

T.C.
ERZİNCAN BİNALİ YILDIRIM ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
YAPAY ZEKA VE ROBOTİK ANABİLİM DALI

DİKKAT MEKANİZMASI TABANLI GELİŞTİRİLMİŞ WAVE-U-NET MİMARİSİ
KULLANILARAK KONUŞMA İYİLEŞTİRME

Ali GÜNDÜZ

Danışman: Doç. Dr. İsmail AKGÜL

TEZ JÜRİ ÜYELERİ

Prof. Dr. Rıdvan SARAÇOĞLU

Doç. Dr. İsmail AKGÜL

Dr. Öğr. Üyesi Yıldız AYDIN

YÜKSEK LİSANS TEZİ

ERZİNCAN, 2026

© 2026 [Ali GÜNDÜZ]. Tüm hakları saklıdır.

Kabul ve Onay Sayfası

Doç. Dr. İsmail AKGÜL danışmanlığında, Ali GÜNDÜZ tarafından hazırlanan bu çalışma 25.06.2026 tarihinde aşağıdaki jüri tarafından Yapay Zeka ve Robotik Anabilim Dalı'nda Yüksek Lisans Tezi olarak oybirliği ile kabul edilmiştir.

Başkan : Prof. Dr. Rıdvan SARAÇOĞLU İmza:

Üye : Doç. Dr. İsmail AKGÜL İmza:

Üye : Dr. Öğr. Üyesi Yıldız AYDIN İmza:

Bu tez Enstitü Yönetim Kurulunun / / 20.... tarih ve/..... sayılı kararı ile onaylanmıştır.

Prof. Dr. Kemal Volkan ÖZDOKUR
Enstitü Müdür V.

Not: Bu tezde kullanılan özgün ve başka kaynaklardan yapılan bildirişlerin, şekil ve tabloların kaynak olarak kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere tabidir.

Bilimsel Etięe Uygunluk Sayfası

“Dikkat mekanizması tabanlı geliştirilmiş Wave-U-Net mimarisi kullanılarak konuşma iyileştirme” isimli “Yüksek Lisans” tezim tarafımca intihal tespit programı ile incelenmiştir. Buna göre tezimde bilimsel etik ihlali ve intihal olarak nitelendirilebilecek herhangi bir durum olmadığını taahhüt ederim.

Bu çalışmadaki tüm bilgilerin, akademik ve etik kurallara uygun bir biçimde elde edildiğini; aynı zamanda bu kural ve davranışların gerektirdiğı gibi, bu çalışmanın özünde olmayan tüm materyal ve sonuçları tam olarak aktardığımı ve referans gösterdiğimi beyan ederim. 25/06/2026

(İmza)

Ali GÜNDÜZ

ÖZET

DİKKAT MEKANİZMASI TABANLI GELİŞTİRİLMİŞ WAVE-U-NET MİMARİSİ KULLANILARAK KONUŞMA İYİLEŞTİRME

Ali GÜNDÜZ

Yüksek Lisans Tezi

Erzincan Binali Yıldırım Üniversitesi, Fen Bilimleri Enstitüsü,

Yapay Zeka ve Robotik Anabilim Dalı

Danışman: Doç. Dr. İsmail AKGÜL

2026,75 sayfa

Gürültülü ortamlarda konuşma sinyallerinin kalitesini ve konuşmanın anlaşılabilirliğini artırmayı amaçlayan konuşma iyileştirme, derin öğrenmeyle hızla gelişen bir araştırma alanıdır. Ancak literatürdeki çalışmalar ağırlıklı olarak İngilizce veri setleri için geliştirilmiş; Türkçe veriye odaklanan çalışmalar oldukça sınırlı kalmıştır. Bu tezde, kanal dikkati ve öz-dikkat mekanizmalarıyla güçlendirilmiş geliştirilmiş Wave-U-Net mimarisi önerilmiş ve model, zaman ile frekans alanını birlikte denetleyen çok bileşenli bir kayıp fonksiyonu (EnhancedLoss) ile eğitilmiştir. Çalışma üç aşamada yürütülmüştür: önce model bileşenlerinin katkısı Türkçe veri üzerinde ablasyon analiziyle incelenmiş; ardından model İngilizce veri seti (VCTK), Türkçe veri seti (TSCD) ve bu iki veri setinin birleşiminden oluşan karma veri seti üzerinde, eğitim/doğrulama bölmeleri üç kez rastgele yeniden oluşturularak ve sabit test kümesi üzerinde yedi nesnel metrik (PESQ, STOI, ESTOI, SI-SNR, CSIG, CBAK, COVL) üzerinden değerlendirilmiştir. Ablasyon sonuçları, dikkat modüllerinin ancak algısal kayıp fonksiyonuyla birlikte eğitildiğinde katkı sağladığını göstermiştir. Model üç koşulun tamamında ölçülebilir iyileşme üretmiş; PESQ değeri Türkçe’de 1,48’den 3,54’e, İngilizce’de 1,29’dan 2,69’a ve karma koşulda 1,36’dan 3,04’e yükselmiştir. Karma koşulda kalite metrikleri tek dilli koşullara yakın kalmıştır. Bulgular, Türkçe veri üzerinde geliştirilen modelin İngilizce ve çok dilli veriye aktarılabilirliğini desteklemekte ve tek bir modelin çok dilli konuşma iyileştirme çalışmalarında da kullanılabileceğini göstermektedir.

Anahtar Kelimeler: Konuşma iyileştirme, derin öğrenme, Wave-U-Net, dikkat mekanizması, VCTK veri seti, Türkçe veri seti

ABSTRACT

SPEECH ENHANCEMENT USING AN IMPROVED WAVE-U-NET ARCHITECTURE BASED ON ATTENTION

Ali GÜNDÜZ

Master's Thesis

Erzincan Binali Yıldırım University, Institute of Science and Technology,

Department of Artificial Intelligence and Robotics

Advisor: Assoc. Prof. Dr. İsmail AKGÜL

2026, 75 pages

Speech enhancement, which aims to improve the quality and intelligibility of speech in noisy environments, has advanced rapidly with deep learning. However, studies in the literature have predominantly been developed for English datasets, while research focusing on Turkish data has remained quite limited. In this thesis, an Improved Wave-U-Net architecture with channel attention and self-attention mechanisms is proposed, and the model is trained with a multi-component loss function (EnhancedLoss) that jointly supervises the time and frequency domains. The study was carried out in three stages: first, the contribution of the model components was examined through an ablation analysis on Turkish data; then, the model was evaluated on the English (VCTK), Turkish (TSCD), and mixed datasets using seven objective metrics, with the training/validation splits randomly regenerated three times and the test set kept fixed. The ablation results showed that the attention modules contributed only when trained together with the perceptual loss function. The model produced measurable improvements across all three conditions; PESQ increased from 1.48 to 3.54 for Turkish, 1.29 to 2.69 for English, and 1.36 to 3.04 for the mixed condition, with the mixed-condition metrics remaining close to the monolingual ones. The findings support the transferability of the Turkish-developed model to English and multilingual data and indicate that a single model can be used in multilingual speech enhancement studies.

Keywords: Speech enhancement, deep learning, Wave-U-Net, Attention mechanism, VCTK dataset, Turkish dataset

TEŞEKKÜR

Yüksek lisans çalışmam boyunca bilgi ve deneyimiyle bana yol gösteren, değerli görüşleri ve sabırlı yaklaşımıyla bu tezin her aşamasına katkı sunan danışmanım Doç. Dr. İsmail AKGÜL hocama içtenlikle teşekkür ederim.

Eğitim hayatım boyunca maddi ve manevi desteğini hiçbir zaman esirgemeyen, bugünlere gelmemde en büyük emeğe sahip olan aileme minnettarım.

Bu yoğun süreçte gösterdiği anlayış, sabır ve desteğiyle her an yanımda olan, varlığıyla bana güç veren eşime ayrıca teşekkür ederim.

Ali GÜNDÜZ

Haziran, 2026

İÇİNDEKİLER

ÖZET	i
ABSTRACT	ii
TEŞEKKÜR	iii
İÇİNDEKİLER.....	iv
TABLolar DİZİNİ.....	vii
ŞEKİLLER DİZİNİ	viii
SİMGELER VE KISALTMALAR DİZİNİ	x
1.GİRİŞ.....	1
1.1. Problemin Durumu	1
1.2. Araştırmanın Amacı	2
1.3. Araştırmanın Kapsamı.....	2
1.4. Araştırmanın Önemi	2
1.5. Sınırlılıklar.....	3
1.6. Çalışmanın Katkıları.....	3
2. KAVRAMSAL ÇERÇEVE VE İLGİLİ ÇALIŞMALAR	4
2.1. İlgili Çalışmalar	4
2.1.1. Geleneksel konuşma iyileştirme yöntemleri.....	4
2.1.2. Derin öğrenme tabanlı konuşma iyileştirme.....	5
2.1.3. Wave-U-Net ve dikkat tabanlı mimariler	6
2.1.4. Değerlendirme ölçütleri ve veri setlerinin tarihsel gelişimi	8
2.1.5. Türkçe dili üzerine konuşma iyileştirme çalışmaları.....	9
2.1.6. Literatürdeki boşluk ve bu çalışmanın konumu	10
2.2. Kuramsal Temeller	11
2.2.1. Ses sinyallerinin temelleri	11
2.2.1.1. Ses sinyalinin tanımı ve özellikleri.....	11
2.2.1.2. Zaman ve frekans alanı gösterimleri	12
2.2.1.3. Konuşma sinyalinin bileşenleri	14
2.2.2. Akustik gürültü modellemesi ve sinyal-gürültü oranı	14
2.2.2.1. Akustik gürültü türleri ve modellemesi	14
2.2.2.2. Sinyal-gürültü oranı	16
2.2.2.3. Konuşma iyileştirme yaklaşımları	16
2.2.3. Derin öğrenme bileşenleri	17

2.2.3.1. Evrişimli sinir ağları (CNN)	17
2.2.3.2. Kodlayıcı-çözücü (encoder-decoder) ve U-Net mimarileri.....	18
2.2.3.3. Wave-U-Net mimarisi	19
2.2.3.4. Dikkat mekanizmaları (Attention Mechanisms)	20
2.2.3.5. Artık bağlantılar ve yığın normalizasyonu	22
2.2.3.6. Aktivasyon fonksiyonları	23
2.2.4. Nesnel değerlendirme metrikleri	24
2.2.4.1. Perceptual Evaluation of Speech Quality (PESQ).....	24
2.2.4.2. STOI ve ESTOI	24
2.2.4.3. Scale-Invariant Signal-to-Noise Ratio(SI-SNR)	25
2.2.4.4. Kompozit ölçütler (CSIG, CBAK, COVL)	25
2.2.5. Genel kayıp fonksiyonları ve optimizasyon teknikleri.....	26
2.2.5.1. Zaman alanı kayıp fonksiyonları	26
2.2.5.2. Frekans alanı kayıp fonksiyonları.....	27
2.2.5.3. Optimizasyon teknikleri	27
3. YÖNTEM	29
3.1. Önerilen Model: Geliştirilmiş Wave-U-Net	29
3.1.1. Model mimarisine genel bakış ve kanal yapısı.....	29
3.1.2. Aşağı-Örnekleme (DownBlock) ve yukarı-örnekleme (upblock) blokları.....	32
3.1.2.1. Aşağı-Örnekleme bloğu (DownBlock).....	32
3.1.2.2. Yukarı-Örnekleme bloğu (UpBlock).....	33
3.1.3. Artık blok (ResidualBlock) tasarımı	34
3.1.4. Kanal dikkati ve öz-dikkat entegrasyonu	35
3.1.4.1. Kanal dikkati mekanizması (ChannelAttention)	35
3.1.4.2. Öz-Dikkat mekanizması (SelfAttention)	36
3.1.4.3. Darboğaz katmanı yapısı (Bottleneck)	38
3.1.5. Önerilen karma kayıp fonksiyonu (EnhancedLoss)	38
3.2. DeneySEL Düzenek ve Veri Hazırlama	40
3.2.1. İngilizce (VCTK) ve Türkçe (TSCD) veri setleri.....	41
3.2.1.1. İngilizce veri seti: VCTK Corpus.....	41
3.2.1.2. Türkçe veri seti: Turkish Speech Command Dataset (TSCD)	41
3.2.2. Gürültü entegrasyonu ve SNR seviyelerinin belirlenmesi	42
3.2.2.1. Akustik gürültü kaynakları	42
3.2.2.2. Sinyal-Gürültü oranı (SNR) seviyeleri ve modelleme	43

3.2.2.3. Sentezleme ve karıştırma dağılımı	43
3.2.3. Veri ön işleme, parçalama ve doldurma stratejileri	44
3.2.3.1. Ses dosyası okuma ve standardizasyon	44
3.2.3.2. Sabit uzunluklu parçalama ve doldurma	44
3.2.4. Çoklu rastgele bölme ve karma eğitim düzeni	45
3.2.4.1. Veri bölme oranları.....	45
3.2.4.2. Çoklu rastgele bölme protokolü	45
3.2.4.3. Karma eğitim yapılandırması	45
3.2.5. Eğitim hiperparametreleri, donanım ve yazılım altyapısı	46
3.2.5.1. Hiperparametre yapılandırması	46
3.2.5.2. Donanım ve yazılım altyapısı	46
3.2.6. Nesnel başarımlar değerlendirme protokolü.....	47
4. BULGULAR	48
4.1. Model Bileşenlerinin ve Kayıp Fonksiyonlarının Ablasyon Analizi	48
4.2. Dil Bazlı ve Çok Dilli Eğitim Performansı	49
4.2.1. İngilizce veri seti sonuçları.....	49
4.2.2. Türkçe veri seti sonuçları	52
4.2.3. Karma veri seti sonuçları.....	54
4.2.4. Genel dil performansı karşılaştırması	56
4.3. En İyi Model Ağırlıklarının Diller Arası Performans Kıyası	58
4.4. Rastgele Bölünme Kararlılığı Analizi	60
4.5. Spektrogram ve Dalga Formu Analizi.....	60
5. TARTIŞMA VE SONUÇLAR.....	64
5.1 Tartışma.....	64
5.1.1. Modelin performansının yorumlanması ve metriklerin analizi	64
5.1.2. Model bileşenlerinin ve dil faktörünün değerlendirilmesi	65
5.1.3. Çalışmanın kısıtlamaları ve sınırlılıkları	66
5.1.4. Bulguların literatür ile karşılaştırılması.....	66
5.2 Sonuçlar.....	69
5.2.1. Genel değerlendirme.....	69
5.2.2. Geleceğe yönelik öneriler.....	69
KAYNAKÇA	71

TABLÖLAR DİZİNİ

Tablo 1. Önerilen Geliştirilmiş Wave-U-Net mimarisinin kanal yapısı.....	30
Tablo 2. Orijinal Wave-U-Net ile önerilen modelin karşılaştırması	30
Tablo 3. Geliştirilmiş Wave-U-Net modelinin detaylı eğitilebilir parametre dağılımı	31
Tablo 4. Çoklu çözünürlüklü STFT kaybında FFT boyutlarının gerekçeleri.....	40
Tablo 5. İngilizce ve Türkçe veri setlerinin karşılaştırmalı özellikleri.....	42
Tablo 6. Deneylerde kullanılan SNR seviyeleri ve akustik anlamları.....	43
Tablo 7. Eğitim, doğrulama ve test setlerinin veri dağılımı	45
Tablo 8. Geliştirilmiş Wave-U-Net modeli eğitim hiperparametreleri	46
Tablo 9. Deneysel çalışmalarda kullanılan donanım altyapısı	46
Tablo 10. Kullanılan yazılım ve kütüphane detayları.....	46
Tablo 11. Model bileşenlerinin ve kayıp fonksiyonlarının birleşik ablasyon sonuçları..	48
Tablo 12. İngilizce rastgele oluşturulmuş veri setlerinin detaylı sonuçları.....	50
Tablo 13. Türkçe rastgele oluşturulmuş veri setlerinin detaylı sonuçları.....	52
Tablo 14. Karma dillerden rastgele oluşturulmuş veri setlerinin detaylı sonuçları.....	54
Tablo 15. Üç farklı veri setinin ortak genel performans ve kazanım özeti	56
Tablo 16. En iyi veri seti sonuçlarına göre diller arası karşılaştırma ve kazanımlar.....	58
Tablo 17. PESQ değerlerinin rastgele bölmeler arasındaki kararlılığı ve değişim katsayıları	60
Tablo 18. Geliştirilen modelin literatürdeki çalışmalarla bağlamsal karşılaştırması	67

ŞEKİLLER DİZİNİ

Şekil 1. Bir konuşma sinyalinin zaman alanı ve frekans alanı gösterimi	13
Şekil 2. Temiz ve gürültülü bir konuşma sinyalinin spektrogram (frekans alanı) gösterimi	15
Şekil 3. Temiz ve gürültülü bir konuşma sinyalinin dalga formu (zaman alanı) gösterimi	16
Şekil 4. CNN kullanarak yapılan konuşma iyileştirme	18
Şekil 5. U-Net mimarisi.....	19
Şekil 6. Wave-U-Net mimarisi	20
Şekil 7. Dikkat mekanizması mimarisi.....	21
Şekil 8. Kanal dikkat mekanizması mimarisi	22
Şekil 9. Artık (residual) bağlantı bloğu	23
Şekil 10. Önerilen Geliştirilmiş Wave-U-Net mimarisinin katman boyutları ve veri akışı	32
Şekil 11. DownBlock'un iç yapısı	33
Şekil 12. UpBlock'un iç yapısı	34
Şekil 13. ResidualBlock yapısı.....	35
Şekil 14.Önerilen modelde kullanılan dikkat mekanizmalarının şematik gösterimi: (a) kanal dikkati mekanizması, (b) öz-dikkat mekanizması.	37
Şekil 15. Ablasyon sonuçlarının PESQ grafiği	49
Şekil 16. İngilizce veri seti rastgele bölme bazlı PESQ değerleri	51
Şekil 17. İngilizce veri seti rastgele bölme bazlı SI-SNR değerleri	51
Şekil 18. Türkçe veri seti rastgele bölme bazlı PESQ değerleri.....	53
Şekil 19. Türkçe veri seti rastgele bölme bazlı SI-SNR değerleri.....	53
Şekil 20. Karma veri seti rastgele bölme bazlı PESQ değerleri	55
Şekil 21. Karma veri seti rastgele bölme bazlı SI-SNR değerleri	55
Şekil 22. Dil bazlı PESQ karşılaştırması (kazanç Δ).....	57
Şekil 23. Dil bazlı SI-SNR karşılaştırması (kazanç Δ).....	57
Şekil 24. En iyi model Data 3 veri seti ile diller arası PESQ karşılaştırması	59
Şekil 25. En iyi model Data 3 veri seti ile diller arası SI-SNR karşılaştırması	59
Şekil 26. İngilizce örnek temiz, gürültülü ve iyileştirilmiş sinyallerin spektrogramları .	61
Şekil 27. İngilizce örnek temiz, gürültülü ve iyileştirilmiş sinyallerin dalga formları....	61
Şekil 28. Türkçe örnek temiz, gürültülü ve iyileştirilmiş sinyallerin spektrogramları....	62

Şekil 29. Türkçe örnek temiz, gürültülü ve iyileştirilmiş sinyallerin dalga formlar 62

SİMGELER VE KISALTMALAR DİZİNİ

α	SI-SNR hesabında temiz sinyale uygulanan ölçek katsayısı
λ	EnhancedLoss bileşenlerinin ağırlık katsayıları ($\lambda_1, \lambda_2, \lambda_3$)
s	Temiz (referans) konuşma sinyali
$\hat{s}$	Model tarafından iyileştirilen (kestirilen) konuşma sinyali
y	Gürültülü giriş sinyali ($y = s + n$)
n	Toplamsal gürültü sinyali
N	STFT'de pencere/FFT uzunluğu (512, 1024, 2048)
dB	Desibel
Hz / kHz	Frekans birimi (Hertz / kilohertz)
AdamW	Adaptive Moment Estimation with Weight Decay (ağırlık sönümlü optimize edici)
ASR	Automatic Speech Recognition (Otomatik Konuşma Tanıma)
BN	Batch Normalization (Yığın Normalizasyonu)
CA	Channel Attention (Kanal Dikkati)
CBAK	Composite measure for Background distortion (arka plan gürültüsü kompozit ölçütü)
CMGAN	Conformer-based Metric GAN
CNN	Convolutional Neural Network (Evrışimli Sinir Ağı)
Conv1d	One-Dimensional Convolution (Bir Boyutlu Evrışım)
COVL	Composite measure for Overall quality (genel kalite kompozit ölçütü)
CRN	Convolutional Recurrent Network
C SIG	Composite measure for Signal distortion (sinyal bozulması kompozit ölçütü)
DCCRN	Deep Complex Convolution Recurrent Network
DEMAND	Diverse Environments Multichannel Acoustic Noise Database
DNS	Deep Noise Suppression (Challenge)
ImprovedWaveUNet	Geliştirilmiş Wave-U-Net
ESTOI	Extended Short-Time Objective Intelligibility
FFT	Fast Fourier Transform (Hızlı Fourier Dönüşümü)
GAN	Generative Adversarial Network (Çekişmeli Üretici Ağ)
GAP	Global Average Pooling (Küresel Ortalama Havuzlama)
GMP	Global Max Pooling (Küresel Maksimum Havuzlama)

IBM	Ideal Binary Mask (İdeal İkili Maske)
IRM	Ideal Ratio Mask (İdeal Oran Maskesi)
ISTFT	Inverse Short-Time Fourier Transform (Ters Kısa Zamanlı Fourier Dönüşümü)
ITU-T	International Telecommunication Union – Telecommunication Standardization Sector
L1	Ortalama Mutlak Hata tabanlı kayıp fonksiyonu
LeakyReLU	Leaky Rectified Linear Unit (sızıntılı doğrultulmuş doğrusal birim)
LLR	Log-Likelihood Ratio (Logaritmik Olabilirlik Oranı)
LSTM	Long Short-Term Memory (Uzun Kısa Süreli Bellek)
MAE	Mean Absolute Error (Ortalama Mutlak Hata)
MetricGAN	Metric Generative Adversarial Network
MLP	Multi-Layer Perceptron (Çok Katmanlı Algılayıcı)
MMSE	Minimum Mean Square Error (En Küçük Ortalama Kare Hata)
MOS	Mean Opinion Score (Ortalama Görüş Puanı)
MSE	Mean Squared Error (Ortalama Kare Hata)
PESQ	Perceptual Evaluation of Speech Quality
ReLU	Rectified Linear Unit (Doğrultulmuş Doğrusal Birim)
RNN	Recurrent Neural Network (Tekrarlayan Sinir Ağı)
SA	Self-Attention (Öz-Dikkat)
SDR	Signal-to-Distortion Ratio (Sinyal-Bozulma Oranı)
SE	Speech Enhancement (Konuşma İyileştirme)
SEGAN	Speech Enhancement Generative Adversarial Network
SI-SDR	Scale-Invariant Signal-to-Distortion Ratio
SI-SNR	Scale-Invariant Signal-to-Noise Ratio
SNR	Signal-to-Noise Ratio (Sinyal-Gürültü Oranı)
SSNR	Segmental Signal-to-Noise Ratio (Parçalı Sinyal-Gürültü Oranı)
STFT	Short-Time Fourier Transform (Kısa Zamanlı Fourier Dönüşümü)
STOI	Short-Time Objective Intelligibility (Kısa Zamanlı Nesnel Anlaşılabilirlik)
TSCD	Turkish Speech Command Dataset
TTS	Text-to-Speech (Metinden Konuşmaya)
U-Net	U-shaped Network (U biçimli kodlayıcı-çözücü ağı)
VAD	Voice Activity Detection (Ses Etkinliği Algılama)

VCTK	Voice Cloning Toolkit Corpus (CSTR VCTK konuşma veri seti)
Wave-U-Net	Dalga formu (zaman alanı) tabanlı U-Net mimarisi
WSS	Weighted Spectral Slope (Ağırlıklı Spektral Eğim)

1. GİRİŞ

İnsanlık tarihi boyunca, bireyler arasındaki bilgi aktarımı ve haberleşme gereksinimi doğrultusunda çeşitli yöntemler geliştirilmiştir. İlk çağlarda ateşin bulunmasıyla duman ve benzeri ilkel araçlarla başlayan haberleşme çalışmaları, teknolojinin gelişimiyle birlikte günümüzde telefon, telsiz ve bilgisayar gibi gelişmiş dijital sistemler aracılığıyla gerçekleştirilmektedir. Mobil iletişim altyapılarının ve teknolojilerinin hızla gelişmesi, kablosuz telefon kullanımını insanlar için öncelikli haberleşme kanalı haline getirmiştir. İletişimin mobil ve dinamik ortamlara taşınması, farklı akustik ve sinyal koşullarında birtakım teknik sorunları da beraberinde getirmiştir. Bu sorunlar arasında baz istasyonu kapsama alanı dışında kalma veya sinyal kaybına bağlı kesintilerin yanı sıra, özellikle konuşma ortamındaki çeşitli çevresel gürültüler yer almaktadır. Sinyal gücüne ve kapsama alanına dayalı altyapısal problemler, mobil baz istasyonu ağlarının genişletilmesi ve iletişim protokollerinin iyileştirilmesiyle çözülmeye çalışılmıştır. Ancak trafik, insan kalabalığı veya cihaz uğultusu gibi çeşitli çevresel gürültü kaynakları sesli iletişimi olumsuz etkilemektedir. Bu durum, insanlar açısından konuşmanın kalitesini ve anlaşılabilirliğini düşürmekle kalmayıp, otomatik konuşma tanıma (ASR) ve ses tabanlı dijital asistan sistemlerinin başarımını da doğrudan ve olumsuz yönde etkilemektedir. Konuşma iyileştirme çalışmaları, gürültülü bir ses sinyalinden temiz konuşmayı kestirerek ses kalitesini ve anlaşılabilirliğini artırmayı hedefleyen temel bir araştırma alanıdır. Geçmişte bu amaca yönelik olarak geleneksel filtreleme ve istatistiksel sinyal işleme tabanlı yöntemler kullanılırken, günümüzde yapay zeka modellerinin gelişmesi ve özellikle derin öğrenmenin yükselişiyle birlikte, gürültülü sesi doğrudan temiz sese dönüştürebilen derin öğrenme modelleri konuşma iyileştirme uygulamalarında başarıyla kullanılmaya başlanmıştır.

1.1. Problemin Durumu

Konuşma iyileştirme literatürü, derin öğrenme teknolojilerinin entegrasyonu ile birlikte son derece aktif bir araştırma alanı haline gelmiştir. Ancak mevcut çalışmalarda kullanılan yöntemler ve yapay zeka model tasarımları, ağırlıklı olarak İngilizce veri setleri (TIMIT, VCTK ve VoiceBank-DEMAND gibi) üzerinde geliştirilmekte, eğitilmekte ve değerlendirilmektedir. Türkçe dili özelinde, otomatik konuşma tanıma (ASR) ve metinden konuşma sentezleme (TTS) alanlarında güçlü akademik çalışmalar bulunmasına karşın, Türkçe ses verisine odaklanan ve bu dilin kendine has akustik özelliklerini temel alan konuşma iyileştirme araştırmaları

literatürde sınırlı sayıda çalışmaya sahiptir. Türkçe; sondan eklemeli morfolojik yapısı, ünlü uyumu kuralları ve İngilizce gibi Batı dillerinden köklü biçimde ayrılan fonetik özellikleri nedeniyle özgün bir araştırma alanı teşkil etmektedir. Bu durum, Türkçe ses sinyallerinin gürültüden arındırılması sürecinde mevcut modellerin başarımı ve diller arası aktarılabilirliği konusunda belirgin bir literatür boşluğunun varlığına işaret etmektedir.

1.2. Araştırmanın Amacı

Bu çalışmanın amacı; Türkçe ses verisini temel alan, dikkat mekanizmalarıyla güçlendirilmiş Wave-U-Net mimarisi geliştirmek ve bu modelin diğer dillere aktarılabilirliğini değerlendirmeyi amaçlamaktadır. Çalışma, bu amaç doğrultusunda üç temel araştırma sorusu etrafında kurgulanmıştır:

- Temel Wave-U-Net mimarisine entegre edilen kanal dikkati, öz-dikkat ve artık bağlantı gibi yapısal bileşenlerin model performansına olan katkısının belirlenmesi,
- Türkçe veri seti üzerinde optimize edilen en iyi model yapılandırmasının, Türkçe ve İngilizcede performansının nasıl değiştiğinin ortaya konması,
- Aynı modelin, iki farklı dilin (Türkçe ve İngilizce) birleşiminden oluşan karma eğitim senaryosunda, çok dilli (multilingual) bir ortamda nasıl bir başarı sergileyeceğinin incelenmesidir.

1.3. Araştırmanın Kapsamı

Bu çalışma; İngilizce akıcı konuşma verileri için VCTK (Voice Cloning Toolkit) veri setini, Türkçe Konuşma Komutları Veri Setini (Turkish Speech Command Dataset- TSCD) ve bu iki veri setinin birleşiminden oluşan karma dil veri setini kapsamaktadır. Geliştirilen modelin başarımı yedi farklı nesnel kalite ve anlaşılabilirlik metriği (PESQ, STOI, ESTOI, SI-SNR, CSIG, CBAK, COVL) üzerinden değerlendirilmektedir. Ayrıca önerilen model mimarisindeki bileşenlerin (kanal dikkati, öz-dikkat, artık bağlantı) tekil katkıları, Türkçe veri seti üzerinde analiz edilerek kapsam dahilinde incelenmektedir.

1.4. Araştırmanın Önemi

Bu çalışma, Türkçe konuşma iyileştirme alanındaki sınırlı literatüre katkı sunmayı hedeflemektedir ve dikkat mekanizmalı zaman alanında çalışan modellerin diller arası

genellenebilirliğine ilişkin deneysel bulgular sunmaktadır. Türkçe odaklı geliştirilen bir modelin İngilizce ve çok dilli verilere aktarılabilmesi, tek bir modelin çok dilli ses işleme sistemlerinde kullanılabilir olabileceğine işaret etmektedir.

