T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DERİN ÖĞRENME İLE SES İYİLEŞTİRİLMESİ

MUSTAFA ERSEVEN

YÜKSEK LİSANS TEZİ
ELEKTRONİK VE HABERLEŞME MÜHENDİSLİĞİ ANABİLİM DALI
HABERLEŞME PROGRAMI

DANIŞMAN
DR. ÖĞR. ÜYESİ BÜLENT BOLAT

İSTANBUL, 2018

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

DERİN ÖĞRNME İLE SES İYİLEŞTİRME

Mustafa ERSEVEN tarafından hazırlanan tez çalışması 10.12.2018 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Elektronik ve Haberleşme Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Dr. Öğr. Üyesi Bülent BOLAT

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Dr. Öğr. Üyesi Bülent BOLAT

Yıldız Teknik Üniversitesi _____

Prof. Dr. Mehmet Serdar Ufuk TÜRELİ

Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üyesi Gökalp TULUM

İstanbul Arel Üniversitesi _____

ÖNSÖZ

Yüksek lisans eğitimim boyunca desteklerini esirgemeyen herkese ve Sayın Hocam Bülent Bolat'a teşekkür ederim. Bu tezi sevgili aileme ithaf ediyorum.

Aralık, 2018

Mustafa ERSEVEN

İÇİNDEKİLER

SİMGE LİSTESİ

f	Olasılık yoğunluk fonksiyonu
m	Mean
r	İlişki fonksiyonu
c	Varyans fonksiyonu
ξ	Minimum mean square error
Δ	Gradient
a	Öğrenme katsayısı
γ	Momentum katsayısı
$E\{.\}$	Beklenen değer operatörü
θ	Değişkenler vektörü
$L\{.\}$	Log-olabilirlik

KISALTMA LİSTESİ

ADÖ	Algısal Doğrusal Öngörü
DA	Dereceli Alçalma
DÖ	Derin Öğrenme
DSA	Derin Sinir Ağı
ESA	Evrişimsel Sinir Ağı
GAD	Geniş Anlamda Durağanlık
GSM	Global System for Mobile Communications
GYA	Geri Yayılım Algoritması
İSA	İleri Beslemeli Sinir Ağı
KFD	Kesikli Fourier Dönüşümü
KKAD	Konuşma Kalitesinin Algısal Değerlendirilmesi
KSFD	Kısa Süreli Fourier Dönüşümü
KSNN	Kısa Süreli Nesnel Netlik
LGS	Logaritmik Güç Spektrumu
LSTM	Long Short Term Memory
MFKK	Mel Frekansı Kepstrum Katsayıları
ReLU	Rectified Linear Unit
SDA	Stokastik Dereceli Alçalma
SDT	Sonlu Dürtü Tepkili
SGO	Sinyal Gürültü Oranı
TSA	Tekrarlamalı Sinir Ağı
YSA	Yapay Sinir Ağı
Z-F	Zaman-Frekans

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

DERİN ÖĞRENME İLE SES İYİLEŞTİRİLMESİ

Mustafa ERSEVEN

Elektronik ve Haberleşme Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Dr. Öğr. Üyesi Bülent BOLAT

Günümüzde ses sinyalini kullanan sayısız cihaz ve uygulama vardır; örneğin haberleşme uygulamaları, müzik sistemleri ve biyomedikal cihazlar gibi. Özellikle telsiz ve cep telefonlarının kullanımının yaygınlığı düşünülürse, bu konu gerek askeri gerekse sivil alanda önemli bir yer tutmaktadır. Bahsi geçen bu uygulama alanlarının ortak problemlerinden biri ses sinyali üzerinde oluşan gürültüdür. Sinyali etkileyen gürültü kaynakları ile her alanda karşılaşılabilir. Ticari bağlamda kullanıcılara daha iyi bir hizmet sunulması amacıyla, akademik bağlamda ise istatistiksel sinyal işleme alanında ilgi çekici bir konu olması nedeniyle bu problem üzerinde sayısız çalışma yapılmıştır.

Bu çalışmada ise problem özel olarak konuşma sinyalinin iyileştirilmesiyle ile sınırlandırılmıştır. Konuşma sinyalinin özellikle haberleşme bağlamında önemli uygulamaları olması konuyu ilgi çekici kılmıştır. Gerek telsiz gerekse GSM haberleşmesi sırasında oluşan arka plan gürültüsünün giderilmesi önemlidir. Arka plan gürültüsünün giderilmesiyle konuşma sinyalinin, kısa süreli nesnel netlik (KSNN) (short time objective intelligibility) ve konuşma kalitesinin algısal değerlendirilmesi (KKAD) (perceptual evaluation of speech quality) ölçütleri iyileştirilir. Böylece dinleyici tarafına arka plandaki gürültüden arındırılmış ve anlaşılır bir konuşma sunulmuş olur. Bu çalışmada, arka plan gürültüsü olarak gevezelik (babble) gürültüsü seçilmiştir. Yöntem olarak ise, makine öğrenmesinin bir dalı olan derin öğrenmenin (deep learning) altında bulunan derin sinir ağları kullanılmıştır. Derin sinir ağı olarak ise en popülerinden biri olan evrişimsel sinir ağının (ESA) probleme uygun bir mimarisi önerilmiştir.

