önemli bir yer tutmaktadır. Modelin kendi öğrenimi gerçekleştirdiği aşama dikkat sarma aşamasıdır. Her karaktere karşılık gelen ses burada öğrenilmektedir. Ek olarak sentezleme sırasında kod çözücü modülün ses kodlayıcıya gönderdiği hecelerin takibini de yapmaktadır. Örneğin 'selam' metinsel ifadesini sentezlemek isteniyorsa, 'se' hecesi seslendirildikten sonra devamında gelecek hecenin takibi dikkat sarma aşamasında yapılmaktadır. Kod çözücü aşaması ise dikkat sarmadan tarafından gönderilen seslendirilecek hece bilgisine göre seslendirilecek hecenin spektrogramlarını üreterek ses kodlayıcıya iletmektedir. Ses kodlayıcı ise gelen spektrogramları gerçek ses sinyallerine dönüştürmektedir.

Derin öğrenme tabanlı konuşma sentezleme yöntemiyle gerçekleştirilen modeli eğitmek oldukça maliyetlidir. Bu eğitimin en önemli parçası yeterli bir veri kümesinin olmasıdır. Derin öğrenme tabanlı konuşma sentezleme teknikleri için korpus (derlem) şeklinde veri kümeleri hazırlanmaktadır. Bu veri kümelerinde dikkat edilen özellikler konuşma dili, konuşma süreleri, konuşmacının cinsiyeti, gibi bilgilerdir. Türkçe üzerine hazırlanmış derlemlerin konuşma tanıma sistemleri üzerine ve istatiksel parametrik konuşma sentezleme sistemleri üzerine uygulandığı görülmektedir [35, 36].

3.3. Veri Kümesi

Derin öğrenme tabanlı konuşma sentezleme yöntemi için metin ve ses dosyalarının eşleştirilmesi şeklinde hazırlanmış derlem tipi veri kümeleri hazırlanmaktadır. Bu veri kümelerinde dikkat edilmesi gereken özellikler Çizelge 3.1'de verilmiştir.

Çizelge 3.1. Veri kümesi için özellik/birim çizelgesi

Özellik	Birimi
Konuşmanın Dili	Türkçe, İngilizce, İspanyolca vs.
Konuşmanın Süresi	Saat
Konuşmanın Kategorisi	Metin konusu

Çizelge 3.1. (devamı) Veri kümesi için özellik/birim çizelgesi

Özellik	Birimi
Konuşmacının Cinsiyeti	Kadın/Erkek
Konuşmacının Yaşı veya Yaş Düzeyi	Çocuk, yaşlı, 18 ve üzeri veya Genç
Konuşmacının Şive Türü	İstanbul Türkçesi, Ege Şivesi vs.
Konuşma Parçasının Ortalama Süresi	Saniye
Konuşma Parçasının Minimum Süresi	Saniye
Konuşma Parçasının Maksimum Süresi	Saniye
Benzersiz Kelime/Cümle Sayısı	Adet
Toplam Kelime/Cümle Sayısı	Adet

Derlem tipinde hazırlanmış veri kümesinin içeriği, günlük konuşmalar, edebi bir metin, haber bülteni içeriği, birbirini izleyen bağımsız metinler veya mantıksal bağlantısı olmayan bir dizi bağımsız cümle olabilmektedir. Bu konuların çeşitlendirilerek özenle seçilmesi gerekmektedir. Bu içerikteki toplam kelime sayısı ve benzersiz kelime sayısı veri kümesinin daha dengeli bir yapıda olmasını sağlayacaktır. Özellikle üretken dil yapısına sahip Türkçe için daha fazla benzersiz kelimenin seslendirilmesi gerekmektedir. Konuşmanın toplam uzunluğu yine veri kümesi için en önemli noktalardan birisidir. Ayrıca konuşma her bir konuşma parçasının uzunluğu geliştirilen model yapısında kullanılacak batch size (bir yinelemede kullanılan eğitim örneklerinin sayısı) ve çevrim (epoch) (model eğitim adımlarının her biri) süresi üzerinde doğrudan etkiye sahiptir. Metin ve ses kaydı olarak birebir eşleştirilen konuşmalar kısa konuşma parçaları halinde saklanmalıdır.

Veri kümesinde yer alan konuşma içeriğinin yanı sıra konuşmacı hakkındaki bilgiler de geliştirilecek olan konuşma sentezleme modelinin başarısını etkilemektedir. Konuşmacının cinsiyeti, konuştuğu dil ve kullandığı şive sentezlenen konuşmanın doğallığı ve anlaşılırlığına etki etmektedir. Literatürde aynı dil için hazırlanan veri kümelerinde farklı cinsiyete sahip konuşmacılardan faydalanılmıştır. Konuşmacının kullandığı şive türü sentezlenecek konuşmanın da aynı şive yapısına sahip olmasına neden olacaktır. Bu nedenle olabildiğinde şive kullanımından kaçınılmalıdır. Literatürde aynı dilin farklı bölgelerdeki

konuşma yapısına göre farklı veri kümeleri hazırlanmıştır.

Yukarıda belirtilen özelliklere ek olarak veri kümesinde çeşitli bazı çalışmalarında yapılması gerekmektedir. Derlem tipi hazırlanmış veri kümeleri hazırlanırken veri toplamanın yanında toplanan veri üzerinde yapılan işlemlerde çok önemlidir. Konuşma sentezleme sistemleri belirli bir alana özgü bir çalışma ise veri kümesi konuşma sentezleme sistemleri için alana özgü en iyi örnekleri içerecek şekilde hazırlanmalıdır. Ancak büyük ölçekli bir çalışma ise Türkçe için doğal kesintisiz bir konuşma sentezi üretmek için veri kümesi, dile ait en iyi örnekleri içeren ve büyük ölçekli hazırlanmalıdır. Veri kümesinde yanlış telaffuz edilmiş veya yanlış eşleştirilmiş ifadeler yer almamalıdır. Kayıttaki gürültülerin ve sessiz kısımların temizlenmiş olması gerekir. Kayıttaki ses kalitelerindeki farklılıklar bu yöntem ile gerçekleştirilen sistemlerin başarısını etkilemektedir. Konuşma parçaları sürelerinin dağılımı belirli istatiksel yöntemler kullanılarak yapılmalıdır. Veri kümesi, toplam uzunluğu konuşma sentezinin yapılacağı dile göre değişkenlik göstermektedir. Toplam uzunluğu kısa olan veri kümeleri aşırı öğrenmeyi önlemek ve modeli doğru bir şekilde eğitmek için yetersizdir. Hazırlanan veri kümesiyle ilgili dil için sık kullanılan sözcükler dışında, telaffuzu zor olan kelimeler içerecek şekilde hazırlanmalıdır [37].

İstanbul Türkçesi konuşan bir konuşmacı tarafından seslendirilen Türkçe veri kümesi hazırlanırken Türkçe atasözleri, haber bültenleri, raporlar, kitap bölümleri, dini, edebi ve tarihi yazılar, dünya ülkelerinin isimleri, Türkiye Cumhuriyeti'nin illeri, ilçeleri, çeşitli mahalle ve cadde isimleri, Türkiye Cumhuriyeti'nin çeşitli kurum ve kuruluşlarına ait isimler, sayılar ve mantıksal bağlantısı olmayan bir dizi bağımsız cümleler kullanılmıştır. Hazırlanan metin verisi daha önce profesyonel eğitim almış gönüllü bir konuşmacı tarafından seslendirilmiştir. Konuşma kayıtları Audio Technica marka stüdyo tipi kardioid kondansatör özelliğine sahip bir mikrofon yardımı elde edilmiştir. Kayıt esnasında çevre gürültüsü ve ses patlamalarını azaltmak için bir pop filtresi kullanılmıştır. Bu mikrofonun tercih edilmesindeki en önemli özellikler kullanımının basit, gürültüsüz ve çok yönlü ses alabilmesidir. Kayıtların alınması için ofis ortamında ses yalıtımı yapılarak bir stüdyo hazırlanmıştır. Elde edilen veri kümesine ait istatiksel özellikler Çizelge 3.2'de verilmiştir.

Çizelge 3.2. Türkçe veri kümesi için özellik birim çizelgesi

İstatiksel Özellikler	Türkçe Veri Kümesi
Toplam Kelime Sayısı	109826 adet
Benzersiz Kelime Sayısı	35050 adet
Toplam Karakter Sayısı	745011 adet
Toplam Konuşma Uzunluğu	12 saat 38 dakika 59 saniye
Toplam Konuşma Parçası Sayısı	8480 adet
Ortalama Konuşma Parçası Uzunluğu	5,19 saniye
Minimum Konuşma Parçası Uzunluğu	0,54 saniye
Maksimum Konuşma Parçası Uzunluğu	9,85 saniye
Konuşma Parçası Başına Düşen Ortalama Kelime Sayısı	12,95
Konuşmacının Cinsiyeti	Erkek
Konuşmacının Yaş Düzeyi	Genç
Konuşmacının Ağız Türü	İstanbul Türkçesi

Hazırlanan veri kümesi toplam 109826 adet kelimeden oluşmaktadır. 109826 kelimenin 35050 tanesi benzersiz kelimelerdir. Toplam 12 saat 38 dakika 59 saniye olan bu veri kümesi 8480 adet konuşma parçasından oluşmaktadır. Minimum uzunluğu 0,541 saniye ve maksimum uzunluğu 9,859 saniye olan konuşma parçalarının ortalama uzunluğu 5,19 saniyedir. Konuşma parçası başına düşen kelime sayısı yaklaşık 13 adettir. Konuşmanın yanında daha önce de belirtildiği gibi konuşmacı hakkındaki bilgiler de veri kümesinin başarısını etkilemektedir. Bu metin genç, gönüllü bir erkek konuşmacı tarafından seslendirilmiştir. Konuşmacı, daha önceden hazırlanan metinleri İstanbul Türkçesi kullanarak seslendirmiştir.

Türkçe konuşma sentezleme sistemleri için hazırlanan veri kümeleriyle ilgili literatürde, noktalama işaretlerindeki ve kısaltmalardaki telaffuz farklılıkları, nümerik ifade kombinasyonlarının okuma alışkanlıkları ve ses değişimlerinin göz ardı edildiği belirtilmektedir [38]. Bu nedenle, bu çalışmanın kapsamında, elde edilen konuşma kayıtları

5 farklı kullanıcı tarafından kontrol edilmiştir.

Elde edilen kayıtlar, kontrol amaçlı olarak 4 farklı kullanıcıya eşit olarak bölünmüştür. Kullanıcılar, bütün kontrolleri gerçekleştirdikten sonra son kontrolcü, kontrol edilen tüm kayıtları birleştirmiş ve tekrar kontrol işlemini gerçekleştirmiştir. Bu yöntem, elde edilen konuşma kayıtlarının doğruluğunu artırmak ve farklı kullanıcıların farklı perspektiflerle kayıtları kontrol etmesini sağlamak amacıyla kullanılmıştır. Bu şekilde, elde edilen konuşma kayıtları daha kapsamlı bir kontrol sürecinden geçmiş ve daha yüksek kalitede sonuçlar elde etmek için çaba harcanmıştır.

Veri kümesinin kontrol işlemlerinde konuşma içerisindeki sessizlik anları konuşma kaydından silinmiştir. Bu işlemin ana nedeni konuşma sentezleme sistemlerinin geliştirilmesinde kullanılan modellerde ses ve metin eşleştirmesinin yapılmasıdır. Ses-metin hizalamaları yapılırken ses kayıtlarının başında, sonunda ve içerisinde yer alan sessizlik anları model geliştirilirken gerçekleştirilen hizalamaları olumsuz etkilemektedir. Bu nedenle çalışma kapsamında elde edilen ses kayıtlarındaki sessizlik anları kontrol edilerek ilgili kayıt içerisinden çıkarılmıştır.

Sessizlik anlarının tespiti ve kaldırılması için, durağan ve aralıklı konuşmalardaki sessizlik bilgisi öncelikle belirlenmektedir. Bu işlemde, her bir konuşma parçasındaki ses dalgalarının örneklem frekansı ve genliği kullanılmaktadır. Belirli bir frekans ve genlik değerinin altında olan seslerin sessizlik olduğu varsayılmaktadır. Eşik değerinin altına düşüldüğü ilk andan itibaren bir zaman sayacı kullanılmaktadır. Bu sayacın başlama zamanından itibaren sürekli olarak arttırılmaktadır. Sayacın belirli bir sınır değerine ulaşmasıyla sessizlik tespit edilmektedir.

Son adımda, sayacın başlama zamanından bitiş zamanına kadar olan kısımdaki konuşma içeriği kayıttan silinmektedir. Bu sayede sessizlik süresi boyunca kaydedilen ses bilgisi ortadan kaldırılmış olmaktadır. Bu yöntem, ses dalgalarındaki frekans ve genlik değerlerini analiz ederek sessizlik anlarını tespit ederek ve bu sessizlik anlarını kayıttan çıkarılmaktadır. Bu sayede konuşma kaydı daha akıcı hale getirilebilmektedir.

Otomatik olarak gerçekleştirilen sessizliklerin tespiti ve kaldırılması işleminin yanı sıra gerçek kullanıcılar tarafından sessizlik kontrolü gerçekleştirilmiştir. Kontrol işlemlerinde

ses kayıtlarının başlangıcı ve sonlarındaki sessizlik anları silinmiştir. Böylelikle geliştirilecek olan modellerde sessizlik anları dikkate alınmayacaktır. Metin-ses hizalamalarında zaman değişimi olmayacak ve model parametreleri herhangi bir sapmaya uğramayacaktır.

Kontrol işlemlerinde sadece sessizlik anları değil aynı zamanda metinlerin okunuşları da kontrol edilmiştir. Gerçek kullanıcılar tarafından gerçekleştirilen bu işlem sayesinde metinlerin telaffuzlarının doğruluğu kontrol edilmiştir. Sadece kelimelerin okunuşlarına değil aynı zamanda rakamlar, sayılar ve kısaltmaların da okunuşlarına da dikkat edilmiştir. Yanlış veya eksik telaffuzlar veri kümesinden çıkarılmıştır.

Türkçedeki kısaltmalar üzerine hem eğitim hem de eğitim sonrasında giriş metni üzerinde çalışabilecek ek bir liste hazırlanmıştır. Bu listede yer alacak kısaltmalar için Türk Dil Kurumu'nun belirlediği kısaltmalardan yararlanılmıştır. Kısaltmaların veri kümesine eklenmesinde yazılışları değil okunuşları dikkate alınmıştır. Böylelikle ses-metin hizalamasında karakter bazlı değil fonem temelli hizalamada kullanılabilecek bir veri sunulmuştur.

Nümerik ifade kombinasyonları için veri kümesinde seslendirilen metnin içine çeşitli nümerik ifadeler eklenmiştir. Nümerik ifadelerin veri kümesinde eklenmesi işleminde nümerik ifadelerin yazılışı değil okunuşları dikkate alınmıştır. Ayrıca para birimlerini ve bu birimlerin okunuşları veri kümesine eklenmiştir. Böylelikle hesaplama ve para birimleri ile ilgili konuşma sentezlerinin daha gürbüz sonuçlar vereceği ön görülmektedir.

Veri kümesinin kontrol işlemlerinde son işlem olarak noktalama işaretleri gözden geçirilmiştir. Metinlerin okunuşlarındaki vurguları net olarak ortaya koyabilmek ve konuşma sentezinde gerekli vurguları elde edebilmek için noktalama işaretleri tek tek kontrol edilmiştir. Bu kontrol işlemlerinde özellikle virgül ve nokta işaretlerine dikkat edilmiştir.

3.4. GlowTTS Mimarisi

GlowTTS, CoquiAI tarafından geliştirilmiş bir konuşma sentezleme modelidir. GlowTTS, girdi olarak doğal dil metinlerini alan ve gerçekçi, doğal konuşma sinyalleri üreten bir

modeldir. GlowTTS, temel olarak iki aşamadan oluşmaktadır. İlk aşama metin özetleme aşaması olarak adlandırılmaktadır. Bu aşamada, girdi olarak verilen metin sinyali, bir metin özetleme modeli kullanılarak işlenmektedir. Bu özetleme modeli, metin sinyalini bir özellik dizisi haline dönüştürmektedir. Bu özellik dizisi, konuşma sinyali üretmek için kullanılan bir formata dönüştürmektedir. Bu formata, "mel spektrogram" denir. Mel spektrogram, frekans spektrumunun zamana karşı değişimini görselleştiren bir formattır.

Diğer aşama ise Flow modeli aşaması olarak adlandırılmaktadır. Bu aşamada, özetlenmiş metin özelliği, bir "Flow" modeli kullanılarak ses sinyaline dönüştürülmektedir. Flow, bir tür yapay sinir ağıdır ve bir özellik dizisi alarak, belirli bir dağılıma uygun şekilde rastgele örneklem alarak ses sinyalini üretmektedir. Bu şekilde, özetlenmiş metin özelliği, gerçekçi ve doğal bir konuşma sinyaline dönüştürülmektedir. GlowTTS, birçok farklı konuşma stilini ve tonunu kapsayabilen geniş bir veri kümesiyle eğitilmektedir. Bu nedenle, model, farklı konuşma stillerini ve farklı konuşmacıların seslerini sentezleyebilmektedir. Ayrıca, model, kısa metinlerden uzun konuşmalar oluşturmak için kullanılabilmekte ve sınırlı veriyle bile iyi sonuçlar verebilmektedir.

Metin ve konuşma temsilleri arasında tekdüze ve atlamasız bir hizalama koşuluyla ilişkilendirilmiş bir mel spektrogram üretmek amacıyla tasarlanan GlowTTS için bu bölümde öncelikle eğitim ve çıkarım süreçleri formülleri verilmektedir. Daha sonra da bu süreçler için mimariler açıklanmaktadır.

GlowTTS [39], mel spektrogramların koşullu dağılımını $P_X(x|c)$ metni $P_Z(z|c)$ üzerinden dönüştürülmektedir. Akış temelli bir çözücü olan f_{dec}: $z \rightarrow x$ kullanarak modellemektedir. Burada x ve c, sırasıyla giriş mel spektrogramı ve metin dizisini ifade etmektedir. Değişkenlerin dönüşümü kullanarak, verilerin tam log-olasılığı aşağıdaki şekilde hesaplanabilmektedir:

$$\log P_X(x|c) = \log P_Z(z|c) + \log\left|\frac{\det\left(\partial f_{dec}^{1}(x)\right)}{\partial x}\right| \tag{3.3}$$

Veri ve öncül dağılımlarını ağ parametreleri θ ve hizalama fonksiyonu A ile parametrelendirilmektedir. Öncül dağılım P_Z, izotropik çok değişkenli Gauss dağılımıdır ve öncül dağılımın tüm istatistikleri, μ ve σ, metin kodlayıcı $f_{kodlayıcı}$ tarafından elde

edilmektedir. Metin kodlayıcı, metin koşulunu $c = c_{1:Tmetin}$, istatistiklere $\mu = \mu_{1:Tmetin}$ ve $\sigma = \sigma_{1:Tmetin}$ şeklinde eşlemektedir. Burada T_{metin} metin girişinin uzunluğunu göstermektedir. Formülde, hizalama fonksiyonu A, konuşmanın gizli temsilinin dizinini, $f_{kodlayıcı}$'den gelen istatistiklerin dizinine eşleyen bir haritalamadır: $A(j) = i$ ise $z_j \sim N(z_j ; \mu_i, \sigma_i)$. GlowTTS'nin metin girişini atlamaması veya tekrarlamamasını sağlamak için hizalama fonksiyonu A'nın artan ya da azalan fonksiyon, yani tekdüze (monotonic) ve örten fonksiyon (surjective) olduğunu varsayılmaktadır. Ardından, öncül dağılım aşağıdaki gibi ifade edilebilmektedir:

$$\log P_Z(z|c; \theta, A) = \sum_{j=1}^{T_{mel}} \log \mathcal{N}\left(z_j;\ \mu_{A(j)}, \sigma_{A(j)}\right), \tag{3.4}$$

burada T_{mel} giriş mel spektrogramının uzunluğunu belirtmektedir.