1.5. Sınırlılıklar

Bu tez çalışmasının birtakım deneysel ve yöntemsel sınırlılıkları bulunmaktadır. İlk olarak, Türkçe deneylerinde kullanılan veri setinin izole sözcük kayıtlarından oluşması, günlük hayattaki akıcı ve sürekli konuşma yapısını tam olarak temsil edememektedir. İkinci olarak, deneyler sırasında ses dosyalarına eklenen gürültü türleri ve gürültü oranları (SNR seviyeleri) belirli çevresel senaryolarla sınırlandırılmıştır. Üçüncü olarak, önerilen model yalnızca çevrim dışı modda test edilmiş ve gerçek zamanlı kullanım için gerekli olan gecikme süresi (latency) analizleri yapılmamıştır. Son olarak, elde edilen sonuçlar yalnızca bilgisayar tabanlı nesnel metriklerle (PESQ, STOI vb.) değerlendirilmiş olup, gerçek insan dinleyicilerle yapılması gereken öznel ses kalitesi testleri (Mean Opinion Score - MOS) çalışma kapsamı dışında tutulmuştur.

1.6. Çalışmanın Katkıları

Bu tez çalışmasının literatüre sağladığı başlıca akademik katkılar şu şekilde özetlenebilir:

- Kanal dikkati, öz-dikkat ve zaman-frekans alanlarını birleştiren karma kayıp fonksiyonu (EnhancedLoss) ile güçlendirilmiş, Türkçe ses verisi kullanılarak özgün bir Wave-U-Net mimarisinin geliştirilmesi,
- Model mimarisine eklenen her bir yapısal bileşenin (artık bağlantılar, kanal dikkati, öz-dikkat) modele olan etkisinin ablasyon analiziyle deneysel olarak ölçülmesi,
- Türkçe, İngilizce ve Karma olmak üzere üç farklı dil senaryosunda, sabit veri seti kullanılarak modelin kararlılığının sistematik olarak değerlendirilmesi,
- Türkçe diline odaklanarak geliştirilen bir modelin, farklı fonetik ve morfolojik yapılarla sahip dillere (İngilizceye) aktarılabilirliğine ilişkin deneysel kanıtlar sunulmasıdır.

2. KAVRAMSAL ÇERÇEVE VE İLGİLİ ÇALIŞMALAR

Bu bölüm; konuşma iyileştirme konusundaki geçmiş araştırmaları ve bu tezin dayandığı teorik bilgileri açıklamaktadır. Bölüm kapsamında ilk olarak, geçmişten günümüze ses temizlemede kullanılan yöntemler incelenmiştir. Daha sonra, ses sinyallerinin bilgisayar ortamındaki özellikleri, gürültülerin yapısı ve yapay zeka modellerinin arkasındaki matematiksel temeller teorik olarak ele alınmıştır.

2.1. İlgili Çalışmalar

Bu bölümde, konuşma iyileştirme alanında yapılan geçmiş bilimsel çalışmalar sırasıyla ele alınmaktadır. İnceleme; ilk dönemlerde kullanılan klasik sinyal işleme yöntemlerinden başlayarak modern yapay zeka tekniklerine, oradan da bu tezde kullanılan Wave-U-Net ve dikkat mekanizmalı modellere doğru ilerlemektedir. Ayrıca, bu alanda ses kalitesini ölçmek için kullanılan bilgisayar tabanlı puanlama sistemleri (metrikler) ve test edilen ses veri setleri açıklanmaktadır. Son olarak, Türkçe ve dünyada az çalışılan diğer diller üzerindeki araştırmalar incelenerek literatürdeki eksiklikler gösterilmekte ve yapılan çalışmanın katkıları açıklanmaktadır.

2.1.1. Geleneksel konuşma iyileştirme yöntemleri

Gürültü temizleme problemi, yapay zeka ve derin öğrenme teknolojileri gelişmeden önce, matematiksel ve istatistiksel sinyal işleme yöntemleriyle çözülmeye çalışılmıştır. Bu klasik yaklaşımların en bilineni Boll (1979) tarafından ele alınan, sesteki gürültünün tahmin edilerek gürültülü sestem matematiksel olarak çıkartılmasına dayanan spektral çıkarma (spectral subtraction) yöntemidir. Bu yöntem basit ve bilgisayarı zorlamayan bir yapıya sahip olsa da temizleme işlemi sırasında seste "müzikal gürültü" (musical noise) adı verilen yapay cızırtılara ve bozulmalara neden olmaktadır. Bu sorunu azaltmak amacıyla Ephraim ve Malah (1984), daha gelişmiş bir matematiksel modele dayanan kısa zamanlı spektral genlik kestiricisini (Minimum Mean-Square Error Short-Time Spectral Amplitude - MMSE-STSA) önermiştir. İstatistiksel modellemeye dayanan bu yöntem, rahatsız edici cızırtıları belirgin biçimde bastırarak çalışan başarılı bir yaklaşım olmuştur.

Yine bu dönemin bir diğer temel aracı olan Wiener filtreleme gibi süzgeç teknikleri de yaygın olarak kullanılmıştır. Geleneksel yöntemlerin kapsamlı bir derlemesi Loizou (2013) kitabında incelenebilir. Geleneksel olarak adlandırılan bu eski yöntemlerin ortak sorunu; gürültünün zamanla değişmeyen yapıda olduğunu varsaymalarıdır. Bu nedenle, gürültü ile asıl konuşma sesinin frekanslarının birbirine karıştığı durumlarda (düşük SNR seviyelerinde) bu klasik yöntemlerin ses temizleme başarısı ciddi şekilde düşmektedir

2.1.2. Derin öğrenme tabanlı konuşma iyileştirme

Derin öğrenme teknolojilerinin gelişmesiyle birlikte konuşma iyileştirme üzerine yapılan çalışmalar büyük bir değişim geçirmiştir. Yapay zeka modelleri, ses verilerindeki karmaşık zaman ve frekans örüntülerini doğrudan kendi kendine öğrenebilir hale gelmiştir. Wang ve Chen (2018), derin öğrenme tabanlı konuşma iyileştirme yöntemlerini geniş bir derleme çalışmasında bir araya getirerek, alandaki farklı yaklaşımları sınıflandırmış ve araştırmacılar için kapsamlı bir özet sunmuştur.

Derin öğrenmenin ilk dönemlerinde geliştirilen yöntemler genellikle frekans alanında çalışmıştır. Xu vd. (2014) tarafından sunulan haritalama tabanlı yaklaşım; modelin, gürültülü ses sinyalinin özelliklerini doğrudan temiz ses sinyalinin özelliklerine eşleştirmesini (haritalamasını) temel almaktadır. Maskeleme tabanlı yaklaşımlarda ise sesin kendisine dokunmak yerine; gürültüyü sestten ayıklamak amacıyla ideal ikili maske (Ideal Binary Mask- IBM) veya ideal oran maskesi (Ideal Ratio Mask- IRM) gibi özel filtreleme maskeleri tahmin edilmeye çalışılmaktadır.

Ses sinyallerinin zamana bağlı değişimini (yani ses dalgasının akışını) takip edebilen yapay sinir ağlarından yararlanan Weninger vd. (2015), uzun-kısa süreli bellek (Long Short-Term Memory- LSTM) tabanlı modellerin, gürültüye dayanıklı konuşma tanıma sistemlerinin başarısını nasıl artırdığını göstermiştir. Park ve Lee (2016) ise evrişim katmanlarını kullanan tam evrişimli (fully convolutional) bir ağ mimarisi önermiştir. Benzer şekilde Tan ve Wang (2018), evrişimli ve yinelemeli katmanları bir araya getiren karmaşık bir yapı (Convolutional Recurrent Network- CRN) tasarlayarak sesin gerçek zamanlı olarak temizlenmesini hedeflemiştir. Erken dönem frekans alanı yöntemlerinin önemli bir kısmı genlik spektrogramına odaklanmış, faz bilgisini ise çoğunlukla sınırlı biçimde ele almıştır.

Ses dalgalarının faz bilgisini kayıpsız koruma ve sesi başka bir formata (örneğin spektrograma) dönüştürme ihtiyacını ortadan kaldırma arayışı, doğrudan ham dalga formu (ses dalgası) üzerinde çalışan zaman alanı modellerini öne çıkarmıştır. Pascual vd. (2017), üretici çekişmeli ağlar (Generative Adversarial Networks- GAN) mimarisine dayanan SEGAN modelini önererek, ham ses dalgası üzerinde uçtan uca ses temizleme işlemini ilk gerçekleştiren araştırmacılardan biri olmuştur. Luo ve Mesgarani (2019) tarafından geliştirilen Conv-TasNet, zaman alanı üzerinde maskeleye yöntemini kullanarak ses kaynaklarını birbirinden ayırmada başarılı sonuçlar elde etmiştir. Benzer şekilde Défossez vd. (2020), doğrudan ses dalgası üzerinde çalışan ve gerçek zamanlı ses temizleme yapabilen DEMUCS mimarisini sunmuştur. Yapay zekayı doğrudan kalite puanlarını (PESQ gibi) artıracak şekilde eğitmeyi amaçlayan Fu vd. (2019) ise, ses kalitesi ölçüm metriklerini hedef alan MetricGAN yaklaşımını önermiştir.

Son dönemdeki bilimsel çalışmalar, sesin faz bilgisini matematiksel olarak doğrudan hesaplayabilen karmaşık değerli (complex-valued) yapay sinir ağlarına odaklanmıştır. Hu vd. (2020) tarafından önerilen derin karmaşık evrişimli yinelemeli ağ (Deep Complex Convolutional Recurrent Network- DCCRN) mimarisi, faz bilgisini hesaba katan başarılı bir gürültü temizleme performansı sergilemiş ve katıldığı INTERSPEECH 2020 yarışmasında öne çıkmıştır. Bunun yanı sıra, dikkat mekanizmaları ve Transformer tabanlı modeller de ses alanında hızla yaygınlaşmıştır. Örneğin, Kim ve Seo (2021) zaman alanında çalışan ve ses dalgalarını işleyen SE-Conformer modelini, Subakan vd. (2021) ise sesleri birbirinden ayırma amacıyla geliştirdikleri SepFormer modelini tanıtmıştır. Daha yakın bir tarihte Richter vd. (2023), difüzyon tabanlı üretici modelleri kullanarak hem gürültü temizleme hem de sesin yankısını giderme işlemlerini bir arada ele almıştır. Alandaki bu hızlı gelişim ve yönelimler, literatürdeki çeşitli özet ve derleme çalışmalarında da ayrıntılı bir biçimde ele alınmaktadır (Pirklbauer vd., 2023; Xie ve Tan, 2025).

2.1.3. Wave-U-Net ve dikkat tabanlı mimariler

Bu çalışmanın temelini oluşturan mimari temel olarak, ilk olarak biyomedikal görüntü bölütleme görevleri için önerilen U-Net mimarisine dayanmaktadır (Ronneberger vd., 2015). U-Net'in simetrik kodlayıcı-çözücü (encoder-decoder) yapısı ve bu iki bölümü birbirine bağlayan atlama bağlantıları (skip connections), alt-örnekleme sırasında kaybolan ince detayların çözücü katmanlarda korunmasını sağlayarak farklı alanlardaki birçok modele ilham kaynağı olmuştur. Stoller vd. (2018), bu başarılı mimariyi zaman alanındaki tek boyutlu ses

sinyallerine uyarlayan Wave-U-Net modelini önermiştir. Geleneksel U-Net modelinde iki boyutlu (2B) görüntüler üzerinde çalışan evrişim ve havuzlama katmanları, Wave-U-Net modelinde yerini tek boyutlu (1B) evrişimlere bırakmıştır. Böylece model; kısa zamanlı Fourier dönüşümü (STFT) gibi zaman-frekans alanı dönüşümlerine ve dolayısıyla ara temsil veya faz tahmini hesaplamalarına ihtiyaç duymadan, ham dalga formu (raw waveform) üzerinde uçtan uca (end-to-end) ses kaynağı ayrıştırma işlemini gerçekleştirebilmektedir. Macartney ve Weyde (2018) ise bu mimariyi doğrudan konuşma iyileştirme görevine uyarlayarak Wave-U-Net'in bu görevdeki başarımını ve kararlılığını artırmıştır.

Zaman alanında çalışan bu modellerin gürültü temizleme performansını daha da geliştirmek amacıyla dikkat (attention) mekanizmalarının mimariye entegre edilmesi, bu tezin odaklandığı temel araştırma yönlerinden birini oluşturmaktadır. Giri vd. (2019), Wave-U-Net mimarisine dikkat mekanizmasını dahil ederek VCTK veri setinde yürüttükleri deneylerde nesnel konuşma kalitesi ölçütlerinde belirgin iyileşmeler elde etmiştir. Araştırmacılar ayrıca, modelin son katmanında üretilen dikkat maskesinin yumuşak bir ses etkinliği saptayıcısı (Voice Activity Detector - VAD) gibi davranarak konuşma olan ve olmayan bölümleri başarıyla ayırt edebildiğini göstermiştir. Benzer şekilde Xu vd. (2023), kanal dikkati ile zaman-frekans dikkatini birleştiren bir U-Net varyasyonu önermiş ve tek kanallı konuşma iyileştirmede kayda değer bir başarımlı artışı raporlamıştır. Söz konusu dikkat modüllerinin kuramsal temeli; özellik haritalarındaki kanallar arasında adaptif ağırlıklandırma yapan sıkıştır ve uyar (squeeze-and-excitation) bloklarına (Hu vd., 2018) ve sinyal içindeki uzun menzilli zamansal bağımlılıkları modelleyen öz-dikkat (self-attention) mekanizmalarına (Vaswani vd., 2017) dayanmaktadır.

Bu tez çalışması, yukarıda özetlenen literatürü temel almakla birlikte, mevcut yaklaşımlardan üç temel noktada ayrılarak özgün bir konum elde etmektedir:

1. Dil Odağı ve Aktarılabilirlik: Literatürdeki öncü çalışma olan Giri vd. (2019) yalnızca İngilizce konuşma verileriyle sınırlı kalırken, bu çalışmada odak doğrudan Türkçe konuşma sinyalleri üzerine kurulmuş ve modelin diller arası (Türkçe ve İngilizce) aktarılabilirlik özellikleri kapsamlı bir şekilde incelenmiştir.
2. Hibrit Dikkat Mekanizması: Tek bir dikkat türüne odaklanmak yerine, özellik haritasının kanal boyutundaki önemli bilgileri öne çıkaran kanal dikkati ile zamansal boyuttaki uzun menzilli ilişkileri yakalayan öz-dikkat mekanizmaları hibrit olarak kullanılmıştır.
3. Algısal Odaklı Çok Bileşenli Kayıp Fonksiyonu: Modelin eğitimi için standart zaman alanı kayıpları yerine; zaman alanı hatasını, ölçekten bağımsız sinyal-gürültü oranını (SI-SNR)

ve çok çözünürlüklü spektral (STFT) büyüklük ile faz terimlerini birleştiren algısal odaklı, çok bileşenli özgün bir kayıp fonksiyonu benimsenmiştir.

2.1.4. Değerlendirme ölçütleri ve veri setlerinin tarihsel gelişimi

Konuşma iyileştirme sistemlerinin başarısını ölçmek ve modelleri objektif bir şekilde karşılaştırmak amacıyla tarihsel süreçte standartlaşmış bir değerlendirme ölçüt kümesi ve test veri setleri geliştirilmiştir. Erken dönem çalışmalarda sıklıkla kullanılan geleneksel Sinyal-Gürültü Oranı (Signal-to-Noise Ratio-SNR) ve Segmental SNR (SSNR) gibi matematiksel ölçütlerin, insan işitme sisteminin algısal kalitesi ve konuşma anlaşılabilirliği ile düşük korelasyon gösterdiği belirlenmiştir. Ayrıca bu metriklerin sinyal genliğindeki ölçek değişimlerine karşı aşırı duyarlı olması, araştırmacıları insan algısını modelleyen daha gelişmiş nesnel ölçütler aramaya yöneltmiştir.

Bu arayışın ilk önemli çıktılarından biri, algısal konuşma kalitesini ölçmek amacıyla ITU-T P.862 standardında tanımlanan Perceptual Evaluation of Speech Quality(PESQ) metriği olmuştur (Rix vd., 2001). PESQ, telefon ağları ve kodeklerdeki bozulmaları insan algısına benzer şekilde modelleyerek -0,5 ile 4,5 arasında bir skor üretmektedir. PESQ'nun gürültü temizleme kalitesini genel hatlarıyla ölçmesine karşın, sinyal bozulması ve arka plan gürültüsünü ayrı ayrı değerlendirememesi üzerine Hu ve Loizou (2008), insan dinleme testlerindeki öznel kalite puanlarıyla (Mean Opinion Score - MOS) yüksek korelasyon gösteren kompozit ölçütleri önermiştir. Bu ölçütler; konuşma sinyalindeki bozulmayı hedefleyen Composite Signal Distortion(CSIG), arka plan gürültüsünün rahatsız ediciliğini ölçen Composite Background Noise Intrusiveness(CBAK) ve genel kaliteyi temsil eden Composite Overall Quality(COVL) değerlerinden oluşmaktadır.

Konuşmanın sadece kalitesini değil, aynı zamanda anlaşılabilirliğini de nesnel olarak ölçebilmek amacıyla Taal vd. (2011) tarafından kısa zamanlı nesnel anlaşılabilirlik ölçütü Short-Time Objective Intelligibility(STOI) geliştirilmiştir. Temiz ve gürültülü sinyallerin kısa zamanlı temporal zarfları arasındaki korelasyonu temel alan STOI, özellikle durağan olmayan gürültü koşullarında yüksek doğruluk sunmuştur. Jensen ve Taal (2016) ise zaman-frekans maskelemesinin ve değişken gürültülerin zarf korelasyonları üzerindeki etkisini normalize ederek düşük SNR koşullarında daha kararlı sonuçlar veren genişletilmiş Extended STOI(ESTOI) metriğini literatüre kazandırmıştır. Derin öğrenme tabanlı zaman alanı

modellerinin yaygınlaşmasıyla birlikte, sinyal genliğindeki ölçek değişimlerinden etkilenmeyen ve model eğitiminde kayıp fonksiyonu olarak doğrudan optimize edilebilen Scale-Invariant Signal-to-Noise Ratio (SI-SNR) metriği de standart bir değerlendirme aracı olarak kabul görmüştür (Le Roux vd., 2019).

Modellerin eğitildiği ve test edildiği veri setleri de benzer bir tarihsel gelişim göstermiştir. Erken dönem modelleri kısıtlı ve yapay olarak oluşturulmuş veri setlerinde eğitilirken, zamanla çok konuşmacılı ve yüksek kayıt kalitesine sahip İngilizce veri seti Voice Cloning Toolkit (VCTK, Yamagishi vd., 2019) alanın temel açık kaynaklı verilerinden biri haline gelmiştir. Çevresel gürültülerin çeşitliliğini yansıtmak amacıyla geliştirilen Diverse Environments Multi-channel Acoustic Noise Database (DEMAND) gürültü veri tabanının (Thiemann vd., 2013) VCTK konuşma kayıtlarıyla belirli SNR seviyelerinde birleştirilmesiyle oluşturulan VoiceBank-DEMAND veri seti (Botinhao vd., 2016), konuşma iyileştirme modellerinin karşılaştırmalı performans değerlendirmesinde fiili standart haline gelmiştir. Yakın dönemde ise, derin öğrenme modellerinin büyük veri gereksinimini karşılamak ve gerçek dünya senaryolarını daha iyi modellemek amacıyla, çok daha geniş konuşmacı ve gürültü havuzuna sahip INTERSPEECH Derin Gürültü Bastırma (Deep Noise Suppression-DNS) yarışması veri setleri (Reddy vd., 2020) geliştirilerek literatürün kullanımına sunulmuştur.

2.1.5. Türkçe dili üzerine konuşma iyileştirme çalışmaları

Konuşma iyileştirme alanındaki bilimsel çalışmaların büyük bir bölümü, üzerinde çok fazla araştırma yapılan ve çok miktarda ses verisine sahip İngilizce gibi zengin kaynaklı diller üzerine kuruludur. Büyük ölçekli ve çok dilli ses veri havuzları bu eşitsizliği gidermeyi hedeflese de (Ardila vd., 2020), Türkçe gibi görece daha az kaynağa sahip diller için yapay zeka eğitiminde kullanılan etiketli ses verisi miktarı ve doğrudan bu göreve yönelik eğitilmiş modeller henüz yeterli düzeyde değildir. Türkçe özelinde yapılan çalışmalar incelendiğinde, araştırmaların çok büyük bir kısmının ses kayıtlarını metne çeviren otomatik konuşma tanıma (ASR) ve metinleri yapay sesle okuyan konuşma sentezi (TTS) teknolojilerinde yoğunlaştığı görülmektedir. Ren vd. (2023) ses dalgalarını kelimelere eşleştiren hibrit CTC/Dikkat (Connectionist Temporal Classification / Attention) mimarileri ile gelişmiş dil modelleri, Türkçede konuşma tanımaya yönelik çalışma yapmıştır. Benzer şekilde, Türkçe için büyük ölçekli ilk ses veri tabanları oluşturulmuş (Salor vd., 2007) ve Türkçeyle akraba olan Kazakça gibi diğer Türki diller için de konuşma sentezine yönelik veri setleri literatüre kazandırılmıştır (Abilbekov vd., 2024).

ASR ve TTS alanındaki bu gelişmelere karşılık, Türkçede seslerin gürültüden arındırılmasını ve konuşma kalitesinin artırılmasını amaçlayan konuşma iyileştirme çalışmalarının sayısı oldukça azdır. Bu alandaki literatür, genellikle sınırlı kapsamdaki çalışmalardan oluşmaktadır. Bu tezin deneysel aşamasındaki Türkçe deneylerinde, günlük konuşma komutlarını içeren ve izole sözcük kayıtlarından oluşan Türkçe Konuşma Komutları Veri Seti kullanılmıştır (Kurtkaya, 2025). Sondan eklemeli kelime yapısı (kelimelerin sonuna gelen eklerle uzaması), ünlü uyumu kuralları ve İngilizceden tamamen farklı olan fonolojik yapısıyla Türkçe, gürültü temizleme modellerinin başarısını test etmek için oldukça özgün bir alan sunmaktadır.

Türkçe konuşma iyileştirme üzerine yapılan sınırlı çalışmalar arasında, Ayata ve Horasan (2024) Conformer tabanlı özel bir üretici yapay zeka modeli olan CMGAN ile difüzyon (diffusion) tabanlı iki farklı konuşma iyileştirme modelinin performansını hem Türkçe hem de İngilizce ses kayıtları üzerinde karşılaştırmıştır. Elde edilen bulgular, CMGAN modelinin her iki dilde de gürültü temizlemede daha başarılı olduğunu göstermiştir. Ancak dikkat çekici bir şekilde, Türkçe veri setinde ulaşılan en yüksek ses kalitesi skoru (PESQ 2,66), İngilizce veri setindeki skordan (PESQ 3,39) gerisinde kalmıştır. Araştırmacılar, bu farkın dillerin yapısından ziyade veri setlerinin kayıt ortamlarındaki ve kullanılan mikrofonlardaki farklılıklardan kaynaklandığını belirtmiştir. Bu nedenle, diller arasında yapılacak doğrudan bir sayısal karşılaştırmanın yanıltıcı olabileceğini ve her dilin kendi veri koşulları içinde değerlendirilmesi gerektiğini önemle vurgulamışlardır.

2.1.6. Literatürdeki boşluk ve bu çalışmanın konumu

Yukarıda sunulan literatür incelemesi, konuşma iyileştirme alanında iki temel gerçeği ortaya koymaktadır. Birincisi, konuşma iyileştirme literatürünün son derece zengin ve hızla gelişen bir yapıda olmasına karşın, geliştirilen yöntemlerin büyük bir çoğunluğunun İngilizce veri seti örnekleri üzerinde tasarlanması ve değerlendirilmesidir. İkincisi ise, Türkçe için son yıllarda otomatik konuşma tanıma (ASR) ve konuşma sentezi (TTS) çalışmaları geliştirilmiş olmasına rağmen, doğrudan Türkçe ses verilerine odaklanan konuşma iyileştirme ve gürültü temizleme çalışmalarının çok az olmasıdır. Bu iki durumun kesişim noktası, literatürde doldurulması gereken önemli bir bilimsel boşluğa işaret etmektedir.

Bu tez çalışması, söz konusu boşluğu doldurmayı ve Türkçe konuşma iyileştirme alanındaki çalışmalara katkı sunmayı amaçlamaktadır. Çalışmanın odağı Türkçe üzerine kurulmuş, kanal dikkati ve öz-dikkat mekanizmalarıyla güçlendirilmiş özgün bir Wave-U-Net mimarisi geliştirilmiştir. Deneysel süreçte, en iyi model yapılandırması Türkçe konuşma verileri üzerinde belirlendikten sonra, modelin İngilizce ve Karma veri setlerindeki performansı incelenerek diller arası aktarılabilirlik başarısı ölçülmüştür. Böylece bu çalışma, hem Türkçe konuşma iyileştirmeye yönelik kısıtlı literatüre özgün bir katkı sunmakta hem de dikkat tabanlı zaman alanı modellerinin farklı diller arasındaki genellenebilirliğine dair önemli bulgular sunmaktadır.

Bununla birlikte, Ayata ve Horasan (2024) tarafından dile getirilen; farklı kayıt koşullarına sahip veri setleri arasında yapılacak doğrudan sayısal karşılaştırmanın yanıltıcı olabileceği uyarısı bu çalışmada da benimsenmiş ve diller arası elde edilen bulgular doğrudan bir üstünlük yarışından ziyade, modelin diller arası aktarılabilirlik potansiyelinin bir göstergesi olarak yorumlanmıştır.

2.2. Kuramsal Temeller

Konuşma iyileştirme çalışmaları; ses sinyalinin fiziksel ve matematiksel özellikleri, gürültü modellemesi, derin öğrenme mimari bileşenleri ve optimizasyon kriterleri üzerine inşa edilmektedir. Bu bölümde, tez çalışmasında önerilen modelin ve deneysel yöntemin üzerine kurulduğu kuramsal altyapı detaylandırılmaktadır.

2.2.1. Ses sinyallerinin temelleri

2.2.1.1. Ses sinyalinin tanımı ve özellikleri

Ses, akustik bir ortamda yayılan basınç dalgalanmalarının insan işitme sistemi tarafından algılanmasıyla oluşan fiziksel bir olgudur. İnsan sesinin fiziksel yansıması olan konuşma sinyali ise, akciğerlerden gelen havanın ses tellerini titreştirmesi ve ardından insan ses yolu tarafından şekillendirilmesiyle üretilen, durağan olmayan ve zamanla sürekli değişen karmaşık bir akustik sinyaldir (Loizou, 2013). Bir konuşma sinyalinin temel akustik özellikleri arasında; ses tellerinin saniyedeki temel titreşim sıklığını temsil eden temel frekans (F_0), ses yolunun

rezonans kanallarını gösteren formant frekansları ve sesin genel karakterini belirleyen spektral zarf yer almaktadır.

Fiziksel ortamda sürekli ve analog yapıda olan sürekli zamanlı bir ses sinyali $x(t)$, bilgisayar ortamında dijital olarak işlenebilmek amacıyla analog-dijital dönüştürücü (Analog-to-Digital Converter- ADC) aracılığıyla belirli bir örnekleme frekansında f_s sayısal formata dönüştürülür (Oppenheim ve Schafer, 2013). Nyquist-Shannon örnekleme teoremine göre, analog bir sinyalin sayısal forma kayıpsız dönüştürülebilmesi ve spektral örtüşme bozulmasının önlenmesi için örnekleme frekansının, sinyaldeki en yüksek frekans bileşeninin en az iki katı olması gerekmektedir $f_s \geq 2f_{max}$. Konuşma iyileştirme uygulamalarında genellikle 8 kHz (dar bant- narrowband) ve 16 kHz (geniş bant- wideband) örnekleme frekansları tercih edilmektedir (Rabiner ve Schafer, 2011). Bu tez çalışmasında, konuşma kalitesinin ve anlaşılabilirliğinin daha yüksek frekans detaylarını koruyabilmesi amacıyla tüm ses kayıtlarında 16 kHz örnekleme frekansı benimsenmiştir.

2.2.1.2. Zaman ve frekans alanı gösterimleri

Sayısal ses sinyalleri, zaman eksenindeki genlik değişimlerini gösteren dalga formu (waveform) olarak temsil edilmektedir. Ancak ses sinyallerinin zaman içindeki frekans dağılımlarını ve spektral değişimlerini analiz edebilmek için zaman-frekans alanı gösterimleri çok daha fazla bilgi sunar (Oppenheim ve Schafer, 2013). Bu amaçla başvurulacak temel yöntemler ve dönüşümler aşağıda açıklanmaktadır:

Kısa Zamanlı Fourier Dönüşümü (Short-Time Fourier Transform- STFT): Zamanla değişen frekans bileşenlerini incelemek amacıyla, sinyali örtüşen kısa zaman çerçevelerine (frames) bölerek her çerçevenin Hızlı Fourier Dönüşümünü (Fast Fourier Transform- FFT) hesaplayan bir yöntemdir (Rabiner ve Schafer, 2011). Sayısal bir konuşma sinyali için STFT ifadesi Denklem (1) ile tanımlanmaktadır:

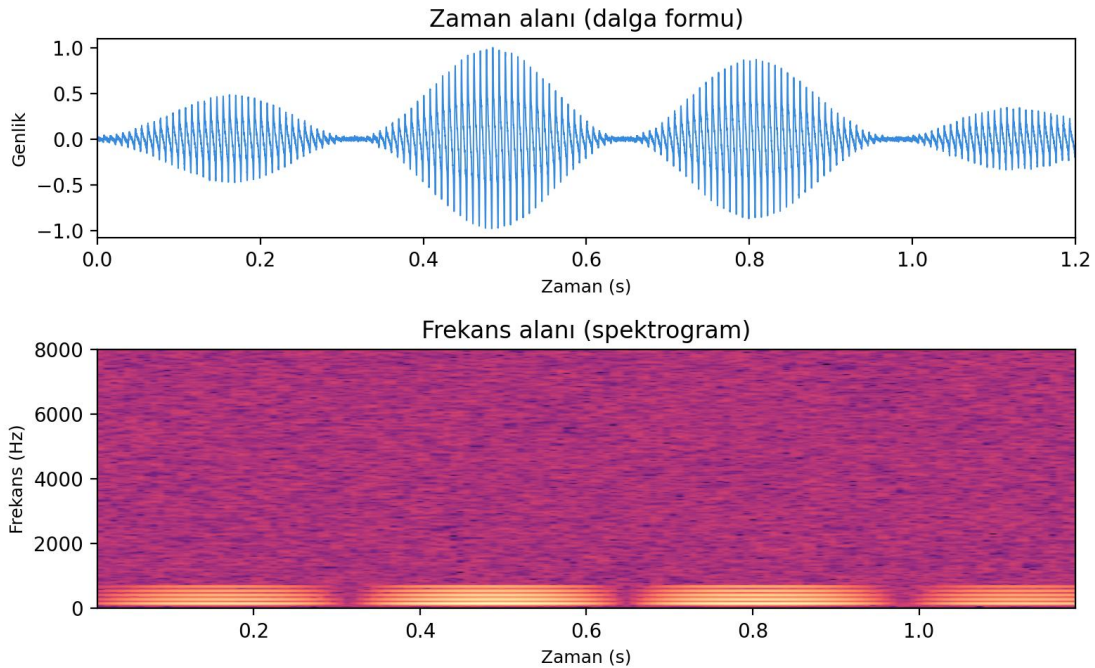
$$X(m, k) = \sum_{n=-\infty}^{\infty} x[n] \cdot w[n - mH] \cdot e^{-j\frac{2\pi}{N}kn} \quad (1)$$

Burada $w[n]$ analizde kullanılan pencere fonksiyonunu (örneğin Hamming veya Hanning penceresi), H çerçeveler arasındaki kaydırma adımını (hop size), N FFT boyutunu (frekans kanallarının sayısını), m zaman çerçeve indeksini ve k ise frekans indeksini temsil etmektedir.