Gevezelik (babble) gürültüsünün istatistiksel modelinin çıkarılması zor olduğundan dolayı, istatistiksel özellikleri taban alan klasik yöntemler (Wiener süzgeci, minimum ortalama karesel hata tabanlı spektral genlik kestirimcisi gibi) kullanılamamaktadır. Bu problemin üstesinden gelmek amaclı, gürültülü konuşma spektrumları ile temiz konuşma spektrumları arasında doğrusal olmayan haritalama yapabilen sinir ağları popüler yöntem olmuştur. Özellikle derin öğrenme (DÖ) görüntü ve ses işleme uygulamalarında kendine geniş yer bulmuştur. Evrişimsel sinir ağının ses için frekans ilişkisini de işlemesi ve diğer derin öğrenme yöntemlerine nazaran daha az değişken gerektirmesi konuşma işleme konusunda ESA'yı daha popüler bir yöntem haline getirmiştir.

Bu çalışmada, konuşma iyileştirme için regresyon temelli bir ESA mimarisi oluşturulmuştur. İki farklı konuşmacıya ait 460 farklı cümle kullanılmıştır. Temiz konuşma sinyallerine gevezelik gürültüsü eklenmiş ve farklı sinyal gürültü oranlarında (SGO) veri setleri oluşturulmuştur. Temiz ve gürültülü sinyallerin kısa süreli Fourier dönüşümü (KSFD) katsayıları elde edilmiş, ardından bu katsayıların genlik bilgilerinden logaritmik güç spektrumu (LGS) katsayıları hesaplanmıştır. Gürültülü sinyalin faz bilgisi, genlik spektrumu iyileştirildikten sonra sesi sentezlemek için saklanmıştır. Önerilen ESA eğitilmiştir. İyileştirilen LGS katsayıları ve saklanan faz ile ses tekrar sentezlenmiş ardından KKAD ve KSNN ölçütleri ile değerlendirilmiştir.

Anahtar Kelimeler: Evrişimsel sinir ağları, konuşma iyileştirme, logaritmik güç spektrumu, regresyon modeli.

ABSTRACT

VOICE ENHANCEMENT BY DEEP LEARNING

Mustafa ERSEVEN

Department of Electronics and Communications Engineering

MSc. Thesis

Adviser: Dr. Bülent BOLAT

Today, there are countless devices and applications that use audio signals; such as communication apps, music systems and biomedical devices. Especially when considering the prevalence of the use of wireless and mobile phones, it keeps these issues, both military and civilian areas in an important place. One of the common problems of these applications is the noise generated on the audio signal. Noise sources affecting the signal can be encountered in every field. Numerous studies have been carried out on this problem because it is an interesting subject in the field of statistical signal processing in the academic context in order to provide better service to users in a commercial context.

In this study, the problem was limited by the enhancement of the speech signal. The fact that the speech signal has important applications in the context of communication has made it interesting. It is important to remove the background noise generated during both wireless and GSM communications. By eliminating background noise, short time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) are improved. In this way, clear speech is provided to the listener that free of background noise. In this study, babble noise was chosen as background noise. As a method, deep neural networks under deep learning have been used. As a deep neural network, an appropriate architecture of the convolutional neural network (CNN), one of the most popular, has been proposed.

Because the statistical modeling of babble noise is difficult to extract, classical methods based on statistical features (such as the Wiener filter, the minimum mean square error-based spectral amplitude estimator) cannot be used. To overcome this problem, neural networks which can perform non-linear mapping between noisy speech spectra and clean speech spectra have been popular. Especially deep learning has found wide space in image and sound processing applications.

In this study, CNN architecture based on regression was created for speech enhancement. 460 different sentences of two different speakers were used. Speech noise was added to the clean speech signals and data sets were generated at different signal to noise ratios (SNR). Short time Fourier transform (STFT) coefficients of the clear and noisy signals were obtained and then the logarithmic power spectrum (LPS) coefficients were calculated from the amplitude information of these coefficients. The phase information of the noisy signal was stored to synthesize the signal after the amplitude spectrum was enhanced. Proposed CNN is trained. Speech signal were re-synthesized from enhanced LPS coefficients and the stored phase and then evaluated by PESQ and STOI criteria.

Keywords: Convolutional neural network, speech enhancement, logarithmic power spectrum, regression model.