Amaçlanan, Denklem 3.5'teki gibi verinin log olasılığını maksimize eden parametreleri θ ve hizalamayı yani A'yı bulmaktır. Ancak, global çözümü bulmak hesaplama açısından mümkün değildir. Hesaplanamama sorununu ele almak için, parametrelerin ve hizalamanın arama alanını azaltarak hedef iki ardışık probleme bölünmektedir. Bunlar, mevcut parametreler θ'ye göre en olası tekdüze hizalama $A*$'yı Denklem 6'daki gibi arama ve parametreleri θ'yi güncellemek için log-olasılığı maksimize eden $\log P_x(x|c;\ \theta,\ A*)$. Uygulamada, bu iki problemi yinelemeli bir yaklaşımla ele alınmaktadır. Her eğitim adımında, önce $A*$ bulunur ve ardından θ gradyan inişi kullanarak güncellenmektedir. Değiştirilmiş amaç, Denklem 3.5'in evrensel çözümünü garanti etmemektedir. Ancak evrensel çözümün iyi bir alt sınırını sağlamaktadır.

$$\max_{\theta, A} \quad L(\theta, A) \quad = \quad \max_{\theta, A} \quad \log P_x(x|c; A, \theta) \tag{3.5}$$

$$A* = \quad \operatorname{argmax}_{A} \quad \log Px\ (x|c; A, \theta) = \quad \operatorname{argmax}_{A} \quad \sum_{j=1}^{T_{mel}} \log \mathcal{N}\left(z_j;\ \mu_{A(j)}, \sigma_{A(j)}\right) \tag{3.6}$$

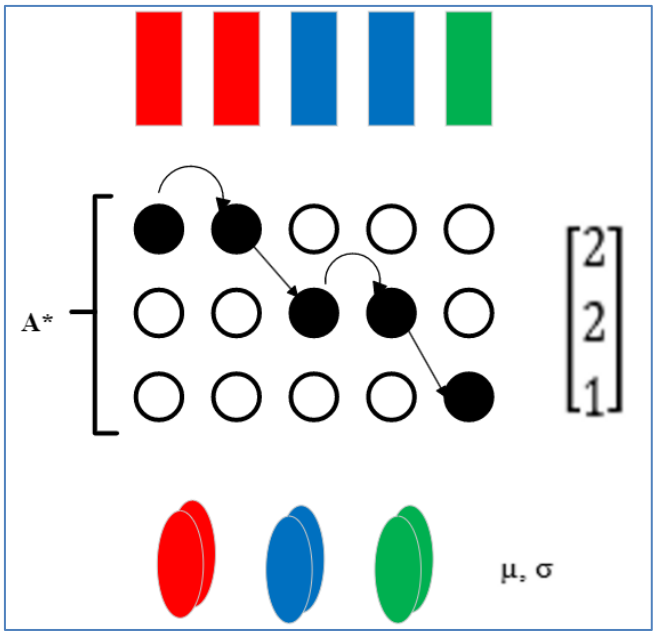
Hizalama arama problemini çözmek için, bir hizalama arama algoritması olan Monotonik Hizalama Araması (MHA: Monotonic Alignment Search) kullanılmıştır. Aşağıda MHA hakkında daha detaylı bilgi verilecektir. En olası monotonik hizalama $A*$'yi tahmin etmek için çıkarımda, aynı zamanda Denklem 3.7'deki gibi hizalama $A*$'den hesaplanan süre etiketiyle eşleşmesi için süre tahmin edici f_{ste} eğitilmektedir. Süre tahmin edici metin

kodlayıcının üzerine ekleniyor ve logaritmik alanda ortalama karesel kayıp hatası (OKKH) ile eğitilmektedir. Ayrıca, maksimum olabilirlik hedefini etkilememek için süre tahmin edicinin girdisine geri yayılım sırasında gradyanı kaldıran gradyanların durdurulması (gd: stop gradient) operatörü uygulanmaktadır. Süre tahmin edici için kayıp Denklem 3.8'de açıklanmaktadır.

$$d_i = \sum_{j=1}^{T_{mel}} 1_{A*(j)=i} \ , i = 1, \dots, n, T_{Metin} \quad (3.7)$$

$$L_{sınıflandırıcı} = OKKH\ (f_{ste}(gd[f_{kodlayıcı}(c)]), d) \quad (3.8)$$

Çıkarım sürecinde, önceki dağılımın istatistikleri ve hizalama, metin kodlayıcı ve süre tahmin edici tarafından tahmin edilmektedir. Ardından, önceki dağılımdan bir gizli değişken örnekleme yapılmaktadır ve bir mel spektrogram, gizli değişkenin akış tabanlı çözücü aracılığıyla dönüştürülerek paralel olarak sentezlenmektedir.



Şekil 3.8. Monotonik hizalama araması

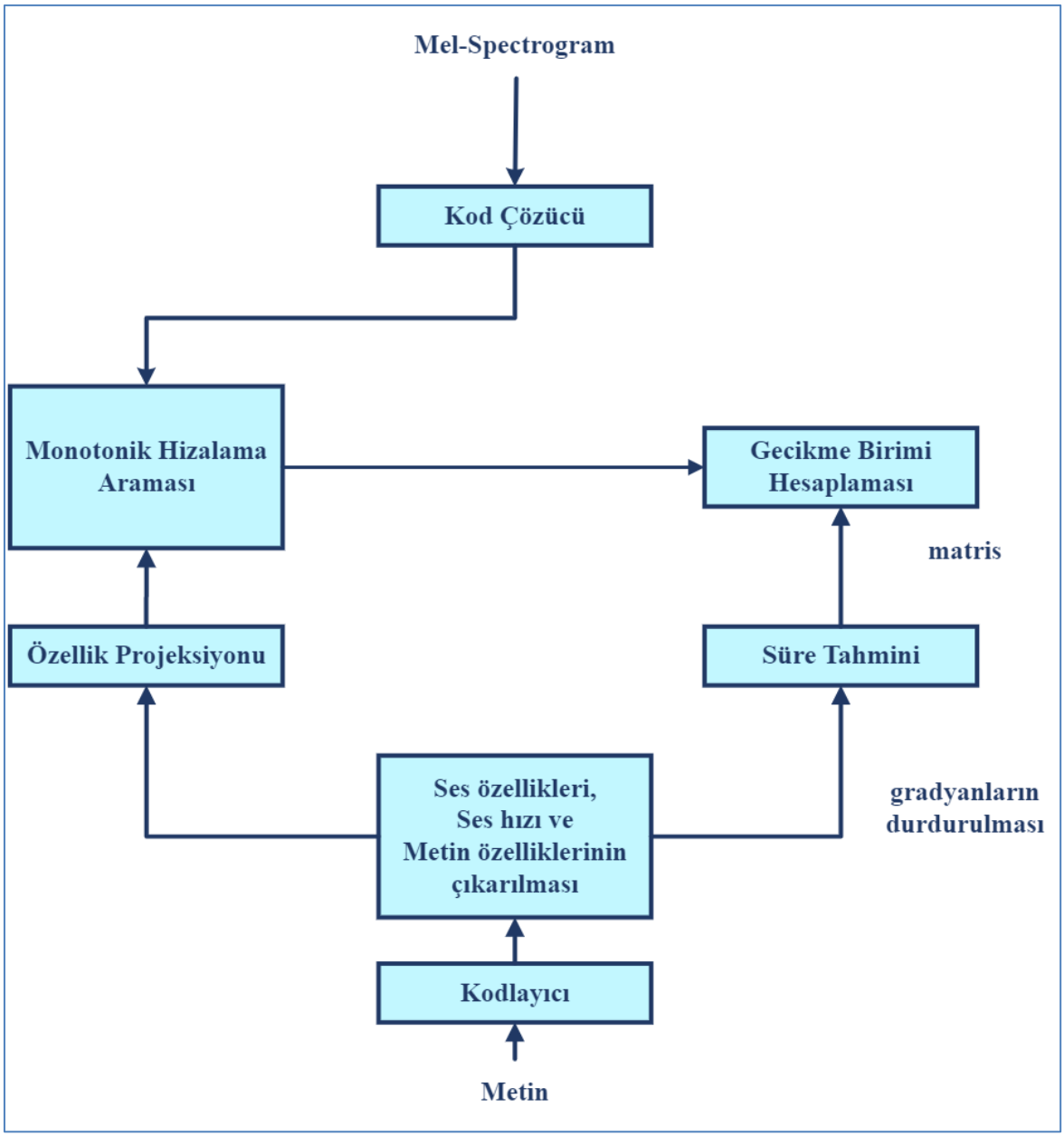
MHA, gizli değişken ile önceki dağılımın istatistikleri arasında en olası tekdüze hizalamayı aramaktadır. Bu istatistikler sırasıyla giriş konuşma ve metinden elde edilmektedir. Şekil 3.8'de MHA yapısı gösterilmektedir. İlk olarak, kısmi hizalamalar üzerinde özyinelemeli bir çözüm türetilmekte ve ardından tüm hizalama bulunmaktadır.

$Q_{i,j}$, önceki dağılımın istatistiklerinin ve gizli değişkenin i'nci ve j'nci elemanlarına kadar kısmen verildiği durumda maksimum log-olasılıktır. Denklem 3.9'da olduğu gibi, $Q_{i,j}$, $Q_{i-1,j-1}$ ve $Q_{i,j-1}$ ile özyinelemeli olarak formüle edilmektedir. Çünkü eğer kısmi dizilerin son elemanları z_j ve $\{\mu_i, \sigma_i\}$ hizalanmışsa, önceki gizli değişkeni z_{j-1}, tekdüzelik ve örten fonksiyon sağlamak için ya $\{\mu_{i-1}, \sigma_{i-1}\}$ ya da $\{\mu_i, \sigma_i\}$ ile hizalanmış olmalıdır.

$$Q_{i,j} = \max_{A} \sum_{k=1}^{j} \log \mathcal{N}\left(z_k;\ \mu_{A(k)}, \sigma_{A(k)}\right) = \max\left(Q_{i-1,j-1} Q_{i,j-1}\right) + \log \mathcal{N}\left(z_j;\ \mu_i, \sigma_i\right) \quad (3.9)$$

T_{Metin}'den T_{Mel}'e kadar Q değerleri iteratif olarak hesaplanmaktadır. Benzer şekilde, en olası hizalama $A*$, özyinelemeli ilişki, Denklem 3.9'a göre hangi Q değerinin daha büyük olduğunu belirleyerek elde edilebilmektedir. Bu nedenle, tüm Q değerleri önbelleğe alınarak dinamik programlama ile $A*$ verimli bir şekilde bulunabilmektedir; $A*$ değerlerinin hepsi hizalamanın sonundan, $A*(T_{mel}) = T_{metin}$, geri izlenir. Algoritmanın zaman karmaşıklığı $O(T_{metin} \times T_{mel})$'dir. Ayrıca, çıkarım sırasında MAS'a ihtiyaç duyulmamaktadır, çünkü hizalamayı tahmin etmek için süre tahmin edici kullanılmaktadır.

Bu kapsamda da yukarıda anlatılan süreçler için 2 adet mimari ele alınmıştır. Bu modellerden bir tanesi eğitim süreci için, diğeri ise çıkarım süreci için geliştirilmiştir. Şekil 3.9'da eğitim süreci için, Şekil 3.10'da ise çıkarım süreci için mimari gösterilmiştir.



Şekil 3.9. GlowTTS'in eğitim süreci mimarisi

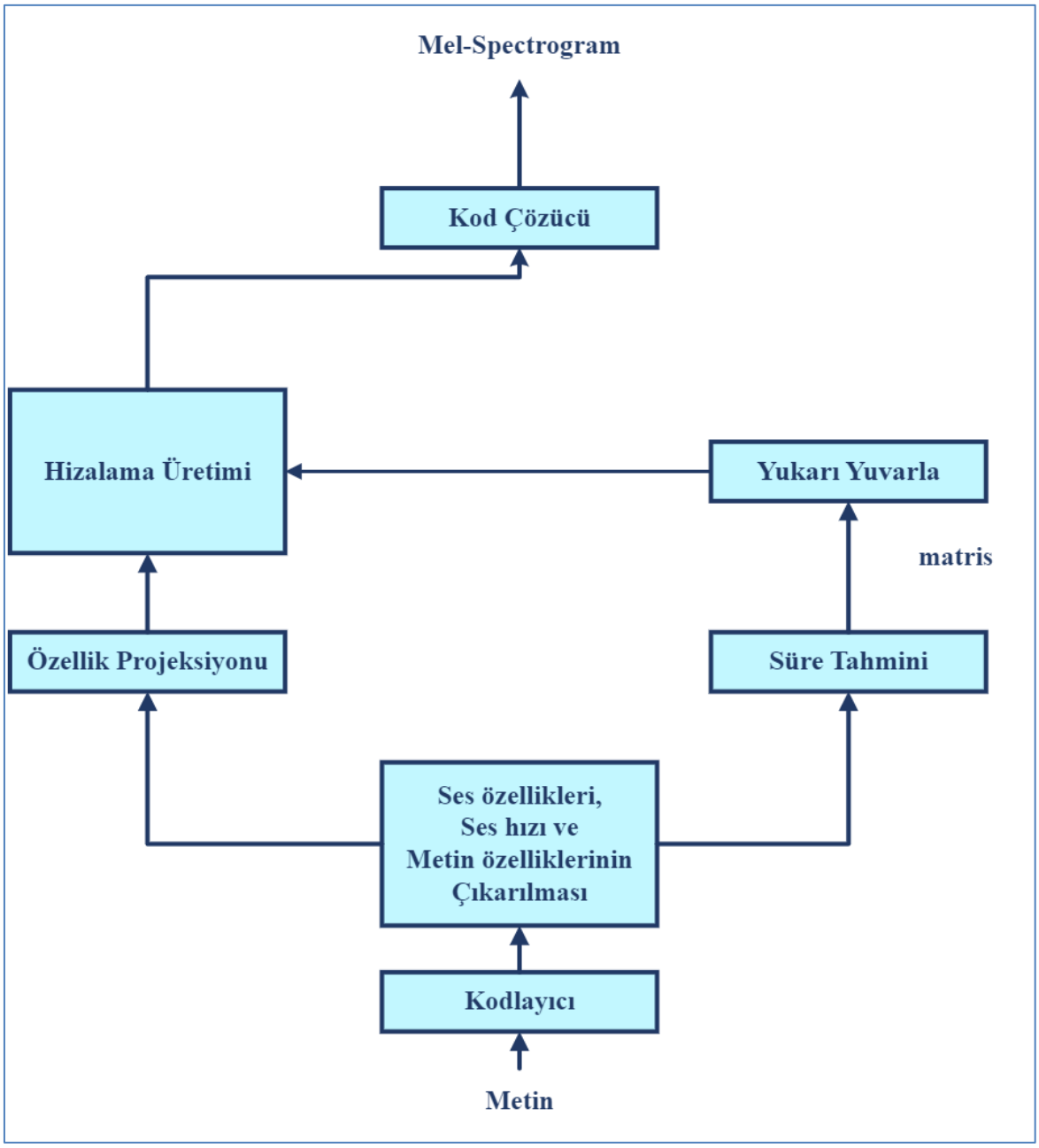
GlowTTS [39], doğal insan sesi sentezlemek için kullanılan bir uçtan uca bir konuşma sentezleme sistemi mimarisidir. Bu mimari, birçok bileşenden oluşmaktadır ve eğitim süreci oldukça karmaşıktır. Mimarinin başlangıç noktası, girdi olarak bir dizi karakter alan bir metin dizisidir. Metin önce bir karakter gömme katmanı tarafından işlenir ve ardından bir dizi dönüştürücü katman kullanılarak derin bir temsile dönüştürülmektedir. Bu temsil, bir kodlayıcı olarak adlandırılan bir önceden eğitilmiş modelden geçirilmektedir.

Kodlayıcının çıkışı, bir gradyanların durdurulması (stop_gradient) işlemi ile işlenmektedir.

Bu, kodlayıcının güncellenmemesini sağlayarak, eğitim sırasında aşırı uyum sorunlarını engellemektedir. Daha sonra, bir süre tahmin edicisi (duration predictor) kullanılarak her karakterin ses çıkışının ne kadar uzun sürdüğü tahmin edilmektedir. Bu tahmin, kodlayıcının çıkışından bir lineer sınıflandırıcı ($L_{sınıflandırıcı}$) kullanılarak elde edilmektedir.

Bir sonraki adım, girdinin spektrogramını oluşturmak için bir projeksiyon katmanı kullanmaktır. Projeksiyon katmanı, kodlayıcının çıktısını bir dizi özellik vektörüne dönüştürmektedir. Bu özellik vektörleri, bir monotonik hizalama arama algoritması ile işlenmektedir ve bir sinyal sentezleyicisi (decoder) kullanılarak ses sinyaline dönüştürülmektedir.

Son olarak, sentezlenen ses sinyali bir mel spektrograma dönüştürülmekte olup, kayıp hesaplamak için gerçek ses sinyaliyle karşılaştırılmaktadır. Kayıp, tüm parametrelerin (kodlayıcı, tahmin edici, projeksiyon, sentezleyici) kayıplarının toplamıdır ve geriye yayılım kullanılarak tüm parametreler optimize edilmektedir. Bu şekilde, eğitim verileri üzerinde iteratif bir şekilde çalışarak, modele doğal insan sesi sentezleme becerisi kazandırılmaktadır.