Spektrogram: STFT çıktısının mutlak genlik değerinin karesi alınarak $(|X(m, k)|^2)$ elde edilen, sesin enerjisinin zaman ve frekans eksenindeki dağılımını gösteren iki boyutlu bir zaman-frekans görselleştirmesidir. Konuşma sinyallerindeki formant geçişlerini, harmonikleri ve gürültü profillerini görselleştirmek amacıyla yaygın olarak kullanılmaktadır.

Mel Ölçeği (Mel Scale): İnsan işitme sisteminin frekans algılama yapısını modelleyen logaritmik bir frekans ölçeğidir. İnsan kulağı, düşük frekanslardaki (örneğin 100 Hz ile 500 Hz arası) küçük frekans değişimlerine karşı son derece hassasken, yüksek frekanslardaki (örneğin 5 kHz ile 10 kHz arası) değişimleri ayırt etmede daha az duyarlıdır. Mel ölçeği, bu doğrusal olmayan algıyı taklit ederek düşük frekanslarda yüksek çözünürlük, yüksek frekanslarda ise daha düşük çözünürlük sağlar (Rabiner ve Schafer, 2011). Hertz cinsinden verilen bir frekans değerinin (f) Mel ölçeğindeki karşılığı (m) Denklem (2) ile hesaplanmaktadır:

$$m = 2595 \cdot \log_{10} \left(1 + \frac{f}{700} \right) \quad (2)$$



Şekil 1. Bir konuşma sinyalinin zaman alanı ve frekans alanı gösterimi

Şekil 1'de bir konuşma sinyalinin zaman alanındaki dalga formu ile frekans alanındaki gösterimi bir arada verilerek iki temsil biçimi gösterilmektedir.

2.2.1.3. Konuşma sinyalinin bileşenleri

Akustik açıdan konuşma sinyali, üretim mekanizmalarına göre temel olarak iki ana kategoriye ayrılmaktadır (Loizou, 2013):

- **Sesli Bölümler:** Akciğerlerden gelen hava akımının ses tellerini belirli bir frekansta titreştirmesiyle üretilen, periyodik veya yarı-periyodik dalga yapısına sahip seslerdir. Bu sesler, konuşmadaki formant frekanslarını ve temel perde frekansını (F_0) belirgin biçimde taşır. Türkçedeki tüm ünlü harfler (a, e, ı, i, o, ö, u, ü) ve bazı yumuşak ünsüzler bu gruptadır.
- **Sessiz Bölümler:** Ses tellerinin titreşmediği, havanın ses yolundaki dar geçitlerden geçerken oluşturduğu türbülans ile üretilen, gürültü benzeri (aperiyodik) seslerdir. Spektral enerjileri genellikle yüksek frekans bölgelerinde yoğunlaşmıştır. Türkçedeki sert ünsüzler (f, s, t, ş, ç, p, k, h) bu kategoriye örnek oluşturmaktadır.

2.2.2. Akustik gürültü modellemesi ve sinyal-gürültü oranı

2.2.2.1. Akustik gürültü türleri ve modellemesi

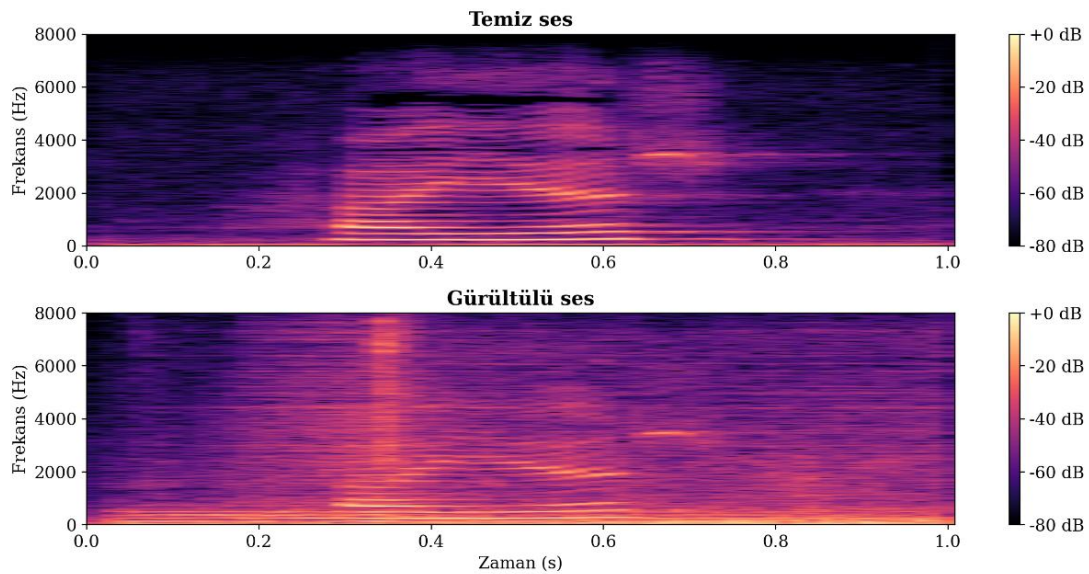
Gerçek dünya ortamlarında üretilen ve iletilen konuşma sinyalleri, çevredeki çeşitli istenmeyen akustik kaynaklardan gelen gürültülerin etkisiyle bozulmaya uğramaktadır. Konuşma iyileştirme algoritmalarının tasarımı ve başarımı, bozucu gürültü kaynağının spektral ve zamansal özelliklerine doğrudan bağlıdır. Literatürde gürültüler, akustik karakterlerine göre temel olarak şu şekilde sınıflandırılmaktadır (Loizou, 2013):

- **Eklemeli (Additive) Gürültü:** Konuşma sinyali ile gürültü kaynağının akustik olarak birbirini etkilemeden, doğrusal bir şekilde toplandığını varsayan modeldir. Gürültülü konuşma sinyali $y(t)$, temiz konuşma sinyali $x(t)$ ve gürültü sinyali $n(t)$ bileşenlerinin toplamı olarak Denklem (3) ile modellenmektedir:

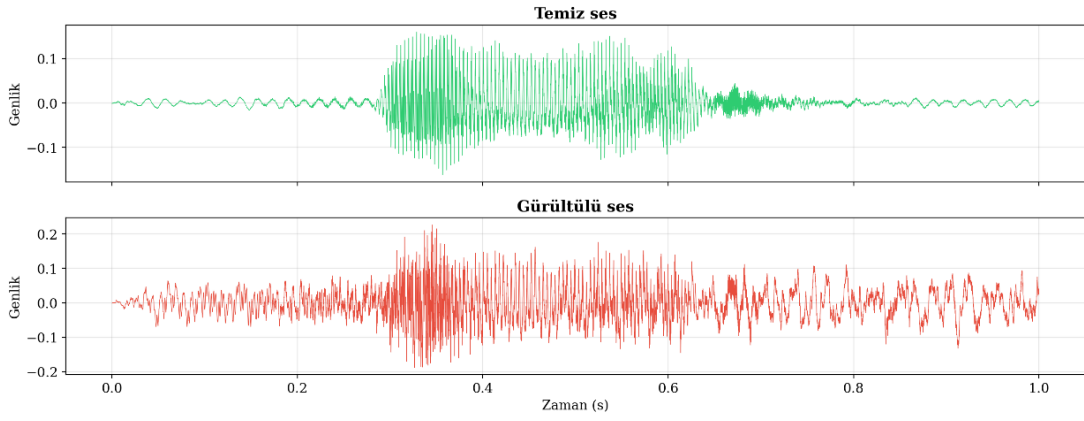
$$y(t) = x(t) + n(t) \quad (3)$$

Bu doğrusal toplamsal ilişki, konuşma iyileştirme alanındaki algoritmaların büyük bir çoğunluğunda temel varsayım olarak kabul edilmektedir ve bu tez çalışmasında da benimsenmiştir.

- **Çevresel ve Durağan Gürültüler:** Zamanla spektral özellikleri çok fazla değişmeyen (durağan - stationary), fan, klima, motor veya sürekli rüzgar gürültüsü gibi kaynaklardır. Bu tür gürültülerin spektral profilleri görece kolay tahmin edilebildiğinden, geleneksel yöntemlerle temizlenmesi daha kolaydır.
- **Kalabalık-Uğultu (Babble) Gürültüsü:** Çok sayıda insanın aynı anda konuştuğu kalabalık ortamlardaki (kafe, ofis, vb.) uğultudan kaynaklanan gürültü türüdür. Spektral özellikleri temiz konuşma sinyaliyle örtüştüğü ve zamanla sürekli değişen (durağan olmayan - non - stationary) bir yapı sergilediği için temizlenmesi en güç gürültü türlerinden biridir.
- **Beyaz Gürültü (White Noise):** Tüm frekans bandı boyunca eşit güç spektral yoğunluğuna sahip, istatistiksel olarak rastgele ve durağan bir gürültüdür. Teorik analizlerde ve modellerin sınırlarını test etmede referans gürültü olarak kullanılır.
- **Renkli Gürültüler:** Spektral yoğunluğu frekansla değişen pembe (pink) veya kahverengi (brownian) gürültü gibi, düşük frekanslarda daha yüksek enerjiye sahip gürültü türleridir.



Şekil 2. Temiz ve gürültülü bir konuşma sinyalinin spektrogram (frekans alanı) gösterimi



Şekil 3. Temiz ve gürültülü bir konuşma sinyalinin dalga formu (zaman alanı) gösterimi

Şekil 2'de temiz ve gürültülü bir sinyalin spektrogramları karşılaştırılarak gürültünün frekans içeriği üzerindeki etkisi gösterilmektedir. Şekil 3'te aynı sinyalin temiz ve gürültülü dalga formları verilerek gürültünün zaman alanındaki etkisi görülmektedir.

2.2.2.2. Sinyal-gürültü oranı

Bozulmaya uğramış gürültülü bir ses sinyalindeki gürültü seviyesini sayısal olarak ifade etmek ve konuşma kalitesini derecelendirmek amacıyla kullanılan temel fiziksel ölçüt Sinyal-Gürültü Oranı (Signal-to-Noise Ratio - SNR) değeridir. Desibel (dB) cinsinden ifade edilen SNR, temiz konuşma sinyalinin gücünün (P_s), gürültü sinyalinin gücüne (P_n) oranının logaritmik karşılığıdır ve Denklem (4) ile hesaplanmaktadır:

$$\text{SNR} = 10 \cdot \log_{10} \left(\frac{P_s}{P_n} \right) = 10 \cdot \log_{10} \left(\frac{\sum_t x^2(t)}{\sum_t n^2(t)} \right) \quad (4)$$

Burada $x(t)$ temiz konuşmayı, $n(t)$ ise gürültü sinyalini göstermektedir. Yüksek SNR değerleri gürültünün az olduğu temiz sinyalleri, düşük veya negatif SNR değerleri ise gürültünün konuşma enerjisinden daha yüksek olduğu zorlu akustik ortamları ifade etmektedir.

2.2.2.3. Konuşma iyileştirme yaklaşımları

Gürültülü sinyalden ($y(t)$) temiz konuşmayı ($x(t)$) elde etmeyi amaçlayan konuşma iyileştirme yaklaşımları tarihsel gelişimlerine ve matematiksel altyapılarına göre iki grupta incelenmektedir:

- a. **Geleneksel sinyal işleme yöntemleri:** Gürültünün istatistiksel özelliklerinin yavaş değiştiği varsayımına dayanan, spektral çıkarma (Boll, 1979), Wiener filtreleme veya minimum ortalama karesel hata (MMSE) tahmincilerine (Ephraim ve Malah, 1984) dayanan yöntemlerdir. Hesaplama maliyetleri düşüktür ancak durağan olmayan gürültülerde yetersiz kalırlar ve müzikal gürültü (musical noise) adı verilen yapay akustik bozulmalara yol açarlar.
- b. **Derin öğrenme tabanlı yöntemler:** Gürültülü sinyal ile temiz sinyal arasındaki doğrusal olmayan karmaşık ilişkileri veri üzerinden öğrenen yapay sinir ağı modelleridir. Frekans alanında çalışan maskeleye ve haritalama tabanlı modeller ile doğrudan ses dalgası üzerinde çalışan zaman alanı (uçtan uca) modellerini kapsar. Bu modeller, durağan olmayan gürültülerin temizlenmesinde geleneksel yöntemlere göre üstünlük sağlamaktadır.

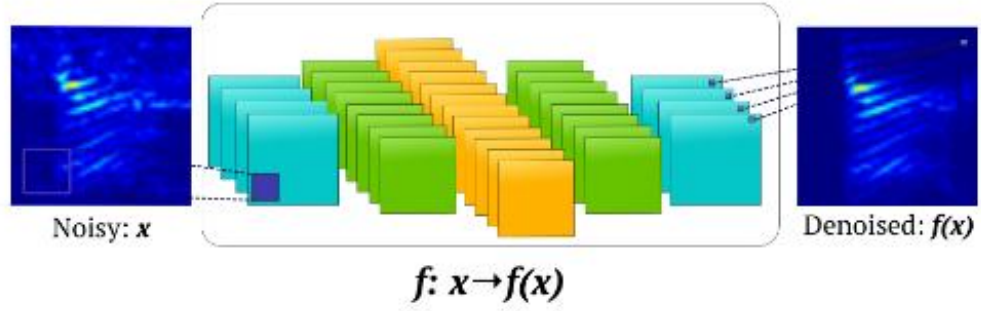
2.2.3. Derin öğrenme bileşenleri

2.2.3.1. Evrişimli sinir ağları (CNN)

Evrişimli Sinir Ağları (Convolutional Neural Networks- CNN), özellikle veri içindeki yerel örüntüleri ve uzamsal/zamansal ilişkileri öğrenmede güçlü sonuçlar veren derin öğrenme mimarileridir (Goodfellow vd., 2016). Görüntü işlemede iki boyutlu (2B) evrişimler kullanılırken, ses dalgası gibi zaman alanındaki tek boyutlu dizilerin işlenmesinde tek boyutlu (1B) evrişim katmanları tercih edilmektedir. 1B evrişim işlemi, girdi sinyalinin üzerinden öğrenilebilir ağırlıklara sahip bir filtrenin kaydırılmasıyla gerçekleştirilir. Girdi sinyali x ve filtre w için 1B evrişim matematiksel olarak Denklem (5) ile tanımlanmaktadır:

$$(x * w)[n] = \sum_{k=-K}^K x[n - k] \cdot w[k] \quad (5)$$

Burada toplam sembolünün sınırlarından dolayı $2K + 1$ toplam filtre (kernel) boyutunu, n zaman indeksini ve w ise filtrenin öğrenilebilir ağırlık katsayılarını temsil etmektedir. 1B evrişimin ses işlemedeki temel avantajları; zaman eksenindeki yerel akustik özellikleri yakalayabilmesi, parametre paylaşımı sayesinde hesaplama yükünü azaltması ve sinyalin zamansal kaymalarına karşı dayanıklı olmasıdır. Şekil 4'te CNN tabanlı bir konuşma iyileştirme yaklaşımının genel yapısı sunulmaktadır (Park ve Lee, 2016).

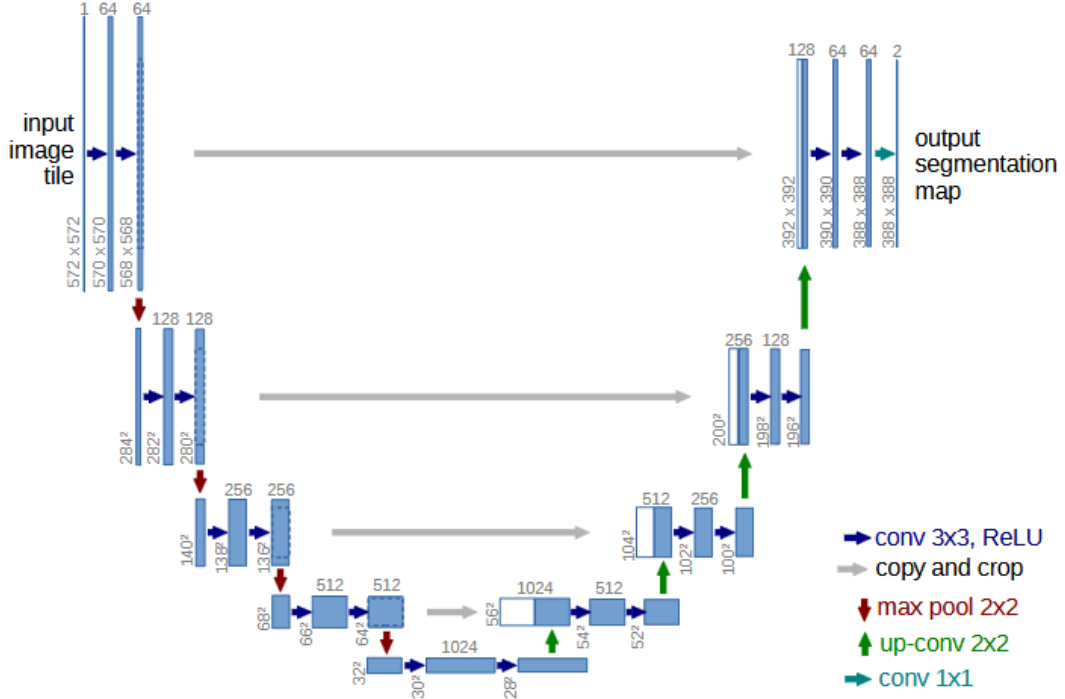


Şekil 4. CNN kullanarak yapılan konuşma iyileştirme (Park, S. R. ve Lee, J.,2016).

2.2.3.2. Kodlayıcı-çözücü (encoder-decoder) ve U-Net mimarileri

Kodlayıcı-çözücü (encoder-decoder) mimarileri, girdi sinyalini hiyerarşik katmanlar aracılığıyla işleyerek önce düşük boyutlu ve yüksek soyutluk derecesine sahip bir darboğaz (bottleneck) temsiline sıkıştıran, ardından bu temsilden çıktı sinyalini adım adım yeniden inşa eden yapılardır.

- **Kodlayıcı (Encoder):** Girdi sinyalini aşamalı olarak alt-örnekleme (downsampling) işlemine tabi tutar. Her aşamada zamansal çözünürlük azaltılırken, çıkarılan özellik haritalarının (feature maps) kanal sayısı artırılır. Bu süreç, sinyaldeki kaba ve yüksek seviyeli anlamsal özelliklerin öğrenilmesini sağlar.
- **Çözücü (Decoder):** Kodlayıcıdan gelen soyut darboğaz özelliklerini alır ve aşamalı olarak üst-örnekleyerek (upsampling) orijinal zamansal çözünürlüğe geri getirir.
- **U-Net ve Atlama Bağlantıları (Skip Connections):** Standart kodlayıcı-çözücü yapısında, alt-örnekleme adımlarındaki bilgi kayıpları nedeniyle çıktı sinyalinde ince detaylar (yüksek frekanslı bileşenler, geçici sesler vb.) kaybolmaktadır. Ronneberger vd. (2015) tarafından önerilen U-Net mimarisi, kodlayıcıdaki her seviyenin özellik haritasını çözücüdeki karşılık gelen seviyeye doğrudan aktaran atlama bağlantılarını (skip connections) sunarak bu problemi çözmüştür. Bu bağlantılar, çözücü katmanların hem soyut özellikleri hem de kodlayıcıdan gelen yüksek çözünürlüklü ham fiziksel detayları bir arada kullanarak sinyali çok daha kaliteli bir şekilde yeniden oluşturmasını sağlar.



Şekil 5. U-Net mimarisi (Ronneberger vd., 2015)

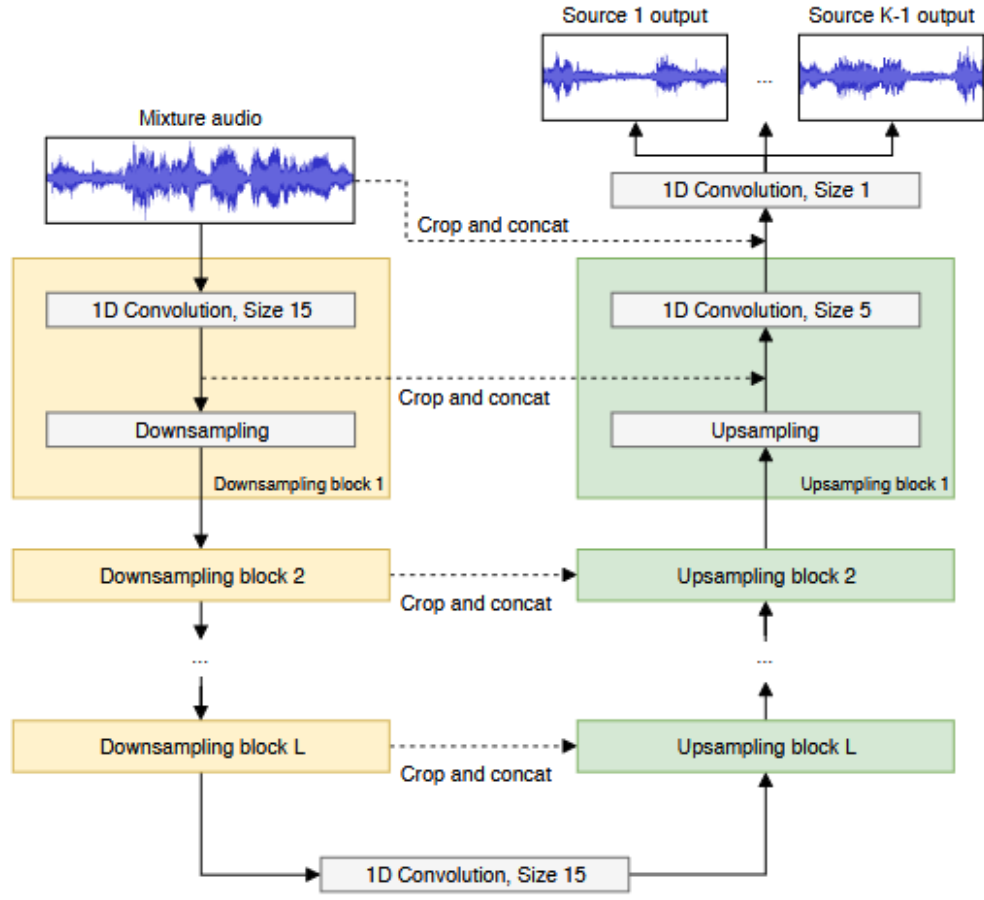
Şekil 5'te biyomedikal bölütleme için önerilen U-Net'in kodlayıcı-çözücü yapısı gösterilmektedir.

2.2.3.3. Wave-U-Net mimarisi

Wave-U-Net, orijinal U-Net mimarisinin doğrudan zaman alanındaki ham ses dalgaları üzerinde çalışabilecek şekilde uyarlanmış versiyonudur (Stoller vd., 2018). Geleneksel U-Net modelleri spektrogram gibi iki boyutlu zaman-frekans temsilleri üzerinde 2B evrişimlerle çalışırken, Wave-U-Net 2B katmanların tamamını 1B evrişim ve havuzlama katmanlarıyla değiştirmiştir.

Böylece model, ses sinyalini spektrograma dönüştürmek için gereken Kısa Zamanlı Fourier Dönüşümü (STFT) ve temiz ses sinyalini sentezlemek için gereken Ters STFT (ISTFT) işlemlerini devreden çıkarır. Bu durum, STFT tabanlı yöntemlerde karşılaşılan faz tahmini veya fazın yeniden yapılandırılması gereksinimini azaltmakta; modelin ham dalga formu üzerinden uçtan uca öğrenme yapmasına olanak sağlamaktadır. Böylece Wave-U-Net, genlik ve faz bileşenlerini ayrı bir ara temsil üzerinden ele almak yerine, konuşma sinyalinin zaman alanındaki yapısını doğrudan modelleyerek temiz konuşma sinyalini yeniden üretmektedir.

Şekil 6'da U-Net'in ham dalga formuna uyarlanmış hali olan Wave-U-Net mimarisi verilmektedir.



Şekil 6. Wave-U-Net mimarisi (Stoller vd., 2018)

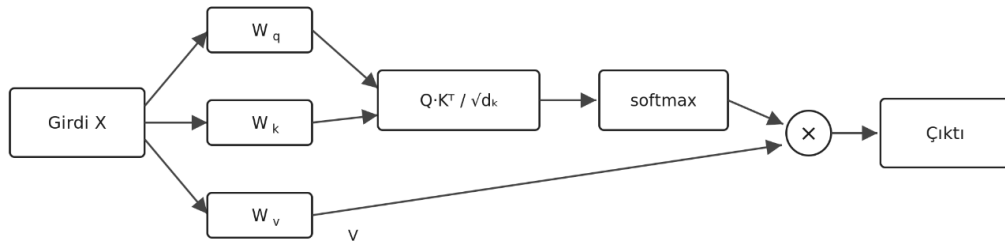
2.2.3.4. Dikkat mekanizmaları (Attention Mechanisms)

Dikkat mekanizmaları, derin öğrenme modellerinin girdi özellik haritalarındaki belirli bölgelere veya kanallara dinamik olarak daha fazla odaklanmasını sağlayan ağırlıklandırma yapılarıdır. Bu çalışmada iki tür dikkat mekanizması kullanılmıştır:

Öz-Dikkat (Self-Attention): Bir dizideki her bir zaman adımının, dizideki diğer tüm zaman adımlarıyla olan ilişkisini (uzun menzilli zamansal bağımlılıkları) modellemek için kullanılır (Vaswani vd., 2017). Evrişimli katmanların sınırlı alıcı alan kısıtını aşan öz-dikkat mekanizması, sorgu (Q), anahtar (K) ve değer (V) matrisleri üzerinden Denklem (6) ile hesaplanmaktadır:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

Burada d_k anahtar vektörlerinin boyutunu, $\sqrt{d_k}$ ise gradyanların kararlılığını sağlamak için kullanılan ölçekleme faktörünü temsil etmektedir. Öz-dikkat, konuşmadaki formant geçişleri gibi uzun zamansal bağımlılıkların korunmasında kritik rol oynar.



Şekil 7. Dikkat mekanizması mimarisi (Vaswani vd., 2017)

Şekil 7'de sorgu, anahtar ve değer matrisleri üzerinden işleyen öz-dikkat mekanizması gösterilmektedir.

Kanal Dikkati (Channel Attention): Özellik haritasındaki kanalların (farklı filtrelerin çıkışlarının) önem derecelerini dinamik olarak belirleyen ve adaptif olarak yeniden ölçekleyen bir mekanizmadır. Bu yapının temeli, kanal bazlı ağırlıklandırmayı ilk öneren Squeeze-and-Excitation (SE) bloklarına dayanır (Hu vd., 2018). Bu çalışmada, SE bloğunun yalnızca global ortalama havuzlama kullanan yapısı genişletilerek, CBAM'de önerildiği gibi hem global ortalama (GAP) hem de global maksimum (GMP) havuzlama işlemleri birlikte kullanılmıştır (Woo vd., 2018). Model, özellik haritasını zamansal boyutta bu iki havuzlama işlemiyle "sıkıştırarak" kanal bazlı küresel istatistikler elde eder. Ardından bu istatistikleri paylaşımlı iki katmanlı bir yapay sinir ağından (MLP) geçirip toplar; sigmoid aktivasyonu ile $[0, 1]$ aralığında kanal ağırlıkları üretir ve bu ağırlıkları orijinal özellik haritasıyla çarparak önemli kanalları öne çıkarır, gürültülü kanalları ise bastırır.

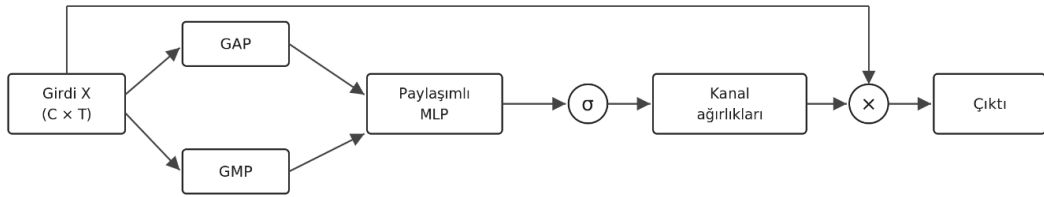
Girdi özellik haritası X olmak üzere, kanal dikkat katsayılarını (s) üreten işlem matematiksel olarak Denklem (7) ile tanımlanmaktadır:

$$s = \sigma \left(MLP(GAP(X)) + MLP(GMP(X)) \right) \quad (7)$$

Burada global ortalama havuzlama GAP , global maksimum havuzlama GMP , paylaşımlı iki katmanlı yapay sinir ağı MLP ve sigmoid aktivasyon fonksiyonu σ temsil edilmektedir. Elde edilen dikkat katsayıları (s), orijinal özellik haritası (X) ile kanal boyutunda çarpılarak Denklem (8) ile nihai ölçeklenmiş çıktı X' elde edilir:

$$X' = s \otimes X \quad (8)$$

Buradaki çarpım işareti ($\otimes$), kanal bazlı çarpmayı ifade etmektedir.