BÖLÜM 1

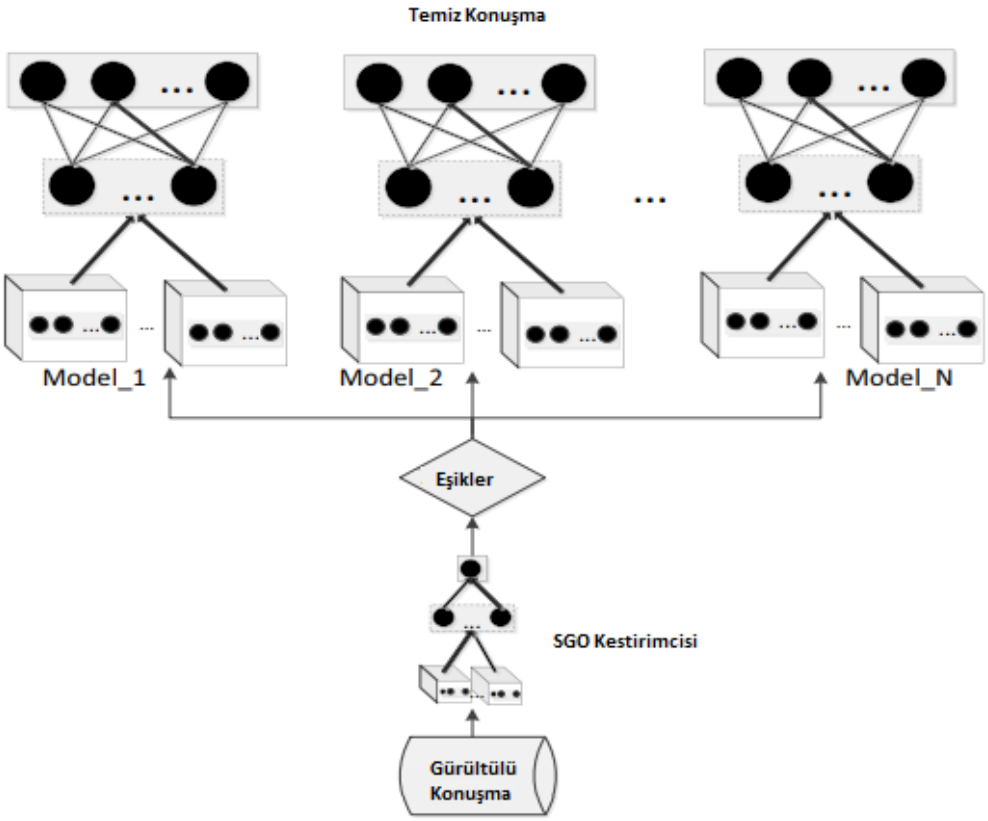
GİRİŞ

1.1 Literatür Özeti

Literatürde konuyla ilgili geçmiş zamanlardan bu yana birçok çalışma mevcuttur. Takdir edileceği üzere derin öğrenmeyle ilgili çalışmalar son yıllarda ortaya çıkmıştır. Daha önceki çalışmalarda süzgeçler ve kestirimciler ile çalışılmıştır. Ayrıca, yanlızca genlik spektrumunun değil, faz spektrumunun da iyileştirilmesinin önemli olduğunu vurgulayan çalışmalar mevcuttur. Bu kısımda kısaca derin öğrenme dışındaki çalışmalardan örnekler verdikten sonra asıl odaklandığımız derin öğrenmeyle ilgili çalışmalar geniş olarak aktarılacaktır. Ephraim vd. [1] tarafından 1984 yılında, konuşma sinyalinin kısa süreli spektral genliğinin, konuşmanın algılanmasındaki önemi üzerinde durulmuştur. Minimum ortalama karasel hata tabanlı spektral genlik kestirimcisi kullanan bir sistem önerilmiştir. Ardından, Wiener süzgeci ve spektral cıkarma (spectral subtraction) algoritmasına dayanan diğer yaygın yöntemler ile karşılaştırılmıştır. Konuşma ve gürültü spektral bileşenleri istatistiksel olarak bağımsız Gauss rasgele değişkenleri olarak modellenmiştir. Gerkmann vd. [2] tarafından 2012 yılında, gürültülü konuşma sinyalinden gürültüsüz konuşmanın faz spektrumunu kestirmek için bir yöntem önerilmiştir. Daha önceki çalışmalarda faz spektrumunun göz ardı edildiği vurgulanmış yalnızca genlik spektrumunun iyileştirildiği daha sonra gürültülü fazın konuşmanın tekrar oluşturulmasında kullanıldığı belirtilmiştir. Önerilen yöntemle beraber faz spektrumunun da iyileştirilmesinin konuşmanın iyileştirilmesine katkı sağladığı gösterilmiştir.

Literatüde sinir ağlarını kullanan çalışmalar genlik spektrumu üzerine yoğunlaşmıştır. Xu vd. [3] tarafından 2015 yılında, derin sinir ağlarına (DSA) dayanan gürültülü ve temiz konuşma sinyalleri arasında bir eşleşme fonksiyonu bulmak süretiyle konuşmayı iyileştirmek için gözetimli bir yöntem önerilmiştir. Ağ, güçlü bir modelleme özelliği olsun diye doğrusal olmayan bir regresyon mimarisi şeklinde tasarlanmıştır. DSA tabanlı konuşma iyileştirme sisteminin, regresyon modelinin aşırı yumuşatma problemini hafifletmek için küresel varyans eşitlemesi ve DSA'ların genelleme kabiliyetini daha da iyileştirmek için gürültüden haberdar eğitim stratejileri de dahil olmak üzere çeşitli teknikler önerilmiştir.