Şekil 3.10. GlowTTS'in çıkarım süreci mimarisi

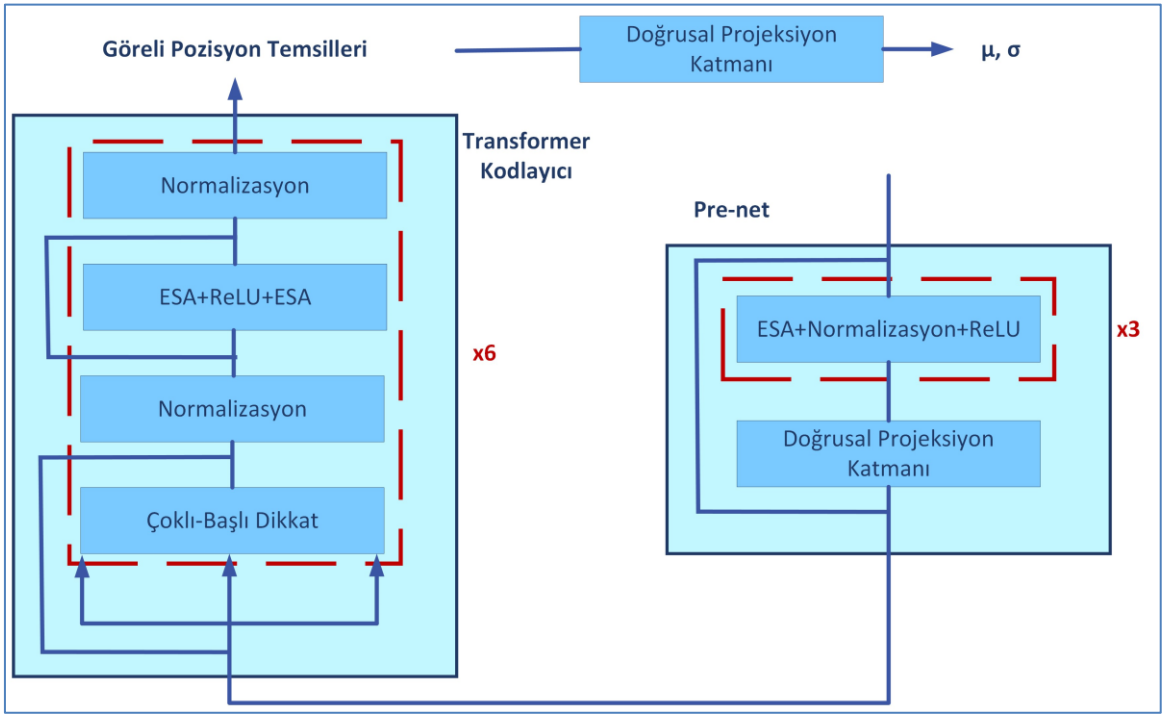
GlowTTS [39], daha önce de bahsedildiği gibi doğal insan konuşmayı sentezlemek için kullanılan bir uçtan uca konuşma sentezleme mimarisidir. Bu mimari, birçok bileşenden oluşmaktadır ve çıkarım süreci de tıpkı eğitim süreci gibi oldukça karmaşıktır. Çıkarım süreci, bir girdi metin dizisinin, sentezlenmiş bir ses sinyaline dönüştürülmesini içermektedir. Girdi metin dizisi, bir karakter gömme katmanı tarafından işlenir ve ardından bir dizi dönüştürücü katman kullanılarak derin bir temsile dönüştürülmektedir.

Temsil, bir kodlayıcı olarak adlandırılan bir önceden eğitilmiş modelden geçirilmektedir.

Kodlayıcının çıkışı, bir süre tahmin edicisi tarafından işlenmektedir. Tahmin edici, her karakterin ses çıkışının ne kadar uzun sürdüğünü tahmin etmektedir ve ardından bir tavan fonksiyonu (ceil) kullanılarak tahmin edilen süre tam sayıya yuvarlanmaktadır. Daha sonra, girdinin spektrogramını oluşturmak için bir projeksiyon katmanı kullanılmaktadır. Projeksiyon katmanı, kodlayıcının çıktısını bir dizi özellik vektörüne dönüştürmektedir. Bu özellik vektörleri, bir hizalama oluşturma (alignment generation) algoritması ile işlenmekte ve bir sinyal sentezleyicisi (decoder) kullanılarak ses sinyaline dönüştürülmektedir.

Son olarak, sentezlenen ses sinyali bir mel spektrograma dönüştürülmekte olup, çıkış olarak sunulmaktadır. Bu işlem, kayıp hesaplamak için gerçek ses sinyaliyle karşılaştırılan eğitim verilerindeki işlemle benzerdir. Mimari, çıkarım işlemi sırasında önceden eğitilmiş parametreleri kullanmaktadır. Bu şekilde, modele, eğitim verileri üzerinde kazandığı doğal insan sesi sentezleme yeteneğini kullanarak, girdi metin dizilerinden doğal insan sesi sentezleyebilme yeteneği kazandırılmaktadır.

Yukarıdaki mimarilerde bahsedilen kodlayıcının mimarisi Şekil 3.11'de, süre tahmin edicinin mimarisi Şekil 3.12'de ve Şekil 3.13'te ise kod çözücü mimarisi verilmektedir.



Şekil 3.11. Kodlayıcı yapısı

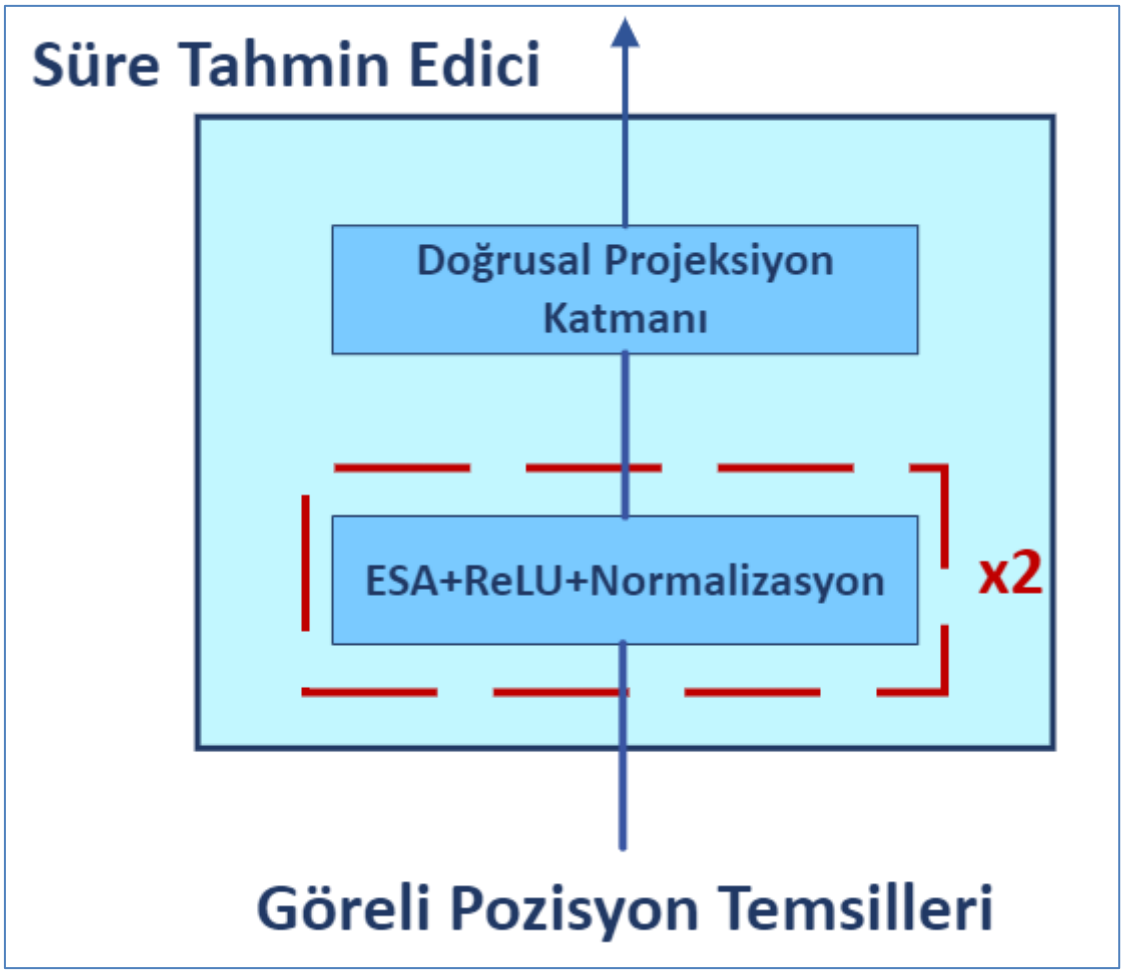
GlowTTS'de kodlayıcı [39], metin temsili oluşturmak için önemli bir bileşendir. Metin dizisini girdi olarak almaktadır ve bu metin verilerini yüksek boyutlu bir vektör temsiline dönüştürerek içerdiği bilgileri öğrenmektedir. Ardından, elde edilen temsil, ses sentezi işlemi için kod çözücü bölümüne (decoder) iletilmektedir. Kodlayıcı yapısında ana olarak Ön-Öniçe Bağlantı Kodlayıcı (Encoder Pre-Net), Transformer Kodlayıcı ve Doğrusal Projeksiyon Katmanı içermektedir.

Ön-Öniçe Bağlantı Kodlayıcı, metin verilerini işlemeye başlamadan önce, metin temsillerini hazırlamak için kullanılmaktadır. Bu katmanda, metin verilerinin uygun bir formata getirilmesi ve ön işlemlerin yapılması sağlanmaktadır. Ayrıca, aktivasyon fonksiyonları ve diğer ön işlemler bu katmanda uygulanmaktadır.

Transformer Kodlayıcısı, kodlayıcının temel kısmı oluşturmaktadır. Transformer mimarisi kullanılarak tasarlanmıştır. Transformer, dil işleme ve doğal dil işleme modellerinde başarılı bir şekilde kullanılan bir yapıdır. Bu mimari, dikkat mekanizmasını kullanarak metindeki uzak bağlantıları daha etkin bir şekilde ele almaktadır. GlowTTS'de, Transformer kodlayıcısında pozisyon kodlaması yerine göreli pozisyon temsilleri (relative position representations) kullanılmaktadır. Göreli pozisyon temsilleri, metindeki kelimeler arasındaki ilişkileri daha etkin bir şekilde yakalamayı sağlamaktadır.

Doğrusal Projeksiyon Katmanı, kodlayıcının çıkışında bir doğrusal projeksiyon katmanı eklenmektedir. Bu katman, kodlayıcının çıktısını latent temsil formatına dönüştürmektedir ve önceden dağılımın istatistiklerini tahmin etmek için gerekmektedir.

Kodlayıcı, metni temsili bir gizli vektöre dönüştürdükten sonra, bu vektörü çözücüye ileterek ses sentezi işlemini gerçekleştirir. Bu şekilde, kodlayıcı ve çözücü birlikte çalışarak, metinden mel-spektrogramlarına dönüştürme süreci tamamlanır ve sonuç olarak ses sinyali elde edilir. GlowTTS'deki kodlayıcının bu yapısı, yüksek kaliteli ve doğal ses sentezini gerçekleştiren temel bir bileşen olarak önemlidir.



Şekil 3.12. Süre tahmin edici mimarisi

GlowTTS'nin [39] süre tahmin edici (duration predictor) bileşeni, metin tabanlı konuşma sentezleme modeli olan GlowTTS'nin önemli bir parçasıdır. Süre tahmin edici, metindeki her bir konuşma parçasının süresini tahmin etmek amacıyla kullanılmaktadır ve bu bilgi, konuşma akışını daha hassas bir şekilde kontrol etmeyi ve sonuç olarak ses sentezini geliştirmeyi sağlamaktadır.

Süre tahmin edici, iki adet ESA katman, ReLU aktivasyon fonksiyonu, katman normalizasyon ve dropout işlemleri ile bir doğrusal projeksiyon katmanından oluşmaktadır. Bu katmanlar, metin temsili üzerinden konuşma parçalarının sürelerini tahmin etmek için kullanılmaktadır.

GlowTTS'nin kodlayıcısından elde edilen metin temsili, süre tahmin ediciye aktarılmaktadır. Bu temsil, her bir konuşma parçasının süresini tahmin etmek için bir evrişimli sinir ağında

işlenmektedir. Evrişimli katmanlar, metindeki farklı konuşma parçalarını tanımak için kullanılan özelliklerin çıkarılmasına yardımcı olmaktadır.

Sonrasında, elde edilen özellikler ReLU aktivasyon fonksiyonu ile etkin hale getirilmektedir ve katman normalizasyonu uygulanmaktadır. Bu işlem, özelliklerin daha iyi öğrenilmesini ve daha istikrarlı bir eğitim sürecini desteklemektedir. Daha sonra sönümleme işlemi uygulanmaktadır. Sönümleme, ESA'nın aşırı öğrenimini azaltmak ve daha iyi genelleştirme yapmasını sağlamak için kullanılan bir tekniktir.

Tüm bu işlemlerin ardından, doğrusal projeksiyon katmanı, özelliklerin konuşma sürelerine dönüştürülmesini sağlamaktadır. Doğrusal projeksiyon katmanı, önceki işlemlerle elde edilen özelliklerden faydalanarak metindeki her bir konuşma parçasının süresini tahmin etmektedir.

Bu süre tahminleri, sonraki aşamalarda ses sentez işlemi için kullanılmaktadır. Örneğin, metinde daha uzun süreli konuşma parçalarına daha fazla zaman ayrılarak daha yavaş bir konuşma hızı elde edilebilmektedir. Süre tahmin edicinin katmanları, GlowTTS modelinin esnek ve etkili bir şekilde metni ses haline dönüştürmesini ve konuşma akışını daha doğru bir şekilde kontrol etmesini sağlamaktadır.

Şekil 3.13. Kod çözücü mimarisi

GlowTTS'nin kod çözücüsü [39], akış tabanlı bir çözücüdür ve metinden mel spektrograma dönüşümü gerçekleştirilmektedir. Bu çözücü, metin tabanlı konuşma sentezleme modeli GlowTTS'nin önemli bir bileşenidir ve metni ses sinyaline dönüştürerek konuşma sentezi yapmaktadır.

Çözücü, metin temsilini almaktadır ve bunu mel spektrograma dönüştürmek için akış bloklarından oluşan bir yapıyı kullanmaktadır. Her bir akış bloğu, aktivasyon normalizasyon katmanı, ters çevrilebilir 1x1 evrişimli sinir ağı katmanı ve afin bağlantı katmanından oluşmaktadır. WaveGlow'dan farklı olarak, çözücüde yerel koşullandırma (local conditioning) yöntemi kullanılmamaktadır. Ayrıca, hesaplama verimliliği için mel spektrogram çerçeveleri 80 kanallı olacak şekilde zaman boyutu boyunca ikiye bölünmektedir ve bu yarılar 160 kanallı bir özellik haritasına birleştirilmektedir.

Kod çözücü, mel spektrogramı sıkıştırmaktadır. Mel spektrogram çerçeveleri sıkıştırılarak, özelliklerin boyutu azaltılmaktadır ve hesaplama verimliliği artırılmaktadır. Bu işlem sırasında, kanal boyutu ikiye katlanarak zaman adımları yarıya düşmektedir. Eğer zaman adımlarının sayısı tek ise, son öğe basitçe yok sayılmaktadır, çünkü ses kalitesini etkilemeyecek kadar kısa bir süreyi temsil etmektedir.

Aktivasyon normalizasyon katmanları, giriş verilerinin dağılımını düzenlemektedir ve daha iyi öğrenmeyi sağlamaktadır. Kod çözücüde kullanılan 1x1 evrişimli sinir ağı katmanı, ters çevrilebilir bir katman olup giriş verilerini dönüştürmektedir ve akış modelinin temelini oluşturmaktadır. Afin bağlantı katmanı, giriş verilerini dönüştürerek çıktıyla girişin log olasılık oranını hesaplamaktadır. Kod çözücü, son olarak mel spektrogramın olasılık dağılımını elde ettiğinde, çıktıyı girişin boyutuna eşit hale getirmektedir ve böylece mel spektrogramın sentezi tamamlanmaktadır.

GlowTTS'nin kod çözücüsü, akış tabanlı yapısı sayesinde metin temsilinden mel spektrogramın olasılık dağılımını ve ses sentezini etkin bir şekilde gerçekleştirmektedir. Bu kod çözücü, esnek ve etkili bir konuşma sentezleme modeli sunmaktadır ve yüksek kaliteli ses sentezi elde edilmesine imkan vermektedir.

Oluşturulan mel spektrogramlarının sese dönüştürülmesi için bu sistemlere ek olarak “Flow” modeli adı verilen sistemler kullanılmaktadır. Bu aşamada, özetlenmiş metin özelliği, bir "Flow" modeli kullanılarak ses sinyaline dönüştürülmektedir. Flow, bir tür yapay sinir ağıdır ve bir özellik dizisi alarak, belirli bir dağılıma uygun şekilde rastgele örneklem alarak ses sinyalini üretmektedir. Bu şekilde, özetlenmiş metin özelliği, gerçekçi ve doğal bir konuşma sinyaline dönüştürülmektedir.

Bu çalışma kapsamında da Python'ın sunduğu kütüphaneler ile bir kod geliştirilmiştir. Bu kapsamda "TTS" kütüphanesinin sunduğu yetenekler vasıtası eğitilmiş modelin çalıştırılması sonucu sentezlemiş sesler üretilmiştir. Model dosyası ile yapılandırma dosyası da mel spektrogramların sese dönüştürülmesinde çok önemli bir yer tutmaktadır. TTS kütüphanesinin belirli bir kodlayıcı kullanılarak mel spektrogramlarını ses dönüşümü gerçekleştirmesi için, kodlayıcının doğru şekilde yapılandırılması ve gerekli parametrelerin sağlanması gerekmektedir.

3.5. Çalışmada Kullanılan Araçlar

Bu çalışma, Windows 11 Pro işletim sistemine sahip olan bir bilgisayarda gerçekleştirilmiştir. Bilgisayarın özellikleri şunlardır: 11th Gen Intel(R) Core (TM) i7-11800H @ 2.30GHz işlemci, 16 GB RAM ve NVIDIA GeForce RTX 3060 6GB ekran kartı. Çalışma sırasında, çeşitli araçlardan yararlanılmıştır. İşlevsel ve kullanıcı dostu bir dil olan Python tercih edilmiş olup, Python programlama dilinin 3.9.13 versiyonu kullanılmıştır. Python, son dönemde popüler hale gelmiş ve birçok büyük şirket tarafından da tercih edilen bir programlama dilidir. Bu dil kullanılarak geliştirme ve değerlendirme süreçleri gerçekleştirilmiştir.

Ayrıca, Python dilindeki geliştirme sürecinde çeşitli kütüphanelerden faydalanılmıştır. Bu kütüphaneler, Python'un geniş ekosistemi içinde bulunan ve çeşitli fonksiyonları sağlayan hazır paketlerdir. Bu kütüphaneler, çalışmanın daha verimli ve etkili bir şekilde yürütülmesine yardımcı olmuştur.

Kullanılan kütüphanelerin yanı sıra çeşitli programlarda mevcuttur. Bu bölümde kullanılan programlar ve kütüphanelerden bahsedilecektir.

3.5.1. Audacity

Veri kümesinin ses kayıtları alındıktan sonra, ses dosyaları üzerinde çeşitli düzenlemeler yapma ihtiyacı ortaya çıkmaktadır. Bu düzenlemeler arasında sessizlik alanlarını işlemek önemlidir. Bu amaçla, Dominic Mazzoni tarafından Carnegie Mellon Üniversitesi'nde yüksek lisans eğitimi sırasında geliştirilen Audacity isimli program kullanılmıştır.