Şekil 8. Kanal dikkat mekanizması mimarisi (Hu vd., 2018; Woo vd., 2018)

Şekil 8'de ortalama ve maksimum havuzlamayı birleştiren kanal dikkat mekanizması verilmektedir.

2.2.3.5. Artık bağlantılar ve yığın normalizasyonu

Artık Bağlantılar (Residual Connections): Çok derin yapay sinir ağlarının eğitiminde karşılaşılan gradyan kaybolması problemini çözmek amacıyla geliştirilmiş yapılardır (He vd., 2016). Girdi x ve katman fonksiyonu $F(x)$ olmak üzere, artık blok çıktısı Denklem (9) ile ifade edilir:

$$y = F(x) + x \quad (9)$$

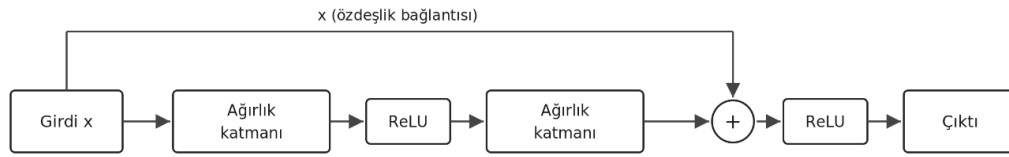
Bu yapı sayesinde gradyanlar, geriye yayılım esnasında doğrudan x üzerinden engelsiz bir şekilde alt katmanlara akabilir ve eğitimin kararlı sürdürülmesini sağlar.

Yığın Normalizasyonu (Batch Normalization): Ara katman çıktılarının ortalamasını sıfır, varyansını ise bir olacak şekilde normalize ederek içsel eşdeğişken kayması problemini azaltan bir tekniktir (Ioffe ve Szegedy, 2015). Bir B mini-yığınınındaki girdi x için normalizasyon işlemi Denklem (10) ve (11) ile gerçekleştirilir:

$$\hat{x} = \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \quad (10)$$

$$y = \gamma \hat{x} + \beta \quad (11)$$

Burada μ_B ve σ_B^2 yığın ortalama ve varyansını, ϵ sayısal kararlılık için eklenen küçük bir sabiti, γ ve β ise modelin eğitim sırasında öğrendiği ölçekleme ve kaydırma parametrelerini temsil eder.



Şekil 9. Artık (residual) bağlantı bloğu (He vd., 2016)

Şekil 9'da gradyan akışını kolaylaştıran artık bağlantı bloğunun yapısı gösterilmektedir.

2.2.3.6. Aktivasyon fonksiyonları

Modellerin doğrusal olmayan karmaşık ilişkileri öğrenebilmesi için kullanılan aktivasyon fonksiyonları şunlardır:

- **LeakyReLU:** Negatif bölgede sıfır olmayan gradyan üreterek yapay sinir hücrelerinin "ölmesi" (dying ReLU) problemini önleyen fonksiyondur. Çalışmada negatif eğim katsayısı $\alpha = 0,2$ olarak belirlenmiştir ve Denklem (12) ile gösterilmektedir:

$$f(x) = \max(\alpha x, x) \quad (12)$$

- **Sigmoid:** Çıktıyı $[0,1]$ aralığına sıkıştıran ve özellikle dikkat kapısı (attention gating) yapılarında ağırlık üretmek için kullanılan fonksiyondur (Denklem 13):

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (13)$$

- **ReLU:** Negatif girdileri sıfırlayan, pozitif girdileri ise doğrudan aktaran standart doğrusal olmayan fonksiyondur (Denklem 14):

$$f(x) = \max(0, x) \quad (14)$$

2.2.4. Nesnel değerlendirme metrikleri

2.2.4.1. Perceptual Evaluation of Speech Quality (PESQ)

PESQ, konuşma kalitesinin algısal değerlendirmesi için ITU-T P.862 standardında tanımlanmış, insan işitme algısını modelleyen uluslararası bir metriktir (Rix vd., 2001). Temiz referans sinyal ile iyileştirilmiş sinyal arasındaki farkı psikofiziksel bir işitme modeli (frekans ve zaman hizalama, kulak salyangozu filtrelemesi, gürültü maskeleyme vb.) üzerinden işleyerek $-0,5$ ile $4,5$ arasında objektif bir kalite skoru üretir. Yüksek PESQ skorları, ses kalitesinin yüksek olduğunu gösterir. Bu tezde PESQ'nun geniş bant (P.862.2) sürümü kullanılmıştır.

2.2.4.2. STOI ve ESTOI

Short-Time Objective Intelligibility (STOI): Konuşma anlaşılabilirliğini ölçmek amacıyla geliştirilmiş bir metriktir (Taal vd., 2011). Sinyali kısa zaman çerçevelerine böldükten sonra $1/3$ -oktav bantlarında süzerek temporal (zamansal) zarfları hesaplar. Temiz ve iyileştirilmiş sinyalin bu zamansal zarfları arasındaki korelasyonu hesaplayarak 0 ile 1 arasında bir anlaşılabilirlik değeri sunar.

Extended STOI (ESTOI): STOI'nin genişletilmiş versiyonudur (Jensen ve Taal, 2016). Standart STOI'den farklı olarak, zarflar arasındaki korelasyonu hesaplamadan önce spektral normalizasyon uygular. Bu normalizasyon, özellikle çok değişkenli gürültü koşullarında ve düşük sinyal-gürültü oranlarında (düşük SNR) anlaşılabilirlik tahmin doğruluğunu belirgin şekilde artırır.

2.2.4.3. Scale-Invariant Signal-to-Noise Ratio (SI-SNR)

Zaman alanında çalışan konuşma iyileştirme modellerinde, model çıktısının genlik ölçeğini değiştirerek (sesi kısarak veya açarak) kayıp fonksiyonlarını aldatmasını önlemek amacıyla geliştirilmiş bir metriktir (Le Roux vd., 2019). SI-SNR, iyileştirilmiş sinyali ($\hat{s}$), temiz referans sinyal (s) doğrultusundaki hedef bileşen (s_{target}) ve bu doğrultuya dik olan gürültü bileşeni (e_{noise}) olarak ikiye ayırır. Matematiksel hesaplama Denklem (15), (16) ve (17) ile tanımlanmaktadır:

$$s_{\text{target}} = \frac{\langle \hat{s}, s \rangle}{\|s\|^2} s \quad (15)$$

$$e_{\text{noise}} = \hat{s} - s_{\text{target}} \quad (16)$$

$$\text{SI-SNR} = 10 \cdot \log_{10} \left(\frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2} \right) \quad (17)$$

SI-SNR desibel (dB) cinsinden hesaplanır ve yüksek değerler daha başarılı bir gürültü arındırma performansını gösterir.

2.2.4.4. Kompozit ölçütler (CSIG, CBAK, COVL)

Hu ve Loizou (2008) tarafından geliştirilen insanların sesleri dinleyerek verdiği öznel Mean Opinion Score (MOS) değerlendirme puanlarını en yakın şekilde tahmin etmek üzere geliştirilmiş doğrusal regresyon tabanlı bileşik metriklerdir. 1 ile 5 aralığında değer alan bu metrikler şu şekildedir:

- **Composite Signal Distortion (CSIG):** Temizlenen sesteki konuşma bozulmasının derecesini tahmin eder:

$$\text{CSIG} = 3,093 - 1,029 \cdot \text{LLR} + 0,603 \cdot \text{PESQ} - 0,009 \cdot \text{WSS} \quad (18)$$

- **Composite Background Noise (CBAK):** Kalan arka plan gürültüsünün rahatsız edicilik düzeyini ölçer:

$$\text{CBAK} = 1,634 + 0,478 \cdot \text{PESQ} - 0,007 \cdot \text{WSS} + 0,063 \cdot \text{SSNR} \quad (19)$$

- **Composite Overall Quality (COVL):** Sinyal ve gürültü kalitesini birleştirerek genel ses kalitesini tahmin eder:

$$\text{COVL} = 1,594 + 0,805 \cdot \text{PESQ} - 0,512 \cdot \text{LLR} - 0,007 \cdot \text{WSS} \quad (20)$$

Formüllerde kullanılan alt metrikler; Doğrusal Öngörü Katsayıları farkını temsil eden Log-Likelihood Ratio (LLR), Bark ölçeğindeki Weighted Spectral Slope (WSS) ve segment bazlı Segmental SNR (SSNR) değerleridir.

2.2.5. Genel kayıp fonksiyonları ve optimizasyon teknikleri

2.2.5.1. Zaman alanı kayıp fonksiyonları

Zaman alanında çalışan modellerde, tahmin edilen ses dalga formu ile temiz referans dalga formu arasındaki farkı doğrudan minimize etmek için kullanılan kayıplardır.

- **L1 Kaybı (Mean Absolute Error - MAE):** Tahmin ve hedef sinyal örnekleri arasındaki mutlak farkların ortalamasıdır (Denklem 21):

$$\mathcal{L}_1 = \frac{1}{N} \sum_{n=1}^N |s[n] - \hat{s}[n]| \quad (21)$$

- **L2 Kaybı (Mean Squared Error - MSE):** Hataların karesini aldığı için büyük hataları aşırı derecede cezalandıran klasik bir kayıptır. Ancak ses sinyallerindeki uç değerlere (aykırı

gürültülere) karşı aşırı hassas olması ve sesi pürüzsüzleştirme (blurring/muffled ses) eğilimi nedeniyle konuşma iyileştirmede L1 kaybı daha kararlı sonuçlar vermektedir.

$$\mathcal{L}_2 = \frac{1}{N} \sum_{n=1}^N (s[n] - \hat{s}[n])^2 \quad (22)$$

2.2.5.2. Frekans alanı kayıp fonksiyonları

Ses sinyallerinin spektral detaylarını ve harmonik yapılarını korumak amacıyla Kısa Zamanlı Fourier Dönüşümü (STFT) üzerinden hesaplanan kayıplardır. Çok çözünürlüklü spektral kayıplar hem ince frekans detaylarını hem de kaba zaman zarflarını optimize etmek için farklı analiz penceresi boyutlarında eş zamanlı hesaplanır.

- **Spektral Büyüklük Kaybı ($\mathcal{L}_{\text{mag}}$):** Sinyallerin genlik spektrogramları arasındaki L1 farkıdır (Denklem 23):

$$\mathcal{L}_{\text{mag}} = \frac{1}{T} \sum_{t=1}^T \| |Y_t| - |\hat{Y}_t| \|_1 \quad (23)$$

- **Spektral Faz Kaybı ($\mathcal{L}_{\text{phase}}$):** Tahmin edilen ve referans sinyalin STFT faz açıları arasındaki farkın kosinüs mesafesidir (Denklem 24):

$$\mathcal{L}_{\text{phase}} = \frac{1}{T \cdot F} \sum_{t=1}^T \sum_{f=1}^F (1 - \cos(\angle Y_{t,f} - \angle \hat{Y}_{t,f})) \quad (24)$$

2.2.5.3. Optimizasyon teknikleri

AdamW Algoritması: Klasik Adam optimizasyon algoritmasının, L2 düzenlemesi le öğrenme oranını birbirinden ayırıştıran gelişmiş bir versiyonudur (Loshchilov ve Hutter, 2019). Geleneksel Adam algoritmasında ağırlık azaltma işlemi gradyan güncellemesiyle doğrudan ilişkiliyken, AdamW'de ağırlık azaltma katsayısı gradyandan bağımsız olarak doğrudan parametrelere uygulanır. Bu durum, derin ağlarda aşırı öğrenmeyi daha etkin engeller ve genelleme performansını artırır.

Öğrenme Oranı Çizelgelemesi (ReduceLROnPlateau): Eğitim sırasında doğrulama kaybını izleyen ve kayıpta belirli bir epoch süresi boyunca (patience) iyileşme gözlenmediğinde öğrenme oranını belirli bir faktörle düşüren çizelgeleme yöntemidir. Bu yöntem, optimizasyon algoritmasının yerel minimumlardan kaçınmasına ve eğitimin son aşamalarında daha küçük adımlarla hassas ağırlık ayarı yapmasına olanak tanır.

Erken Durdurma (Early Stopping): Doğrulama kaybının önceden belirlenen bir epoch limiti boyunca iyileşmediği durumlarda eğitimi otomatik olarak sonlandıran koruyucu bir tekniktir. Bu sayede modelin eğitim verisine aşırı uyum sağlaması (overfitting) engellenir ve gereksiz hesaplama maliyetinin önüne geçilir.

3. YÖNTEM

Bu bölümde, tez çalışmasında önerilen dikkat mekanizmaları ile güçlendirilmiş geliştirilmiş Wave-U-Net mimarisi, model bileşenleri, veri akışı ve optimizasyon sürecinde kullanılan karma kayıp fonksiyonu detaylandırılmaktadır.

3.1. Önerilen Model: Geliştirilmiş Wave-U-Net

Bu tez çalışmasında geliştirilen model, zaman alanındaki ham ses dalgalarını doğrudan işleyen uçtan uca bir yapıya dayanmaktadır. Mimarinin temeli olarak Wave-U-Net (Stoller vd., 2018); ham dalga formu üzerinde çalışarak spektrogram tabanlı yöntemlerde karşılaşılan faz bilgisi kaybı ve faz kestirimi gereksinimini büyük ölçüde azaltması, U-Net'ten devraldığı kodlayıcı–çözücü yapısı ve atlama bağlantıları sayesinde alt-örnekleme sırasında zayıflayabilen ince zamansal detayları koruyabilmesi ve çok ölçekli zamansal bağlamı öğrenebilmesi, ayrıca konuşma iyileştirme görevinde daha önce olumlu sonuçlar vermiş olması (Macartney ve Weyde, 2018) gibi nedenlerden dolayı tercih edilmiştir.

Bu temel üzerine kurulan tasarım; kanal dikkati, öz-dikkat, artık bloklar ve yığın normalizasyonu gibi derin öğrenme bileşenleri eklenerek yeniden yapılandırılmıştır. Söz konusu geliştirmelerle, modelin yerel öznitelikleri yakalama kapasitesinin güçlendirilmesi ve gürültülü sinyallerdeki karmaşık zaman–frekans ilişkilerinin daha kararlı biçimde öğrenilmesi amaçlanmıştır.

3.1.1. Model mimarisine genel bakış ve kanal yapısı

Önerilen mimari, hiyerarşik bir kodlayıcı–çözücü (encoder-decoder) yapısına dayanmaktadır. Kodlayıcı (encoder), girdi sinyalinin zamansal çözünürlüğünü hiyerarşik olarak azaltırken kanal boyutunu artırır; çözücü (decoder) ise bu soyut özellikleri alarak zamansal çözünürlüğü adım adım orijinal boyutuna geri döndürür. Kodlayıcı ve çözücü arasındaki atlama bağlantıları (skip connections), alt-örnekleme sırasında kaybolabilecek yüksek frekanslı ince detayların ve zaman alanındaki faz bilgisinin çözücüye doğrudan aktarılmasını sağlar (Ronneberger vd., 2015). Mimarideki kanal yapısı ve katmanlar arasındaki boyut değişimleri Tablo 1'de özetlenmiştir.

Tablo 1. Önerilen geliştirilmiş Wave-U-Net mimarisinin kanal yapısı

Katman Adı	Giriş Kanalı	Çıkış Kanalı	Çıktı Boyutu (Zaman x Kanal)
Girdi Sinyali	—	—	16000×1
DownBlock 1	1	32	8000×32
DownBlock 2	32	64	4000×64
DownBlock 3	64	128	2000×128
DownBlock 4	128	256	1000×256
Darboğaz (Bottleneck)	256	512	1000×512
UpBlock 1	512	256	2000×256
UpBlock 2	256	128	4000×128
UpBlock 3	128	64	8000×64
UpBlock 4	64	32	16000×32
Son Katman (Final)	32	1	16000×1

Geliştirilmiş mimari, orijinal Wave-U-Net modeliyle karşılaştırıldığında birçok yapısal yenilik barındırmaktadır. Bu yapısal farklar ve tasarım tercihleri Tablo 2'de karşılaştırmalı olarak sunulmuştur.

Tablo 2. Orijinal Wave-U-Net ile önerilen modelin karşılaştırması

Özellik	Orijinal Wave-U-Net (Stoller vd., 2018)	Önerilen Model (Geliştirilmiş Wave-U-Net)
Dikkat Mekanizması	Bulunmamaktadır	Kanal Dikkati ve Öz-Dikkat Entegrasyonu
Normalizasyon	Bulunmamaktadır	Yığın Normalizasyonu (Batch Normalization)
Artık Bağlantılar	Bulunmamaktadır	Blok içi Artık Bağlantılar (He vd., 2016)
Kayıp Fonksiyonu	Ortalama Karesel Hata (MSE)	Zaman-Frekans Karma Kaybı (Enhanced Loss)
Alt-Örnekleme	Adımlı Evrişim (Strided Conv)	Maksimum Havuzlama (MaxPool1d)
Üst-Örnekleme	Doğrusal İnterpolasyon	Transpoze Evrişim (ConvTranspose1d)
Filtre Çekirdek Boyutu	5	15

Önerilen geliştirilmiş Wave-U-Net modelinin toplam eğitilebilir parametre sayısı tam olarak 27.043.362 adettir. Parametrelerin modelin ana mimari bloklarına göre dağılımı incelendiğinde; kodlayıcı bloklarının toplamda yaklaşık 3,28 M parametre, darboğaz katmanının yaklaşık 18,16 M parametre, çözücü bloklarının yaklaşık 5,60 M parametre ve son dalga formu dönüştürme

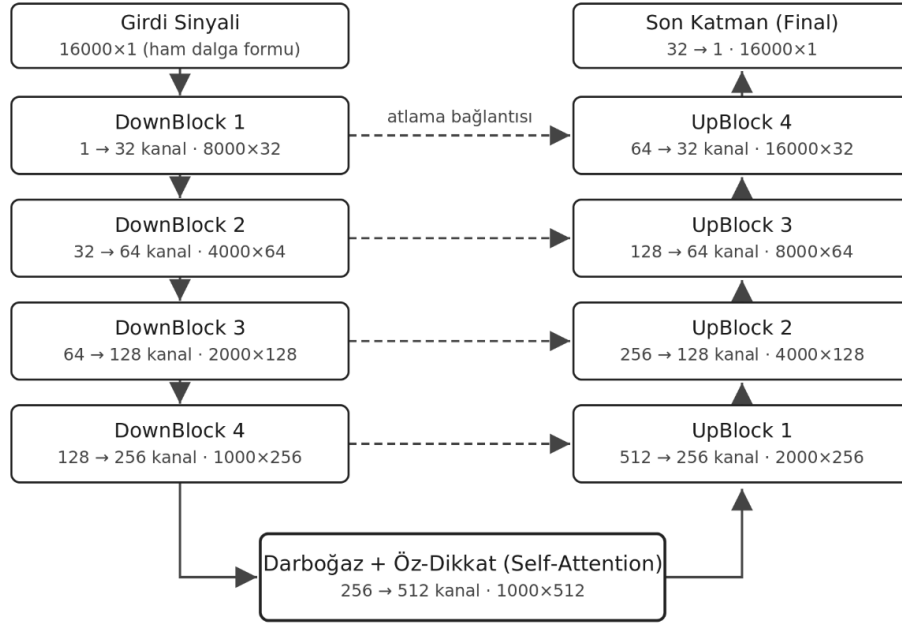
katmanının ise yaklaşık 1,57 K parametre içerdiği görülmektedir. Darboğaz katmanının, toplam parametrelerin yaklaşık %67,16'sını kapsayacak şekilde son derece yüksek bir parametre yoğunluğuna sahip olması; 512 kanal boyutundaki derin evrişim filtrelerinin, iki adet ardışık artık bloğun ve uzun menzilli ilişkileri modelleyen öz-dikkat projeksiyon matrislerinin bu bölgede konumlandırılmış olmasından kaynaklanmaktadır.

Önerilen modelin hiyerarşik katman yapısı ve her bir alt bileşenin matematiksel parametre hesaplama detayları Tablo 3'te sunulmuştur. Şekil 10'da ise önerilen geliştirilmiş Wave-U-Net mimarisinin katmanları, kanal sayıları ve veri akışı bir arada sunulmuştur.

Tablo 3. Geliştirilmiş Wave-U-Net modelinin detaylı eğitilebilir parametre dağılımı

Ana Blok	Alt Bileşen / Katman	Kanal Geçiş	Formül Hesaplama ($C_{in} \times C_{out} \times K + Bias$)	Parametre Sayısı
Kodlayıcı (Encoder)	DownBlock 1	1→32	$(1 \times 32 \times 15 + 32) + BN + ResBlock(32)$	31.744
	DownBlock 2	32→64	$(32 \times 64 \times 15 + 64) + BN + ResBlock(64)$	155.200
	DownBlock 3	64→128	$(64 \times 128 \times 15 + 128) + BN + ResBlock(128)$	619.648
	DownBlock 4	128→256	$(128 \times 256 \times 15 + 256) + BN + ResBlock(256)$	2.476.288
Darboğaz (Bottleneck)	Giriş Evrişimi + BN	256→512	$(256 \times 512 \times 15 + 512) + (512 \times 2)$	1.967.616
	Öz-Dikkat (Self-Attention)	512→512	$Q(512 \times 64 + 64) + K(512 \times 64 + 64) + V(512 \times 512 + 512) + \gamma$	328.321
	Artık Blok 1 (Residual)	512→512	$2 \times (512 \times 512 \times 15 + 512) + 2 \times BN + CA(512)$	7.932.928
	Artık Blok 2 (Residual)	512→512	$2 \times (512 \times 512 \times 15 + 512) + 2 \times BN + CA(512)$	7.932.928
Çözücü (Decoder)	UpBlock 1	512→256	$ConvT(512 \times 256 \times 2) + Conv(512 \times 256 \times 15) + BN + ResBlock(256)$	4.213.248
	UpBlock 2	256→128	$ConvT(256 \times 128 \times 2) + Conv(256 \times 128 \times 15) + BN + ResBlock(128)$	1.053.952
	UpBlock 3	128→64	$ConvT(128 \times 64 \times 2) + Conv(128 \times 64 \times 15) + BN + ResBlock(64)$	263.808
	UpBlock 4	64→32	$ConvT(64 \times 32 \times 2) + Conv(64 \times 32 \times 15) + BN + ResBlock(32)$	66.112
Son Katman (Final)	Final Conv 1	32→16	$(32 \times 16 \times 3 + 16)$	1.552
	Final Conv 2	16→1	$(16 \times 1 \times 1 + 1)$	17
TOPLAM				27.043.362

Kısaltmalar: BN: Yığın Normalizasyonu (Batch Normalization), ResBlock: Artık Blok, ConvT: Transpoze Evrişim (ConvTranspose1d), CA: Kanal Dikkati (ChannelAttention).



Şekil 10. Önerilen Geliştirilmiş Wave-U-Net mimarisinin katman boyutları ve veri akışı

3.1.2. Aşağı-Örnekleme (DownBlock) ve yukarı-örnekleme (Upblock) blokları

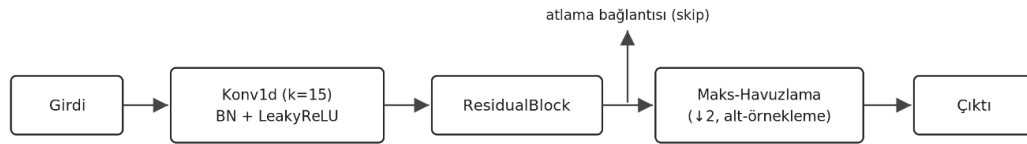
3.1.2.1. Aşağı-Örnekleme bloğu (DownBlock)

Kodlayıcıda yer alan her bir DownBlock, girdi özellik haritasından yerel zaman özneteliklerini çıkarırken zamansal çözünürlüğü yarıya indirir. Bir DownBlock sırasıyla şu katmanlardan oluşur:

- 1. Bir Boyutlu Evrişim (Conv1d) Katmanı:** Büyük bir çekirdek genişliği (kernel size = 15, padding = 7) kullanılarak girdi uzunluğu korunur. Büyük çekirdek boyutu, 16 kHz örnekleme frekansında yaklaşık 1 ms'lik bir zamansal pencereye denk gelerek, konuşmadaki geçici akustik değişimlerin ve patlayıcı ünsüzlerin yerel bağlamını yakalamada kritik rol oynar.
- 2. Yığın Normalizasyonu (BatchNorm1d):** Eğitim boyunca dahili eşdeğişken kaymasını engeller ve gradyan akışını dengeler (Ioffe ve Szegedy, 2015).
- 3. LeakyReLU Aktivasyon Fonksiyonu:** Negatif bölgedeki hücrelerin sönmesini önlemek amacıyla $\alpha=0.2$ sızıntı katsayısıyla uygulanır.

4. **Artık Blok (ResidualBlock):** Özellik çıkarım derinliğini artıran kanal dikkatli artık öğrenme bloğudur.
5. **Maksimum Havuzlama (MaxPool1d):** Filtre boyutu ve adım uzunluğu 2 olarak ayarlanmış havuzlama katmanı, zaman eksenindeki baskın özellikleri koruyarak çözünürlüğü yarıya indirir.

DownBlock iki adet çıktı üretir. İlki havuzlama katmanından geçerek bir sonraki seviyeye aktarılan alt-örnekleme özelliğidir. İkincisi ise çözücüdeki karşılık gelen yukarı-örnekleme bloğuna aktarılmak üzere havuzlama işleminden hemen önce ayrılan atlama bağlantısı çıktısıdır. Şekil 11'de kodlayıcıda kullanılan DownBlock'un iç yapısı gösterilmektedir.

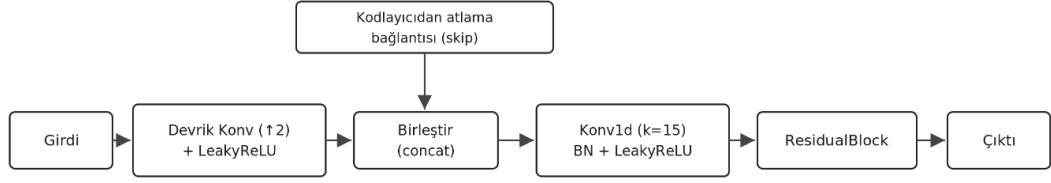


Şekil 11. DownBlock'un iç yapısı

3.1.2.2. Yukarı-Örnekleme bloğu (UpBlock)

Çözücüde yer alan her bir UpBlock, özellikleri daha yüksek zamansal çözünürlüklere taşımak ve kodlayıcıdan gelen ham fiziksel detayları sentezlemekle görevlidir. Şekil 12'de çözücüde kullanılan UpBlock'un iç yapısı verilmektedir. Blok yapısı şu şekildedir:

1. **Transpoze Evrişim (ConvTranspose1d):** Filtre boyutu ve adım uzunluğu 2 olan bu katman, zamansal boyutu iki katına çıkarırken kanal sayısını yarıya indirir. Sabit interpolasyon yöntemleri yerine transpoze evrişimin tercih edilmesinin nedeni, modelin veri setine özgü yukarı-örnekleme filtrelerini öğrenerek daha az yapay ses üretmesini sağlamaktır.
2. **LeakyReLU Aktivasyonu ($\alpha=0.2$):** Doğrusal olmayan temsil yeteneğini korur.
3. **Özellik Birleştirme (Concatenation):** Transpoze evrişim çıktısı ile kodlayıcıdan atlama bağlantısıyla gelen özellik haritası kanal boyutunda uç uca birleştirilir.
4. **1 Boyutlu Evrişim (Conv1d) Katmanı:** Çekirdek genişliği 15 ve padding = 7 olan katman, birleştirilmiş kanalları işleyerek hedef kanal boyutuna indirger.
5. **Yığın Normalizasyonu ve LeakyReLU ($\alpha=0.2$) Katmanları**
6. **Artık Blok (ResidualBlock):** Özellik sentezinin doğrusal olmayan kapasitesini artırır.



Şekil 12. UpBlock'un iç yapısı

3.1.3. Artık blok (ResidualBlock) tasarımı

Eğitim esnasında gradyanların kaybolmasını önlemek ve derin katmanlardaki bilgi akışını optimize etmek için DownBlock ve UpBlock yapılarının içerisine özelleştirilmiş artık bloklar entegre edilmiştir (He vd., 2016). Klasik artık blok tasarımlarından farklı olarak bu tez çalışmasında, artık bağlantı eklenmeden önce kanalların önem derecelerini dinamik olarak düzenleyen bir kanal dikkati mekanizması bloğa dahil edilmiştir. Şekil 13'te ResidualBlock'un ve içerdiği kanal dikkati modülünün yapısı gösterilmektedir.

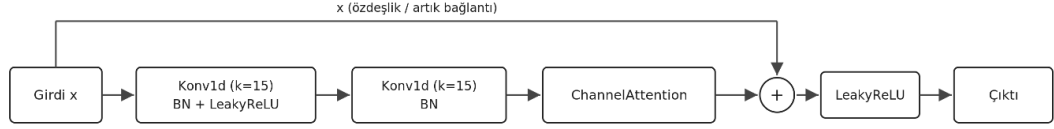
ResidualBlock'un iç yapısı sırasıyla şu şekildedir:

1. **Evrişim Katmanı 1:** 1B Evrisim (kernel=15, padding=7)
2. **Yığın Normalizasyonu 1**
3. **LeakyReLU Aktivasyonu ($\alpha=0.2$)**
4. **Evrişim Katmanı 2:** 1B Evrisim (kernel=15, padding=7)
5. **Yığın Normalizasyonu 2**
6. **Kanal Dikkati Modülü (ChannelAttention):** Özellik haritasındaki gürültülü kanalları bastırıp konuşma kanallarını öne çıkarır.
7. **Artık Toplama:** Kanal dikkati çıktısı ile bloğun orijinal girdisi toplanır.
8. **LeakyReLU Aktivasyonu ($\alpha=0.2$):** Bloğun nihai çıktısını üretir.