Park vd. [4] tarafından 2016 yılında, işitme cihazlarındaki gürültülü sesin iyileştirilmesi üzerinde durulmuştur. Gürültülü konuşma spektrumları ile gözetimli öğrenme (supervised learning) yoluyla temiz konuşma spektrumları arasında bir haritalama yaparak sorunu çözmeye çalışılmıştır. Tam bağlı ağlardan daha az parametre içeren evrişimsel sinir ağları (convolutional neural networks) önerilmiştir. Önerilen mimari yedeklemeli evrişimsel kodlayıcı kod çözücüdür, böylece daha az parametre kullanarak ilgili mimarinin gömülü sistemler içinde uygun olabileceği gösterilmiştir. Fu vd. [5] tarafından 2016 yılında, konuşma iyileştirme için SGO bilinen ESA önerilmiştir. Genelleme kabiliyetini ve doğruluğu artırmak için iki aşamalı bir algoritma önerilmiştir. Şekil 1.1'de görüldügü gibi, algoritmanın ilk kısmında gürültülü konuşmanın SGO kestirilmiştir. Ardından ikinci kısımda ise bulunan SGO'ya uygun ESA modeli seçilerek iyileştirme işlemi yapılmıştır. Gao vd. [6] tarafından 2016 yılında, DSA dayalı konuşma iyileştirme için ilerici öğrenme (progressive learning) modeli sunulmuştur. Gürültülü konuşma sinyalini temiz konuşma sinyaline haritalama problemini, sistem karmaşıklığını azaltmak ve performansı artırmak amaçlı alt problemlere ayrılmıştır. Deneysel sonuçlar, ilerici öğrenmenin düşük SGO ortamlarında konuşma kalitesi ve KSNN değerini etkili bir şekilde geliştirdiği ve DSA tabanlı sistem ile karşılaştırıldığında model parametrelerinin % 50 oranında azaltabildiği gösterilmiştir.



Şekil 1. 1 İki aşamalı SGO kestirimcili ESA mimarisi [5]

Wang vd. [7] tarafından 2017 yılında, görsel ve işitsel akışları bir araya getiren ESA önerilmiştir. Görsel ve işitsel veriler ilk olarak ayrı ayrı ESA'lar kullanılarak işlenir. Daha sonra iyileştirilmiş konuşma ve yeniden yapılandırılmış görüntüler oluşturmak için bir ortak ağa geçer. Bu çalışmada görsel bilginin, konuşma iyileştirme sürecine entegre edilmesinin performansı artıracağı vurgulanmıştır. Fu vd. [8] tarafından 2017 yılında, ham dalga formuna dayalı konuşma iyileştirme için tamamen evrişim katmanlarından oluşan ağ mimarisi önerilmiştir. Önerilen mimaride dalga formunda giriş ve dalga formunda çıkış mevcuttur. Bu noktada LGS katsayılarılarını kullanan diğer yöntemlerden farklılık gösterir. ESA'larında bulunan tamamen bağlı katmanın, konuşma sinyalinin yüksek frekans bileşeni doğru modelleyemediğinden dolayı yalnızca evrişim katmanlarından oluşan bir mimari kullanılmıştır. Deneysel sonuçlarda tamamen bağlı katman barındıran mimarilere göre ham dalga formunda daha başarılı olduğu gösterilmiştir. Karşılaştırma için KSNN ve KKAD ölçütleri kullanılmıştır. Xu vd. [9] tarafından 2017 yılında, konuşma iyileştirmede doğrudan kullanılan LGS katsayıları yanı sıra, mel frekansı kepstral katsayıları ve ideal ikili maske (ideal binary mask) gibi ikincil

öznitelikleri kullanmak için bir yardımcı yapı önerilmiştir. Önerilen yapı derin sinir ağı mimarisine entegre edilerek, tüm değişkenlerin ortak optimizasyonu sağlanmıştır. Önerilen ortak tahmin modelinde, ikincil öznitelikler sayesinde doğrudan tahmin yapan LGS katsayılarına kısıtlamalar getirmekte ve potansiyel olarak öğrenmesini geliştirmektedir. Böylece yeniden yapılandırılan konuşmanın kalitesinin geliştirildiği gösterilmiştir. Sun vd. [10] tarafından 2017 yılında, konuşma iyileştirme için uzun kısa süreli hafızalı yineleyen sinir ağı (long short term memory recurrent neural network-LSTM-RNN) modeli önerilmiştir. Gürültülü ve temiz konuşma sinyallerinin özniteliklerini doğrudan haritalama için bir LSTM-RNN tabanlı regresyon yaklaşımı sunulmuş ve uzun süreli akustik bağlamı modellemede derin sinir ağı tabanlı regresyon tekniklerinden daha etkili olduğu gösterilmiştir. Söz konusu olan yöntemler arasında farklı SGO'larda kapsamlı karşılaştırmalar yapılmıştır.

1.2 Tezin Amacı

Sesin iyileştirilmesi, dinleyicilerin algılama kalitesi artırmak için kullanıldığı direk uygulamaların yanı sıra, konuşma tanıma (speech recognition) ya da konuşmacı tanıma (speaker recognition) gibi uygulamaların ön işlemi olarak da kullanılır. Gerek doğrudan uygulandığı, gerekse herhangi bir uygulamanın alt bileşeni olması durumları için ortak amaç KKAD ve KSNN değerlerinin artırılmasıdır.

Bu çalışmanın temel amacı konuşma iyileştirme konusunda işlevsel olduğu kadar basit bir ağ mimarisi oluşturmaktır. Gürültülü bir ortamda, örneğin kafe ya da restoran gibi, farklı konuşmacılarında olduğu ve bu farklı konuşmalarının arka planda gevezelik gürültüsü oluşturduğu bir senaryo düşünülür. Kurulan regresyon temelli ESA mimarisi ile farklı SGO'ya sahip gürültülü konuşmaların aynı ağda iyileştirilmesi amaçlanır. Böylece önerilen ağın farklı SGO'larındaki performansı görülmüş olacaktır. İyileşme ölçütü olarak KKAD ve KSNN değerleri incelenecektir. Temel amaca hizmet etmesi bağlamında, ağ girişine uygulanan LGS katsayılarının frekans-zaman ilişkisi gözlenecektir. Böylece uygun ağ girişi boyutları ve devamında evrişim katmanında kullanılacak filtrelerin boyutları belirlenecektir.