Audacity, ücretsiz ve açık kaynaklı bir ses düzenleme yazılımıdır. Bu yazılım, kullanıcıların ses kayıtlarını düzenlemelerine, kesmelerine, birleştirmelerine, efektler eklemelerine ve birçok başka işlem yapmalarına olanak sağlamaktadır. Audacity, çeşitli ses dosyası formatlarını destekler ve aynı zamanda kayıt yapmak için kullanıcı dostu bir arayüz sunmaktadır. Windows, Mac OS X ve Linux gibi birçok işletim sistemi üzerinde kullanılabilmektedir. Kullanıcı dostu bir arayüze sahip olması ve birçok dilde kullanılabilmesi, müzik kaydı, sesli kitaplar ve diğer sesli materyaller için popüler bir seçenek olmasını sağlamaktadır.

Audacity, veri kümesindeki ses kayıtlarını işlemek için kullanılan etkili bir araçtır. Sessizlik alanlarını belirlemek, kayıtları düzenlemek ve istenen özelliklere sahip ses dosyalarını oluşturmak için bu yazılımdan yararlanmak önemlidir.

3.5.2. Anaconda

Anaconda, açık kaynak kodlu bir Python dağıtımıdır ve birçok Python paketi ve kütüphanesini bir arada bulundurmaktadır. Anaconda'nın amacı, veri bilimi, makine öğrenimi, derin öğrenme, bilimsel hesaplama, büyük veri işleme ve veri görselleştirme gibi alanlarda çalışan araştırmacıların, öğrencilerin ve uygulayıcıların işlerini kolaylaştırmaktır. Anaconda, Python dilinin yanı sıra birçok popüler veri bilimi paketi ve kütüphanesini de içermektedir. Bu paketler arasında NumPy, pandas, Matplotlib, Seaborn, Scikit-learn, TensorFlow, Keras, PyTorch, Spyder ve Jupyter Notebook gibi birçok araç yer almaktadır. Anaconda, bu paketlerin kolayca yüklenmesine ve yönetilmesine olanak tanıyan bir paket yöneticisi olan Conda'yı da içermektedir.

Anaconda, farklı işletim sistemleri için önceden derlenmiş sürümleri sunmakta ve çoğu paketi kendi yönetimi altında tutmaktadır. Bu nedenle, farklı paketlerin uyumluluğu konusunda bir risk taşımamaktadır. Anaconda ayrıca, ortak veri depolama alanları, geliştirme ortamı ve paylaşım için uygun araçlar da sunmaktadır.

Anaconda, veri analizi ve model oluşturma üzerine çalışan araştırmacıların, bu sürecini kolaylaştıran bir araçtır. Ayrıca, birçok popüler paketi bir arada sunması nedeniyle, paket uyumluluğu ve yönetimindeki kolaylığı sayesinde, birçok kullanıcının tercih ettiği bir araçtır.

Bu çalışma kapsamında da paket ve ortam yöneticiliği için Anaconda uygulamasından faydalanılmıştır. Bununla birlikte geliştirme ortamlarından biri olan Spyder, Anaconda aracılığı ile kullanılmıştır.

3.5.3. Kullanılan kütüphaneler

TTS kütüphanesi gelişmiş bir konuşma sentezleme kütüphanesidir. İçerisinde 37 farklı dil veya lehçe barındıran bu kütüphanede 62 adet hazır konuşma sentezleme modeli bulunmaktadır. Buna ek olarak 6 farklı dil için toplam 16 adet ses kodlayıcı model vardır. Hem konuşma sentezleme modellerinin hem de ses kodlayıcı modellerin için çoklu dil modelleri mevcuttur.

Türkçe için GlowTTS adı verilen bir model hazırlanmıştır. Bu model daha önce derin öğrenme tabanlı konuşma sentezleme yönteminde belirtilen çalışma mantığı ile aynı yapıda çalışmaktadır. Bu çalışma kapsamında da eldeki veri kümesi her biri 261 adım olan 500 çevrimlik model ile eğitilmiştir.

Bu eğitim süreci için ek olarak birçok kütüphaneden faydalanılmıştır. Bunlardan bir kısmı Çizelge 3.3'te verilmiştir.

Çizelge 3.3. Kullanılan kütüphaneler

Kütüphane Adı	İşlev
numpy	Çok büyük dizi, matris işlemlerinde kullanılan matematiksel bir Python kütüphanesidir.
scipy	Bilimsel ve teknik hesaplamaları için kullanılan bir Python kütüphanesidir.
torch	İçinde birçok kütüphane barındıran, bilimsel hesaplama için kullanılan makine öğrenme algoritmaları için kapsamlı destek sunan yapıdır.
torchaudio	Görme ve doğal işleme gibi alanlarda kullanılan derin öğrenme kütüphanesi olan PyTorch ile kullanılan ses ve sinyal işleme kütüphanesidir.

Çizelge 3.3. (devamı) Kullanılan kütüphaneler

Kütüphane Adı	İşlev
soundfile	Ses dosyalarının dizi olarak yazılmasını sağlayabiliyor. Buna ek olarak ses verisinin wav formatında kaydedilmesine olanak sağlıyor.
librosa	Ses dosyası formatıyla çalışarak ses analizi için kullanılan bir kütüphanedir.
numba	Python programlama dilinde yazılmış bir kodun çalışma hızını artırarak C gibi makine koduna yakın kodlarla yazılmış programla dillerinin çalışma hızlarına yaklaştırır.
anyascii	Her karakterin bir sayı karşılığının olduğu bir standart olan Unicode karakterleri ASCII karakterlere dönüştüren bir kütüphanedir.
pyyaml	Anlaşılması zor olan diğer biçimlendirme dillerinin yerine türetilen YAML dosya işlemlerinin Python içinde kullanılmasına olanak sağlayan kütüphanedir.
flask	Web servisler üzerinde çalışmak için kullanılan Python çerçevesidir.
pandas	Zaman etiketli serileri ve sayısal tabloları işlemek için bir veri yapısı oluşturan ve oluşturduğu yapıyı çeşitli işlemler ile işler-analiz eden bu kütüphane, temel olarak veri işlemesi ve analizi için kullanılır.
matplotlib	Yapılan matematiksel ve teknik işlemler sonucunda görsel çıktılar almak için kullanılan bir kütüphanedir.
trainer	Derin öğrenme kütüphanesi olan PyTorch'ta kullanılan genel amaçlı model eğitim kütüphanesidir.
subprocess	Python programı dışında çalışan sistem komutlarını yönetmek için kullanılmaktadır. Bu kütüphane mel spektrogramların konuşma sinyallerine dönüştürülmesinde sistem dışına çalışan bir komutun çalıştırılmasında kullanılmıştır.
sys	Sistem düzeyinde işlemler gerçekleştirmek ve Python programının çalışma ortamıyla etkileşimde bulunmak için çok yönlü bir araç sağlamaktadır. Bu çalışmada da Python programının sistemle ilgili özellikleri kullanmasını veya kontrol etmesini gerektiren durumlarda kullanılmıştır.
TTS	Metinden konuşma sentezleme işlemlerini gerçekleştirmek için tasarlanmış bir kütüphanedir. Bu kütüphane, metin girdisini alarak, gerçekçi, insan benzeri konuşma sinyallerini üretebilmektedir. Bu çalışma kapsamında eğitim sonunda oluşturulmuş model dosyası, yapılandırma dosyası ile bir çıktı dosyası oluşturulmuştur ve sentezlenmiş ses üretilmiştir.

4. KONUŞMA SENTEZLEME SİSTEMİNİN DENEYSEL SÜRECİ VE BULGULAR

Bu çalışmada, doğal konuşmadaki özellikleri sentezlenen konuşmalara kazandırmak amacı güdülmüştür. Hedef, geliştirilen sistemlerin gerçekçi ve anlaşılır bir şekilde konuşma üretebilmesini sağlamak, etkili bir iletişim aracı sunmak ve insanlar arasında iletişimde yeni fırsatlar yaratmaktır. Bu amaç doğrultusunda, derin öğrenme tabanlı uçtan uca konuşma sentezleme sistemleri kullanılarak bir çalışma gerçekleştirilmesi planlanmıştır.

Derin öğrenme tabanlı uçtan uca konuşma sentezleme sistemleri, konuşma sentezleme sürecinin tüm aşamalarını içeren bütünsel bir yaklaşım sunmaktadır. Bu sistemler, metinden konuşmaya dönüşümün yanı sıra, dil özelliklerini doğru bir şekilde modellendirme, akustik özellikleri hassas bir şekilde sentezleme ve doğal bir konuşma akışı oluşturma gibi unsurları içermektedir.

Bu çalışmanın hedeflerinden biri, geliştirilen sistemlerin gerçekçi bir şekilde konuşma üretebilmesidir. Bu, konuşma sentezleme sisteminin doğal konuşmadaki vurguları, tonlamaları, duraklamaları ve diğer dilin özgünlüklerini başarılı bir şekilde yakalaması anlamına gelmektedir. Böylece, sentezlenen konuşmalar daha insan gibi ve anlaşılır hale gelmektedir.

Ayrıca, bu çalışmanın amacı etkili bir iletişim aracı sağlamaktır. Sentezlenen konuşmaların anlaşılır olması, iletişimin daha etkili ve verimli olmasını sağlamaktadır. İnsanlar arasındaki iletişimde sesli konuşma önemli bir rol oynamaktadır ve doğal bir konuşma üretebilen sistemler, iletişimde yeni fırsatlar ve kullanım alanları sunabilmektedir.

Bu çalışma, konuşma sentezleme sürecinin her aşamasını kapsayan derin öğrenme tabanlı uçtan uca bir yaklaşımı benimsemektedir. Bu, doğal konuşmadaki özellikleri başarılı bir şekilde sentezlenmiş konuşmalara yansıtmayı hedeflemektedir. Bu sayede, sentezlenen konuşmalar gerçekçi, anlaşılır ve etkili iletişim sağlayabilen bir araç haline gelmektedir.

Derin öğrenme tabanlı uçtan uca konuşma sentezleme sistemi geliştirilirken daha önce de bahsedildiği gibi GlowTTS mimarisinden faydalanılmıştır. CoquiAI tarafından geliştirilmiş

bir konuşma sentezleme modeli olan GlowTTS, girdi olarak doğal dil metinlerini alan ve gerçekçi, doğal konuşma sinyalleri üreten bir modeldir. GlowTTS, temel olarak iki aşamadan oluşmaktadır. İlk aşama metin özetleme aşaması olarak adlandırılmaktadır. Diğer aşama ise Flow modeli aşaması olarak adlandırılmaktadır.

Özetleme aşamasında, girdi olarak verilen metin sinyali, bir metin özetleme modeli kullanılarak işlenmektedir. Bu özetleme modeli, metin sinyalini bir özellik dizisi haline dönüştürmektedir. Bu özellik dizisi, konuşma sinyali üretmek için kullanılan bir format olan mel spektrograma dönüştürülmektedir. Mel spektrogram, frekans spektrumunun zamana karşı değişimini görselleştiren bir formattır.

Flow aşamasında, özetlenmiş metin özelliği, bir "Flow" modeli kullanılarak ses sinyaline dönüştürülmektedir. Flow, bir tür yapay sinir ağıdır ve bir özellik dizisi alarak, belirli bir dağılıma uygun şekilde rastgele örneklem alarak ses sinyalini üretmektedir. Bu şekilde, özetlenmiş metin özelliği, gerçekçi ve doğal bir konuşma sinyaline dönüştürülmektedir. GlowTTS, birçok farklı konuşma stilini ve tonunu kapsayabilen geniş bir veri kümesiyle eğitilmektedir. Bu nedenle, model, farklı konuşma stillerini ve farklı konuşmacıların seslerini sentezleyebilmektedir. Ayrıca, model, kısa metinlerden uzun konuşmalar oluşturmak için kullanılabilmekte ve sınırlı veriyle bile iyi sonuçlar verebilmektedir.

Gerçekleştirim mimarisi GlowTTS olan bu çalışmanın öncesinde ve sonrasında birçok deneysel çalışma yapılmıştır. Konuşma sentezleme sistemlerinin geliştirilmesi sürecinde özetle 6 deneysel süreçten geçilmektedir. Bunlar; veri kümesi toplama, veri kümesi hazırlama, model oluşturma, veri kümesi bölme, model eğitimi ve son olarak model doğruluğunun değerlendirilmesi olarak sıralanabilmektedir.

4.1. Veri Kümesi Toplama

Uçtan uca konuşma sentezleme sistemleri, doğrudan ses özellikleriyle eşleştirerek öğrenen derin öğrenme modelleri kullanarak verileri işlerler ve bu sistemler, konuşma sentezleme alanında en gelişmiş teknolojilerden biridir. Derin öğrenme tabanlı bir uçtan uca konuşma sentezleme sistemi geliştirmek için, veri toplama süreci tamamlanmış bir veri kümesi gerekmektedir.

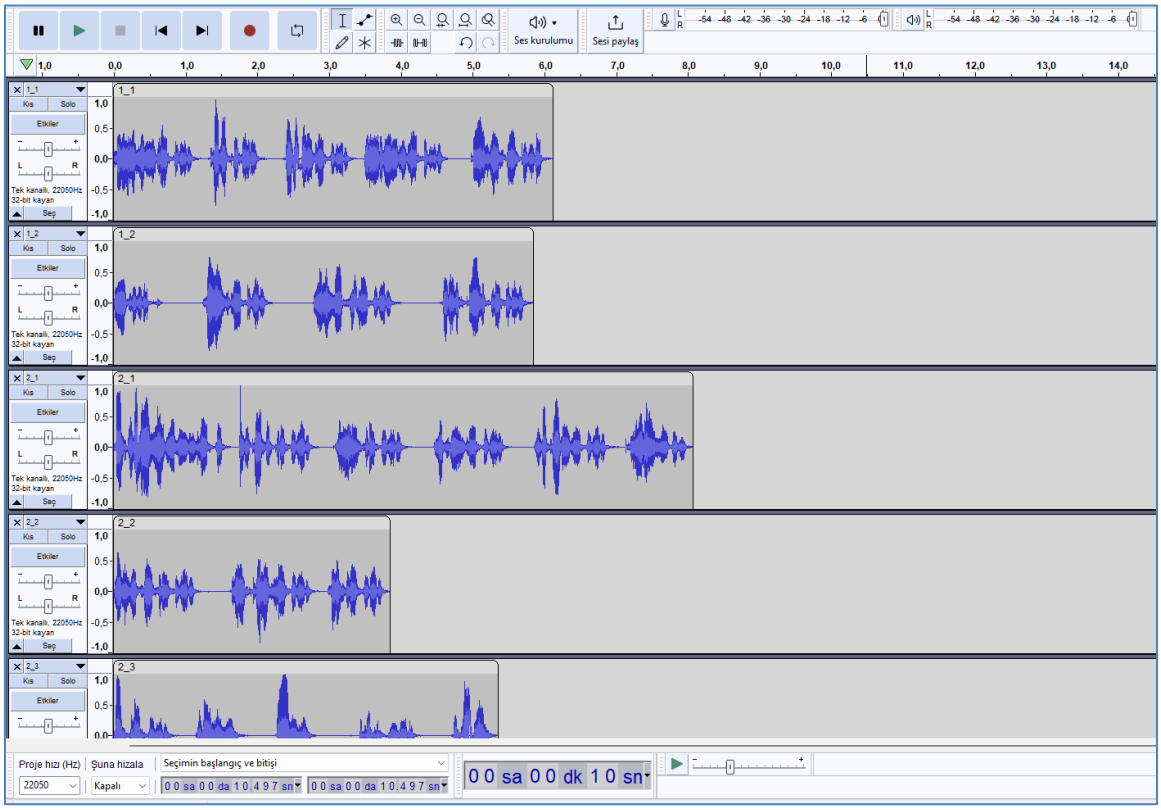
Bir veri kümesi toplanırken ilk olarak amaç ve hedefler belirlenmesi gerekmektedir. Başlangıçta, konuşma sentezleme sisteminin hedeflerini ve kullanım amacını netleştirmek önemlidir. Bu aşamada, dil seçimi ve hedeflenen konuşma sentezleme türü gibi faktörleri belirlemek gerekmektedir. Bu kapsamda bu çalışmada Türkçe konuşma sentezleme sistemi geliştirilmesi amaçlanmış ve hedeflenmiştir.

Bu süreçte yapılan bir sonraki işlem veri kümesi tasarımıdır. Veri kümesi tasarımı, toplanacak verilerin türünü ve kapsamını belirlemeyi içermektedir. Farklı konuşma üzerine yapılan çalışmalarda çeşitli konuşmacılar, farklı yaş grupları, aksanlar ve konuşma tarzları gibi faktörleri içeren çeşitlilik sağlanmalıdır. Konuşma sentezleme sistemlerinde bu tarz tasarımların başarı getireceği düşünülmemektedir. Bu kapsamda çeşitlilikten ziyade bu çalışmada bir genç erkek konuşmacı tarafından İstanbul Türkçesi ile seslendirilmiş bir veri seti tasarlanmıştır.

Sürecin bir sonraki aşaması kaynak belirlenmesidir. Veri toplamak için uygun kaynakları belirlemek önemlidir. Bu kaynaklar, doğal konuşma verileri içeren çeşitli kaynakları içerebilir, örneğin, radyo programları, podcast'ler (internette hazırlanan veya sonradan internete kopyalanan yayın), ses kayıtları içeren veri kümeleri, kişisel kayıtlar veya bir metne bağlı olarak profesyonel bir şekilde seslendirme yoluyla hazırlanabilmektedir. Bu kapsamda da veri toplama yöntemleri de belirlenmelidir. Türkçe veri kümesi hazırlanırken Türkçe atasözleri, haber bültenleri, raporlar, kitap bölümleri, dini, edebi ve tarihi yazılar, dünya ülkelerinin isimleri, Türkiye Cumhuriyeti'nin illeri, ilçeleri, çeşitli mahalle ve cadde isimleri, Türkiye Cumhuriyeti'nin çeşitli kurum ve kuruluşlarına ait isimler, sayılar ve mantıksal bağlantısı olmayan bir dizi bağımsız cümleler kullanılmıştır. Konuşma kayıtları Audio Technica marka stüdyo tipi kardioid kondansatör özelliğine sahip bir mikrofon yardımı elde edilmiştir. Kayıt esnasında çevre gürültüsü ve ses patlamalarını azaltmak için bir pop filtresi kullanılmıştır.

Belirlenen metne ve sağlanan teknik şartlar doğrultusunda veri toplama sürecine geçilmiştir. Bu aşamada, veriler kaydedilip ve belirlenen formatlarda saklanmaktadır. Belirtilen teknik şartlarda toplanan veri kümesindeki tek kanallı, örnekleme hızı 22050 Hz ve bit derinliği 32-bit kayan olan ses verileri .wav formatında saklanmıştır. Kanal, örnekleme hızı ve bit derinliği ifadesi, bir ses dosyasının teknik özelliklerini tanımlayan bir terimdir.

Bir ses dosyası, sol ve sağ kanallar gibi birden fazla kanala sahip olabilmektedir. Tek kanallı ifadesi, ses dosyasının yalnızca bir kanala sahip olduğunu ifade etmektedir. Örnekleme hızı, sesin analog formattan sayısal formata dönüştürülmesi sırasında alınan örnekleme noktalarının sayısını ifade etmektedir. Yani, 22050 Hz ifadesi, saniyede 22,050 örnekleme noktasının alındığı anlamına gelmektedir. Bit derinliği, her bir ses örneğinin temsil edildiği bit sayısını ifade etmektedir. 32-bit kayan ifadesi, ses örneklerinin kayan noktalı ondalık formatında 32 bit olarak temsil edildiğini göstermektedir.