Blok içindeki işlemlerin matematiksel ifadesi Denklem (25) ile tanımlanmaktadır:

$$y = \text{LeakyReLU}(\text{ChannelAttn}(\text{BN}(\text{Conv2}(\text{LeakyReLU}(\text{BN}(\text{Conv1}(x)))))) + x) \quad (25)$$

Burada x bloğun girdisini, y ise çıktı özellik haritasını temsil etmektedir.



Şekil 13. ResidualBlock yapısı

3.1.4. Kanal dikkati ve öz-dikkat entegrasyonu

3.1.4.1. Kanal dikkati mekanizması (ChannelAttention)

Özellik haritalarında bulunan kanalların (filtre çıkışlarının) her biri farklı akustik özellikleri veya frekans bölgelerini temsil etmektedir. Gürültülü ortamlarda bazı kanallar yalnızca arka plan gürültüsünü taşıırken, bazıları temiz konuşma bilgisini içerir. Modelin konuşma taşıyan kanallara odaklanmasını sağlamak amacıyla Squeeze-and-Excitation (SE) yapısına (Hu vd., 2018) dayanan bir kanal dikkati mekanizması kullanılmıştır.

Modül, zaman eksenini boyunca hem global ortalama havuzlama (GAP) hem de global maksimum havuzlama (GMP) işlemlerini bir arada uygulayarak daha zengin ve ayırt edici kanal istatistikleri çıkarır (Woo vd., 2018). İşlem adımları şu şekildedir:

1. Girdi özellik haritası $X \in \mathbb{R}^{C \times L}$ (C : kanal sayısı, L : zamansal uzunluk) zaman boyutu boyunca sıkıştırılarak global istatistik vektörleri elde edilir:

$$z_{\text{avg}} = \text{GAP}(X) \in \mathbb{R}^{C \times 1} \quad (26)$$

$$z_{\text{max}} = \text{GMP}(X) \in \mathbb{R}^{C \times 1} \quad (27)$$

2. Elde edilen vektörler, parametreleri paylaşılan iki katmanlı bir yapay sinir ağından (MLP) geçirilir. MLP yapısında sıkıştırma (indirgeme) oranı $r = 8$ olarak belirlenmiştir. Bu oran, parametre maliyetini sınırlandırırken temsil gücünün korunmasını sağlar:

$$s = \sigma(\text{MLP}(z_{\text{avg}}) + \text{MLP}(z_{\text{max}})) \in \mathbb{R}^{C \times 1} \quad (28)$$

Burada σ sigmoid aktivasyon fonksiyonunu, s ise 0 ile 1 arasında değer alan kanal ağırlık vektörünü temsil eder.

3. Son aşamada orijinal özellik haritası ile kanal ağırlık vektörü çarpılarak kanal bazında ölçeklenmiş nihai çıktı elde edilir:

$$\tilde{X} = s \odot X \quad (29)$$

4. Burada $\odot$ işlemi, kanal boyutunda genişletilerek yapılan (broadcast) eleman bazlı çarpmayı temsil etmektedir.

3.1.4.2. Öz-Dikkat mekanizması (SelfAttention)

Evrışimli katmanlar, yapıları gereği yalnızca filtre çekirdeğinin kapsadığı yerel alıcı alanda çalışırlar. Ancak konuşma sinyalindeki formant geçişleri ve hece yapıları gibi uzun menzilli zamansal bağımlılıkların korunması için global bir bağlam analizine ihtiyaç duyulmaktadır (Vaswani vd., 2017). Bu kısıtlamayı aşmak amacıyla modelin en derin noktası olan darboğaz katmanına öz-dikkat mekanizması entegre edilmiştir.

Darboğazda girdi boyutu zamansal olarak en düşük seviyeye indirgendiğinden (4 kat alt-örnekleme sonrası $L = 1000$), öz-dikkat matrisinin hesaplama karmaşıklığı ($\mathcal{O}(L^2)$) makul sınırlar içinde kalmaktadır. Öz-dikkat modülünün çalışma adımları aşağıda açıklanmaktadır:

1. Girdi özellik haritası $X \in \mathbb{R}^{C \times L}$ üç adet 1 boyutlu (1B) evrişim katmanı üzerinden sorgu (Q), anahtar (K) ve değer (V) matrislerine projekte edilir. Hesaplama maliyetini düşürmek ve aşırı öğrenmeyi engellemek amacıyla sorgu ve anahtar projeksiyonlarında kanal boyutu 8 kat azaltılarak $C/8$ yapılmıştır:

$$Q = W_q X \in \mathbb{R}^{\frac{C}{8} \times L} \quad (30)$$

$$K = W_k X \in \mathbb{R}^{\frac{C}{8} \times L} \quad (31)$$

$$V = W_v X \in \mathbb{R}^{C \times L} \quad (32)$$

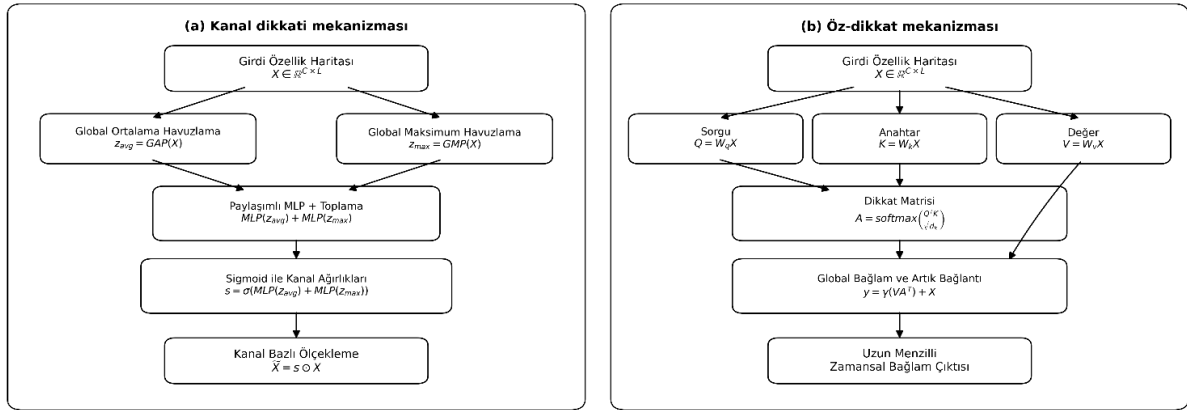
2. Zamansal adımların birbiriyle olan ilişkisini ölçen dikkat matrisi $A \in \mathbb{R}^{L \times L}$ hesaplanır:

$$A = \text{softmax}\left(\frac{Q^T K}{\sqrt{d_k}}\right) \quad (33)$$

3. Burada $d_k = C/8$ anahtar vektörlerinin boyutunu, $\sqrt{d_k}$ ise gradyan kararlılığını sağlayan ölçekleme faktörüdür.
4. Dikkat matrisi ile değer matrisi çarpılarak global bağlam bilgisi sentezlenir. Bu çıktı, öğrenilebilir bir ölçekleme parametresi (γ) ile çarpılarak girdi özellik haritasına artık bağlantı olarak eklenir:

$$y = \gamma(VA^T) + X \quad (34)$$

Öğrenilebilir ölçek parametresi γ , eğitim başlangıcında 0 olarak tanımlanmıştır. Bu sayede model, eğitimin ilk evrelerinde kararlı çalışan artık bağlantı üzerinden öğrenmeye başlar ve eğitim ilerledikçe öz-dikkat mekanizmasını kademeli ve yumuşak bir şekilde sürece dahil eder.



Şekil 14.Önerilen modelde kullanılan dikkat mekanizmalarının şematik gösterimi: (a) kanal dikkati mekanizması, (b) öz-dikkat mekanizması.

Şekil 14'te modelde kullanılan kanal dikkati ve öz-dikkat mekanizmalarının şematik gösterimi verilmektedir.

3.1.4.3. Darboğaz katmanı yapısı (Bottleneck)

Darboğaz katmanı, modelin kodlayıcı ile çözücü arasındaki en derin ve en soyut özelliklerin işlendiği darboğaz bölgesidir. Bu katman sırasıyla şu bileşenlerden oluşmaktadır:

1. 1B Evrişim Katmanı (256 → 512 kanal, kernel = 15, padding = 7)
2. Yığın Normalizasyonu (512 kanal) ve LeakyReLU ($\alpha = 0.2$) aktivasyonu
3. Öz-Dikkat Katmanı (SelfAttention)
4. İki adet ardışık artık blok (ResidualBlock). Yapılan ön deneysel çalışmalarda tek artık bloğun karmaşık gürültü örüntülerini soyutlamada yetersiz kaldığı, iki bloğun ise en ideal başarımlı-maliyet dengesini sunduğu tespit edilmiştir.

3.1.5. Önerilen karma kayıp fonksiyonu (EnhancedLoss)

Zaman alanında doğrudan dalga formu tahmini yapan modellerde, yalnızca tek bir kayıp fonksiyonu kullanılması (örneğin sadece L1 veya L2) algısal ses kalitesinde ve harmonik yapıda bozulmalara yol açmaktadır. Sesi pürüzsüzleştirme ve faz bozulmalarını önlemek amacıyla zaman alanı, frekans alanı ve sinyal kalitesi metriklerini birleştiren çok bileşenli karma bir kayıp fonksiyonu (EnhancedLoss) kullanılmıştır.

Önerilen kayıp fonksiyonunun genel matematiksel ifadesi Denklem (35) ile tanımlanmaktadır:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{time}} + \beta \mathcal{L}_{\text{sisnr}} + \gamma \mathcal{L}_{\text{stft}} \quad (35)$$

Kayıp fonksiyonu bileşenlerinin ağırlıkları $\alpha = 0,25$, $\beta = 0,35$ ve $\gamma = 0,40$ olarak belirlenmiştir. Bu ağırlıklandırma, modelin algısal ses kalitesine en duyarlı olan spektral ve sinyal kalitesi metriklerine öncelik vermesini sağlamaktadır.

Zaman Alanı Kaybı ($\mathcal{L}_{\text{time}}$): Tahmin edilen ses dalga formu $\hat{s}[n]$ ile temiz referans dalga formu $s[n]$ arasındaki dalga şekli benzerliğini zaman düzeyinde optimize etmek için L1 (ortalama mutlak hata) kaybı kullanılmıştır (Denklem 36):

$$\mathcal{L}_{\text{time}} = \frac{1}{N} \sum_{n=1}^N |\hat{s}[n] - s[n]| \quad (36)$$

Burada N sinyalin toplam örnek sayısını (16000) temsil etmektedir.

Ölçekten Bağımsız Sinyal-Gürültü Oranı Kaybı ($\mathcal{L}_{\text{sisnr}}$): Modelin tahmin ettiği sinyalin genlik ölçeğini değiştirerek (sesi kısarak) zaman alanı kayıplarını manipüle etmesini önlemek ve doğrudan gürültü arındırma performansını optimize etmek amacıyla negatif SI-SNR değeri kayıp fonksiyonuna dahil edilmiştir (Le Roux vd., 2019). SI-SNR kaybının hesaplanabilmesi için öncelikle tahmin edilen sinyal, referans sinyal yönündeki hedef bileşen (s_{target}) ve referansa dik olan gürültü bileşeni (e_{noise}) olarak ikiye ayrılır. Sinyallerin sıfır ortalamalı (zero-mean) versiyonları kullanılarak yapılan hesaplama adımları Denklem (37), (38) ve (39) ile verilmiştir:

$$s_{\text{target}} = \frac{\langle \hat{s}, s \rangle}{\|s\|^2} s \quad (37)$$

$$e_{\text{noise}} = \hat{s} - s_{\text{target}} \quad (38)$$

$$\mathcal{L}_{\text{sisnr}} = -10 \log_{10} \left(\frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2} \right) \quad (39)$$

Çoklu Çözünürlüklü STFT Kaybı ($\mathcal{L}_{\text{stft}}$): Ses sinyallerinin harmonik yapısını, formant frekanslarını ve spektral detaylarını korumak amacıyla Kısa Zamanlı Fourier Dönüşümü (STFT) üzerinden hesaplanan frekans alanı kaybıdır. Sinyal analizinde tek bir pencere boyutu kullanılması, zaman ve frekans çözünürlüğü arasındaki belirsizlik ilkesi nedeniyle spektral detayların bir kısmının kaçırılmasına yol açmaktadır. Bu problemi çözmek için üç farklı FFT boyutu ($M = \{512, 1024, 2048\}$) kullanılarak çoklu çözünürlüklü STFT kaybı uygulanmıştır.

Her bir çözünürlük ölçeği için spektral büyüklük kaybı ($\mathcal{L}_{\text{mag}}$) ile spektral faz kaybı ($\mathcal{L}_{\text{phase}}$) eş zamanlı olarak hesaplanır (Denklem 40):

$$\mathcal{L}_{\text{stft}}^{(f)} = \mathcal{L}_{\text{mag}}^{(f)} + 0,1 \cdot \mathcal{L}_{\text{phase}}^{(f)} \quad (40)$$

Burada kullanılan bileşenlerin formülleri Denklem (41) ve (42) ile gösterilmiştir:

$$\mathcal{L}_{\text{mag}}^{(f)} = \frac{1}{T \cdot F} \sum_{t=1}^T \sum_{k=1}^F \left| |X^{(f)}(t, k)| - |\hat{X}^{(f)}(t, k)| \right| \quad (41)$$

$$\mathcal{L}_{\text{phase}}^{(f)} = \frac{1}{T \cdot F} \sum_{t=1}^T \sum_{k=1}^F \left(1 - \cos \left(\angle X^{(f)}(t, k) - \angle \hat{X}^{(f)}(t, k) \right) \right) \quad (42)$$

Burada $X^{(f)}(t, k)$ ve $\hat{X}^{(f)}(t, k)$, f pencere yapılandırmasında sırasıyla temiz ve tahmin edilen sinyallerin STFT çıktıları; T zaman çerçeve sayısını ve F ise frekans kanalı sayısını temsil eder. Spektral faz kaybı, iki sinyalin faz açıları arasındaki kosinüs mesafesini minimize ederek zaman alanındaki faz tutarsızlıklarını ortadan kaldırır.

Nihai çoklu çözünürlüklü STFT kaybı, kullanılan tüm çözünürlüklerin ortalaması alınarak hesaplanır (Denklem 43):

$$\mathcal{L}_{\text{stft}} = \frac{1}{|M|} \sum_{f \in M} \mathcal{L}_{\text{stft}}^{(f)} \quad (43)$$

Kullanılan çoklu FFT boyutlarının fiziksel gerekçeleri ve yakaladıkları akustik özellikler Tablo 4'te açıklanmıştır:

Tablo 4. Çoklu çözünürlüklü STFT kaybında FFT boyutlarının gerekçeleri

FFT Boyutu (f)	Pencere Uzunluğu	Frekans Çözünürlüğü	Zamansal Çözünürlük	Yakalanan Akustik Özellikler
512	32 ms	Düşük (31,25 Hz)	Yüksek (2 ms)	Ani akustik değişimler, geçici sesler ve patlayıcı ünsüzler.
1024	64 ms	Orta (15,63 Hz)	Orta (4 ms)	Genel spektral zarf ve perde periyodunun genel yapısı.
2048	128 ms	Yüksek (7,81 Hz)	Düşük (8 ms)	İnce frekans detayları, harmonikler ve formant kararlılığı.

3.2. Deneysel Düzenek ve Veri Hazırlama

Bu bölümde, önerilen modelin eğitimi, doğrulanması ve test edilmesinde kullanılan deneysel altyapı anlatılmaktadır. Bu kapsamda; Türkçe ve İngilizce veri setleri, akustik gürültülerin

sentezlenmesi, veri ön işleme adımları, çoklu rastgele bölme stratejisi, eğitim hiperparametreleri ile nesnel başarımlar değerlendirme protokolü açıklanmaktadır.

3.2.1. İngilizce (VCTK) ve Türkçe (TSCD) veri setleri

Tez çalışması kapsamında önerilen mimarinin diller arası başarımlarını ve aktarılabilirliğini değerlendirmek amacıyla İngilizce ve Türkçe dillerini kapsayan toplam 70.939 adet ses dosyasından oluşan geniş bir veri havuzu oluşturulmuştur.

3.2.1.1. İngilizce veri seti: VCTK Corpus

İngilizce konuşma kayıtları için literatürde referans kabul edilen ve Edinburgh Üniversitesi Konuşma Teknolojileri Araştırma Merkezi (CSTR) tarafından yayımlanan Voice Cloning Toolkit (VCTK) Corpus kullanılmıştır (Yamagishi vd., 2019). Bu veri setinin tercih edilme nedenleri; kayıtların yankısız oda koşullarında yapılmış olması, farklı yaş ve aksan grubundan çok sayıda konuşmacıyı barındırması ve yüksek sinyal kalitesidir.

- **Toplam Dosya Sayısı:** 44455 adet temiz ses dosyası.
- **Örnekleme Frekansı:** 16 kHz.
- **Akustik Kanal Seçimi:** Akustik yansımaları en aza indirmek için omni-directional karakterdeki "Mic1" mikrofonuyla kaydedilen ses kanalı kullanılmıştır.
- **İçerik Yapısı:** Konuşmacıların gazete metinlerini ve fonetik olarak dengelenmiş paragrafları okuduğu akıcı cümle kayıtlarından oluşmaktadır.

3.2.1.2. Türkçe veri seti: Turkish Speech Command Dataset (TSCD)

Türkçe konuşma kayıtları için açık kaynaklı ve güncel bir veri seti olan "Turkish Speech Command Dataset" tercih edilmiştir (Kurtkaya, 2025). Bu veri seti, Türkçe konuşma tanıma ve iyileştirme çalışmalarında test edilmek üzere izole sözcük yapısında kurgulanmıştır.

- **Toplam Dosya Sayısı:** 26484 adet temiz ses dosyası.
- **Örnekleme Frekansı:** 16 kHz.
- **İçerik Yapısı:** Günlük yaşamda ve akıllı asistan komutlarında sıklıkla kullanılan tekil (izole) Türkçe kelimelerin (örneğin; "aç", "kapat", "yukarı", "aşağı", "evet", "hayır" vb.) farklı konuşmacılar tarafından seslendirilmiş kayıtlarından oluşmaktadır.

Her iki veri setinin yapısal ve akustik özellikleri karşılaştırmalı olarak Tablo 5'te özetlenmiştir:

Tablo 5. İngilizce ve Türkçe veri setlerinin karşılaştırmalı özellikleri

Akustik ve Yapısal Özellik	İngilizce Veri Seti (VCTK)	Türkçe Veri Seti (TSCD)
Toplam Temiz Ses Sayısı	44.455 adet	26.484 adet
Örnekleme Frekansı (Hz)	16.000 (16 kHz)	16.000 (16 kHz)
Ses Kanalı / Formatı	Tek Kanallı (Mono) / WAV	Tek Kanallı (Mono) / WAV
Konuşma Yapısı	Akıcı ve sürekli konuşma (cümleler)	İzole sözcükler (tekil komutlar)
Kaynak / Yayıncı	CSTR, Edinburgh Üniversitesi	Kurtkaya, 2025
Dilsel Karakter	Vurgu-zamanlı (stress-timed) dil yapısı	Sondan eklemeli (agglutinative) dil yapısı

3.2.2. Gürültü entegrasyonu ve SNR seviyelerinin belirlenmesi

3.2.2.1. Akustik gürültü kaynakları

Modelin gerçek dünya koşullarındaki gürültü arındırma başarısını simüle etmek amacıyla Microsoft DNS Challenge (Deep Noise Suppression) veri setinden (Reddy vd., 2020) seçilen üç farklı gerçek dünya gürültü türü ile kontrol referansı olarak bir adet sentetik gürültü türü kullanılmıştır. Seçilen gürültülerin özellikleri şu şekildedir:

- 1. Klima Dış Ünitesi Gürültüsü (Gerçek Dünya):** Düşük frekanslı, zamanla spektral özellikleri yavaş değişen durağan karakterde mekanik bir gürültüdür.
- 2. Elektrikli Süpürge Gürültüsü (Gerçek Dünya):** Tüm frekans bandına yayılan, yüksek enerjili ve sürekli gürültü profiline sahiptir.
- 3. Bulaşık Makinesi Gürültüsü (Gerçek Dünya):** Su çarpması ve motor seslerini içeren, orta frekanslarda yoğunlaşan yarı-durağan bir gürültüdür.
- 4. Beyaz Gürültü (Sentetik):** Tüm frekanslarda eşit güç spektral yoğunluğuna sahip, istatistiksel olarak rastgele üretilmiş kontrol referansıdır.

3.2.2.2. Sinyal-Gürültü oranı (SNR) seviyeleri ve modelleme

Modelin farklı gürültü yoğunluklarındaki başarımını test etmek için 0 dB, 5 dB, 10 dB ve 15 dB olmak üzere dört farklı Sinyal-Gürültü Oranı (SNR) seviyesi belirlenmiştir. Bu seviyelerin anlamları Tablo 6'da açıklanmıştır.

Tablo 6. Deneylerde kullanılan SNR seviyeleri ve akustik anlamları

SNR Seviyesi	Gürültü Seviyesi	Akustik Açıklama
0 dB	Çok Yüksek	Konuşma sinyali ile gürültü eşit güçtedir; en zorlu koşuldur.
5 dB	Yüksek	Gürültü konuşmayı belirgin şekilde gölgelemektedir.
10 dB	Orta	Konuşma anlaşılır durumdadır ancak arka plan gürültüsü baskındır.
15 dB	Düşük	Hafif ve rahatsız etmeyen bir arka plan gürültüsü mevcuttur.

Gürültülü seslerin sentezlenmesinde eklemeli (additive) gürültü modeli benimsenmiştir (Denklem 44):

$$y[n] = s[n] + \alpha \cdot d[n] \quad (44)$$

Burada $y[n]$ sentezlenen gürültülü sinyali, $s[n]$ temiz konuşma sinyalini, $d[n]$ gürültü sinyalini ve α ise hedef SNR seviyesini sağlamak için kullanılan ölçekleme katsayısını temsil eder. Ölçekleme katsayısı α , sinyal güçleri üzerinden Denklem (45) ile hesaplanmıştır:

$$\alpha = \sqrt{\frac{P_s}{P_d \cdot 10^{\frac{SNR}{10}}}} \quad (45)$$

Burada P_s temiz konuşma sinyalinin gücünü, P_d ise ham gürültü dosyasının gücünü göstermektedir.

3.2.2.3. Sentezleme ve karıştırma dağılımı

4 farklı gürültü türü ile 4 farklı SNR seviyesinin kombinasyonu ile toplam 16 farklı akustik gürültü koşulu oluşturulmuştur. Veri setlerindeki her bir temiz ses dosyası için bu 16

kombinasyondan biri rastgele seçilerek eşleştirilmiş ve gürültülü karşılıkları üretilmiştir. Rastgele eşleştirme stratejisi, modelin belirli bir gürültü türüne veya SNR seviyesine aşırı uyum sağlamasını engelleyerek genelleme kabiliyetini artırmaktadır.

3.2.3. Veri ön işleme, parçalama ve doldurma stratejileri

3.2.3.1. Ses dosyası okuma ve standardizasyon

Tüm ses kayıtları ön işleme aşamasında torchaudio kütüphanesi kullanılarak bilgisayar ortamına aktarılmıştır. Bu aşamada uygulanan standardizasyon işlemleri şunlardır:

- Farklı örnekleme frekanslarına sahip kayıtlar 16 kHz örnekleme hızına yeniden örneklendirilmiştir (resampling).
- Çok kanallı (stereo) kayıtlar, kanalların matematiksel ortalaması alınarak tek kanallı (mono) biçime dönüştürülmüştür.

3.2.3.2. Sabit uzunluklu parçalama ve doldurma

Evrişimli sinir ağlarının mini-yığın (mini-batch) tabanlı eğitiminde, girdi boyutlarının sabit olması gerekmektedir. Zaman alanında çalışan modelin girdisini standartlaştırmak için tüm ses dosyaları 1 saniyelik (16000 örnek) pencerelere ayrılmıştır.

- **Uzun Sinyaller (≥ 16.000 örnek):** Eğitim esnasında, sinyalin içerisinden rastgele seçilen bir başlangıç noktasından itibaren 16.000 örneklik (1 saniye) bir kesit alınır. Bu rastgele kırpma işlemi, modelin eğitimi boyunca her epoch'ta farklı pencereleri görmesini sağlayarak bir veri artırma (data augmentation) etkisi yaratır.
- **Kısa Sinyaller (< 16.000 örnek):** 16.000 örnek sınırına ulaşamayan kısa seslerin sonuna, sinyal boyutunu tam 16.000'e tamamlayacak şekilde sıfır ekleme (zero-padding) işlemi uygulanmıştır.

3.2.4. Çoklu rastgele bölme ve karma eğitim düzeni

3.2.4.1. Veri bölme oranları

Eğitim, doğrulama ve test setleri oluşturulurken %70 eğitim, %15 doğrulama ve %15 test oranı kullanılmıştır. İngilizce (VCTK), Türkçe (TSCD) ve her iki dilin eşit oranda birleştirilmesiyle oluşturulan karma veri setlerinin dağılım sayıları Tablo 7’de verilmiştir.

Tablo 7. Eğitim, doğrulama ve test setlerinin veri dağılımı

Deney Konfigürasyonu	Eğitim Seti	Doğrulama Seti	Test Seti	Toplam Dosya Sayısı
İngilizce (ENG)	31.119	6.668	6.668	44.455
Türkçe (TUR)	18.540	3.972	3.972	26.484
Karma (MIX)	49.659	10.640	10.640	70.939

3.2.4.2. Çoklu rastgele bölme protokolü

Veri setinin rastgele bölünmesinden kaynaklanabilecek istatistiksel sapmaları önlemek ve modelin başarımlı kararlılığını doğrulamak amacıyla çoklu bölme stratejisi uygulanmıştır:

- Tüm deneylerde test veri seti tamamen sabit tutulmuştur. Amaç her örneği sırayla test etmek değil, ortak ve değişmez bir değerlendirme hedefi üzerinden modelin eğitim rastgeleliğine karşı kararlılığını izole ederek ölçmektir.
- Eğitim ve doğrulama setleri, veri tabanından rastgele seçilerek 3 farklı bölünme kombinasyonu (Data 1, Data 2, Data 3) oluşturulacak şekilde 3 kez tekrarlanmıştır.
- Modeller bu 3 farklı bölünme için bağımsız olarak eğitilmiş; nihai başarımlı ölçümleri bu 3 çalışmanın ortalaması ve standart sapması olarak raporlanmıştır.

3.2.4.3. Karma eğitim yapılandırması

Tez çalışmasında, sadece tek bir dilde eğitilen modellerin sınırlılıklarını aşmak amacıyla İngilizce ve Türkçe verilerin bir arada eğitildiği karma veri seti oluşturulmuştur. Bu veri seti modelin diller arası akustik transfer kabiliyeti ve tek bir modelin iki farklı dilin fonetik özelliklerini (izole kelimeler ve sürekli cümleler) aynı anda öğrenme başarımlı analiz edilmiştir.

3.2.5. Eğitim hiperparametreleri, donanım ve yazılım altyapısı

3.2.5.1. Hiperparametre yapılandırması

Modellerin eğitiminde kullanılan hiperparametrelerin detayları Tablo 8'de özetlenmiştir:

Tablo 8. Geliştirilmiş Wave-U-Net modeli eğitim hiperparametreleri

Hiperparametre	Tercih Edilen Değer
Optimizasyon Algoritması	AdamW
Başlangıç Öğrenme Oranı (Learning Rate)	6×10^{-4}
Yığın Boyutu (Batch Size)	24
Maksimum Dönem (Epoch) Sayısı	100
Öğrenme Oranı Çizelgeleyici (Scheduler)	ReduceLROnPlateau
Çizelgeleyici İzleme Sabrı (Patience)	10 Epoch
Öğrenme Oranı Düşürme Katsayısı (Factor)	0,5
Erken Durdurma Sabrı (Early Stopping)	15 Epoch
Karma Kayıp Katsayıları	$\alpha=0,25$, $\beta=0,35$, $\gamma=0,40$

3.2.5.2. Donanım ve yazılım altyapısı

Tüm eğitim ve test süreçleri yüksek hesaplama gücüne sahip bulut tabanlı sanal bir sunucu üzerinde gerçekleştirilmiştir. Donanım ve yazılım detayları sırasıyla Tablo 9 ve Tablo 10'da sunulmuştur:

Tablo 9. Deneysel çalışmalarda kullanılan donanım altyapısı

Donanım Bileşeni	Sistem Özelliği
Grafik İşlem Birimi (GPU)	L4 GPU
Grafik Belleği (VRAM)	22 GB
İşlemci Platformu	COLAB PRO

Tablo 10. Kullanılan yazılım ve kütüphane detayları

Yazılım / Kütüphane Adı	Sürüm / Kullanım Amacı
Python	3.10 / Ana programlama dili ortamı
PyTorch	2.1.2 / Derin öğrenme mimarisinin kurulması ve eğitimi
torchaudio	Ses dosyalarının okunması, resample ve kanal işlemleri
pesq	Geniş bant PESQ (P.862.2) metrik hesaplamaları
pystoi	STOI ve ESTOI metriklerinin hesaplanması
librosa	Spektrogram görselleştirme ve ses analizi

Yazılım / Kütüphane Adı	Sürüm / Kullanım Amacı
scipy	Sayısal sinyal işleme filtreleri ve LPC katsayı analizleri
matplotlib	Eğitim eğrileri ve dalga formu analiz grafiklerinin çizimi

Tablo 10. (Devamı)

3.2.6. Nesnel başarımlı değerlendirme protokolü

Modelin iyileştirdiği konuşma sinyallerinin kalitesini ve anlaşılabilirliğini değerlendirmek amacıyla nesnel bir test protokolü yürütülmüştür. Test setinde yer alan her bir gürültülü ses dosyası için sırasıyla şu adımlar izlenmiştir:

1. Gürültülü ses dosyası, zaman alanında ham dalga formu olarak modele girdi olarak verilir.
2. Girdi sinyali, eğitim aşamasına paralel olacak şekilde 16.000 örneklik pencereler halinde parça parça işlenir.
3. Model tarafından iyileştirilen pencereler zaman ekseninde uç uca birleştirilir ve orijinal sinyal uzunluğuna göre kırılarak tam uzunlukta iyileştirilmiş sinyal elde edilir.
4. Elde edilen iyileştirilmiş sinyal, temiz referans sinyal ve gürültülü giriş sinyali karşılaştırılarak nesnel metrikler (PESQ, STOI, ESTOI, SI-SNR, CSIG, CBAK ve COVL) hesaplanır.
5. Test kümesindeki tüm dosyalar üzerinden elde edilen metrik skorlarının aritmetik ortalaması ve standart sapma değerleri raporlanır.