1.3 Hipotez

Bu çalışmada, konuşma iyileştirme problemi analiz edilecektir. Klasik yöntemlerden farklı olarak ESA mimarisi kullanılacaktır fakat probleme olan yaklaşım filtreleme mantığından uzak değildir. ESA mimarisine ihtiyaç duyulmasının nedeni gürültünün türüdür. Gevezelik gürültüsü istatistiksel olarak modellenmesi zor bir gürültüdür. Bu yüzden klasik filtreleme yöntemleriyle başarılı sonuçlar vermez. Bu bağlamdan yola çıkarak önerilen ESA mimarisini bir adaptif filtre gibi düşünebiliriz. Bir adaptif filtre tasarlarken, uygun filtre katsayıları filtre çıkış hatasını azaltan; mümkünse en küçük yapan değerler olarak seçilir. Nümerik çözülemeyen bir adaptif filtre denklemi düşünürse (örneğin yule-walker), denklemi çözmek için bir takım optimizasyon algoritmaları kullanılabilir. Aynı mantık çerçevesinde düşünüldüğü takdirde, önerilen ESA mimarisine bir optimizasyon problemi gözüyle bakılabilir. Bu yaklaşım ESA'ların çalışma mantığına uygundur. Ağ, yapısında bulunan katmanlardaki katsayıları değiştirmek suretiyle tamınlanan kayıp fonksiyonunu (loss function) azaltmaya çalışır. Bu işlem için sıklıkla stokastik dereceli alçalma (stochastic gradient descent) optimizasyon algoritması kullanılır. Bu yaklaşım göz önüne alındığında, ESA ile konuşma iyileştirme işlemi yapılırken oluşturulacak mimari için iyi bir başlangıç noktası yakalanmış olacaktır.

Bir başka değinilmesi gereken konu ise, ESA'ları için genelde girişine ham veri (yani herhangi bir öznitelik çıkarılmaksızın) uygulanırken, bu çalışmada LGS katsayıları giriş olarak uygulanacaktır. Bu uygulamanın nedeni yukarıda bahsettiğimiz yaklaşımdır. Yani ağ eğitimini bir optimizasyon problemi gibi düşündüğümüzde en iyi sonucu bulmak için eldeki tüm ön bilgi kullanılmalıdır. Böylece ağın katman sayısı azaltılırken, başarısı istenilen düzeyde tutulabilir. LGS katsayıları, konuşma üzerinden öznitelik çıkarma konusunda yaygın kullanılan yöntemlerden biridir [11]. LGS katsayıları, sinyalin zaman frekans ilişkisine dair bilgi içerdiğinden ve konuşmanın tekrar sentezlenmesi sırasında da kolayca uygulanabildiğinden dolayı tercih edilecektir.

BÖLÜM 2

KONUŞMA SİNYALİNİN ÖZELLİKLERİ

Konuşma iyileştirme konusunda, öncelikle üzerinde çalıştığımız sinyal sınıfının istatistiksel özelliklerini ve fiziksel yapısını irdelememiz gereklidir. Böylece, ilgili sinyal sınıfının güçlü ve zayıf yanlarını rahatca kavrayarak geliştirmekte olduğumuz algoritmarda kolayca neden sonuç ilişkisi kurabiliriz. Bu bölümde, konuşma sinyalinin genel özelliklerine kısaca değinilecektir.

2.1 Konuşma Sinyali

Tarih boyunca canlıların organize olup daha karmaşık toplumlar oluşturmasının temelinde aralarında geliştirdikleri iletişim sistemi yatmaktadır. İnsanlığın büyük medeniyetler kurmasını sağlayan gelişmiş iletişim sistemidir. Günümüzde bulunan haberleşme sistemleri göz önünde bulundurulursa, artık farklı kıtada yaşayan insalar arasında bile iletişim dolayısıyla konuşma yapabilmek mümkün duruma gelmiştir. Konuşma literatürde, iki ya da daha fazla bireyin sözlü bildiri alışverişinde bulunması, sözlü bildirişim, görüşüp danışması şeklinde tanımlanır. Mühendislik boyutunda ise konuşma sinyali, akustik bir sinyal olarak tanımlanabilir. Temelinde sinüzoidal dalgalardan oluşur, farklı frekanslar ve genlikler arasında gidip gelen dalgalardır. Konuşma sinyali istatistiksel olarak durağan olmayan (non-stationary) bir özellik sergilemektedir. Yani, zamanla bir takım istatistiksel özellikleri değişmektedir. Bu durum nedeniyle konuşma sinyali üzerinde bir analiz yapmak güçleşir. İlgili sinyalin durağanlığı ile ilgili analiz yapabilmek için etkili bir yöntem olarak geniş anlamda durağanlık (wide sense stationary) kullanılır. Bu yüzden konuşma sinyali üzerinde çalışırken, sinyali kısa sürelerde pencereleyerek (örneğin, 16ms ya da 32ms gibi)

durağan olduğu kabul edilen uzunluklar üzerinden Fourier analizi yapılabilir (bu yaklaşım ve durağanlık bir sonraki bölümde detaylandırılacaktır).