Şekil 4.1. Veri setindeki seslerin dalga görünümü

Beşinci aşama olarak toplanan verilerin doğrulanması ve temizlenmesi aşamasına geçilmiştir. Bu kapsamda, ses kayıtlarını kontrol ederek kaliteleri değerlendirilmiştir. Veri kümesinde tutarsızlıklar veya hatalar varsa bunları düzeltilmiştir veya veri kümesinden çıkartılmıştır. Bu kapsamda 100'e yakın ses dosyası veri kümesinden silinmiştir.

Son olarak etik ilkelere dikkat edilmiştir. Veri toplama sürecinde etik ilkeler göz önünde bulundurulmuştur. Gizlilik ve kullanıcı izinleri gibi konulara dikkat edilmiş ve toplumun kabul edilebilir standartlarına uyulmuştur.

Özetle bir konuşma sentezleme sistemi geliştirmek için, öncelikle kullanılacak veri kümesinin toplanması gerekmektedir. Bu veri kümesi hem doğru telaffuzlara sahip metinlerin hem de bu metinlerin sesli kayıtlarının bir arada bulunmasından oluşmalıdır. Veri kümesinin kalitesi, sonuçta elde edilecek olan sentez kalitesini belirleyeceğinden, ses kayıtlarının yüksek kaliteli olması önemlidir. Kaliteli ses kayıtları, doğru telaffuzların öğrenilmesinde ve sentezleme sürecinde daha doğru sonuçlar elde edilmesinde etkili olacaktır. Bu nedenle, veri kümesi toplama sürecinde yüksek kaliteli kayıtların kullanımına özen gösterilmelidir.

4.2. Veri Kümesi Hazırlanması

Veri kümesinin toplanması aşamasından sonra, veri kümesinin hazırlanması sürecine geçilmektedir. Toplanan veri kümesi, kullanılacak modele uygun hale getirilmiştir. Bu süreçte, veri kümesinde bulunan metinlerin bazı durumları normalleştirmek için işlemler uygulanmıştır. Metinlerdeki harf büyüklüğü, noktalama işaretleri gibi farklılıkların ortadan kaldırılması için metin normalleştirmesi yapılmıştır. Bu şekilde, metinlerin model tarafından daha iyi işlenebilmesi sağlanmıştır.

Ayrıca, ses dosyaları da modele uygun formata getirilmiştir. Örneğin, stereo formatındaki ses dosyaları monofonik hale getirilmiştir. Böylece, modelin kabul edeceği tek bir kanal üzerinde işlem yapılabilmesi sağlanmıştır.

Bu hazırlık sürecinde, toplamda 109,826 kelimeyi kapsayan ve bunlardan 35,050'si benzersiz olan bir veri kümesi oluşturulmuştur. Bu veri kümesi, 12 saat 38 dakika 59 saniye süresindeki ses dosyalarından oluşmaktadır.

Veri kümesinin hazırlanması aşamasında yapılan işlemler, Bölüm 3.3'te daha detaylı bir şekilde açıklanmaktadır. Bu bölümde, veri normalleştirme, ses dosyası formatlama ve diğer ilgili süreçler hakkında daha kapsamlı bilgiler yer almaktadır.

4.3. Model Oluşturma

Veri kümesinin hazırlanmasının ardından, bir modelin oluşturulması gerekmektedir. Model seçimi, derin öğrenme yöntemlerinden birinin kullanılmasını içermektedir. Örneğin,

WaveNet, Tacotron, GlowTTS ve Transformer gibi farklı derin öğrenme modelleri tercih edilebilmektedir. Ayrıca, TensorFlow, PyTorch veya Keras gibi açık kaynak kodlu kütüphaneler kullanılarak da model oluşturulabilmektedir.

Bu çalışmada, GlowTTS modeli üzerine sistem geliştirilmiştir. GlowTTS, özellikle konuşma sentezleme için tasarlanmış bir derin öğrenme modelidir. GlowTTS modeli, dil özelliklerini akustik özelliklere dönüştürmek için kullanılmaktadır. Bu modelin tercih edilmesi, doğal konuşma üretiminde daha iyi sonuçlar almayı hedeflemektedir.

Modelin oluşturulması, TensorFlow, PyTorch veya Keras gibi açık kaynak kodlu kütüphanelerden biri kullanılarak gerçekleştirilebilmektedir. Bu kütüphaneler, derin öğrenme modellerinin oluşturulması, eğitilmesi ve değerlendirilmesi için gerekli araçları sağlamaktadır. Model oluşturulurken, GlowTTS modelinin mimarisi ve ilgili parametreler dikkate alınmalıdır.

GlowTTS, Coqui AI tarafından geliştirilen bir konuşma sentezleme modelidir. Bu model, doğal dil metinlerini girdi olarak alır ve gerçekçi ve doğal konuşma sinyalleri üretmektedir. GlowTTS'nin çalışma yapısı, Bölüm 3.4'te detaylı bir şekilde açıklanmıştır.

GlowTTS modeli temel olarak iki aşamadan oluşmaktadır. İlk aşama, metin özetleme aşaması olarak adlandırılmaktadır. Bu aşamada, verilen metin sinyali bir metin özetleme modeli kullanılarak işlenmektedir. Metin özetleme modeli, metin sinyalini bir özellik dizisine dönüştürmektedir. Bu özellik dizisi, konuşma sinyali üretmek için kullanılan bir formata dönüştürülmektedir. Bu formata "mel spektrogram" denir. Mel spektrogram, frekans spektrumunun zamana karşı değişimini görselleştiren bir formattır. GlowTTS'nin metin özetleme aşaması, metnin temsilini bir özellik dizisine dönüştürerek konuşma üretiminin temelini oluşturmaktadır.

Diğer aşama ise Flow modeli aşaması olarak adlandırılmaktadır. Bu aşamada, özetlenmiş metin özelliği, bir "Flow" modeli kullanılarak ses sinyaline dönüştürülmektedir. Flow, bir tür yapay sinir ağıdır ve bir özellik dizisi alarak, belirli bir dağılıma uygun şekilde rastgele örneklem alarak ses sinyalini üretmektedir. Bu şekilde, özetlenmiş metin özelliği, gerçekçi ve doğal bir konuşma sinyaline dönüştürülmektedir. GlowTTS, birçok farklı konuşma stilini ve tonunu kapsayabilen geniş bir veri kümesiyle eğitilmektedir. Bu nedenle, model, farklı

konuşma stillerini ve farklı konuşmacıların seslerini sentezleyebilmektedir. Ayrıca, model, kısa metinlerden uzun konuşmalar oluşturmak için kullanılabilmekte ve sınırlı veriyle bile iyi sonuçlar verebilmektedir.

Bu çalışmada, Python dilinde bulunan TTS kütüphanesinin kullanılarak bir kod geliştirilmiştir. TTS kütüphanesi, eğitilmiş bir modelin çalıştırılmasıyla sentezlenmiş sesler üretme yeteneği sunmaktadır. Model dosyası ve yapılandırma dosyası, mel spektrogramlarının sese dönüştürülmesinde önemli bir role sahiptir.

TTS kütüphanesi, belirli bir kodlayıcı kullanılarak mel spektrogramlarını ses dönüşümünü gerçekleştirmektedir. Bu nedenle, kodlayıcının doğru şekilde yapılandırılması ve gerekli parametrelerin sağlanması gerekmektedir. Kodlayıcı, metni mel spektrogramlarına dönüştüren bir işlemdir ve ses sentezlemesi için önemli bir adımdır.

Yapılandırma dosyası, kodlayıcının çalışması için gerekli olan ayarları içermektedir. Bu dosya, kodlayıcının hangi parametrelerle çalışacağını belirlemektedir. Örneğin, dil seçimi, hız ayarları, tonlama vb. gibi parametreler yapılandırma dosyasında belirtilmektedir.

Mel spektrogramlarının sese dönüştürülmesi, TTS kütüphanesi tarafından eğitilmiş bir modelin kullanılmasıyla gerçekleştirilmektedir. Bu model, mel spektrogramlarını ses sinyallerine dönüştüren bir dönüştürücü işlevi görmektedir. Model dosyası, bu dönüştürme işlemi için kullanılmaktadır ve doğru şekilde yapılandırılması önemlidir.

Sonuç olarak, Python dilindeki TTS kütüphanesi kullanılarak mel spektrogramları seslere dönüştüren bir kod geliştirilmiştir. Model dosyası ve yapılandırma dosyası, doğru şekilde yapılandırılmış ve gerekli parametreler sağlanmıştır. Bu sayede, TTS kütüphanesi aracılığıyla konuşma sentezleme gerçekleştirilebilmiştir.

4.4. Veri Kümesi Bölme

Modelin oluşturulmasından sonra, modelin performansını değerlendirmek ve doğruluğunu ölçmek için bir test veri kümesi oluşturulmuştur. Bu test veri kümesi, ayrılmış olan veri kümesinin belirli bir bölümünü temsil etmektedir.
Test veri kümesi, toplamda 2518 kelime içermekte olup, bu kelimelerin 1820'si benzersizdir.

Böylelikle, farklı kelime öbeklerinin model tarafından nasıl işlendiğini değerlendirebilmek için çeşitlilik sağlanmıştır. Ayrıca, test veri kümesi, 18 dakika 44 saniye uzunluğunda bir süreyi temsil etmektedir.

Test veri kümesi, modelin performansını nesnel bir şekilde değerlendirmek için kullanılmaktadır. Model, bu veri kümesi üzerindeki sentezleme işlemlerini gerçekleştirmektedir ve çıktıları incelenmektedir. Bu şekilde, modelin sentezleme kalitesi, kelime doğruluğu ve genel performansı değerlendirilmektedir.

Test veri kümesi, modelin gerçek dünya kullanım senaryolarında ne kadar başarılı olduğunu değerlendirmek için önemlidir. Modelin bu veri kümesi üzerinde yüksek bir doğruluk ve performans göstermesi, genel olarak modelin başarısını yansıtabilmektedir.

4.5. Model Eğitimi

Daha sonra eğitim sürecine gerçekleştirilmiştir. Bu süreçte eğitim iterasyonlarının sayısının belirlenmesini sağlayan “epochs”, bir öğrenme iterasyonunda eğitim verilerinin kaç örneğini işleyeceğini belirleyen “batch_size” gibi çeşitli parametrelerin belirlenmesine ihtiyaç vardır. Bu parametreler ve ne işe yaradıklarına dair bilgi Çizelge 4.1’de verilmiştir.

Çizelge 4.1. Model eğitim özellikleri ve açıklamaları

Özellik Adı	Girilen Değer	İşlevi
batch_size	32	Bir öğrenme iterasyonunda eğitim verilerinin kaç örneğini işleyeceğini belirler. Burada 32 örnek bir kerede işlenecektir.
eval_batch_size	16	Eğitim sonrası değerlendirme yaparken kaç örneğin kullanılacağını belirler. Burada 16 örnek kullanılacaktır.
num_loader_workers num_eval_loader_workers	0	Veri yükleyicileri tarafından kullanılacak işçi sayısını belirler. 0 değeri, verilerin ana işlemci tarafından yüklenmesi anlamına gelir.

Çizelge 4.1. (devamı) Model eğitim özellikleri ve açıklamaları

Özellik Adı	Girilen Değer	İşlevi
run_eval	True	Eğitim sonrası değerlendirme yapılıp yapılmayacağını belirler.
test_delay_epochs	-1	Eğitim sırasında, eğitim iterasyonları arasındaki gecikmeyi belirler. -1 değeri, test edilmediği anlamına gelir.
epochs	500	Eğitim iterasyonlarının sayısını belirler.
text_cleaner	basic_turkish_cleaners	Veri öncesi işleme adımlarında kullanılacak olan bir dizi metin temizleyici işlevi belirler.
use_phonemes	True	Fonetik işlemler için, öğrenme için kullanılacak öğelerin harfler yerine fonemler olup olmayacağını belirler.
phoneme_language	tr	Fonetik işlem için kullanılacak dilin hangisi olduğunu belirler.
phoneme_cache_path		Fonetik işlem sırasında önbellek dosyalarının kaydedileceği konumu belirler.
print_step	5	Her kaç eğitim iterasyonunda bir ilerleme durumu yazdırılacağını belirler.
print_eval	False	Değerlendirme sonrasında sonuçların yazdırılıp yazdırılmayacağını belirler.
mixed_precision	True	İşlemci verimliliği için kullanılan bir özelliktir. Bu, daha düşük hassasiyetli veri türlerinin kullanılmasına izin verir.
num_squeeze	2	Çıktı katmanlarındaki boyut azaltmanın kaç kez gerçekleştirileceğini belirler.

Çizelge 4.1. (devamı) Model eğitim özellikleri ve açıklamaları

Özellik Adı	Girilen Değer	İşlevi
cudnn_enable	True	NVIDIA ekran kartları için, derin öğrenme işlemlerinin hızlandırılması için kullanılan bir kütüphanedir.
output_path		Eğitim sonrası model ve sonuçlarının kaydedileceği konumu belirler.
datasets		Eğitim verilerinin yüklenmesi için kullanılacak veri kümesini belirler.

Çizelge 4.1'deki özellikler kullanılarak eğitim süreci yaklaşık 23 günde tamamlanmıştır. Bu eğitim sürecinden sonra ise değerlendirme sürecine geçilmiştir.

4.6. Model Doğruluğunun Değerlendirilmesi

Model eğitimi tamamlandıktan sonra, modelin doğruluğunu test veri kümesi üzerinde değerlendirilmesi gerekmektedir. Değerlendirme yapılması için doğallık ve anlaşılırlık değerlendirilmesinde bulunulması gerekmektedir. Bu değerlendirmelerin yapılması için öznel ve nesnel ölçü yöntemleri kullanılmaktadır. Öznel yöntemlerde genellikle OGP veya A/B testi kullanıldığı görülmektedir. Nesnel yöntemlerde ise Ortalama Görüş Puanı-Dinleme Kalite Hedefi (OGP-DKH) altında birçok yöntem vardır. Bunlar; Sanal Konuşma Kalitesi Nesnel Dinleyici (SKKND: Virtual Speech Quality Objective Listener), Cross Correlation (Çapraz Korelasyon), Konuşma Kalitesinin Algısal Değerlendirmesi (KKAD: Perceptual Evaluation of Speech Quality), Mel Sepstral Bozukluğu (MSB: Mel Cepstral Distortion) ve Kök Ortalama Kare Hatası (KOKH: Root Mean Square Error) yöntemleri ile değerlendirmelerin yapıldığı görülmüştür.

Literatürde, çeşitli sinyal işleme yöntemlerini değerlendirmek için yaygın olarak ITU-T P.800 standardı kullanılan bir değerlendirme tekniği olan OGP kullanılmaktadır. OGP, istatistiksel olarak anlamlı sonuçlar elde etmek için çeşitli dinleyici deneklere ihtiyaç duymaktadır. Deneyler genellikle kontrol edilebilen bir ortamda belirli akustik özelliklere

ve donanıma sahip bir düzenek kullanılarak gerçekleştirilmektedir. Her denek aynı talimatlar veya etkiler altında işlem yapmaktadır [40].

Ayrıca, bazı çalışmalarda OGP yerine Bozulma Ortalama Görüş Puanı (BOGP) veya Karşılaştırmalı Ortalama Görüş Puanı (KOGP) gibi başarı ölçütleri kullanıldığı da görülmektedir. BOGP, bozulmaların puanlanması amacıyla kullanılırken, KOGP karşılaştırmalı değerlendirme yapmak istendiği durumlarda tercih edilmektedir [13, 41]. Bu değerlendirme yöntemleri, sinyal işleme yöntemlerinin performansını öznel bir şekilde değerlendirmek için kullanılmaktadır. Ancak, doğru sonuçlar elde etmek için uygun denek sayısı, kontrollü bir ortam ve standart talimatlar gibi faktörlere dikkat etmek önemlidir.

A/B testi, öznel değerlendirme yöntemlerinden biridir ve iki seviyeli bir faktörün değerlendirilmesinde kullanılmaktadır. A/B testi, konuşma sentezleme sistemlerinin farklı sürümlerini karşılaştırmak için kullanılan bir yöntemdir. Bu yöntem, doğal konuşma özelliklerinin değerlendirilmesinde, kullanıcı tercihlerinin belirlenmesinde ve geliştirilen sistemlerin kalitesinin analizinde önemli bir araçtır. Bu yöntemde, bir kontrol grubu (sürüm A) ve bir işlem grubu (sürüm B) oluşturulmaktadır. Kontrol grubu genellikle mevcut veya varsayılan sürümü temsil ederken, işlem grubu ise test edilen değişikliği içermektedir [42].

A/B testi, belirli bir değişikliğin etkisini öznel bir şekilde değerlendirmek için kullanılmaktadır. İki farklı grup arasında karşılaştırma yapılmaktadır ve hangi grup daha iyi sonuçlar verdiği istatistiksel olarak analiz edilmektedir. Bu yöntem, konuşma sentezleme sistemlerinin doğallık veya diğer özelliklerine odaklanan çalışmalarda sıklıkla kullanılmıştır [43].

Örneğin, ikili şekilde eşleştirilmiş konuşma sentezleme sistemlerinden biri kontrol grubunu oluştururken, diğeri ise işlem grubunu temsil etmektedir. Ardından, iki grup arasında gerçekleştirilen A/B testi ile hangi sistem daha doğal olarak algılandığı belirlenmeye çalışılmaktadır. Bu test, kullanıcıların tercihlerine dayanarak, daha iyi sonuç veren konuşma sentezleme sistemini belirlemeye yardımcı olmaktadır.

Nesnel değerlendirme yöntemlerinden birçoğu OGP-DKH'nin alt başlığı altında değerlendirilmektedir. Bu kapsamda ses sinyallerinin OGP-DKH değerlerini hesaplamak için kullanılan birçok yöntemden biri olan çapraz korelasyon yöntemi [44] iki sinyal

arasındaki benzerlik ölçüsünü bulmaya yaramaktadır. Ses kalitesinin nesnel olarak ölçülmesi için bir model olan SKKND [45], iki sinyal arasındaki benzerlik özelliklerini çıkararak ve bu özellikleri kullanarak sinyaller arasındaki farkı hesaplayarak çalışmaktadır. SKKND, iki adımda çalışmaktadır. İlk adımda, sinyallerin özellikleri, spektral içerikleri, zamansal özellikleri ve diğer özellikler gibi birçok faktöre göre analiz edilir ve bir dizi özellik vektörü olarak temsil edilmektedir. İkinci adımda, SKKND, özellik vektörlerini kullanarak sinyaller arasındaki farkı hesaplamaktadır. Fark ölçüsü olarak kesit frekansı mesafe ölçümü kullanılmaktadır.

KKAD (Perceptual Evaluation of Speech Quality), dar bantlı telefon ağlarının ve konuşma kodlayıcı, kod çözücülerinin uçtan uca konuşma kalitesini nesnel olarak değerlendirmek için tasarlanmış bir yöntemdir [46]. KKAD, kullanıcı deneyimini temel alan ve insan algısını yansıtan bir değerlendirme yapmaktadır.