Değerlendirme protokolünde her metrik için hem gürültülü giriş sinyalinin skoru (noisy baseline) hem de model çıktısının skoru (enhanced) hesaplanmış; aralarındaki farkı gösteren başarımlı kazanımı ($\Delta = \text{enhanced} - \text{noisy}$) modelin başarımlı ölçmek amacıyla temel ölçüt olarak benimsenmiştir.

4. BULGULAR

Bu bölümde, önerilen Geliştirilmiş Wave-U-Net mimarisinin ve EnhancedLoss fonksiyonunun nesnel başarımı sunulmuştur.

Modelin zaman alanında ham dalga formunu işleme yeteneği, nesnel kalite ve anlaşılabilirlik metrikleriyle ölçülmüştür. Tüm ölçümler, modellerin eğitimde görmediği bu sabit test setleri üzerinden alınmıştır.

4.1. Model Bileşenlerinin ve Kayıp Fonksiyonlarının Ablasyon Analizi

Eklenen bileşenlerin (artık bağlantı, kanal dikkati, öz-dikkat) ve EnhancedLoss fonksiyonunun katkısını ölçmek için Türkçe veri seti üzerinde bir ablasyon çalışması yürütülmüştür. Hem geleneksel L1 kaybı hem de önerilen EnhancedLoss, farklı bileşenlerle birlikte değerlendirilip, sonuçlar Tablo 11'de verilmiştir.

Tablo 11. Model bileşenlerinin ve kayıp fonksiyonlarının birleşik ablasyon sonuçları

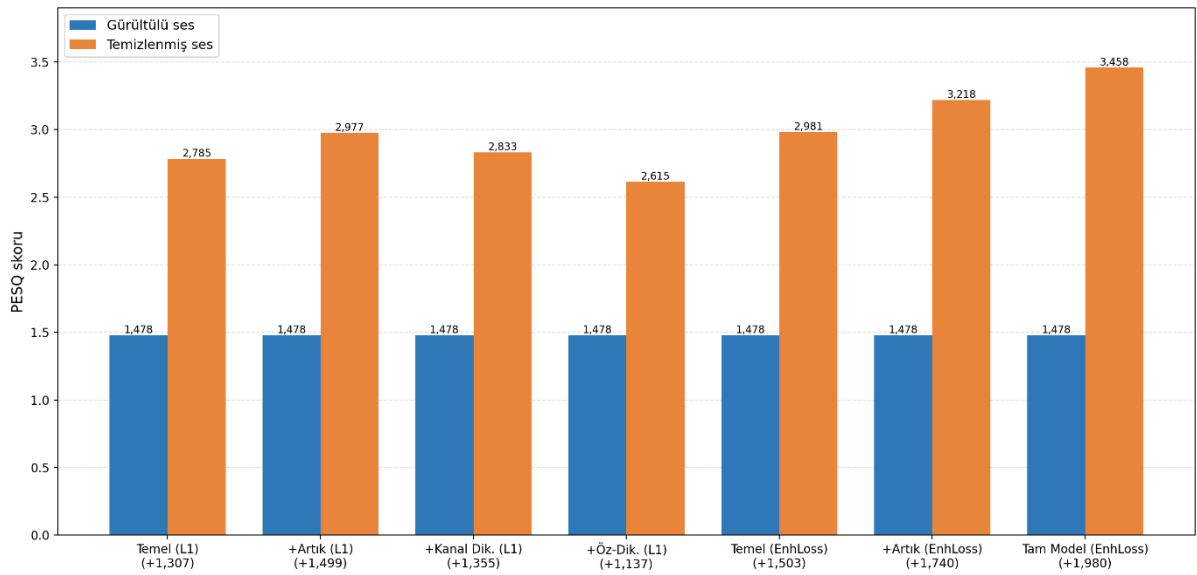
Konfigürasyon	Kayıp Fonksiyonu	Parametre	PESQ	SI-SNR (dB)	Δ PESQ	Δ SI-SNR
Gürültülü (Giriş)	—	—	1,478	7,5	—	—
Temel (Baseline)	L1	12,13 M	2,785	24,21	1,307	16,71
+ Artık Bağlantı	L1	26,54 M	2,977	26,37	1,499	18,87
+ Kanal Dikkati	L1	26,72 M	2,833	25,91	1,355	18,41
+ Öz-Dikkat	L1	27,04 M	2,615	24,17	1,137	16,67
Temel (Baseline)	EnhancedLoss	12,13 M	2,981	27,82	1,503	20,32
+ Artık Bağlantı	EnhancedLoss	26,54 M	3,218	31,25	1,74	23,75
Tam Model	EnhancedLoss	27,04 M	3,458	35,08	1,98	27,58

Tablo 11'deki Δ değerleri gürültülü girişe göre hesaplanmıştır. Ablasyon sonuçları şöyle özetlenebilir:

- Kayıbın tek başına etkisi:** Yalnızca L1 kaybıyla eğitilen taban model 2,785 PESQ ve 24,21 dB SI-SNR verir. Aynı model sadece kayıp EnhancedLoss ile değiştirilerek eğitildiğinde PESQ 2,981'e ($\Delta=1,503$), SI-SNR 27,82 dB'ye yükselir. Kayıp fonksiyonu, ek mimari bileşen olmadan da başarımı artırmaktadır.
- Artık bağlantı:** Her iki kayıpta da başarımı yükseltmektedir. L1 altında PESQ 2,977'ye (+1,499), EnhancedLoss altında 3,218'e (+1,740) çıkar.

- **L1 altında dikkat:** L1 kaybıyla kanal dikkati eklendiğinde başarımlar 2,833'e, öz-dikkat de eklenince 2,615'e geriler. Yani dikkat modülleri L1 altında PESQ'yi düşürür.
- **Tam model + EnhancedLoss:** Tüm modüllerin açık olduğu tam model, EnhancedLoss ile en yüksek değerlere ulaşır: PESQ 3,458 (+1,980), SI-SNR 35,08 dB (+27,58 dB).

Burada Öz-Dikkat ile Tam Model'in parametre sayısının aynı olmasıdır (27,04 M); aralarındaki tek fark kayıp fonksiyonudur. Aynı mimari, L1 altında 2,615 PESQ verirken EnhancedLoss altında 3,458'e çıkar. Şekil 15'te ablasyon deneyindeki PESQ değerleri konfigürasyonlara göre karşılaştırılmaktadır.



Şekil 15. Ablasyon sonuçlarının PESQ grafiği

4.2. Dil Bazlı ve Çok Dilli Eğitim Performansı

Önerilen geliştirilmiş Wave-U-Net modeli sırasıyla İngilizce veri seti (VCTK), Türkçe veri seti (TSCD) ve her iki dilin birleşiminden oluşan karma veri setleri üzerinde değerlendirilmiştir. Her veri seti için rastgele oluşturulmuş Data 1, Data 2 ve Data 3 veri setlerine ait detaylı sonuçlar aşağıda gösterilmiştir.

4.2.1. İngilizce veri seti sonuçları

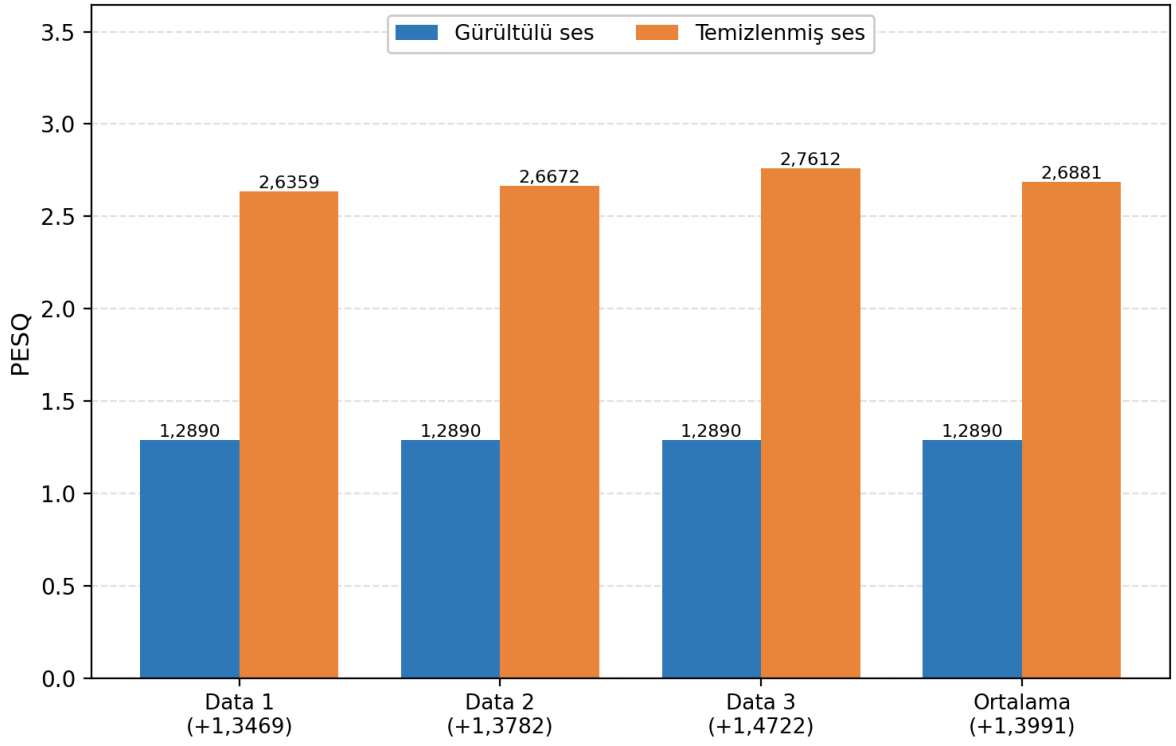
Sürekli konuşma ve akıcı cümlelere sahip İngilizce veri seti kullanılarak eğitilen modelin 3 farklı rastgele oluşturulmuş veri setlerine ait sonuçlar Tablo 12'de detaylandırılmıştır.

Tablo 12. İngilizce rastgele oluşturulmuş veri setlerinin detaylı sonuçları

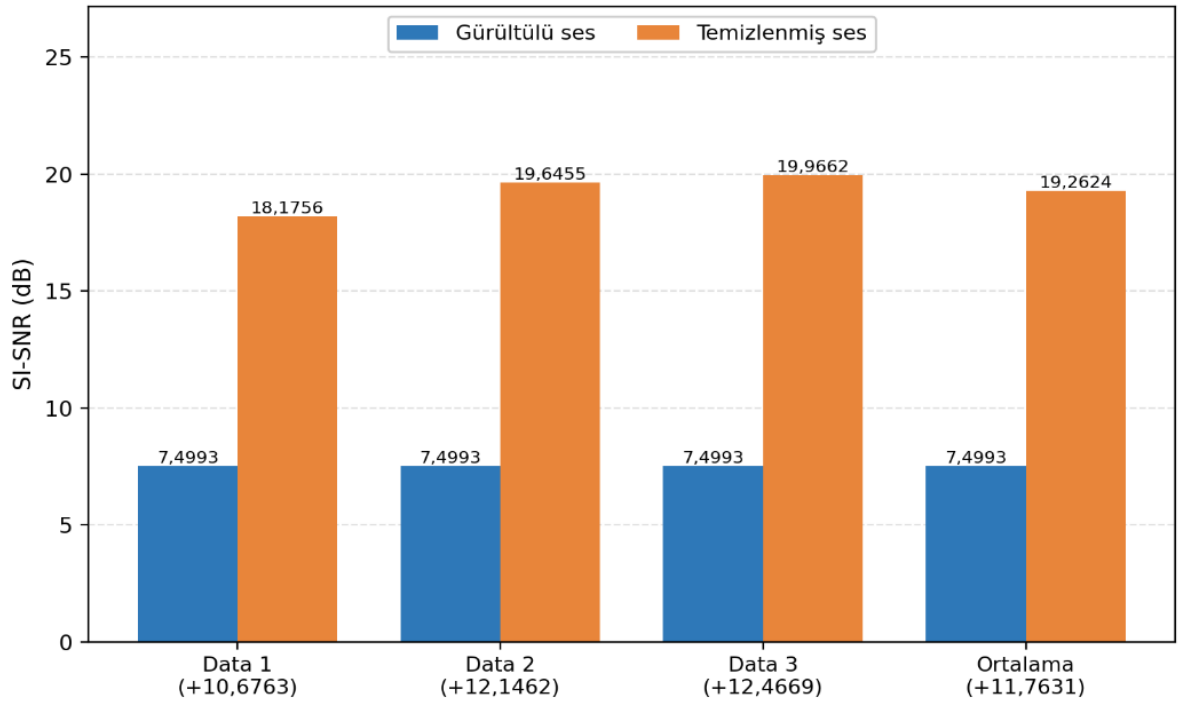
Veri Seti	PESQ	STOI	ESTOI	SI-SNR (dB)	CSIG	CBAK	COVL
Gürültülü (Giriş)	1,2890	0,7262	0,5238	7,4993	2,9732	2,4169	2,5411
Data1	2,6359	0,8534	0,769	18,1756	4,4493	2,7959	4,0188
Data2	2,6672	0,8582	0,7709	19,6455	4,2948	4,139	3,9417
Data3	2,7612	0,8609	0,7764	19,9662	4,4416	2,8129	4,0557
Ortalama	2,6881	0,8575	0,7721	19,2624	4,3952	3,2493	4,0054
± Standart Sapma	±0,053	±0,003	±0,003	±0,780	±0,071	±0,629	±0,047

İngilizce sonuçlarına ilişkin bulgular şu şekildedir:

- Model, gürültülü sinyalin 1,2890 olan ortalama PESQ skorunu 2,6881 seviyesine yükseltip, ortalama +1,399 puanlık bir algısal kalite kazanımı sağlamıştır.
- Zaman alanındaki gürültü arındırma başarısını gösteren SI-SNR metriğinde, gürültülü girişin 7,4993 dB olan değeri ortalama 19,2624 dB'ye çıkarak +11,763 dB kazanım elde edilmiştir.
- Konuşma sinyalindeki bozulmayı ölçen CSIG değeri gürültülü girişteki 2,9732 seviyesinden 4,3952 seviyesine yükselerek +1,422 kazanım elde etmiştir. 4,0 eşiğinin üzerindeki bu yüksek CSIG skoru, modelin gürültüyü temizlerken temiz konuşma sinyalinde herhangi bir kesilme veya yapay bozulma üretmediğini göstermektedir. Arka plan gürültüsünün bastırılma kalitesini ölçen CBAK skoru 2,4169'dan 3,2493 seviyesine çıkmış; genel algısal ses kalitesini öngören COVL değeri ise 2,5411'den 4,0054 seviyesine yani kullanıcı memnuniyeti açısından kabul edilebilir olan "iyi" kalite bandına (>4,0) ulaşmıştır.



Şekil 16. İngilizce veri seti rastgele bölme bazlı PESQ değerleri



Şekil 17. İngilizce veri seti rastgele bölme bazlı SI-SNR değerleri

Şekil 16'da İngilizce veri setinde üç rastgele bölmenin ve ortalamanın PESQ değerleri gösterilmektedir. Şekil 17'de İngilizce veri setinde bölme bazlı SI-SNR değerleri verilmektedir.

4.2.2. Türkçe veri seti sonuçları

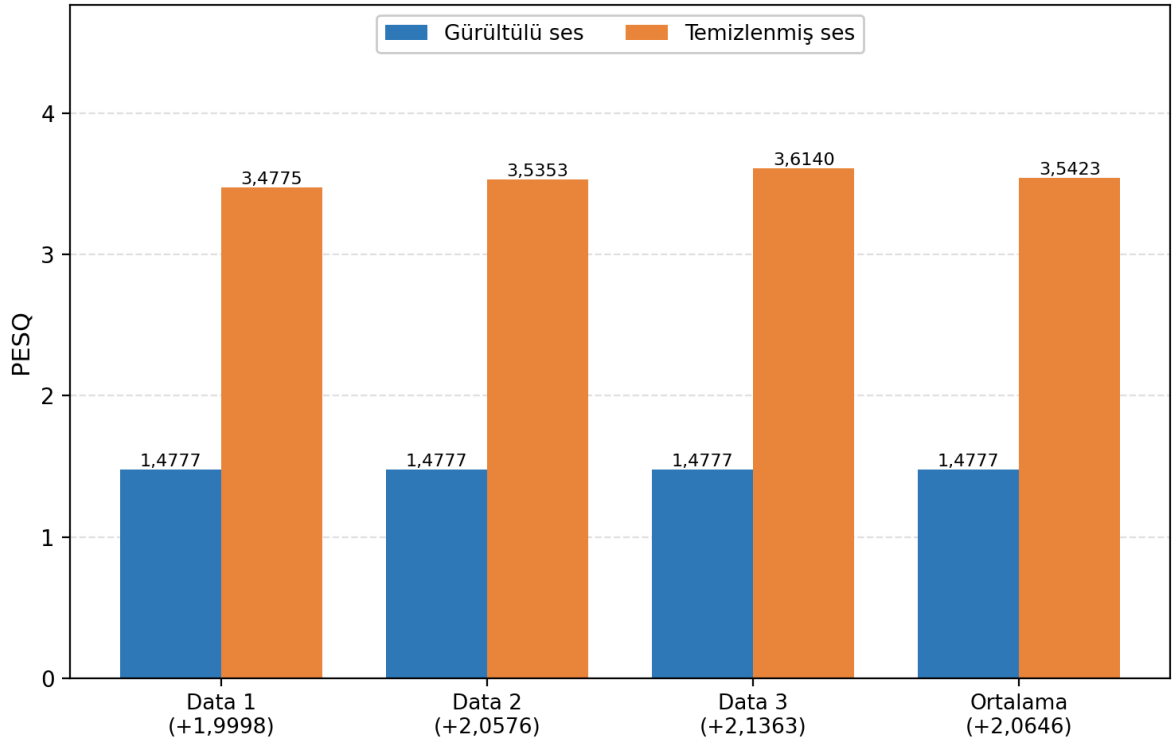
İzole kelime yapısına sahip Türkçe veri seti üzerinde eğitilen model, tüm test senaryoları arasında en yüksek başarımlar ve kazanım oranlarının elde etmiştir. Üç farklı rastgele oluşturulmuş veri setlerine ait sonuçlar Tablo 13'te detaylandırılmıştır.

Tablo 13. Türkçe rastgele oluşturulmuş veri setlerinin detaylı sonuçları

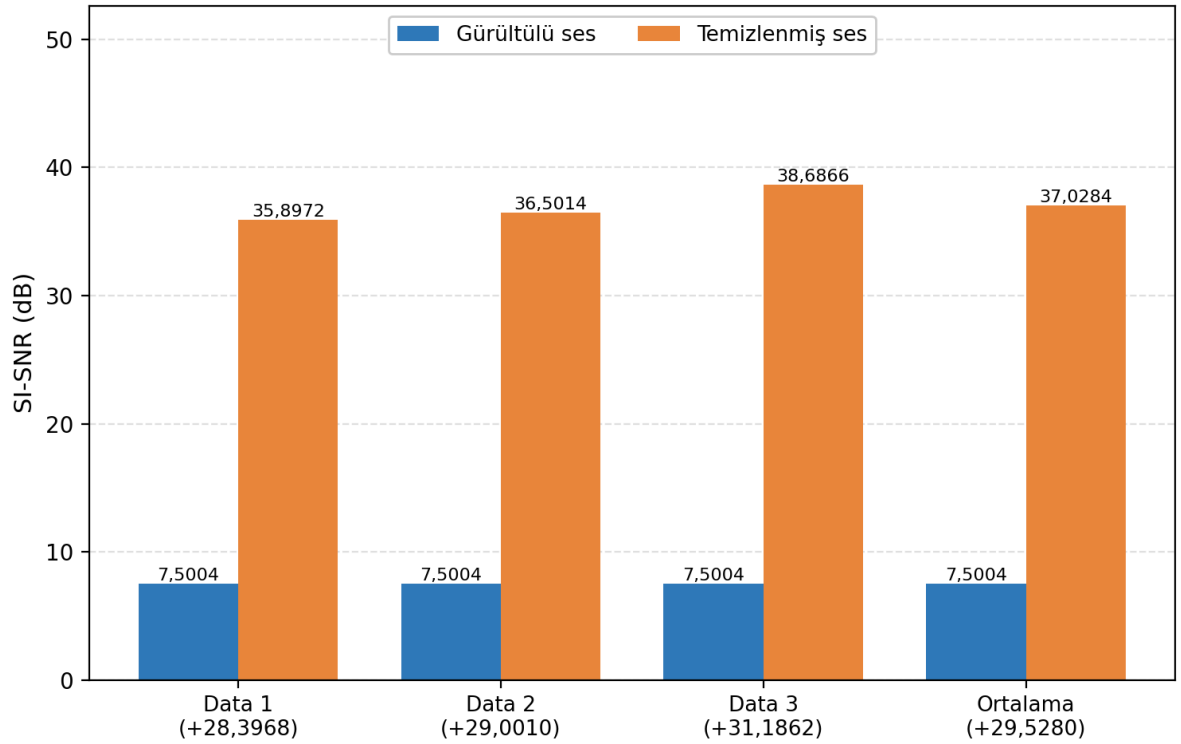
Veri Seti	PESQ	STOI	ESTOI	SI-SNR (dB)	CSIG	CBAK	COVL
Gürültülü (Giriş)	1,4777	0,7751	0,5995	7,5004	3,2848	2,5512	2,8061
Data 1	3,4775	0,8454	0,7926	35,8972	4,3691	3,0197	4,266
Data 2	3,5353	0,8441	0,7923	36,5014	4,395	3,0393	4,298
Data 3	3,614	0,8465	0,7957	38,6866	4,4181	4,2212	4,3192
Ortalama $\pm$ Standart	3,5423	0,8453	0,7935	37,0284	4,3941	3,4267	4,2944
Sapma	$\pm 0,056$	$\pm 0,001$	$\pm 0,002$	$\pm 1,198$	$\pm 0,020$	$\pm 0,562$	$\pm 0,022$

Türkçe sonuçlarına ilişkin bulgular şu şekildedir:

- Gürültülü girişin 1,4777 olan ortalama PESQ değeri, iyileştirme sonrasında 3,5423 seviyesine ulaşarak +2,065 puanlık çok yüksek bir artış göstermiştir.
- SI-SNR metriği ise başlangıçtaki 7,5004 dB değerinden 37,0284 dB seviyesine çıkarak +29,528'lik üstün bir gürültü arındırma kazanımı sağlamıştır.
- Türkçe veri setinde konuşma kalitesindeki iyileşme bileşik metriklerinde de görülür. CSIG skoru 3,2848'den 4,3941'e yükselerek konuşma sinyalinde bozulma başarılı bir sentez kalitesi elde edildiğini gösterir. Gürültü bastırma yeteneğini temsil eden CBAK skoru 2,5512'den 3,4267 seviyesine çıkmış; genel algısal ses kalitesi metriği olan COVL ise 2,8061'den 4,2944 seviyesine ulaşarak +1,488 puanlık net bir artış yakalamıştır. 4,29 seviyesindeki COVL ortalaması, modelin Türkçe konuşma sinyalini insan kulağını rahatsız etmeyecek derecede başarılı temizlediğini göstermektedir.



Şekil 18. Türkçe veri seti rastgele bölme bazlı PESQ değerleri



Şekil 19. Türkçe veri seti rastgele bölme bazlı SI-SNR değerleri

Şekil 18'de Türkçe veri setinde bölme bazlı PESQ değerleri gösterilmektedir. Şekil 19'da Türkçe veri setinde bölme bazlı SI-SNR değerleri verilmektedir.

4.2.3. Karma veri seti sonuçları

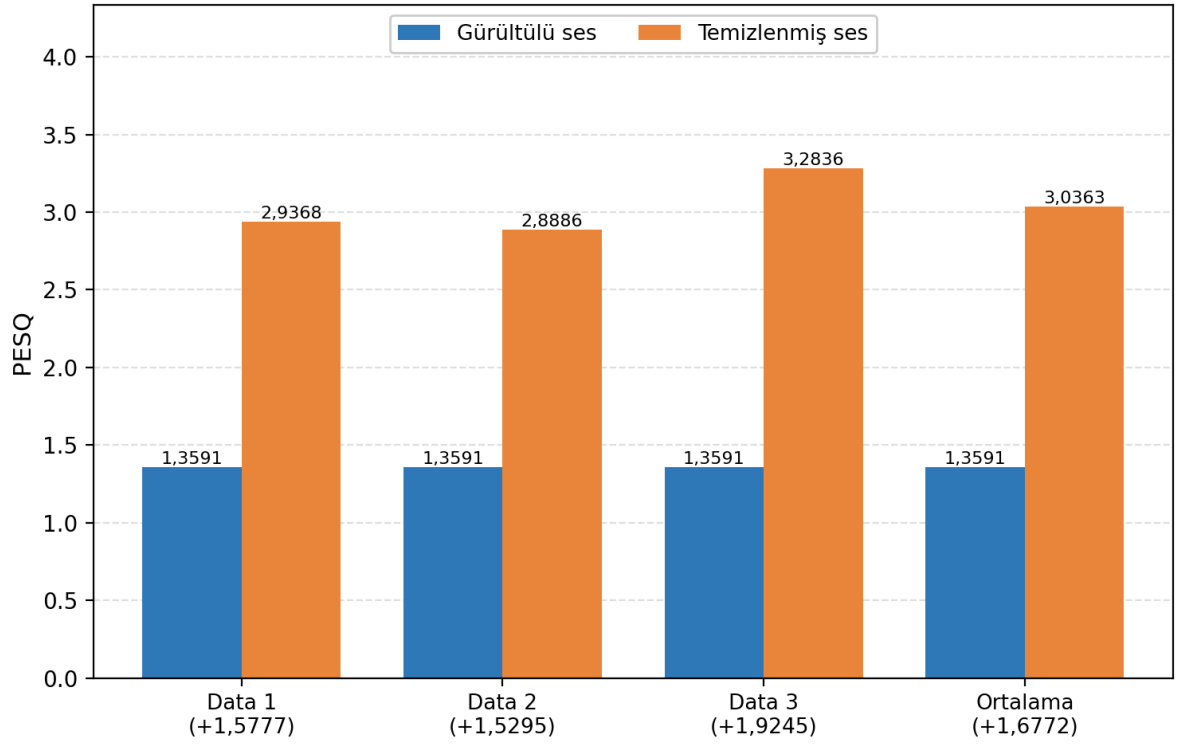
İngilizce sürekli cümleler ile Türkçe izole sözcüklerin eşit oranda birleştirildiği Karma dil yapılandırması, modelin çok dilli ve farklı fonetik yapılarıdaki konuşma sinyallerini işleme yeteneğini test etmek amacıyla kurgulanmıştır. Üç farklı rastgele oluşturulmuş veri setlerine ait sonuçlar Tablo 14'te gösterilmiştir.

Tablo 14. Karma dillerden rastgele oluşturulmuş veri setlerinin detaylı sonuçları

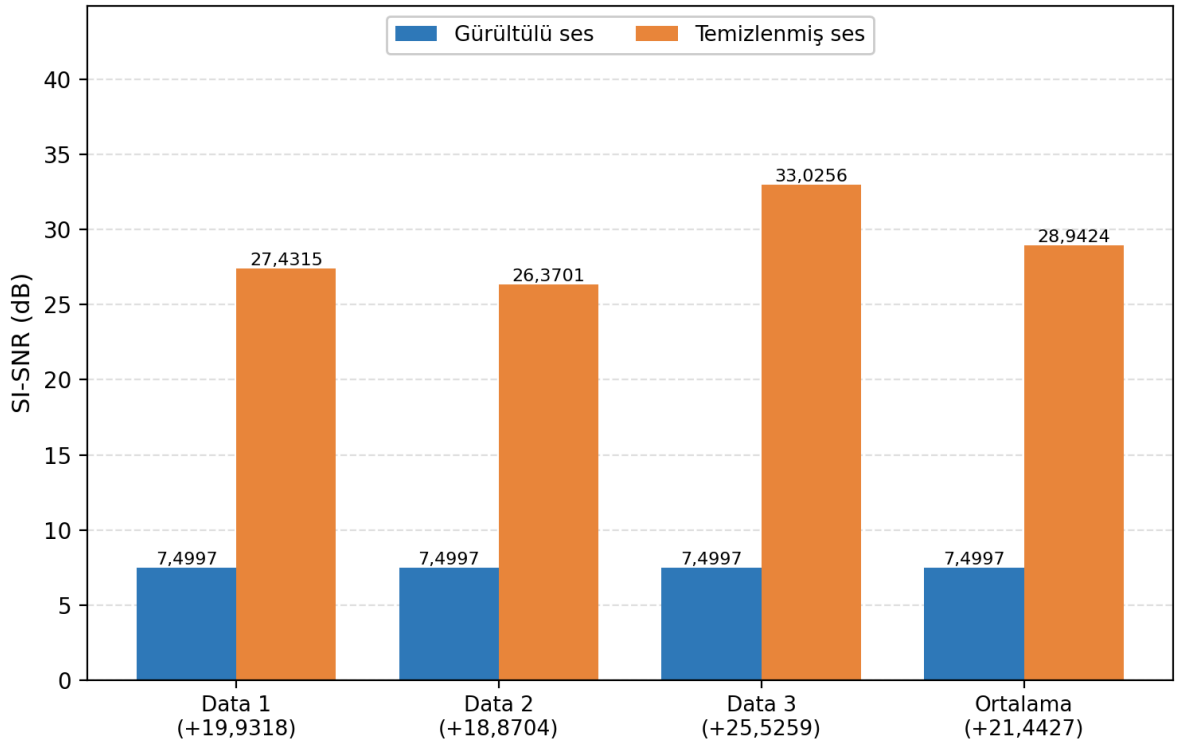
Veri Seti	PESQ	STOI	ESTOI	SI-SNR (dB)	CSIG	CBAK	COVL
Gürültülü (Giriş)	1,3591	0,7445	0,5521	7,4997	3,0887	2,4667	2,6393
Data 1	2,9368	0,8541	0,7793	27,4315	4,3932	2,8980	4,0843
Data 2	2,8886	0,8518	0,7723	26,3701	4,3188	4,0837	4,0180
Data 3	3,2836	0,8481	0,7825	33,0256	4,3902	2,9726	4,2031
Ortalama $\pm$ Standart Sapma	3,0363 $\pm 0,176$	0,8514 $\pm 0,002$	0,7780 $\pm 0,004$	28,9424 $\pm 2,920$	4,3674 $\pm 0,034$	3,3181 $\pm 0,542$	4,1018 $\pm 0,077$

Karma dil sonuçlarına ilişkin bulgular şu şekildedir:

- Çok dilli model, karma test setinde gürültülü girişin 1,3591 olan PESQ değerini ortalama 3,0363 seviyesine yükselterek +1,677 puanlık kazanım elde etmiştir.
- SI-SNR değeri ise 7,4997 dB'den 28,9424 dB seviyesine ulaşarak +21,443 dB'lik bir artış sağlamıştır.
- Karma veri seti için CSIG skoru 4,3674 seviyesine ulaşmış; CBAK skoru ise 2,4667'den 3,3181 seviyesine yükselmiştir. COVL metriği ise 2,6393 olan gürültülü giriş seviyesinden 4,1018 değerine ulaşarak önemli bir gelişim kaydetmiştir. Çok dilli eğitim veri tabanındaki yüksek varyansa rağmen COVL skorunun 4,10 seviyesinde tutunması, modelin iki farklı dilin fonetik özelliklerini öğrenirken sinyal doğallığından ödün vermediğini göstermektedir.