Fiziksel bağlamda ise, konuşma sinyalinin insan anatomisi tarafından nasıl üretildiği ve algılandığını bilmek ön bilgi anlamında faydalıdır. Akciğerlerdeki hava öncelikle gırtlağa gelir. Gırtlak bir takım kas ve bağlarla vokal kıvrımları kontrol eder. Vokal kıvrımlar önemli noktalardan biridir. Çünkü bu kıvrımların titreşmesi sonucu sesler açığa çıkar. Son olarak ise vokal yol, yani burun ve ağız boşluğunun yer aldığı kısım konuşma sinyalini şekillendirir. Bu yaklaşımda ise bu sistem, konuşma sinyalini şekillendiren bir doğrusal süzgeç olarak görülebilir. Bu süzgecin değişkenleri dil, dişler ve dudak olarak düşünülebilir [12],[13]. İlerleyen alt bölümlerde ise KSFD ve durağanlık konuları üzerinde durulacaktır. Akustik özelliklerden ise kısaca bahsedilecektir.

2.2 Kısa Süreli Fourier Dönüşümü Ve Geniş Anlamda Durağanlık

Bu bölümde, çalışmada kullandığımız ve durağan olmayan sinyal grupları için epey faydalı bir dönüşüm olan KSFD ve geniş anlamda durağanlık (GAD) açıklanacaktır. Konuşma sinyalinin istatistiksel özelliklerinden dolayı neden KSFD kullanılması gerektiğini daha iyi içselleştirebilmemiz için öncelikle istatistiksel olarak durağanlığın iyi kavrabilmesi gerekir.

2.2.1 Geniş Anlamda Durağanlık

Rasgele bir sürecin istatistiklerinin zamandan bağımsız olmasına durağanlık ya da istatistiksel zamandan bağımsızlık denir. İlgili sürecin durağanlığına karar vermek için farklı derecelerdeki olasılık yoğunluk fonksiyonlarının zamandan bağımsız olması gerekir. Örneğin, *x[n]* bir rasgele süreç olsun;

$$f_{x(n)}(a) = f_{x(n+k)}(a) \tag{2.1}$$

Şeklinde yazılabilen birinci dereceden olasılık yoğunluk fonksiyonu ile *x[n]* sinyaline "birinci dereceden durağan" denilir. İkinci dereceden birleşik olasılık yoğunluk fonsiyonuda benzer olarak zamandan bağımsız olmalıdır. Bu işlem yüksek dereceden birleşik olasılık yoğunluk fonksiyonlarının kontrol edilmesiyle devam eder.

$$f_{x(n_1),x(n_2)\ldots x(n_L)}(a) = f_{x(n_1+k),x(n_2+k)\ldots x(n_L+k)}(a) \quad (2.2)$$

Eşitlik 2.2'de L'inci dereceden birleşik olasılık yoğunluk fonsiyonunun istatistiksel olarak zamandan bağımsızlığı gösterilmiştir. Böylece tüm dereceler için bulunan durağanlık için sinyale "tam anlamda durağan" (strict sense stationary) denir. Görüldüğü gibi tam anlamıyla durağanlığın hesap zorluğu nedeniyle, GAD sinyal sınıfları üzerinde incelenmesi daha kolay ve yararlı bir yöntemdir. Aşağıda belirtilen 3 koşulu sağlaması durumunda ilgili sinyal için GAD denilir [14].

1) Sürecin ortalaması sabit olmalıdır, $m_x(n)=m_x$

2) Özilişki fonksiyonu $r_x(k,l)$ yalnızca k-l farkına bağlı olmalıdır.

3) Sürecin varyansı sonlu olmalıdır, $c_x(0) < \infty$

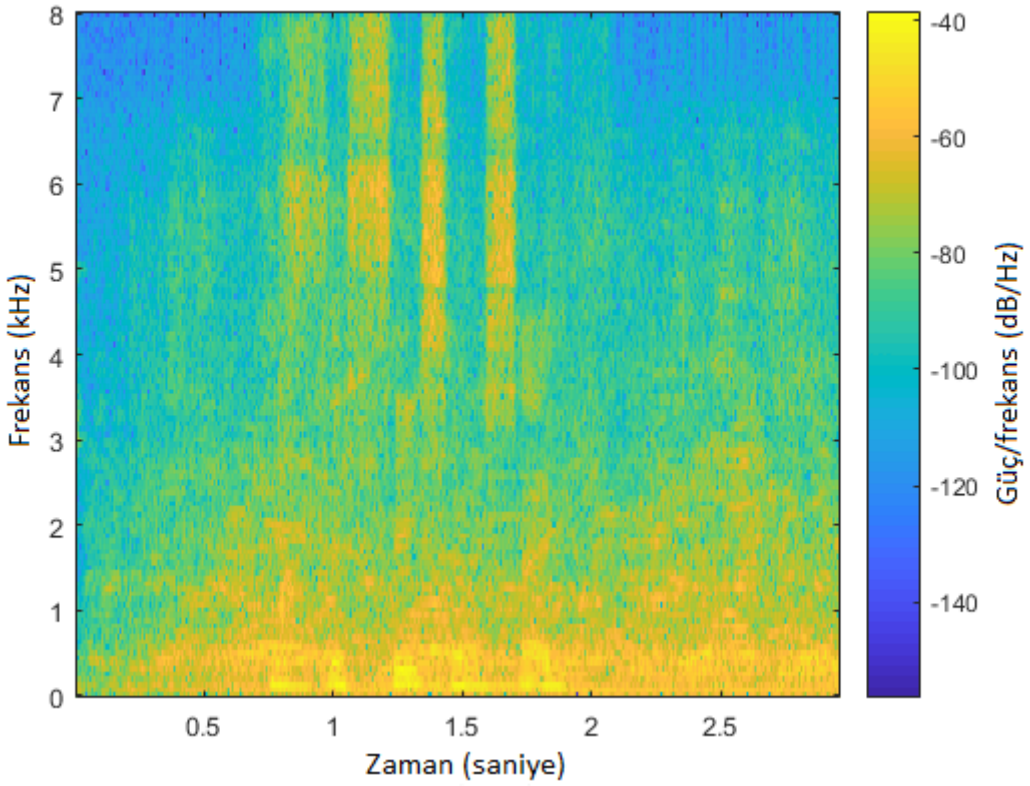
Daha öncede değindiğimiz gibi durağan olmayan sinyal sınıfları üzerinde analiz yapmak zor bir görevdir. Örneğin bir Fourier analizi yapıldığını var sayalım. Kesikli zaman durağan olmayan bir rasgele sinyale kesikli Fourier dönüşümü (KFD) yapılsın. Analiz için kullanılan KFD zaman bilgisini tamamen ortadan kaldırdığı ve yalnızca frekans spektrumu sonuçlarını yansıttığı için durağan olmayan yani istatistiksel özellikleri zamana bağlı olan sinyal kümesi için yanıltıcı sonuçlar oratya çıkaracaktır. Bu yüzden durağan olmayan sinyal sınıfları için geliştirilmiş olan KSFD kullanmak gereklidir.