MSB (Mel Cepstral Distortion), istatistiksel parametrik konuşma sentezleme sistemlerinin değerlendirilmesinde kullanılan bir diğer nesnel değerlendirme yöntemidir [47]. MSB, genellikle iki konuşma çerçevesi arasındaki Öklid mesafesi olarak tanımlanmaktadır. Bu ölçüt, sentezlenmiş ve doğal mel kepstral diziler arasındaki farkı ölçmektedir. Daha küçük bir MSB değeri, sentetik konuşmanın doğal konuşmayı daha iyi yeniden ürettiğini göstermektedir.

KOKH (Root Mean Square Error), bir modelin tahmin ettiği değerler ile gerçek değerler arasındaki farkın standart sapmasını ölçen bir nesnel değerlendirme yöntemidir [47, 48]. KOKH, sentetik konuşmanın kalitesini değerlendirmek için kullanılan bir ölçüt olmakla birlikte, mükemmel bir ölçü olmadığı belirtilmektedir. Ancak genellikle diğer ölçütlerle birlikte bir gösterge olarak kullanılabilmekte ve faydalı sonuçlar verebilmektedir [49].

Bu nesnel değerlendirme yöntemleri, konuşma sentezleme sistemlerinin kalitesini objektif bir şekilde değerlendirmek için kullanılmaktadır. Her bir yöntemin farklı ölçütleri ve metrikleri vardır ve sentezlenmiş konuşmanın gerçek konuşmayı ne kadar iyi taklit ettiğini değerlendirmek için birlikte kullanılmaları tavsiye edilmektedir.
Yapılan çalışma sonucunda bir çeşitli değerlendirme yöntemleri ile deneysel bulgular elde edilmiştir. Bu kapsamda hem öznel değerlendirme yöntemleri hem de nesnel değerlendirme yöntemleri kullanılarak deneysel bulgu elde edilmiştir. Bu çerçevede de öznel değerlendirme

yöntemi olarak 5 erkek ve 5 kadından oluşan toplam 10 kişilik gönüllü dinleyiciye 10 adet doğal cümle ve 10 adette sentezlenmiş cümleyi dinleterek bir OGP elde edilmiştir. Diğer yandan nesnel değerlendirme yöntemi için ise ilk olarak konuşma sinyallerinin özelliklerini kullanarak otomatik olarak elde edilen bir kalite puanı olan OGP-DKH tercih edilmiştir. Başka bir nesnel değerlendirme yöntemi olarak ise konuşma sinyallerinin özelliklerini analiz ederek, insan dinleyiciler tarafından verilen benzer puanların bir benzerini üreten SKKND tercih edilmiştir. Test sürecinde değerlendirme için kullanılan 10 adet cümle belirlenmiştir ve öznel ve nesnel değerlendirmelerin tamamında Çizelge 4.2'de görülen bu cümleler kullanılmıştır.

Çizelge 4.2. Test işlemlerinde kullanılan cümleler ve sıra numaraları

#	Cümle
1	Bu konuda bir düzenleme yapmak laiklik ilkesinin ihlali olarak algılanmaktadır.
2	Ancak ücretsiz izinli olunan dönemde gelir vergisi kesintisi yapılmadığı için asgari geçim indiriminden yararlanılamayacaktır.
3	Buranın işlerini yürütmekle görevli kişileri de daimi temsilci olarak kabul etmek mümkün bulunmamaktadır.
4	Kooperatiflerde hangi durumlarda iktisadi işletme oluşur?
5	Motorlu taşıtlar vergisi yönünden, bağlı bulundukları vergi dairesine müracaat ederler.
6	Adres paylaşım sistemi üzerinden yapılacak adres teyidi, hangi durumlarda ve şekilde yapılacaktı?
7	Türk uyruklu gerçek kişiler için aranan belgeler üzerinden, kayıp tarihinden sonra ise Türk uyruklu olmayan gerçek kişiler kapsamında yapılması gerekmektedir.
8	Ve belli tutarların altında kalan sözleşmeler için basitleştirilmiş tedbir uygulanmasına imkan verilmiştir.
9	Konyalılar çoluk çocukları ile sekiz ay meramda oturup zevk ederek hayattan tat alırlar.
10	Eski plaktı birdenbire mesudum işitmek hevesiyle, gönlüm dolu İstanbul un en özlü sesiyle sandım ki.

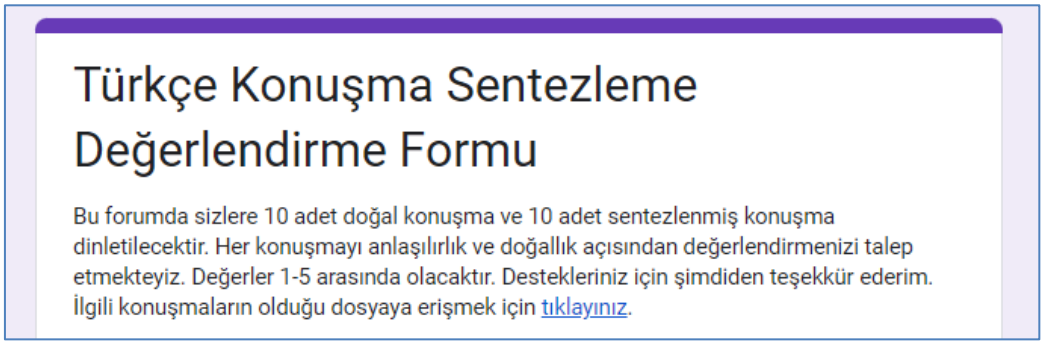
4.6.1. Ortalama görüş puanı deneyi

Ortalama Görüş Puanı (OGP) Deneyi [40], bir konuşma sentezleme sisteminin veya diğer ses işleme sistemlerinin kalitesini değerlendirmek için kullanılan bir ölçüttür. OGP, genellikle bir kullanıcı deneyimi anketi veya değerlendirme süreci aracılığıyla elde edilen subjektif bir değerlendirmedir. Bu değerlendirme, kullanıcıların bir sesin kalitesi, anlaşılabilirliği, doğallığı, netliği ve genel memnuniyeti gibi faktörlere dayanan bir puanlama yapmasını içermektedir.

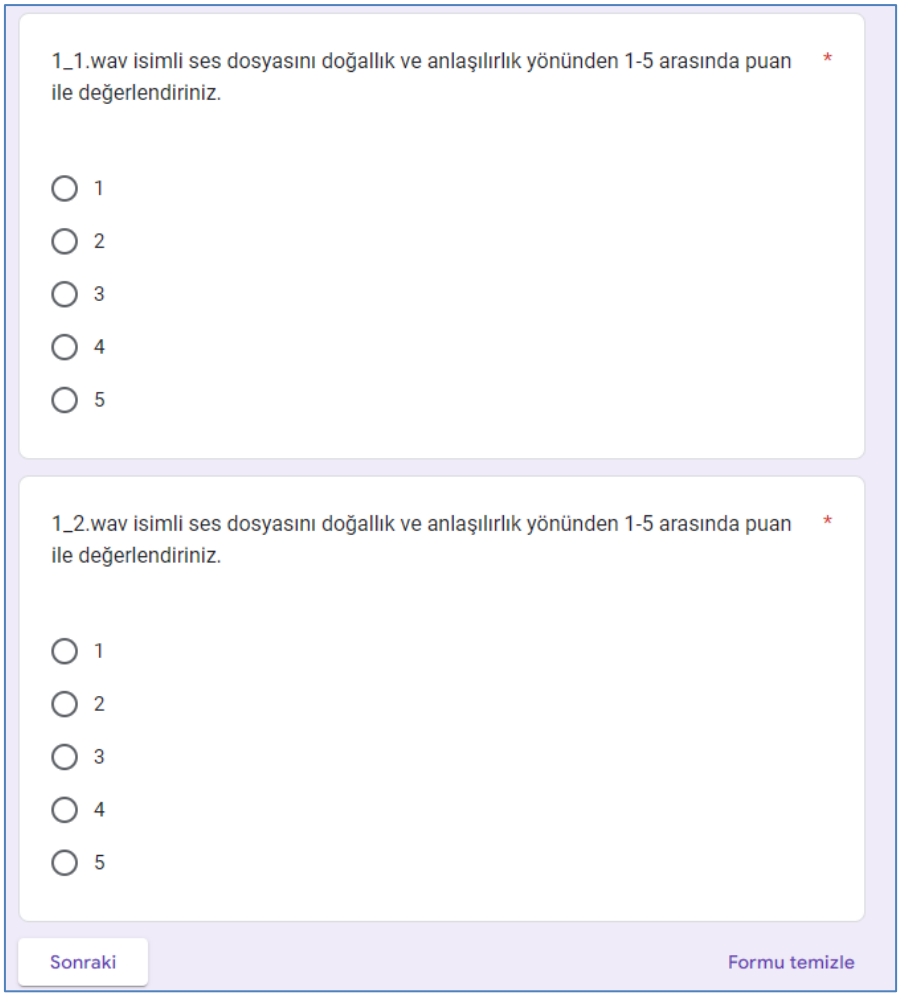
OGP, genellikle bir kullanıcı deneyimi değerlendirme testinde kullanılmaktadır. Bu testlerde, katılımcılara bir dizi ses örneği sunulmaktadır ve her bir örneği bir derecelendirme ölçeği üzerinde puanlamaları istenmektedir. Katılımcılar, sesin kalitesi, anlaşılabilirliği veya diğer özellikleri hakkında yargılarını ifade etmektedirler. Bu puanlar daha sonra toplanmakta ve ortalaması alınarak OGP elde edilmektedir.

OGP değerlendirmesi, akademik araştırmalarda, ses işleme sistemleri üzerinde yapılan çalışmalarda ve endüstriyel uygulamalarda sıkça kullanılan bir ölçüttür. Bu değerlendirme, ses sentezi sistemlerinin performansını öznel bir şekilde karşılaştırmak, iyileştirmek ve optimize etmek için kullanılmaktadır. Ancak, OGP değerlendirmesi tek başına yeterli değildir ve diğer nesnel ölçütlere ve analizlere de başvurulması önemlidir.

Bu çalışma kapsamında ilk olarak öznel değerlendirme ölçütlerinden olan OGP hakkında bilgi verilecektir. OGP'nın hesaplaması için 5'i erkek 5'i kadın olan 10 gönüllü dinleyiciye Çizelge 4.2'de bulunan 10 cümlenin doğal konuşulmuş hali ve sentezlenmiş hali dinletilmiştir.



Şekil 4.2. Değerlendirme ekranı bilgilendirme mesajı



Şekil 4.3. Değerlendirme ekranı puanlama yapısı

Şekil 4.2'de görüldüğü üzere gönüllü dinleyicilere dineleme yapmaları ilgili dosyalar paylaşılmış ve gerekli bilgilendirme mesajları hem yazılı şekilde hem de sözel şekilde belirtilmiştir. Dinleyicilerden ilgili konuşmaları dinleyerek anlaşılırlık ve doğallık açısından değerlendirme yaparak Şekil 4.3'de görüldüğü gibi 1 (çok başarısız) ile 5 (çok başarılı) arasında puan vermeleri istenmiştir. Çizelge 4.3'te ise her bir ses dosyası için dinleyicilerin verdiği puanların ortalaması görünmektedir.

Çizelge 4.3. Gönüllü dinleyiciler tarafından verilen ortalama puanlar

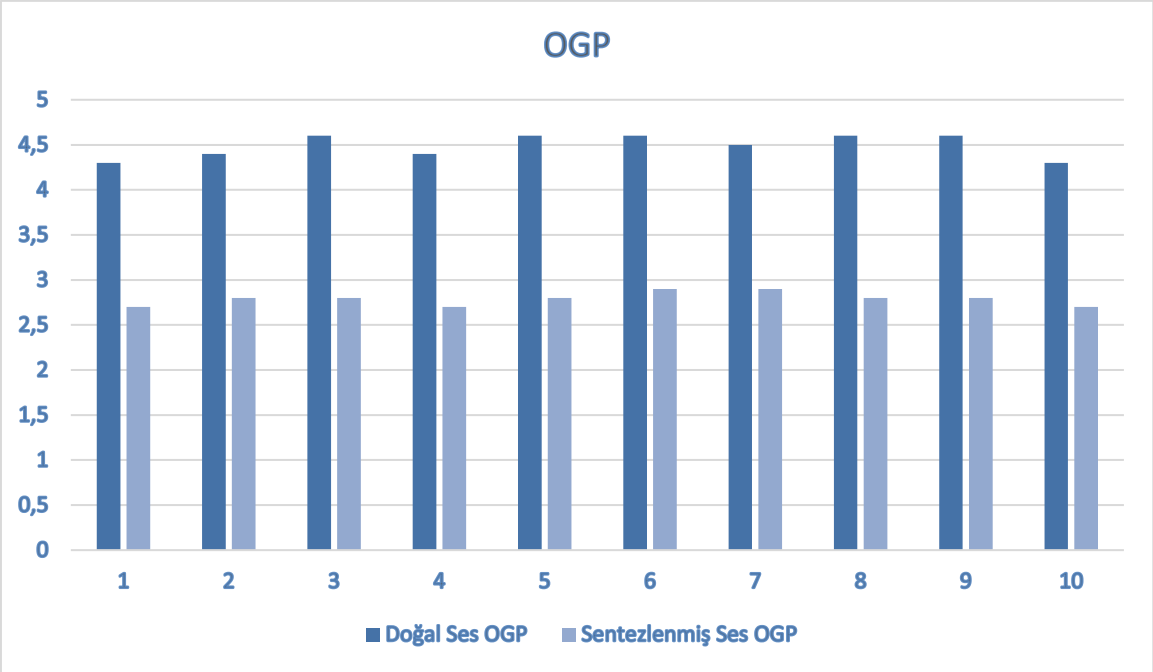
#	Doğal Ses OGP	Sentezlenmiş Ses OGP
1	4,3	2,7
2	4,4	2,8

Çizelge 4.3. (devamı) Gönüllü dinleyiciler tarafından verilen ortalama puanlar

#	Doğal Ses OGP	Sentezlenmiş Ses OGP
3	4,6	2,8
4	4,4	2,7
5	4,6	2,8
6	4,6	2,9
7	4,5	2,9
8	4,6	2,8
9	4,6	2,8
10	4,3	2,7
Ortalama OGP	4,49	2,79

Yapılan değerlendirmeler sonucunda dinleyiciler gerçek konuşmaları içeren 10 ses dosyasına ortalama 4,49 gibi bir OGP vermişlerdir. Aynı dinleyiciler sentezlenmiş konuşmaları içeren 10 ses dosyasına ise 2,79 değerinde bir OGP vermişlerdir.

Bu çalışmada, 40 farklı değerlendiricinin derecelendirme arasındaki uyum düzeyini belirlemek için Kappa istatistiği kullanılmıştır. İki değerlendirici için genellikle kullanılan Kappa istatistiği, Fleiss tarafından birden fazla değerlendirici arasındaki uyumu ölçmek için genelleştirilmiştir [50, 51]. Kappa değerleri, -1 ile +1 arasında değişmektedir ve pozitif değerler, değerlendiriciler arasında rastgele beklenenden daha yüksek bir uyumu göstermektedir. Beş farklı değerlendirme kategorisi için (Zayıf = 0.32, Düşük = 0.53, Orta = 0.68, Yüksek = 0.76, En İyi = 0.87), her biri için Kappa değerleri hesaplanarak değerlendiriciler arasındaki uyum değerlendirilmiştir. Her kategori için pozitif Kappa değerleri, değerlendiriciler arasındaki uyumun rastgele beklenenden daha yüksek olduğunu göstermektedir. En düşük uyum "Zayıf" kategorisinde gözlemlenirken (Kappa = 0.32), en yüksek uyum "En İyi" kategorisinde gözlemlenmiştir (Kappa = 0.87).



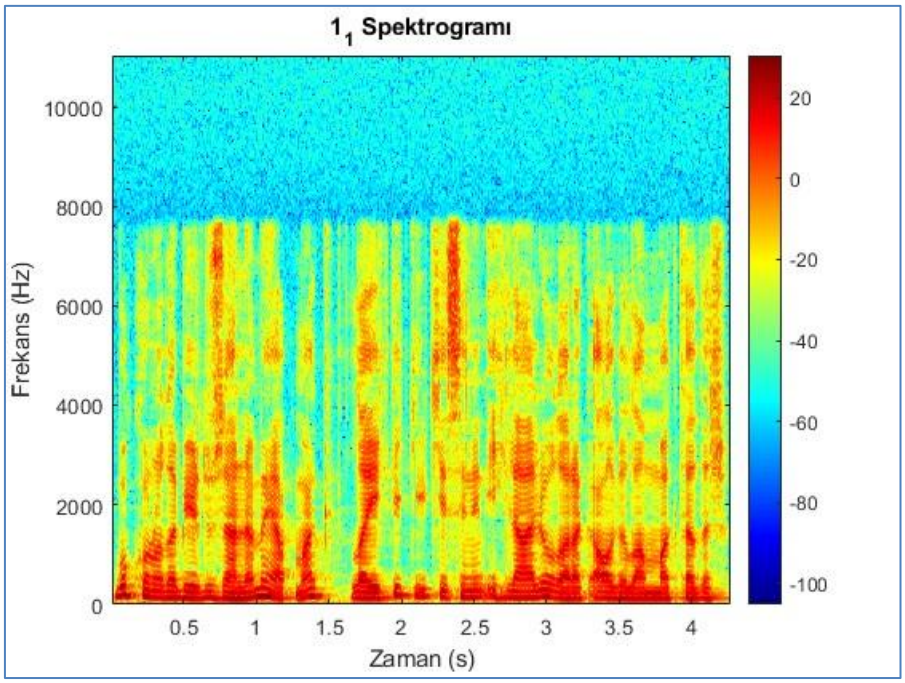
Şekil 4.4. OGP sonuçları grafik görünümü

4.6.2. Spektrogramların değerlendirilmesi

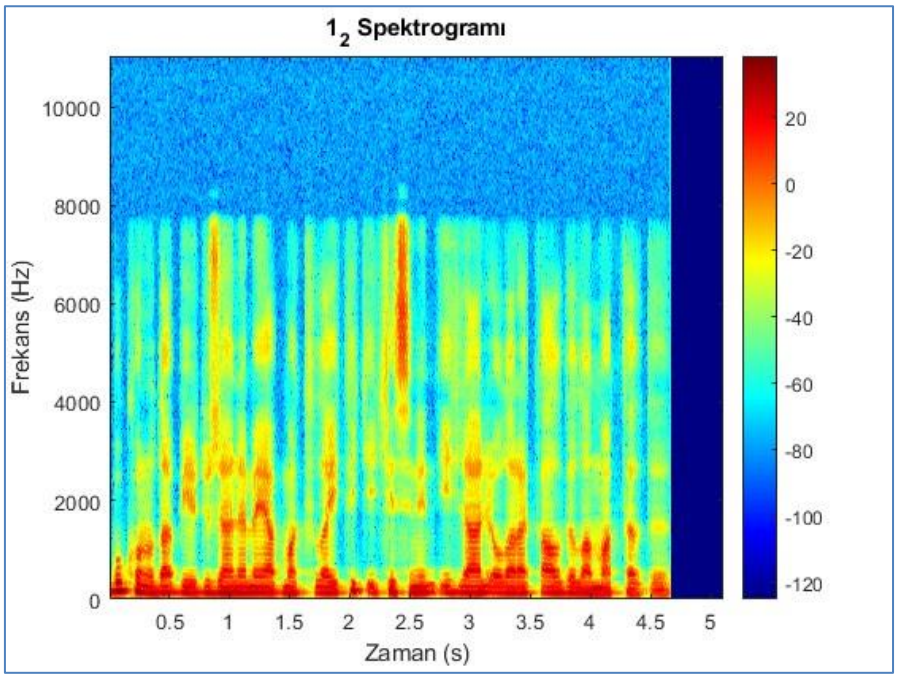
Spektrogramlar, ses sinyallerini zaman, frekans ve şiddet bilgilerini içeren görsel temsiller olarak gösteren grafiklerdir. Sesin frekans bileşenlerini ve zaman içindeki değişimlerini gösterdikleri için, ses dosyalarını karşılaştırmak için kullanılan etkili bir araçtır.