Şekil 20. Karma veri seti rastgele bölme bazlı PESQ değerleri



Şekil 21. Karma veri seti rastgele bölme bazlı SI-SNR değerleri

Şekil 20'de Karma veri setinde bölme bazlı PESQ değerleri gösterilmektedir. Şekil 21'de Karma veri setinde bölme bazlı SI-SNR değerleri verilmektedir.

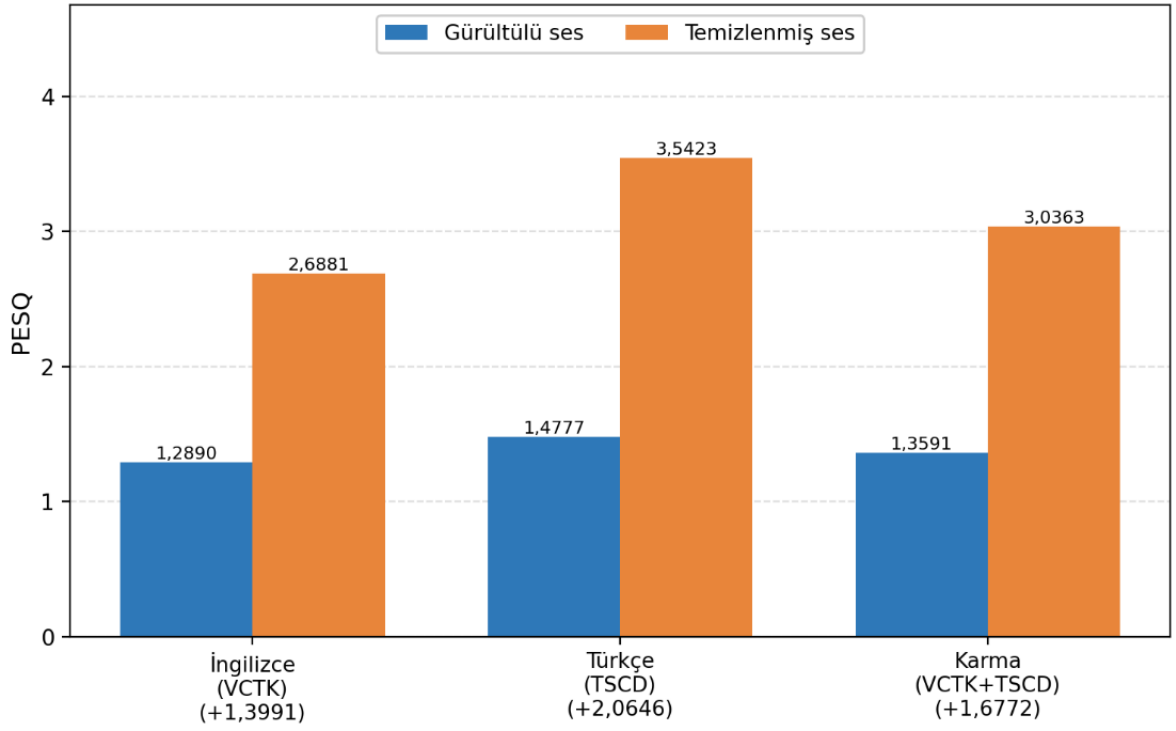
4.2.4. Genel dil performansı karşılaştırması

Üç farklı dil koşuluna ait gürültülü giriş referans değerleri ile 3 bağımsız rastgele veri bölmesinden elde edilen ortalama model başarımları Tablo 4.5'te bir arada sunulmuştur. Bu ortak tablo, satırlarda deney koşullarını ve dilleri, sütunlarda ise nesnel başarımların metriklerini gösterecek şekilde yapılandırılmıştır.

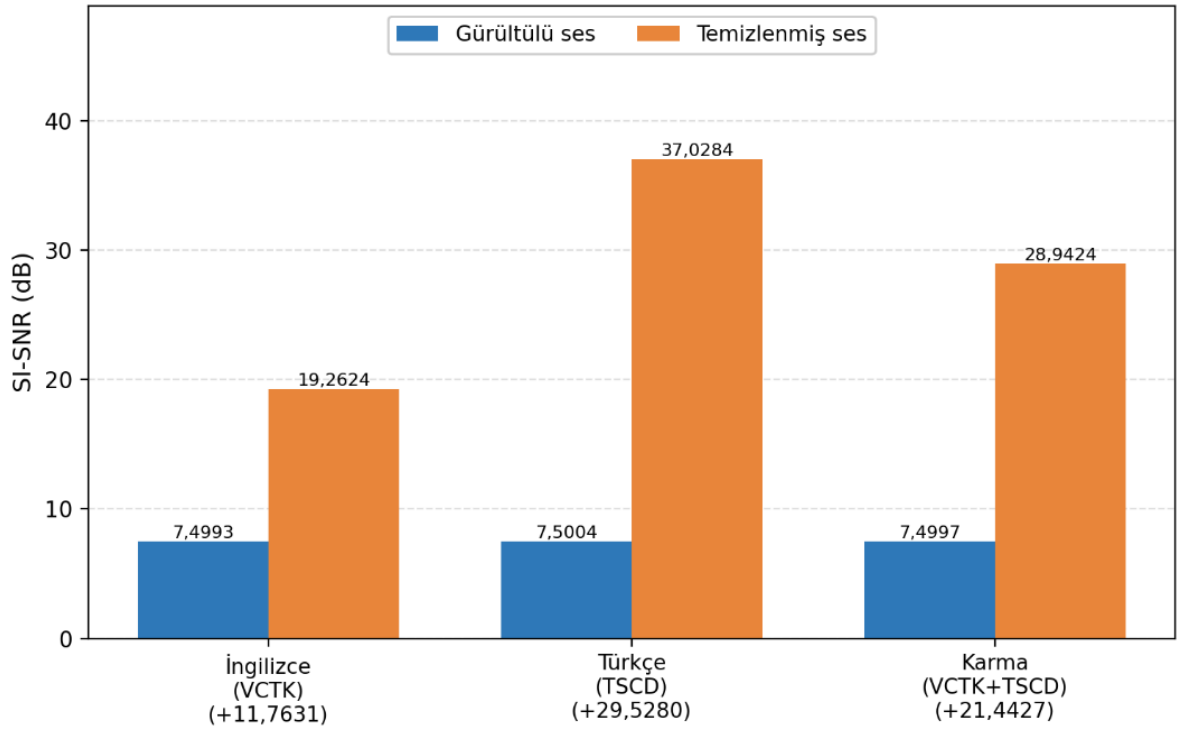
Tablo 15. Üç farklı veri setinin ortak genel performans ve kazanım özeti

Veri Seti	PESQ	STOI	ESTOI	SI-SNR (dB)	CSIG	CBAK	COVL
İngilizce- Gürültülü	1,2890	0,7262	0,5238	7,4993	2,9732	2,4169	2,5411
İngilizce- İyileştirilmiş	2,6881 ±0,053	0,8575 ±0,003	0,7721 ±0,003	19,2624 ±0,780	4,3952 ±0,071	3,2493 ±0,629	4,0054 ±0,047
Türkçe- Gürültülü	1,4777	0,7751	0,5995	7,5004	3,2848	2,5512	2,8061
Türkçe- İyileştirilmiş	3,5423 ±0,056	0,8453 ±0,001	0,7935 ±0,002	37,0284 ±1,198	4,3941 ±0,020	3,4267 ±0,562	4,2944 ±0,022
Karma- Gürültülü	1,3591	0,7445	0,5521	7,4997	3,0887	2,4667	2,6393
Karma- İyileştirilmiş	3,0363 ±0,176	0,8514 ±0,002	0,7780 ±0,004	28,9424 ±2,920	4,3674 ±0,034	3,3181 ±0,542	4,1018 ±0,077

Ortak performans tablosu analiz edildiğinde model, yedi metriğin tamamında ve üç koşulun her birinde ölçülebilir kazanım sağladı. En yüksek iyileştirilmiş değerler Türkçe koşulda elde edildi; üç koşulun başlangıç SI-SNR değeri neredeyse eşitken (~7,5 dB), iyileştirme sonrası İngilizce'de 19,26 dB'ye, Karma'da 28,94 dB'ye, Türkçe'de 37,02 dB'ye çıktığı gözlenmektedir.



Şekil 22. Dil bazlı PESQ karşılaştırması (kazanç Δ)



Şekil 23. Dil bazlı SI-SNR karşılaştırması (kazanç Δ)

Şekil 22'de üç dil koşulunun PESQ değerleri, gürültülü girişe göre kazançlarıyla birlikte karşılaştırılmaktadır. Şekil 23'te üç dil koşulunun SI-SNR değerleri kazançlarıyla birlikte karşılaştırılmaktadır.

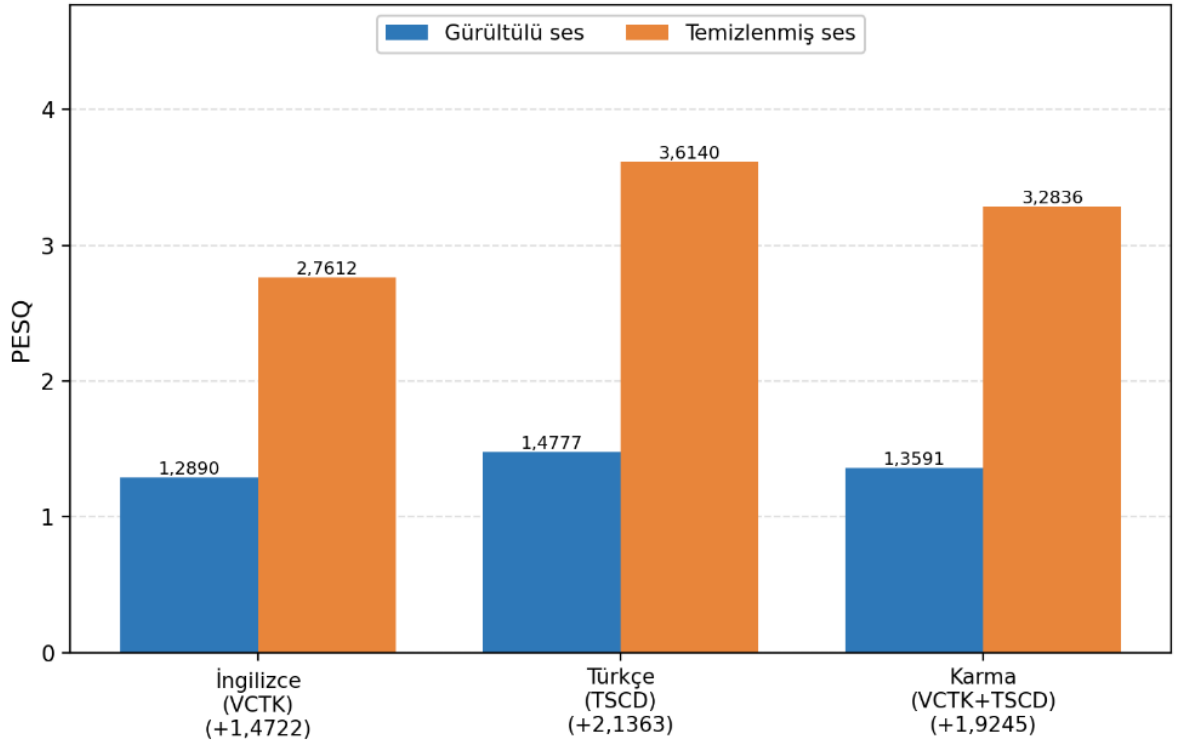
4.3. En İyi Model Ağırlıklarının Diller Arası Performans Kıyası

Her dilden oluşturulan rastgele veri setlerinden en iyi sonucu veren Data 3 veri setinden elde edilen model ağırlıkları, test setleri üzerinde doğrudan birbiriyle kıyaslanarak modelin diller arası diller arası aktarılabilirliği ve başarımı analiz edilmiştir.

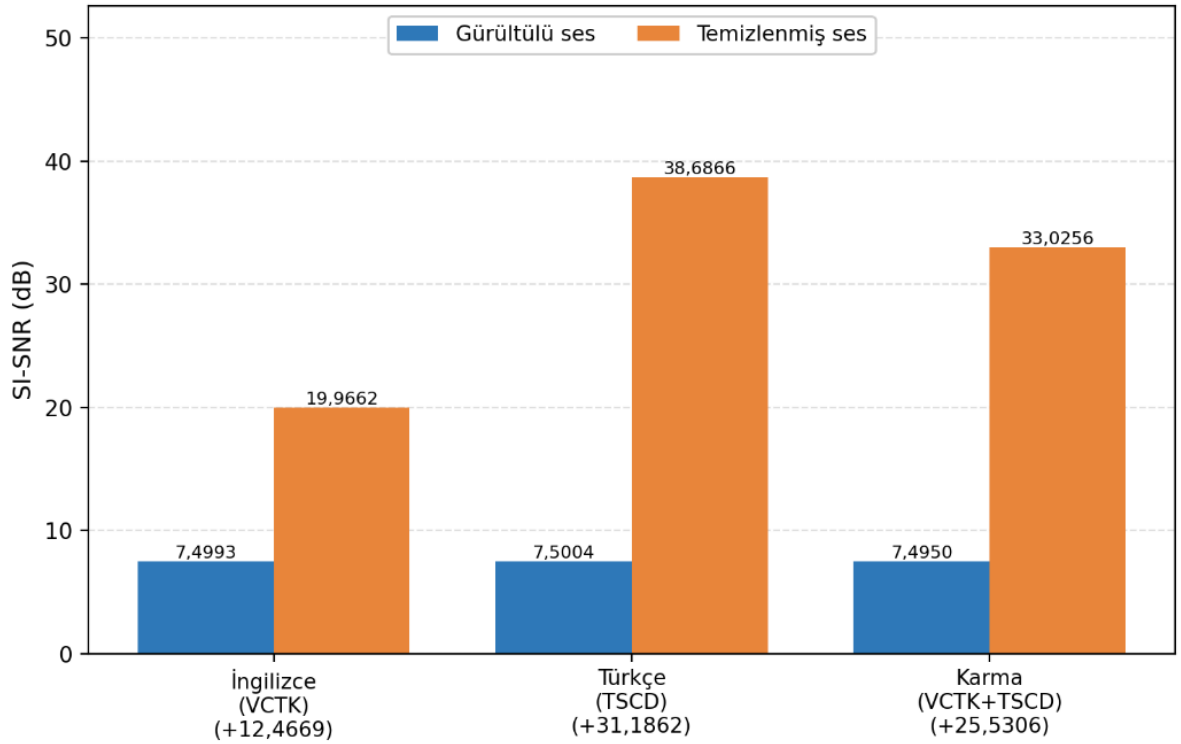
Tablo 16. En iyi veri seti sonuçlarına göre diller arası karşılaştırma ve kazanımlar

Veri Seti (Data 3)	PESQ	STOI	ESTOI	SI-SNR (dB)	CSIG	CBAK	COVL
İngilizce Gürültülü	1,289	0,7262	0,5238	7,4993	2,9732	2,4169	2,5411
İngilizce İyileştirilmiş	2,7612	0,8609	0,7764	19,9662	4,4416	2,8129	4,0557
İngilizce Kazanımı (Δ)	1,4722	0,1347	0,2526	12,4669	1,4684	0,396	1,5146
Türkçe Gürültülü	1,4777	0,7751	0,5995	7,5004	3,2848	2,5512	2,8061
Türkçe İyileştirilmiş	3,614	0,8465	0,7957	38,6866	4,4181	4,2212	4,3192
Türkçe Kazanımı (Δ)	2,1363	0,0714	0,1962	31,1862	1,1333	1,67	1,5131
Karma Gürültülü	1,4211	0,7526	0,5693	7,495	3,2036	2,5247	2,7392
Karma İyileştirilmiş	3,2836	0,8481	0,7825	33,0256	4,3902	2,9726	4,2031
Karma Kazanımı (Δ)	1,8625	0,0955	0,2132	25,5306	1,1866	0,4479	1,4639

En iyi modellerden elde edilen verilere bakıldığında en yüksek PESQ Türkçe modelinde elde edildi (3,6140; +2,136). İngilizce modeli 2,7612'de (+1,472), Karma model 3,2836'da (+1,863) kaldı. CSIG değerleri üç modelde de 4,39–4,44 aralığında birbirine yakındır. CBAK'ta Türkçe (4,2212) belirgin biçimde öndedir; İngilizce (2,8129) ve Karma (2,9726) daha düşüktür. SI-SNR kazanımı Türkçe'de +31,186 dB ile İngilizce'nin (+12,467 dB) iki katından fazladır. Karma model, PESQ ve SI-SNR'de iki tek dilli modelin arasında konumlanır.



Şekil 24. En iyi model Data 3 veri seti ile diller arası PESQ karşılaştırması



Şekil 25. En iyi model Data 3 veri seti ile diller arası PESQ karşılaştırması

Şekil 24'te en iyi bölme (Data 3) ağırlıklarıyla dil koşullarının PESQ karşılaştırması verilmektedir. Şekil 25'te en iyi bölme (Data 3) ağırlıklarıyla dil koşullarının SI-SNR karşılaştırması gösterilmektedir.

4.4. Rastgele Bölünme Kararlılığı Analizi

Data1, Data2, Data3 veri setleri ile eğitilen modelin kararlılığını doğrulamak amacıyla, PESQ standart sapma değerleri ve değişim katsayıları analiz edilmiştir. Karşılaştırma sonuçları Tablo 17'de özetlenmiştir.

Tablo 17. PESQ değerlerinin rastgele bölmeler arasındaki kararlılığı ve değişim katsayıları

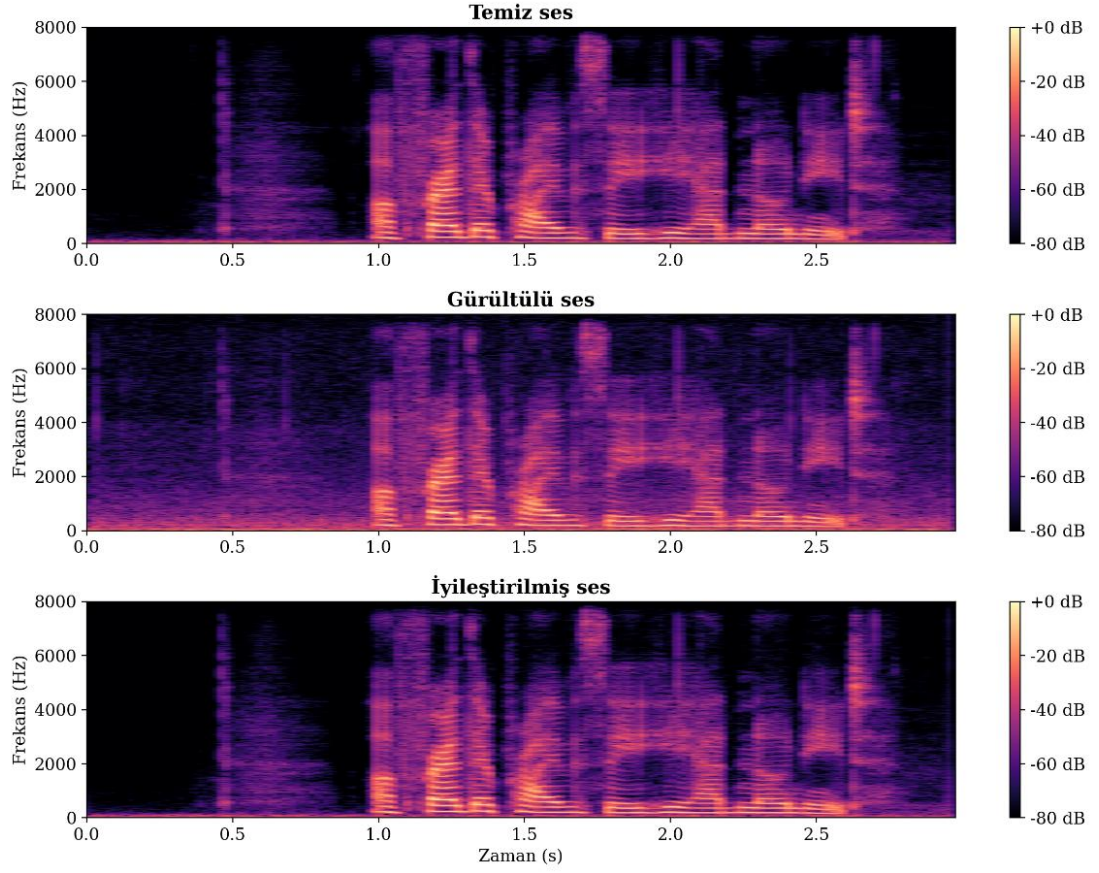
Veri Seti	PESQ Ortalaması	PESQ Standart Sapması	Değişim Katsayısı
İngilizce (VCTK)	2,6881	$\pm 0,053$	1,97%
Türkçe (TSCD)	3,5423	$\pm 0,056$	1,58%
Karma	3,0363	$\pm 0,176$	5,80%

Türkçe koşul %1,58'lik değişim katsayısıyla en kararlı sonucu vermiştir. İngilizce koşul da %1,97 ile benzer bir kararlılık düzeyinde kalmıştır. Karma koşuldaki %5,80'lik oran ise iki farklı dilin ve konuşma türünün bir araya getirdiği akustik çeşitliliğin, rastgele bölme adımlarında yarattığı doğal varyasyonla uyumludur.

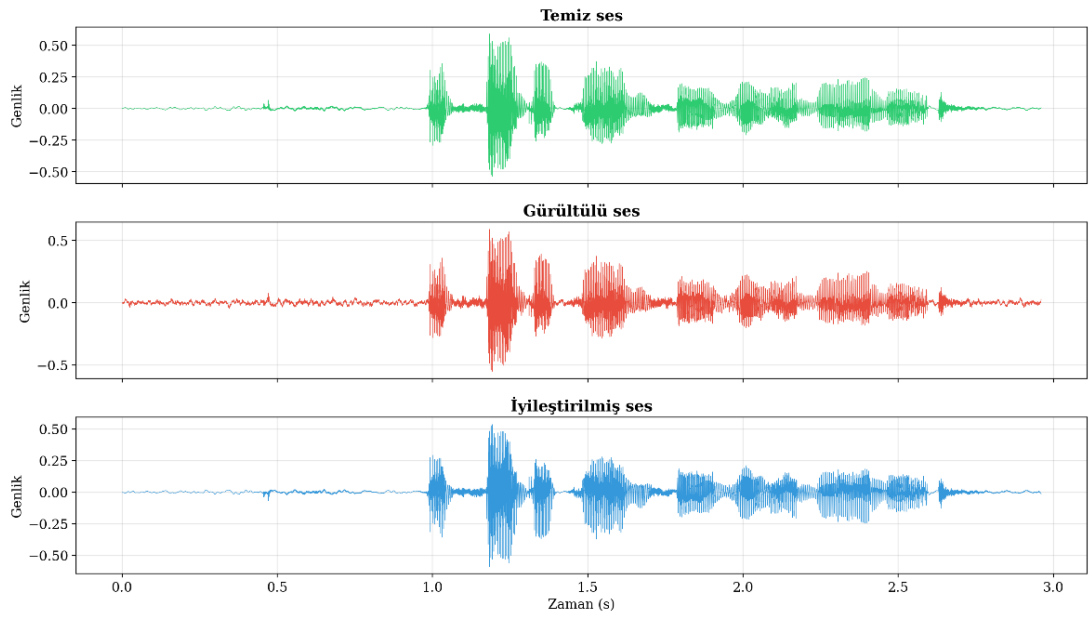
4.5. Spektrogram ve Dalga Formu Analizi

Modelin sinyal üzerindeki etkisini görsel olarak incelemek amacıyla İngilizce test setinden ve Türkçe test setinden birer örnekler seçilmiştir. Bu örneklerin temiz, gürültülü ve iyileştirilmiş halleri spektrogram ve dalga formu üzerinden karşılaştırılmıştır. Gürültülü sinyallerde özellikle sessiz bölgelerin ve yüksek frekans alanlarının gürültüyle kaplandığı görülmektedir. İyileştirilmiş sinyallerde ise bu gürültünün büyük ölçüde bastırıldığı, konuşmaya ait temel yapıların daha belirgin hale geldiği ve dalga formunun temiz sinyale daha yakın bir yapıya ulaştığı gözlemlenmektedir. Türkçe örneğinde, sözcük öncesi ve sonrasındaki sessiz bölgelerin belirgin biçimde temizlenmesi dikkat çekmektedir. Bu durum, Türkçe test koşulunda elde edilen yüksek SI-SNR kazanımıyla uyumlu bir görsel sonuç sunmaktadır. Genel olarak

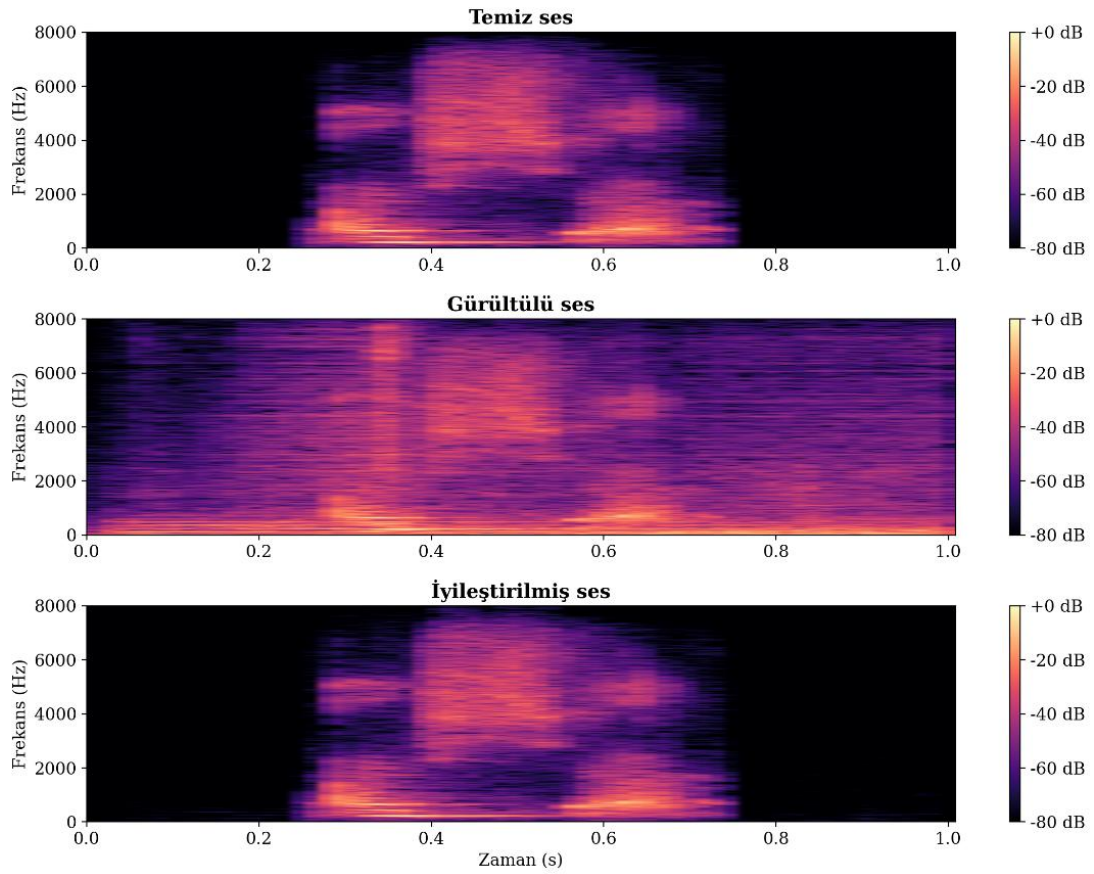
spektrogram ve dalga formu analizleri, önceki bölümlerde metriklerle gösterilen iyileşmenin sinyal düzeyinde de açık biçimde görülebildiğini desteklemektedir.