2.2.2 Kısa Süreli Fourier Dönüşümü

Konuşma sinyalinin durağan olmayan doğasından dolayı, frekans analizi yapılırken faydalanılan yöntem KSFD'dir. KSFD, ilgili sinyalin üzerinde kısa süreli pencereler ile çalışmaktadır. Böylece sinyali yaklaşık olarak durağan varsayıldığı sürelerde ayırıp Fourier analizini yapmış olur. Sinyalin durağan varsayıldığı zaman aralığı pek çok uygulamada 16 ms ile 32 ms arasında değişir. Kesikli sinyalin örnekleme frekansına bağlı olarak uygun sürelere karşılık gelecek pencere uzunlukları seçilir. Örneğin 16 kHz ile örneklenmiş bir sinyal için seçilen 256 örnek uzunluklu pencere 16 ms'ye karşılık gelmektedir. Ardından bir sonraki pencere ile secilen örnek uzunluğunda örtüşecek (overlap) şekilde pencereler sinyal boyunca kaydırılır. Her pencere içerisinde ayrı ayrı KFD uygulanır ve zaman-frekans ilişkisi olan bir çıktı elde edilir. Bu zaman-frekans bilgisini barındıran yapıya spektrogram denir [15].

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x(m)v(n-m)e^{-j\frac{2\pi}{N}\omega m} \quad (2.3)$$

Eşitlik 2.3'de verilen değişkenle; *x(m)* ayrık sinyal, *v(m)* ise pencere fonksiyonudur. Notasyonda da görüldüğü gibi *X(n,ω)* cıktısı iki değişkenlidir. Burada verilen *n* zaman indisi ω ise frekans indisidir. Şekil 2.1'de bir spektrogram örneği olarak bu çalışmada kullandığımız gürültülü konuşma sinyallerinden birinin spektrogramı verilmiştir. KSFD'nin genlik spektrumu $|X(n,\omega)|^2$ şeklinde çizdirilmiştir.



Şekil 2.1 Çalışmada kullanılan gürültülü bir konuşma sinyalinin spektrogramı

Spektrogram yorumlanırken, yatay eksen boyunca değişen zaman eksenine ve dikey eksen boyun verilen frekans değerlerine dikkat edilmelidir. Verilen farklı renkler, şeklin yanındaki skalada belirtilmiş güç değerleridir. Yani herhangi bir zaman aralığında hangi frekans bileşeninin gücünün ne olduğu spektrograma bakılarak yorumlanabilir.

2.3 Akustik Özellikler

Konuşma sinyalinin bir başka faydalı özellik grubuda akustik özelliklerdir. Bu fenomenler insanların sesi nasıl algıladığının üzerinde durduğu için yararlı öznitelikler sunarlar.

2.3.1 Mel Frekansı Kepstrum Katsayıları

Bu yöntem, insanların farklı frekanslardaki sinyalleri nasıl algıladıklarına dayalı bir model sunar. Sunulan Mel ölçeği değerleri, dinleyiciler tarafından algılanan sinyaller ile eşleşmesi şeklinde çalışan kavramsal bir olgudur. Yani, Mel bir haritalama ölçeğidir, bir tonun algılanma sıklığının ölçüsüdür. Eşitlik 2.4 Mel frekansı ile gerçek frekans arasındaki ilişkiyi göstermektedir [13].

$$F_{mel} = \frac{1000}{\log(2)}\left(1+\frac{F_{Hz}}{1000}\right) \tag{2.4}$$

Mel ölçeği deneysel olarak elde edilir. Dinleyiciler 1 kHz referans tonundan başlayarak artırılan sinyalleri dinlerler ve algıladıkları tonları belirtirler.