Spektrogramların kullanılması, ses dosyalarını nesnel olarak karşılaştırmak ve analiz etmek için bir araç sağlamaktadır. Özellikle ses işleme, konuşma tanıma ve sentezleme, müzik analizi gibi alanlarda spektrogramlar yaygın olarak kullanılmaktadır. Spektrogramlar, ses dosyalarının görsel bir temsilini sağladıkları için, insan görsel algısına dayanarak hızlı bir şekilde ses dosyalarının karşılaştırılması ve analiz edilmesi için etkili bir araçtır.

Şekil 4.5 ve 4.6'da 1 numaralı konuşma metni olan "Bu konuda bir düzenleme yapmak laiklik ilkesinin ihlali olarak algılanmaktadır." ifadesinin spektrogramları görülmektedir. MATLAB'da "Signal Processing Toolbox" ile Şekil 4.5'te gerçek konuşma kaydının, Şekil 4.6'da ise sentezlenmiş konuşmanın spektrogramları elde edilmiştir.



Şekil 4.5. Gerçek konuşma kaydının spektrogramı



Şekil 4.6. Sentezlenmiş konuşma kaydının spektrogramı

Spektrogramlar, ses sinyallerinin frekans bileşenlerini ve zaman içindeki değişimlerini görsel olarak temsil eden grafiklerdir. Bir spektrogramda, frekans bileşenleri dikey eksende, zaman ise yatay eksende gösterilmektedir. Renk tonları ise belirli bir frekans bileşeninin gücünü yansıtmaktadır. Yüksek güç, daha parlak renk tonlarıyla temsil edilirken, düşük güç

daha soluk tonlarla temsil edilmektedir.

Şekil 4.5 ve 4.6'daki görüntüler incelendiğinde, iki spektrogramda da benzer frekans bileşenlerinin olduğu gözlemlenmektedir. Ancak, Şekil 4.5'teki spektrogramda, Şekil 4.6'daine kıyasla daha yüksek bir frekans bileşeni olduğu görülmektedir. Bu durum, iki farklı insan sesinin karşılaştırılması sırasında, sesin kaydedildiği ortam, kaydedilme şekli ve kullanılan donanım gibi faktörlere bağlı olarak meydana gelen değişiklikleri yansıtmaktadır.

Spektrogramlar, ses dosyalarının frekans bileşenlerinin dağılımını ve zaman içindeki değişimlerini açıkça göstermektedir. Benzer frekans bileşenlerinin görülmesi, iki sesin belirli frekans aralıklarında benzerlik gösterdiğini göstermektedir. Bununla birlikte, farklı frekans bileşenlerinin gücü ve dağılımı, seslerin karakteristiklerinin farklı olduğunu göstermektedir. Örneğin, Şekil 4.5'teki spektrogramdaki yüksek frekans bileşeni, sesin daha keskin ve zayıf bir bileşen içerdiğini gösterirken, Şekil 4.6'daki daha düşük frekans bileşeni, daha güçlü ve daha yayılmış bir bileşeni yansıtabilmektedir.

Spektrogramlar, ses dosyalarını karşılaştırırken sesin frekans bileşenlerini ve zaman içindeki değişimlerini açıkça gözlememizi sağlamaktadır. Bu da seslerin benzerliklerini ve farklılıklarını değerlendirmemize yardımcı olmaktadır. Ancak, spektrogramlar tek başına bir sesin tüm özelliklerini açıklayamaz ve diğer analiz yöntemleriyle birlikte kullanılması önemlidir.

4.6.3. Çapraz korelasyon

Çapraz korelasyon [44], iki sinyal arasındaki benzerlik veya ilişkiyi ölçmek için kullanılan bir işlemdir. İki sinyalin çapraz korelasyonu, bir sinyalin diğerine göre ne kadar benzediğini veya ilişkili olduğunu belirlemek için kullanılmaktadır. Konuşma sentezleme için çapraz korelasyon, bir konuşma sinyalinin hedeflenen hedef sinyale olan benzerliğini veya ilişkisini değerlendirmek amacıyla kullanılabilmektedir. Çapraz korelasyonun matematiksel ifadesi [51, 52]; $S1 = \{s1_0, s1_1, s1_2, s1_3, \ldots, s1_{N_1-1}\}$ ve $S2 = \{s2_0, s2_1, s2_2, s2_3, \ldots, s2_{N_2-1}\}$ olmak üzere iki ses sinyalidir. Buna göre, $s1_p$ ($p = 0, 1, 2, \ldots, N_1-1$) S1 ses sinyalinin sayısal göstergesi, $s2_p$ ($p = 0, 1, 2, \ldots, N_2-1$) S2 ses sinyalinin sayısal göstergesi olup, N_1 ve N_2 sırasıyla S1 ve S2 ses sinyallerinin gösterge uzunluğudur.

$L_a = \frac{N_a}{M}$ (a = 1, 2) (4.1)

Denklem 4.1'de her bir ses sinyali M eşit parçaya bölünecek şekilde L uzunluktadır. Burada bölme işlemi sonucunda elde edilen değerin tam değeri kullanılmaktadır.

$t^i_a = i*L_a$ (a = 1, 2) (i = 1, 2, 3, …, M-1) (4.2)

t^i_a (a = 1, 2), sırasıyla S1 ses sinyalinin i. segmentinin ilk gösteriminin sayısını ve S2 ses sinyalinin i. segmentinin ilk gösterimlerinin sayısını göstermektedir.

$$r_i(k) = r(i, k) = \frac{\frac{1}{L} \sum_{j=0}^{L-1} s1_{t_1^i} s2_{t_2^i+k+j} - \mu 1_i \mu 2_{i+k}}{\sqrt{\sigma 1_i^2 \sigma 2_{i+k}^2}} \quad (4.3)$$

Denklem 4.3'te S2 sinyalinin i. segmentini k kadar sağa kaydırarak $r_i(k)$ değeri elde edilmektedir. Bu değer elde edilirken L_1 ve L_2 değerlerinden küçük olan değer L değeri olarak alınmaktadır ($L = \min(L_1, L_2)$).

$$\mu 1_i = \frac{1}{L} \sum_{j=0}^{L-1_i} s1_{t_1^i+j} \quad (4.4)$$

$$\mu 2_{i+k} = \frac{1}{L} \sum_{j=0}^{L-1} s2_{t_2^i+k+j} \quad (4.5)$$

$$\sigma 1_i^2 = \frac{1}{L} \sum_{j=0}^{L-1_i} s1_{t_1^i+j} - \mu 1^2 \quad (4.6)$$

$$\sigma 2_{i+k}^2 = \frac{1}{L} \sum_{j=0}^{L-1} s2_{t_2^i+k+j} - \mu 2^2 \quad (4.7)$$

Denklem 4.4, 4.5, 4.6 ve 4.7'de Denklem 4.3'te kullanılan bazı ifadelerin açık hali görülmektedir. μ1 ve μ2 ifadeleri örnekleme ortalamalarını ifade ederken, σ1 ve σ2 ifadeleri örnekleme varyanslarını ifade etmektedir.

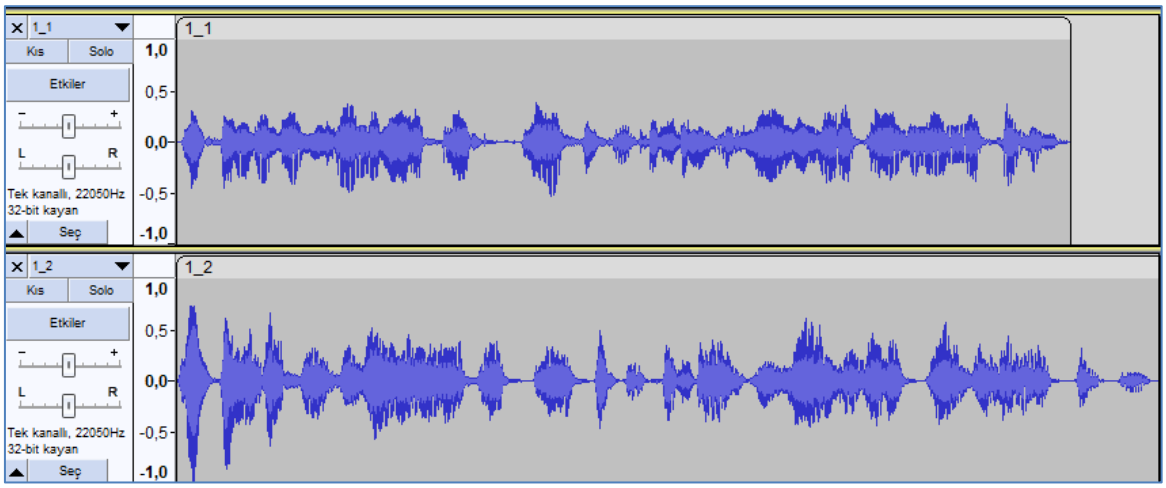
Genel bir ifade ile çapraz korelasyon, iki sinyal arasındaki zaman doğrusu boyunca ortaklık ölçüsünü hesaplamaktadır. Bu, iki sinyal arasındaki örtüşme miktarını ve zamanlama farklarını göstermektedir. Genellikle, bir sinyaldeki belirli bir örneğin diğer sinyaldeki

benzer bir örnekle nasıl ilişkili olduğunu göstermek için kullanılan otomatik bir işlemdir.

Konuşma sentezleme için çapraz korelasyon, sentezlenen konuşma sinyalinin hedeflenen konuşma sinyaline olan benzerliğini değerlendirmek için kullanılabilmektedir. Örneğin, bir konuşma sentezleme sistemini değerlendirmek için, sentezlenen konuşma sinyali ile gerçek konuşma sinyali arasındaki çapraz korelasyonu hesaplayabilmektedir. Yüksek bir çapraz korelasyon değeri, sentezlenen konuşmanın hedeflenen konuşma sinyaliyle daha iyi bir uyum içerdiğini gösterebilmektedir.

Çapraz korelasyon, konuşma sentezleme alanında kullanılan birçok diğer ölçüt ve yöntemle birlikte kullanılmaktadır. Bu, sentezlenen konuşmanın doğallığı, anlaşılabilirliği ve benzeri özelliklerini değerlendirmek için daha kapsamlı bir analiz sağlamaktadır.

Şekil 4.7'de ise bu seslere ait ses dalgaları Audacity isimli programdan alınarak gösterilmiştir.



Şekil 4.7. Gerçek ve sentezlenmiş konuşmaların ses dalgaları

Şekil 4.7'de görülen ses sinyallerinin OGP-DKH değerlerini hesaplamak için kullanılan birçok yöntemden biri olan cross-correlation (çapraz korelasyon) yöntemi [44] kullanılmıştır. Bu işlem yine MATLAB'da "Signal Processing Toolbox" ile yapılmış olup, iki sinyal arasındaki benzerlik ölçüsünü bulmaya yarayan çapraz korelasyon fonksiyonu ile bulunması sağlanmıştır. Bu çalışma sonucu OGP-DKH puanı 51.090703 gibi bir değer elde edilmiştir. Elde edilen OGP-DKH puanı 51.090703, ses kalitesinin ortalama bir seviyede

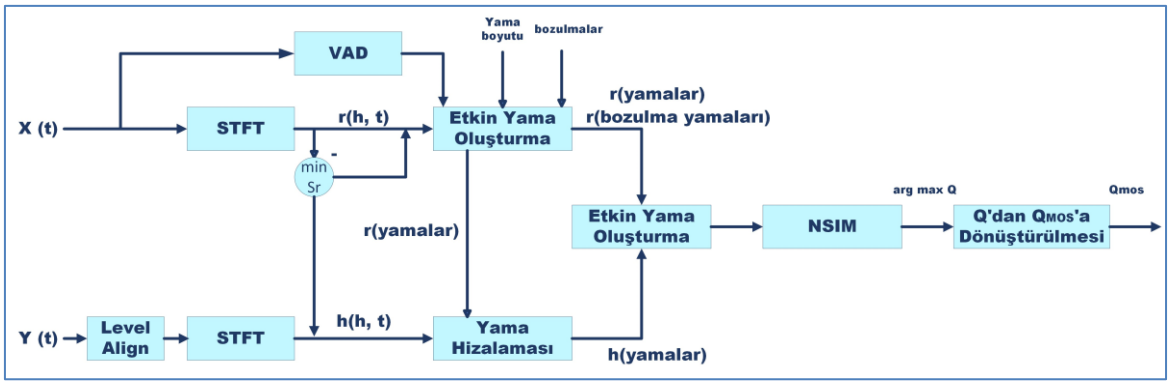
olduğunu göstermektedir. OGP-DKH puanı 0 ila 100 arasında değişen bir ölçektir ve daha yüksek skorlar daha iyi ses kalitesi ile ilişkilidir. Ancak, bir OGP-DKH skoru tek başına yeterli bir değerlendirme yöntemi olmayabilmektedir. Ses kalitesinin belirlenmesinde diğer faktörler de önemli olabilir ve bu faktörlerin göz önünde bulundurulması gerekmektedir.

Ayrıca, sinyallerin görsel olarak da incelendiğinde, benzerliklerin ve farklılıkların olduğu görülmektedir. İki sinyal de aynı konuşmayı içeriyor, ancak bazı küçük farklılıklar da mevcuttur. Örneğin, bazı kelimelerin telaffuzu ve vurgusu farklı olduğu gözle de görülmektedir.

4.6.4. Sanal konuşma kalitesi nesnel dinleyici (SKKND)

Ses kalitesinin nesnel olarak ölçülmesi için bir model olan SKKND [45], iki sinyal arasındaki benzerlik özelliklerini çıkararak ve bu özellikleri kullanarak sinyaller arasındaki farkı hesaplayarak çalışmaktadır. SKKND, iki adımda çalışmaktadır. İlk adımda, sinyallerin özellikleri, spektral içerikleri, zamansal özellikleri ve diğer özellikler gibi birçok faktöre göre analiz edilmektedir ve bir dizi özellik vektörü olarak temsil edilmektedir. İkinci adımda, SKKND, özellik vektörlerini kullanarak sinyaller arasındaki farkı hesaplamaktadır. Fark ölçüsü olarak kesit frekansı mesafe ölçümü kullanılmaktadır. Şekil 4.8'de SKKND genel mimarisi verilmiştir.

SKKND, bu fark skorlarını kullanarak sinyallerin kalitesini ölçen bir kalite skoru üretmektedir. Bu kalite skoru, kullanıcının sağladığı referans sinyali ile test edilen sinyal arasındaki kalite farkını belirlemektedir. SKKND, özellikle ses kalitesinin ölçülmesi konusunda oldukça etkili bir araçtır ve bireysel özellikleri dikkate alarak sentetik sinyallerle test edilebilmektedir.



Şekil 4.8. SKKND genel mimarisi

Şekil 4.8'e göre X ve Y iki ses sinyali arasında referans ve bozulmuş sinyallerin spektrogramlarını oluşturarak, yamalar arasındaki benzerlikleri hesaplayarak ve sonunda Q_{OGP} değerini tahmin ederek konuşma kalitesini hesaplamaktadır. Modelin blok diyagramda gösterilen beş ana işleme aşaması vardır: ön işleme, zaman hizalaması, bozulma tahmini, benzerlik karşılaştırması ve benzerliği objektif kaliteye dönüştüren bir son işlem eşlemedir.

Genel çalışma algoritmasında, ilk olarak, referans sinyali (x) ve bozulmuş sinyali (y) spektrograma dönüştürülmektedir. Spektrogramlardan minimum değerler çıkarılarak işleme yapılmaktadır. Yama boyutu kadar döngü oluşturulmaktadır ve SAA (Voice Activity Detection: Ses Aktivite Algılama) doğru ise referans yamaları ve bozulma yamaları oluşturulmaktadır. Hizalama işlemi gerçekleştirilerek referans yamaları ile bozulmuş sinyal arasındaki uyum hesaplanmaktadır. Tüm referans yamaları, bozulmalar ve hizalama sonuçları üzerinde döngüler oluşturularak benzerlik sonuçlar NBİÖ (Neurogram Similarity Index Measure: Neurogram Benzerlik İndeksi Ölçümü) ile hesaplanmaktadır. Hesaplanan benzerlik sonuçlarında maksimum değerler alınarak sonuçlar elde edilmektedir. Elde edilen benzerlik sonuçları, yama sayısına bölünerek Q değeri hesaplanmaktadır. Son olarak benzerlik sonuçlarını (Q) insan algısıyla uyumlu bir kalite tahminine dönüştüren bir fonksiyon olan maptomos () fonksiyonu ile Q_{OGP} değeri elde edilmektedir. Bu kapsamda Çizelge 4.4'te daha önce kullanılan sesler için elde edilen SKKND değerleri verilmiştir.

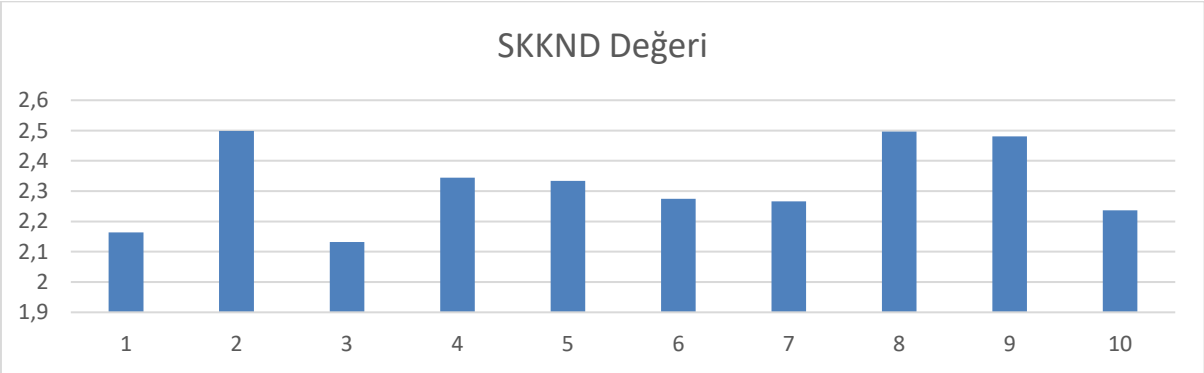
Çizelge 4.4. SKKND sonuçları

#	SKKND Değeri
1	2,16388

Çizelge 4.4. (devamı) SKKND sonuçları

#	SKKND Değeri
2	2,49881
3	2,13251
4	2,34417
5	2,33374
6	2,27508
7	2,26635
8	2,49674
9	2,48059
10	2,23684
Ortalama	2,32287

Google tarafından geliştirilmiş Bazel derleme aracı ile Python'da yazılmış SKKND kodunun çalıştırılmasıyla Tablo 8'deki değerler elde edilmiştir. Bu sonuçların elde edilmesinde Bazel dışında çeşitli TensorFlow, Numpy gibi Python kütüphanelerinden de faydalanılmıştır. İhtiyaç duyulan yardımcı araçların kurulumu sonrasında her bir ses dosyası grubu için SKKND değeri "./bazel-bin/visqol --reference_file DogalSes.wav --degraded_file SentezlenmisSes.wav --verbose" ifadesi ile tek tek elde edilmiştir. Bu işlemler sonucunda elde edilmiş 10 değerin ortalaması olarak 2,323 değerine ulaşılmıştır.