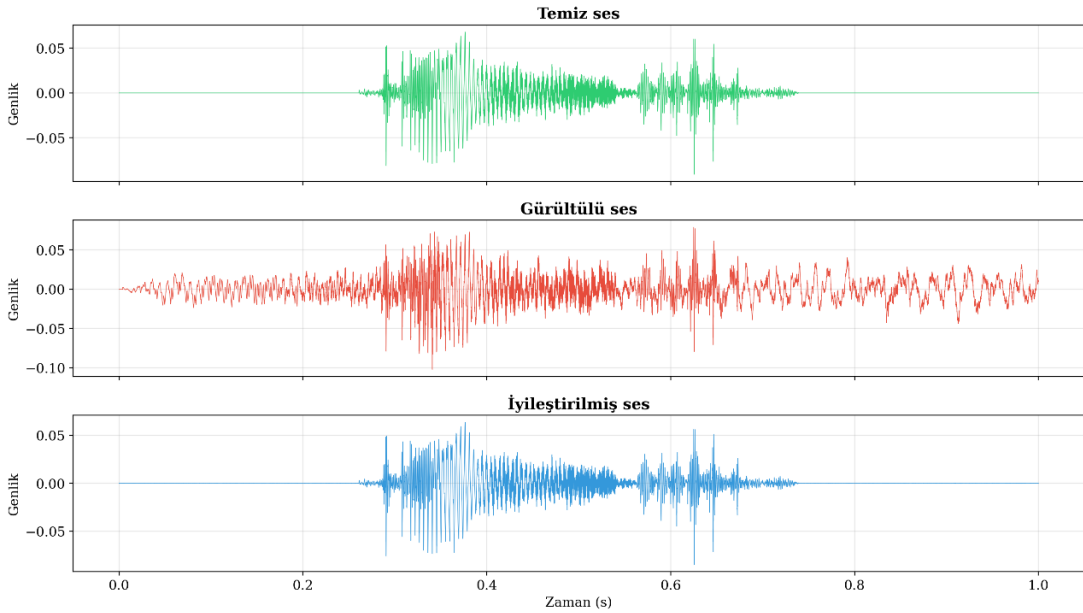
Şekil 26. İngilizce örnek temiz, gürültülü ve iyileştirilmiş sinyallerin spektrogramları



Şekil 27. İngilizce örnek temiz, gürültülü ve iyileştirilmiş sinyallerin dalga formları



Şekil 28. Türkçe örnek temiz, gürültülü ve iyileştirilmiş sinyallerin spektrogramları



Şekil 29. Türkçe örnek temiz, gürültülü ve iyileştirilmiş sinyallerin dalga formları

Şekil 26'da İngilizce bir örneğin temiz, gürültülü ve iyileştirilmiş spektrogramları karşılaştırılmaktadır. Şekil 27'de aynı İngilizce örneğin temiz, gürültülü ve iyileştirilmiş dalga

formları verilmektedir. Şekil 28'de Türkçe bir örneğin temiz, gürültülü ve iyileştirilmiş spektrogramları karşılaştırılmaktadır. Şekil 29'da aynı Türkçe örneğin temiz, gürültülü ve iyileştirilmiş dalga formları gösterilmektedir.

5. TARTIŞMA VE SONUÇLAR

5.1. Tartışma

Önceki bölümde sunulan nesnel metrikler, önerilen geliştirilmiş Wave-U-Net modelinin ses kalitesini ve konuşma anlaşılabilirliğini artırdığını göstermiştir. Bu bölümde söz konusu bulgular yorumlanıp: önce model bileşenlerinin ve kayıp fonksiyonunun performansa katkısı, ardından diller arası genellenebilirlik ve aktarılabilirlik, son olarak sonuçların literatürdeki yeri ele alınmıştır.

5.1.1. Modelin performansının yorumlanması ve metriklerin analizi

İyileştirme sonuçları ITU-T P.862 standardı ve konuşma anlaşılabilirliği teorisi çerçevesinde değerlendirildiğinde modelin pratik başarımı daha net anlaşılmaktadır. ITU-T P.862 standardına göre 1,0–2,0 aralığı "kötü", 2,0–3,0 aralığı "kabul edilebilir", 3,0–4,0 aralığı "iyi" kaliteyi temsil etmektedir. Bu ölçek esas alındığında, Türkçe veri setinde elde edilen 3,54 PESQ ortalaması modelin "iyi" bandında yer aldığını, İngilizce veri setinde ulaşılan 2,69 PESQ ortalamasının ise "kabul edilebilir" kalite seviyesinde bulunduğunu göstermektedir. İngilizce gürültülü sinyal başlangıçta 1,29 PESQ (kötü) seviyesinde olduğundan, aradaki yükseliş kayda değer bir kazançtır. Konuşma anlaşılabilirliğini ölçen STOI metriğinde ise literatürde 0,85 ve üzeri skorlar yüksek anlaşılabilirlik eşiği olarak kabul edilmektedir. Geliştirilen model her üç test koşulunda da (İngilizce: 0,86; Türkçe: 0,85; Karma: 0,85) bu eşiği yakalamış veya aşmıştır. Türkçe veri setinde başlangıç STOI değerinin zaten yüksek olması (0,78) nedeniyle kazanım (+0,07) görece sınırlı görünse de temizleme işlemi sırasında konuşma anlaşılabilirliği başarılı bir şekilde korunmuştur. İnsan görüş puanını (MOS) öngören bileşik metrikler de aynı sonucu desteklemektedir. Sinyal bozulmasını ölçen CSIG üç koşulda da yaklaşık 4,4 düzeyine ulaşmış; 1–5 ölçeğine göre bu yüksek değer, modelin gürültüyü temizlerken konuşmayı bozmadığına işaret etmektedir. Genel kaliteyi öngören COVL Türkçe'de 4,29, Karma'da 4,10 ve İngilizce'de 4,01 ile üç koşulda da 4,0'ın üzerine çıktı. Üç bileşik metrik içinde en düşük değerler CBAK'ta görüldü (3,25–3,43). Bu, modelin arka plan gürültüsünü bastırmada, sinyal kalitesini korumaya kıyasla görece daha zorlandığını gösterir. Ancak elde edilen CBAK değerleri literatürdeki güçlü zaman-alanı modelleriyle uyumludur (Giri vd. 3,35; DEMUCS 3,40; Macartney & Weyde 3,24). SI-SNR, iyileştirilmiş sinyalde temiz konuşmanın kalan gürültü ve bozulmaya göre ne kadar baskın olduğunu dB cinsinden ölçer; değer yükseldikçe sinyal daha temizdir. Modelin

SI-SNR kazanımları Türkçe'de +29,53 dB, Karma'da +21,44 dB ve İngilizce'de +11,76 dB olmuştur. Bu sıralama, gürültü bastırma başarısının veri yapısına güçlü biçimde bağlı olduğunu gösterir. Özellikle Türkçe koşulundaki yüksek değer kısmen izole sözcük yapısından kaynaklanır ve sözcükler arasındaki uzun sessizlik bölgelerinde model neredeyse sıfır enerjili bir çıktı üretebildiğinden SI-SNR belirgin biçimde yükselir. Bu nedenle Türkçe'deki SI-SNR kazanımı mutlak bir üstünlük olarak değil, bu metriğin izole konuşma verisindeki davranışıyla birlikte yorumlanmalıdır.

5.1.2. Model bileşenlerinin ve dil faktörünün değerlendirilmesi

Önerilen geliştirilmiş Wave-U-Net mimarisine eklenen bileşenlerin ve EnhancedLoss kayıp fonksiyonunun performansa katkısı incelendiğinde, mimari modüller ile optimizasyon hedefinin birbirini tamamlayıcı nitelikte olduğu görülmektedir. Zaman alanındaki standart L1 kaybı kullanıldığında, kanal dikkati ve öz-dikkat modüllerinin eklenmesi model başarısını düşürerek PESQ skorunun 2,98'den 2,62 seviyesine gerilemesine yol açmıştır. Bunun olası nedeni, L1 kaybının dalga formunda örnek bazlı hatayı küçültmesi ama faz bilgisini ve ince frekans detaylarını denetleyememesidir; bu hedef altında dikkat modüllerinin eklediği yüksek kapasite algısal kaliteye yansımayaabilir, hatta faz uyumsuzluğu ve aşırı öğrenmeye yol açabilir. Buna karşılık hem zaman hem frekans alanını çok çözünürlüklü spektral terimlerle optimize eden EnhancedLoss fonksiyonu devreye girdiğinde, dikkat modülleri performansı 3,22'den 3,46 PESQ seviyesine yükselttiği görülmüştür. Parametre sayılarına bakıldığında öz-dikkatli tam yapı ile tam model özdeş parametreye sahiptir (27,04 M) ve aralarındaki tek fark kayıp fonksiyonudur. Aynı mimari L1 ile 2,62, EnhancedLoss ile 3,46 değerini vermiştir. Dolayısıyla bu artışı sağlayan şey mimari değil, kaybın algısal yapısıdır. Literatürde Macartney ve Weyde (2018) Wave-U-Net'i konuşma iyileştirmeye uyarlamış, Giri vd. (2019) ise dikkat mekanizması eklemiştir; ancak bu çalışmalar dikkatin katkısını kayıp fonksiyonundan ayırtmaz. Bu tezin ablasyonu, dikkat modüllerinin ancak algısal büyüklük ve faz kısıtlarıyla birlikte eğitildiğinde işe yaradığını göstererek literatüre katkıda bulunmuştur.

Diller arası başarımlar ve aktarılabilirlik boyutunda ise, Türkçe modelinin İngilizce modeline kıyasla daha yüksek kazanımlar elde etmesi fonolojik ve yapısal veri özellikleri çerçevesinde açıklanabilir. Türkçe veri setinin (TSCD) izole kelimelerden oluşması, kelime aralarındaki sessizlik boşluklarının gürültü tahminini kolaylaştırmasını ve zaman alanındaki maskenin daha kararlı öğrenilmesini sağlamıştır. Buna karşılık, sürekli konuşma akışı ve hızlı hece geçişleri

içeren İngilizce veri seti (VCTK) akustik olarak çözülmesi daha zor bir problemdir. İki dilin bu yapısal farklılıklarına rağmen, modelin her iki veri setinde de yüksek performans göstermesi, geliştirilmiş Wave-U-Net mimarisinin diller arası aktarılabilirliğini ortaya koymaktadır.

Karma dil eğitimi bu aktarılabilirliği destekleyen somut bir kanıttır. İki dili bir arada eğiten tek bir model, tek dilli Türkçe performansında sınırlı bir gerilemeye rağmen (3,54'ten 3,04'e PESQ) genel ses kalitesi algısını 4,10 COVL seviyesinde tutmayı başarmıştır. Bu, modelin iki dilin akustik özelliklerini tek bir modelde birleştirebildiğini ve çok dilli bir ses işleme hattında dil başına ayrı model gerektirmeden kullanılabileceğini göstermiştir.

5.1.3. Çalışmanın kısıtlamaları ve sınırlılıkları

Geliştirilen konuşma iyileştirme sisteminin deneysel başarımlarının yanı sıra, gerçek dünya uygulamalarına aktarılmasında dikkate alınması gereken bazı kısıtlamalar bulunmaktadır. İlk olarak, Türkçe deneylerinde kullanılan veri setinin izole sözcük kayıtlarından oluşması önemli bir sınırlılıktır; günlük konuşma dilindeki kelime sınırları arası geçişlerin, vurgu kaymalarının ve tonlama değişimlerinin bu veri setinde yer almaması nedeniyle Türkçe modelindeki yüksek SI-SNR artışının sürekli konuşma diline doğrudan yansımayacağı kabul edilmelidir. İkinci kısıt gürültü çeşitliliğiyle ilgilidir; deneylerde klima, elektrikli süpürge, bulaşık makinesi ve beyaz gürültü olmak üzere dört durağan gürültü türü kullanılmış olup, sokak gürültüsü, kalabalık uğultusu veya müzik gibi frekans karakteristiği hızlı değişen durağan olmayan gürültüler test edilmemiştir. Üçüncü olarak, kodlayıcı-çözücü (U-Net) mimarisi nedensel olmayan (non-causal) bir yapıda olup sinyali bütünsel işlediği için, telefon görüşmesi veya işitme cihazı gibi düşük gecikmeli gerçek zamanlı akış senaryolarında bu haliyle kullanılamamaktadır. Son olarak, model başarımları yalnızca nesnel metriklerle test edilmiş, insan dinleyicilerle öznel Mean Opinion Score (MOS) testleri yapılmamıştır.

5.1.4. Bulguların literatür ile karşılaştırılması

Önerilen geliştirilmiş Wave-U-Net modelinin başarımları, literatürdeki benzer Wave-U-Net varyasyonları, üretici modeller (GAN) ve Türkçe konuşma iyileştirme çalışmalarıyla karşılaştırılmıştır. Farklı veri setleri, SNR aralıkları ve kayıt koşulları kullanıldığından bu karşılaştırmanın amacı önerilen modelin literatürdeki yerini göstermektir.

Tablo 18. Geliştirilen modelin literatürdeki çalışmalarla bağlamsal karşılaştırması

Yöntem	Veri Seti	PESQ (G/Ç/Δ)	STOI (G/Ç/Δ)	Seg-SNR/ SI-SNR (G/Ç/Δ)	CSIG (G/Ç/Δ)	CBAK (G/Ç/Δ)	COVL (G/Ç/Δ)
SEGAN (Pascual vd., 2017)	VoiceBank-DEMAND	1,97 / 2,16 / +0,19	—	1,68 / 7,73 / +6,05 (Seg-SNR)	3,35/3,48/ +0,13	2,44/2,94/ +0,5	2,63/2,80/ +0,17
CNN+GAN (Shah vd., 2018)	VoiceBank-DEMAND	1,97 / 2,34 / +0,37	0,91 / 0,93 / +0,02	—	3,35/3,55/ +0,2	2,44/2,95/ +0,51	2,63/2,92/ +0,29
Improved Wave-U-Net (Macartney & Weyde, 2018)	VoiceBank-DEMAND	1,97 / 2,40 / +0,43	—	1,68 / 9,97 / +8,29 (Seg-SNR)	3,35/3,52/ +0,17	2,44/3,24/ +0,8	2,63/2,96/ +0,33
Attention Wave-U-Net (Giri vd., 2019)	VoiceBank-DEMAND	1,97 / 2,62 / +0,65	—	1,68/10,05/ +8,37	3,35/3,91/ +0,56	2,44/3,35/ +0,91	2,63/3,27/ +0,64
CTFUNet — WB-PESQ, yankılanmalı (Xu, Zhang & Wang, 2023)	DNS Challenge	1,822/3,196/ +1,374	0,866/0,936/ +0,07	9,03/16,36/ +7,33(SI-SDR)	—	—	—
DEMUCS — non-causal (Défossez vd., 2020)	VoiceBank-DEMAND	1,97/3,07/ +1,1	0,92 / 0,95 / +0,03	—	3,35/4,31/ +0,96	2,44/3,4/ +0,96	2,63/3,63/ +1
CMGAN (Ayata & Horasan, 2024)	Özel Türkçe set	— / 2,66 / —	—	— / 20,0 / — (SI-SDR)	—	—	—
Dilated Wave-U-Net (Ali vd., 2020)	VCTK-DEMAND	1,84/2,36/ +0,52	0,92/0,801/ -0,119	22,2/44,31/ +22,11(Seg-SNR)	—	—	—
Önerilen Geliştirilmiş Wave-U-Net	VoiceBank-DEMAND	1,97/ 2,61/ +0,64	0,92/0,93/ +0,01	8,45/19,56/ +11,11	3,35/3,88/ +0,53	2,44/3,36/ +0,92	2,63/3,25/ +0,62
Önerilen Geliştirilmiş Wave-U-Net (İngilizce)	VCTK	1,29 / 2,69 / +1,40	0,73 / 0,86 / +0,13	7,50 / 19,26 / +11,76 (SI-SNR)	2,97 / 4,40 / +1,42	2,42 / 3,25 / +0,83	2,54 / 4,01 / +1,46
Önerilen Geliştirilmiş Wave-U-Net (Türkçe)	TSCD	1,48 / 3,54 / +2,06	0,78 / 0,85 / +0,07	7,50 / 37,03 / +29,53 (SI-SNR)	3,28 / 4,39 / +1,11	2,55 / 3,43 / +0,88	2,81 / 4,29 / +1,49
Önerilen Geliştirilmiş Wave-U-Net (Karma)	VCTK + TSCD	1,36 / 3,04 / +1,68	0,74 / 0,85 / +0,11	7,50 / 28,94 / +21,44 (SI-SNR)	3,09 / 4,37 / +1,28	2,47 / 3,32 / +0,85	2,64 / 4,10 / +1,46

Tablo 18 incelendiğinde şu çıkarımlar yapılmaktadır:

- Farklı veri kümelerinden kaynaklanabilecek belirsizlikleri sınırlamak amacıyla önerilen model, alanın standart veri kümesi olan VoiceBank-DEMAND üzerinde, bu kümenin kendi önceden tanımlı eğitim/test bölünmesi PESQ protokolü kullanılarak değerlendirilmiştir. Bu koşulda model tüm ölçütlerde tutarlı bir iyileşme sağlamıştır: PESQ 1,97' den 2,61'e, STOI 0,92'den 0,93'e, SI-SNR 8,45'ten 19,56 dB'e yükselmiştir. Bileşik ölçütlerde ise CSIG 3,35'ten 3,88'e, CBAK 2,44'ten 3,36'ya ve COVL 2,63'ten 3,25 değerlerine yükselmiştir. Elde edilen sonuçlar, önerilen modelin zaman alanında çalışan Wave-U-Net tabanlı yaklaşımlar arasında rekabetçi bir konumda yer aldığını göstermektedir. PESQ açısından SEGAN (2,16), CNN+GAN (2,34) ve Improved Wave-U-Net (2,40) yöntemlerinin üzerinde, Attention Wave-U-Net (2,62) ile oldukça yakın bir düzeydedir. Bileşik ölçütlerde de Attention Wave-U-Net'in sonuçlarına (CSIG 3,91; CBAK 3,35; COVL 3,27) yakın

değerler gözlenmektedir. Daha güncel ve daha büyük ölçekli bir model olan DEMUCS'un (3,07 PESQ) ise önde olduğu görülmektedir. Aynı veri kümesi ve aynı ölçüm protokolünün kullanılması modelin literatürdeki konumunu ortaya koymaktadır.

- İngilizce konuşma iyileştirme deneylerinde önerilen modelin elde ettiği 2,69 PESQ değeri, zaman alanında ham dalga formu üzerinde çalışan GAN tabanlı SEGAN (2,16), Macartney ve Weyde (2018) tarafından önerilen Improved Wave-U-Net (2,40) ve Attention Wave-U-Net (Giri vd., 2019) (2,62) gibi yaklaşımlara kıyasla rekabetçi bir düzeyde yer almaktadır. Bu eğilim, modelin VoiceBank-DEMAND üzerinde aynı ölçüm protokolüyle elde ettiği 2,61 PESQ değeriyle de tutarlılık göstermektedir. Buna karşın, daha güncel ve daha büyük ölçekli bir zaman-alanı modeli olan DEMUCS'un VoiceBank-DEMAND veri kümesi üzerinde 3,07 PESQ değerine ulaştığı görülmektedir. Kullanılan veri kümeleri ve deneysel koşullardaki farklılıklar dikkate alındığında, önerilen modelin katkısının en yüksek skoru elde etmekten çok, dikkat mekanizmaları ile çok çözünürlüklü EnhancedLoss yapısının oluşturduğu mimari-kayıp sinerjisinin etkinliğini ortaya koyması biçiminde değerlendirilmesi daha uygun olacaktır.
- Türkçe alanındaki güncel çalışmalardan biri olan Ayata ve Horasan (2024) tarafından önerilen CMGAN modeli, özel bir Türkçe veri kümesi üzerinde yaklaşık 2,66 PESQ ve 20 dB SI-SDR değerlerine ulaşmıştır. Önerilen model ise TSCD üzerinde 3,54 PESQ ve 37,03 dB SI-SNR değerleri elde etmiştir. Ancak veri kümeleri, konuşma türleri ve kullanılan ölçütler arasındaki farklılıklar nedeniyle bu sonuçlar doğrudan bir sayısal üstünlük olarak değil, önerilen modelin Türkçe konuşma iyileştirmede yüksek bir başarı potansiyeline işaret etmesi biçiminde yorumlanmalıdır.
- Ayata ve Horasan (2024) çalışmasında da belirtildiği üzere, farklı kayıt ortamları, gürültü türleri ve veri kümeleri doğrudan sayısal karşılaştırmaları güçleştirmektedir. Bu durum, tez kapsamında vurgulanan "izole kelime-sürekli konuşma" ayrımının önemli bir karıştırıcı değişken olduğunu desteklemektedir. Dolayısıyla Türkçe modelde gözlenen yüksek performans artışının yalnızca önerilen mimari ve kayıp fonksiyonu tasarımından değil, aynı zamanda TSCD veri kümesinin izole kelimelerden oluşmasının sağladığı akustik avantajlardan da etkilenmiş olabileceği değerlendirilmektedir.

5.2 Sonuçlar

Bu bölümde, tez çalışması kapsamında geliştirilen dikkat mekanizmalarıyla güçlendirilmiş geliştirilmiş Wave-U-Net mimarisi ve EnhancedLoss fonksiyonundan elde edilen bulgular özetlenmekte, ayrıca gelecekte gerçekleştirilebilecek araştırma yönleri tartışılmaktadır.

5.2.1. Genel değerlendirme

Bu tez çalışmasında, zaman alanında doğrudan ham ses dalgası üzerinde çalışan, kanal dikkati ve öz-dikkat mekanizmaları ile güçlendirilmiş bir geliştirilmiş Wave-U-Net konuşma iyileştirme modeli geliştirilmiş ve model, zaman ve frekans alanı bilgilerini birlikte dikkate alan EnhancedLoss fonksiyonu ile eğitilmiştir. Gerçekleştirilen ablasyon çalışmaları, dikkat mekanizmalarının tek başına sınırlı katkı sağladığını, buna karşın spektral büyüklük ve faz koruma bileşenlerini içeren EnhancedLoss yapısıyla birlikte kullanıldığında daha yüksek performans değerlerine ulaşıldığını göstermiştir.

İngilizce ve Türkçe veri kümeleri üzerinde gerçekleştirilen deneyler, önerilen yaklaşımın farklı dil yapıları altında etkili sonuçlar üretebildiğini ortaya koymuştur. Ayrıca, İngilizce ve Türkçe verilerin birlikte kullanıldığı karma dil modeli deneylerinde elde edilen bulgular, çok dilli konuşma iyileştirme uygulamalarında tek bir modelin kullanılmasının uygulanabilir bir yaklaşım olabileceğini göstermektedir.

Literatürde yer alan Wave-U-Net tabanlı ve GAN tabanlı yöntemlerle yapılan bağlamsal karşılaştırmalar, önerilen modelin özellikle algısal kalite ölçütleri açısından rekabetçi sonuçlar elde ettiğini göstermiştir. Türkçe veri kümesi üzerinde elde edilen yüksek performans değerleri, önerilen mimari ve kayıp fonksiyonu tasarımının etkinliğini desteklemekle birlikte, veri kümesinin izole kelimelerden oluşmasının sağladığı akustik avantajların da bu sonuçlar üzerinde etkili olabileceği değerlendirilmektedir.

5.2.2. Geleceğe yönelik öneriler

Bu tez çalışmasından elde edilen bulgular doğrultusunda, gelecekte gerçekleştirilebilecek bazı araştırma yönleri önerilmektedir.

İlk olarak, önerilen modelin Türkçe konuşma iyileştirme başarımı, TSCD veri kümesine ek olarak sürekli konuşma içeren farklı Türkçe veri kümeleri üzerinde test edilerek daha kapsamlı biçimde değerlendirilebilir.

İkinci olarak, dikkat mekanizmaları ve EnhancedLoss yapısında yapılabilecek iyileştirmeler ile model mimarisi daha da geliştirilebilir. Farklı ağ yapıları ve kayıp fonksiyonlarının performans üzerindeki etkileri gelecekte incelenebilir.

Ayrıca, modelin farklı dil ailelerine ait veri kümeleri üzerinde değerlendirilmesi, yaklaşımın çok dilli konuşma iyileştirme problemlerindeki genellenebilirliğinin daha ayrıntılı olarak incelenmesine katkı sağlayabilir.

Son olarak, daha büyük ve çeşitli veri kümeleri kullanılarak gerçekleştirilecek çalışmaların, modelin gerçek yaşam koşullarındaki performansının daha kapsamlı biçimde değerlendirilmesine katkı sağlayacağı düşünülmektedir.

KAYNAKÇA

- Abilbekov, A., Mussakhajayeva, S., Yeshpanov, R., & Varol, H. A. (2024, May). Kazemotts: A dataset for kazakh emotional text-to-speech synthesis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 9626-9632).
- Ali, M. N., Brutti, A., & Falavigna, D. (2020, September). Speech enhancement using dilated wave-u-net: an experimental analysis. In *2020 27th Conference of Open Innovations Association (FRUCT)* (pp. 3-9). IEEE
- Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., ... & Weber, G. (2020, May). Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 4218-4222).
- Ayata, D., & Horasan, U. (2024, May). GAN and Diffusion Based Speech Enhancement. In *2024 32nd Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.
- Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2), 113-120.
- Botinhao, C. V., Wang, X., Takaki, S., & Yamagishi, J. (2016, September). Investigating RNN-based speech enhancement methods for noise-robust text-to-speech. In *9th ISCA speech synthesis workshop* (pp. 159-165).
- Défossez, A., Synnaeve, G., & Adi, Y. (2020). Real time speech enhancement in the waveform domain. *arXiv preprint arXiv:2006.12847*.
- Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing*, 32(6), 1109-1121.
- Fu, S. W., Liao, C. F., Tsao, Y., & Lin, S. D. (2019, May). Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning* (pp. 2031-2041). PmLR.
- Giri, R., Isik, U., & Krishnaswamy, A. (2019, October). Attention wave-u-net for speech enhancement. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 249-253). IEEE.
- Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2, pp. 1-800). Cambridge: MIT press.

- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
- Hu, Y., & Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1), 229-238.
- Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).
- Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., ... & Xie, L. (2020). DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv preprint arXiv:2008.00264*.
- Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). pmlr.
- Recommendation, I. T. (2001). Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*.
- Jensen, J., & Taal, C. H. (2016). An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11), 2009-2022.
- Kim, E., & Seo, H. (2021, August). SE-conformer: Time-domain speech enhancement using conformer. In *Interspeech* (Vol. 2021, pp. 2736-2740).
- Kurtkaya, M. (2025). *Turkish speech command dataset*. Kaggle. Retrieved July 14, 2025, from <https://www.kaggle.com/datasets/muratkurtkaya/turkish-speech-command-dataset>.
- Le Roux, J., Wisdom, S., Erdogan, H., & Hershey, J. R. (2019, May). SDR—half-baked or well done?. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 626-630). IEEE.
- Loizou, P.C. (2013). *Speech Enhancement: Theory and Practice*, Second Edition (2nd ed.). CRC Press. <https://doi.org/10.1201/b14529>
- Luo, Y., & Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8), 1256-1266.
- Macartney, C., & Weyde, T. (2018). Improved speech enhancement with the wave-u-net. *arXiv preprint arXiv:1811.11307*.

- Oppenheim, A. & Schafer, R. (2013). *Discrete-Time Signal Processing*. (3rd ed.). Pearson International. <https://elibrary.pearson.de/book/99.150005/9781292038155?lang=de>
- Park, S. R., & Lee, J. (2016). A fully convolutional neural network for speech enhancement. *arXiv preprint arXiv:1609.07132*.
- Pascual, S., Bonafonte, A., & Serra, J. (2017). SEGAN: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*.
- Pirklbauer, J., Sach, M., Fluyt, K., Tirry, W., Wardah, W., Moeller, S., & Fingscheidt, T. (2023, September). Evaluation metrics for generative speech enhancement methods: Issues and perspectives. In *Speech Communication; 15th ITG Conference* (pp. 265-269). VDE.
- Rabiner, L., & Schafer, R. (2011). *Theory and applications of digital speech processing*. Pearson.
- Reddy, C. K., Gopal, V., Cutler, R., Beyrami, E., Cheng, R., Dubey, H., ... & Gehrke, J. (2020). The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. *arXiv preprint arXiv:2005.13981*.
- Ren, Z., Yolwas, N., Wang, H., & Slamun, W. (2023). Exploring Turkish Speech Recognition via Hybrid CTC/Attention Architecture and Multi-feature Fusion Network. *arXiv preprint arXiv:2303.12300*.
- Richter, J., Welker, S., Lemercier, J. M., Lay, B., & Gerkmann, T. (2023). Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 2351-2364.
- Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001, May). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)* (Vol. 2, pp. 749-752). IEEE.
- Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing.
- Salor, Ö., Pellom, B. L., Ciloglu, T., & Demirekler, M. (2007). Turkish speech corpora and recognition tools developed by porting SONIC: Towards multilingual speech recognition. *Computer Speech & Language*, 21(4), 580-593.
- Shah, N., Patil, H. A., & Soni, M. H. (2018, November). Time-frequency mask-based speech enhancement using convolutional generative adversarial network. In *2018 Asia-Pacific*

- signal and information processing association annual summit and conference (APSIPA ASC)* (pp. 1246-1251). IEEE.
- Stoller, D., Ewert, S., & Dixon, S. (2018). Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*.
- Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021, June). Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 21-25). IEEE.
- Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on audio, speech, and language processing*, 19(7), 2125-2136.
- Tan, K., & Wang, D. (2018, September). A convolutional recurrent neural network for real-time speech enhancement. In *Interspeech* (Vol. 2018, pp. 3229-3233).
- Thiemann, J., Ito, N., & Vincent, E. (2013, June). The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics* (Vol. 19, No. 1, p. 035081). Acoustical Society of America.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM transactions on audio, speech, and language processing*, 26(10), 1702-1726.
- Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., & Schuller, B. (2015, August). Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *International conference on latent variable analysis and signal separation* (pp. 91-99). Cham: Springer International Publishing.
- Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3-19).
- Xie, Y., & Tan, Z. H. (2025). A survey of deep learning for complex speech spectrograms. *Speech Communication*, 103319.
- Xu, S., Zhang, Z., & Wang, M. (2023). Channel and temporal-frequency attention UNet for monaural speech enhancement. *EURASIP Journal on Audio, Speech, and Music Processing*, 2023(1), 30.

- Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (2014). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM transactions on audio, speech, and language processing*, 23(1), 7-19.
- Yamagishi, J., Veaux, C., & MacDonald, K. (2019). CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92).