Şekil 4.9. SKKND değeri sonuçları grafik görünümü

5. SONUÇ VE ÖNERİLER

Türkçe konuşma sentezleme sistemleri, gerçekçi, doğal ve insan benzeri konuşma sinyalleri üretme amacı güden sistemlerdir. Bu sistemler, iletişim teknolojileri, sanal asistanlar, otomasyon sistemleri gibi birçok alanda yaygın olarak kullanılmaktadır. Ayrıca, dil öğrenme, konuşma tanıma, sese duyarlı arayüzler ve eğitim gibi alanlarda da önemli bir rol oynamaktadır.

Derin öğrenme tabanlı uçtan uca konuşma sentezleme sistemleri, son yıllarda diğer diller için başarılı sonuçlar vermiştir ve Türkçe için de geliştirilmiştir. Ancak literatürde yapılan çalışmalar incelendiğinde, bu tür yöntemlerin henüz Türkçe için uygulanmadığı görülmüştür. Bu nedenle, bu çalışmanın Türkçe konuşma sentezleme alanında öncü bir çalışma olabileceği düşünülmektedir.

Bu çalışmanın sonuçları, Türkçe konuşma sentezleme sistemlerinin geliştirilmesi ve ileriye yönelik çalışmalara temel oluşturması açısından önemlidir. Önerilen derin öğrenme tabanlı uçtan uca konuşma sentezleme yöntemi, Türkçe için uygun bir çözüm sunabilir ve doğal bir konuşma hissiyatını yansıtan, anlaşılır bir ses çıkarma sağlayabilmektedir. Bu çalışma, Türkçe dilindeki konuşma sentezleme araştırmalarına katkıda bulunarak, Türkçe konuşma teknolojilerinin geliştirilmesine yönelik ilerlemelere öncülük edebilmektedir.

Bu çalışma kapsamında, Türkçe konuşma sentezleme sistemlerinin geliştirme süreçleri detaylı bir şekilde ele alınmıştır. Bu sistemler için Türkçe için uygun bir altyapı olan GlowTTS modeli tercih edilmiştir. GlowTTS modeline uygulanmak üzere, İstanbul Türkçesiyle genç bir erkek tarafından seslendirilen bir veri kümesi kullanılmıştır. Bu veri kümesi 12 saat 38 dakika 59 saniye uzunluğunda olup, her biri 261 adımdan oluşan 500 çevrimle GlowTTS modelinde eğitilmiştir.

Eğitim süreci sonucunda, mel spektrogramlar elde edilmiştir. Daha sonra ses kodlayıcısı kullanılarak bu mel spektrogramlar kullanılarak ses dosyaları üretilmiştir. Elde edilen ses dosyalarının doğallık ve anlaşılırlık açısından değerlendirilmesi için öznel ve nesnel değerlendirme yöntemleri kullanılmıştır.

Öznel değerlendirme için, 10 kişilik bir gönüllü dinleyici grubuna 10 adet doğal cümle ve 10 adet sentezlenmiş cümle dinletilerek OGP elde edilmiştir. Nesnel değerlendirme için ise, konuşma sinyallerinin özelliklerini kullanarak otomatik olarak elde edilen OGP-DKH ve konuşma sinyallerinin özelliklerini analiz ederek hesaplanan SKKND değerleri kullanılmıştır.

Yapılan deneysel çalışmalar neticesinde ise doğal konuşmaları içeren grup için öznel olarak ortalama 4,49'luk bir OGP, sentezlenmiş konuşmaları içeren grup için ortalama 2,79'luk OGP elde edilmiştir. Bu değerlendirme yöntemi öznel bir yöntem olduğu için diğer çalışmalardan elde edilen sonuçlar ile direk bu şekilde karşılaştırılması çok mümkün değildir. Ancak bu teknik ile yapılan diğer diller için yapılan çalışmalarda elde edilen sonuçlar ile karşılaştırıldığında doğal konuşmalar ile sentezlenmiş konuşmalar arasında alınan ortalama sonuç arasındaki farkın fazla olduğu görülmektedir.

Çizelge 5.1. Karşılaştırmalı OGP

Çalışma Adı	Kullanılan Yöntem	OGP
Sak, Güngör ve Safkan (2006)	Birleştirmeli	4,2
Bicil (2010)	Birleştirmeli	3,87
Arık (2010)	Birleştirmeli	3,7
Karaahmet (2015)	Birleştirmeli	3,8
Tora, Uslu ve Karaahmet (2017)	Birleştirmeli	3,29
Mevcut Çalışma	Derin Öğrenme Tabanlı	2,79

Çizelge 5.1'de literatürde incelenen çalışmalarda elde edilmiş OGP ve bu çalışmada elde edilmiş OGP görülmektedir. Daha önce yapılan çalışmalarda anlaşılırlık konusunda iyi sonuçlar veren ancak doğallık konusundan yetersiz kalan birleştirmeli konuşma sentezleme sistemi kullanılmıştır. Bu çalışmalarda öznel değerlendirme olarak yapılan deneylerde doğal ses ile elde edilmiş bir OGP ile karşılaşılmamıştır.

Nesnel değerlendirme yöntemi olarak ise doğal konuşma ile sentezlenmiş konuşma arasındaki benzerliği bilimsel olarak hesaplayan birkaç nesnel bilimsel çalışmadan

faydalanılmıştır. Bunlardan biri iki sinyal arasındaki benzerlik ölçüsünü bulmaya yarayan çapraz-korelasyon fonksiyonudur. Bu yöntemden 51,09'luk OGP-DKH puanı elde edilmiştir. Bu sonuç ortalama bir ses kalitesinin ortaya çıkarıldığını ortaya koymaktadır.

Deneysel bulgu elde edilirken kullanılan bir başka nesnel değerlendirme yöntemi ise yine iki sinyal arasındaki benzerlik özelliklerini çıkararak ve bu özellikleri kullanarak sinyaller arasındaki farkı hesaplayarak çalışan SKKND'dir. Bu yöntem sonucu 2,323'lük bir SKKND değeri elde edilmiştir. Bu durumda, 2,323 SKKND puanı, sentezlenmiş konuşmanın doğal konuşmaya göre makul bir kalite düzeyine sahip olduğunu göstermektedir. Ancak, ideal bir sentezleme sisteminden beklenen puanlar daha yüksektir.

Bu deneysel bulgular, çalışmanın potansiyelini göstererek daha fazla çalışmanın gerekliliğini ortaya koymaktadır. Daha iyi sonuçlar elde etmek için çeşitli çalışmaların yapılması öngörülmektedir. Öncelikli olarak, model eğitiminde daha büyük bir veri kümesinin kullanılması, daha iyi bir özellik çıkarma yönteminin benimsenmesi ve daha yüksek kaliteli ses verilerinin kullanılması gibi faktörlerin dikkate alınması önerilmektedir. Ayrıca, modelin eğitim özelliklerinin ayarlanması ve daha uzun süre eğitilmesi gibi iyileştirmeler de potansiyel olarak daha iyi sonuçlar doğurabilmektedir.

Bu bağlamda, mevcut veri kümesinin genişletilmesi veya farklı bir konuşmacı ile daha donanımlı ses verilerinin kullanılması gibi veri kümesi çalışmalarının gelecek çalışmalara daha iyi sonuçlar sağlayabileceği düşünülmektedir. Özellikle GlowTTS üzerine yapılan çalışmanın, çevrim sayısının artırılması gibi eğitim özelliklerinin değiştirilmesiyle sentezlenmiş konuşmanın kalitesinin artabileceği öngörülmektedir.

Son olarak, gelecek çalışmalarda Tacotron, Tacotron2, FastSpeech gibi farklı derin öğrenme tabanlı uçtan uca konuşma sentezleme sistemi modelleri üzerinde çalışmaların yapılması gerektiği düşünülmektedir. Bu modellerin incelenmesi, farklı sentezleme yaklaşımlarının değerlendirilmesi ve performans karşılaştırmalarının yapılması gelecekteki araştırmalara katkı sağlayabilmektedir.

Tüm bu öneriler ve yapılan yorumlar, çalışmanın sınırlamalarını ve ileriye yönelik potansiyel geliştirmeleri vurgulamaktadır. Bu şekilde, gelecekteki araştırmalar, daha iyi konuşma sentezleme sistemleri için gerekli adımları atmak ve genel performansı iyileştirmek

için değerli bir rehber sağlayacaktır.

KAYNAKLAR

1. Visser, E. J., Pak, R. ve Shaw, T. H. (2018). From ‘automation’ to ‘autonomy’: the importance of trust repair in human–machine interaction. *Ergonomics*, 61 (10), 1409-1427.

2. Alwan, A., Narayanan, S., Strope, B. ve Shen, A. (1995). Speech production and perception models and their applications to synthesis, recognition, and coding. *Proceedings of ISSE'95- International Symposium on Signals, Systems and Electronics*, 367–372.

3. Ning, Y., He, S., Wu, Z., Xing, C. ve Zhang L., (2019). A Review of Deep Learning Based Speech Synthesis. *Applied Science*, 9 (19), 4050.

4. Lemmetty, S. (1999). *Review of speech synthesis technology*, Yüksek Lisans Tezi, Helsinki University of Technology Department of Electrical and Communications Engineering, Helsinki.

5. Dudley, H. ve Tarnóczy, T. H. (1949). The Speaking Machine of Wolfgang von Kempelen. *The Journal of the Acoustical Society of America*, 22 (2), 151–166.

6. Dudley, H. (1940). The Carrier Nature of Speech, *The Bell System Technical Journal*, 19 (4), 495–515.

7. Umeda, N. ve Teranishi, R. (1975). The Parsing Program for Automatic Text-to-Speech Synthesis Developed at the Electrotechnical Laboratory in 1968. IEEE Transactions on Acoustics, Speech, and Signal Processing, 23 (2), 183–188.

8. Yilmaz, A. E. (2009). Türkçe Metinden Konuşma Sentezleme Uygulamaları İçin Bir Veri Sözlük Seti ve Yazılım Çerçevesi. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, 24 (4), 735–744.

9. Özüm, İ. Y. (1993). *A Speech Synthesis System for Turkish Language Based on the Concetanation of Phonemes Taken from Speaker*, Yüksek Lisans Tezi, Orta Doğu Teknik Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

10. Eker, B. (2002). *Turkish Text To Speech System*, Yüksek Lisans Tezi, Bilkent Üniversitesi Mühendislik ve Fen Bilimleri Enstitüsü, Ankara.

11. İnternet: MBROLA voices database. *GitHub*. URL: https://github.com/numediart/MBROLA-voices, Son Erişim Tarihi: 09.06.2023.

12. Khan, R. A. ve Chitode, J. S. (2016). Concatenative Speech Synthesis: A Review. *International Journal of Computer Applications*, 136 (3), 1–6.

13. Aşlıyan, R. ve Günel K. (2008). Türkçe metinler için hece tabanlı konuşma sentezleme sistemi. *Akademik Bilişim 2008*, 31–38.

14. Tabet, Y. ve Boughazi, M. (2011). Speech synthesis techniques. A survey. *7th*

International Workshop on Systems, Signal Processing and their Applications WoSSPA 2011, 67–70.

15. Rashad, M. Z., El-Bakry, H. M., Isma'il, I. R. ve Mastorakis, N. (2010). An overview of text-to-speech synthesis techniques. *International Conference on Intelligent Cloud Computing*, 84–89.

16. Govind, D. ve Prasanna, S. R. M. (2013). Expressive speech synthesis: A review. *International Journal of Speech Technology*, 16 (2), 237–260.

17. İlhan Omurca, S. ve ark. (2008). Metin Madenciliği ile Soru Cevaplama Sistemi. *Elektrik- Elektronik ve Bilgisayar Mühendisliği Sempozyumu*.

18. Dutoit, T. (1997). High-quality text-to-speech synthesis: An overview. *Journal Of Electrical And Electronics Engineering Australia*, 17(1), 25-36.

19. Uslu, İ. B. (2010). Metinden Konuşma Sentezleme. TMMOB Elektrik Mühendisleri Odası Ankara Şubesi Haber, 12- 16.

20. Paaß, G. ve Konya, I. (2012). *Modeling, Learning, and Processing of Text Technological Data Structures* (Birinci Baskı). Almanya: Springer Berlin, 221-247.

21. Sasirekha, D. ve Chandra, E. (2012). Text To Speech: a Simple Tutorial. *The International Journal of Soft Computing and Software Engineering*, 2 (1), 275–278.

22. Briscoe, T. (2011). Introduction to linguistics for natural language processing. *Computer Laboratory, University of Cambridge*, 4(25).

23. Kim, B., Lee, G. G., ve Lee, J. H. (2002). Morpheme-Based Grapheme to Phoneme Conversion Using Phonetic Patterns and Morphophonemic Connectivity Information. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 1 (1), 65–82.

24. Sak, H., Saraclar, M., ve Güngör, T. (2010). Morphology-based and sub-word language modeling for Turkish speech recognition. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5402-5405

25. Ward, N., ve Nakagawa, S. (2002, Eylül). Automatic user-adaptive speaking rate selection for information delivery. *7th International Conference on Spoken Language Processing (ICSLP2002- INTERSPEECH 2002).*

26. Prafianto, H., Nose, T., Chiba, Y., ve Ito, A. (2019). Improving human scoring of prosody using parametric speech synthesis. *Speech Communication*, 111, 14-21.

27. Jia, Y., Huang, D., Liu, W., Dong, Y., Yu, S. ve Wang, H. (2008). Text Normalization In Text to Speech System. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4693–4696.

28. Zhou, J., Su, X., Ylianttila, M. ve Riekki, J. (2012). Exploring pervasive service computing opportunities for pursuing successful ageing. *Grid and Pervasive Computing*

Workshops: International Workshops, S3E, HWTS, Doctoral Colloquium, Held in Conjunction with GPC 2011, 73–82.

29. Kidawara, Y., Sumita, E., Kawai, H. ve ark. (2020). *Speech-to-speech Translation* (Birinci Baskı). Singapur: Springer Singapore.

30. Güner E., (2013), *A Hybrid Statistical/Unit-Selection Text-To-Speech Synthesis System For Morphologically Rich Languages*. Yüksek Lisans Tezi, Özyeğin Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

31. Shen J., Pang R., Weiss R.J., Schuster M., Jaitly N., Yang Z., Chen Z., Zhang Y., Wang Y., RJ Skerry-Ryan, Saurous R.A., Agiomyrgiannakis Y. ve Wu Y. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4779-4783.

32. Blackman, R. B., ve Tukey, J. W. (1958). The measurement of power spectra from the point of view of communications engineering—Part I. *Bell System Technical Journal*, 37(1), 185-282.

33. Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1), 51-83.

34. Civaner, O. F. (2018), *Classification of pediatric snoring episodes using deep convolutional neural networks*. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

35. Gürler, B. S. (2014). *Türkçe Konuşma Tanıma Sistemleri İçin Bir Konuşma Veritabanı*. Yüksek Lisans Tezi, Gazi Üniversitesi Bilişim Enstitüsü, Ankara.

36. Gokay, R. ve Yalcin, H. (2019). Improving low resource Turkish speech recognition with data augmentation and TTS. *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)*, 357-360.

37. Dunaev, A. (2019). A Text-to-Speech system based on Deep Neural Networks. Lisans Tezi, Karlsruhe Institute of Technology Institute for Anthropomatics and Robotics, Karlsruhe.

38. Sel, İ., Hanbay, D., ve Karabatak, M. (2011). Beyin Bilgisayar Arayüzleri İçin Türkçe Metinden Konuşma Sentezleme Sistemi. *Fırat Üniversitesi Elektrik-Elektronik Bilgisayar Sempozyumu*, 273-276.

39. Kim, J., Kim, S., Kong, J. ve Yoon, S. (2020). Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search, *Advances in Neural Information Processing Systems*, 8067-8077.

40. Ribeiro, F., Florêncio, D., Zhang, C., ve Seltzer, M. (2011). Crowdmos: An approach for crowdsourcing mean opinion score studies. *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2416-2419.

41. Uslu, B., İlk, H. G., ve Yılmaz, A. E. (2011). Türkçe Metinden Konuşma Sentezlemede Doğallığın Artırılması İçin Öneriler. *Elektrik Mühendisleri Odası Bilimsel Dergi*, 1(2), 95-102.

42. Kohavi, R., ve Longbotham, R. (2017). Online Controlled Experiments and A/B Testing. *Encyclopedia of machine learning and data mining*, 7(8), 922-929.

43. Liu R., Sisman B., Bao F., Gao G. ve Li H., (2020). WaveTTS: Tacotron-based TTS with Joint Time-Frequency Domain Loss, Odyssey *2020 The Speaker and Language Recognition Workshop*, 245- 251.

44. Wang, D., ve Kuo, C. C. J. (2005). On improving the performance of a voice activity detector by cross-correlation analysis. *IEEE Transactions on Speech and Audio Processing*, 13(1), 109-120.

45. Hines, A., Skoglund, J., Kokaram, A. C. ve Harte, N. (2015). ViSQOL: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(13), 1- 18.

46. Cernak M. ve Rusko M. (2005). An Evaluation of Synthetic Speech Using the PESQ Measure. *Proc. European Congress on Acoustics*, 2725-2728.

47. Krishna, G., Tran, C., Han, Y., Carnahan, M., ve Tewfik, A. H. (2020). Speech synthesis using EEG. *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1235-1238.

48. Brassington, G. (2017). Mean absolute error and root mean square error: which is the better metric for assessing model performance?. *EGU General Assembly Conference Abstracts*, 3574.

49. Brandt, E., Zimmerer, F., Andreeva, B., ve Möbius, B. (2017). Mel-Cepstral Distortion of German Vowels in Different Information Density Contexts. *Interspeech 2017*, 2993-2997.

50. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46.

51. Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5), 378.

52. Krasheninnikov, V. R., Armer, A. I., Kuznetsov, V. V., ve Lebedeva, E. Y. (2011). Cross-correlation portraits of voice signals in the problem of recognizing voice commands according to patterns. *Pattern Recognition and Image Analysis*, 21, 192-194.

53. Jin, S. H., Cho, J. U., Lee, D. R., Park, J. H., Kim, H. S., Lee, C. H., Choi J. S. ve Jeon, J. W. (2007). An FPGA-based voice signal preprocessor for the real-time cross-correlation. *2007 International Conference on Control, Automation and Systems*, 793-797.

Gazili olmak ayrıcalıktır