Kepstrum ifadesi ise kısaca, bir zaman sinyalinin logaritmik spektrumunun ters Fourier dönüşümüdür. Ardından, Mel ölçeği ile kepstrumun birleştirilmesi ile Mel frekansı kepstrum katsayıları (MFKK) elde edilir.

2.3.2 Algısal Doğrusal Öngörü

Algısal doğrusal öngörünün (ADÖ) amacı insan duyusunun psikofiziğini daha doğru tanımlamaktır. ADÖ ile çıkarılan öznitelikler kısa süreli konuşma spektrumuna dayanır. Konuşma spektrumunu psikofiziksel dönüşümler ile değiştirir.

2.4 Değerlendirme Yöntemleri

Konuşma iyileştirme algoritmalarının ne kadar başarılı olduğunu söyleyebilmek adına gerekli olan performans ölçütleri mevcuttur. Bunlardan literatürde sıklıkla kullanılan iki tanesine değinilecektir.

2.4.1 Kısa Süreli Nesnel Netlik

KSNN, zaman-frekans (Z-F) ağırlıklı konuşma ile birlikte kullanılmak üzere tasarlanış bir hedef akıllılık ölçüsüdür (Objective Intelligibility Measure). Yani komşu olan Z-F temiz konuşma birimleri ve iyileştirilmiş gürültülü konuşmanın Z-F birimleri arasındaki benzerlik ile ilişkili olduğu varsayımından yola çıkılarak tasarlanmıştır. KSNN hesaplaması üç farklı aşamaya ayrılabilir: Z-F dönüşümü, orta düzeyde anlaşılabilirlik ölçümü ve nihai anlaşılabilirlik ölçümü [16]. Z-F dönüşümü, incelenmekte olan sinyalin, örtüşen hamming pencereleriyle çerçevelere bölünmesiyle gerçekleştirilir. Çerçeveler daha sonra uygun bir şekilde sıfır doldurulmuş ve bir KFD ile frekansa dönüştürülür. Frekans uzayına dönüştürülmüş çerçeveler daha sonra yaklaşık 150 Hz ile 3.8 kHz arasında bir frekans aralığını kapsayan 15 adet bir-üç oktav bandına ayrılır. KSNN, 10 kHz'lik bir örnekleme frekansı varsaymaktadır, bu nedenle bu gereksinimleri karşılamayan sinyaller yeniden örneklenmelidir. Her bir Z-F birimi için orta düzeyde anlaşılabilirlik ölçüsü hesaplanır. Temiz konuşma sinyalinin bir-üç oktav bandları ile iyileştirilmiş konuşma sinyalinin bir-üç oktav bandları arasındaki doğrusal ilişki katsayısı olarak tanımlanır. Nihai anlaşılabilirlik ölçümü, zaman ve frekans boyunca tüm orta düzeyde anlaşılabilirlik ölçümleri arasındaki ortalama olarak hesaplanır. KSNN ölçüsünün -1 ile 1 arasında bir sayı olduğu ve 1'in en yüksek anlaşılabilirlik puanı olduğu anlamına gelir [15], [16].

2.4.2 Konuşma Kalitesinin Algısal Değerlendirilmesi

KKAD, bir haberleşme ağıdan gönderilen konuşma sinyalinin kalitesini değerlendirmek için kullanılır. Temiz ve gürültülü konuşma sinyalinin spektral karşılaştırılmasına dayanır.

$$KKAD = a_0 + a_1 D + a_2 A \tag{2.5}$$

Eşitlik 2.5'de KKAD'nin nasıl hesaplandığı verilmiştir. *D* ve *A* sırasıyla ortalama bozulma ve ortamala asimetrik rahatsızlık değerleridir. Ayrıca a_0, a_1 ve a_2 değişkenleri 4.5, -0.1 ve -0.0309 değerlerine karşılık gelir [28].

BÖLÜM 3

KONUŞMA İYİLEŞTİRME YÖNTEMLERİ

Konuşma iyileştirme problemi uzun yıllardır çalışılmasından dolayı yöntem bakımından zengin bir içeriğe sahiptir. Gerek klasik yöntemler, gerekse derin öğrenmeyle ilgili olan yöntemler için ortak amaç, konuşma sinyali üzerindeki gürültünün mümkünse kaldırılması ya da zayıflatılmasıdır.

3.1 Klasik Yöntemler

Klasik yöntemlerden bahsedildiğinde süzgeçler ve kestirimciler akla gelir. Bu yöntemler sinyalin ve gürültünün iyi modellenmesine bağlıdır. Aksi halde istenen sonuçları vermeleri mümkün olmaz. İlgili sinyalin Eşitlik 3.1'de matematiksel gösterimini yapacak olursak,

$$y(n) = s(n) + g(n) \qquad [3.1]$$

s(n) temiz konuşma sinyalini, *g(n)* gürültüyü ve *y(n)* ise bozulmuş yani gürültülü konuşma sinyalini temsil eder. İstatistiksel olarak *s(n)* sinyali *g(n)* gürültüsünden bağımsızdır. Bu uygulama için oldukça faydalı bir varsayımdır. Çünkü klasik yöntemler, örneğin Wiener süzgeci, çapraz ilişki fonksiyonunu kullanır ve bu bağımsızlık varsayımı işlemlerin basitleşmesi anlamıda faydalı bir kullanım sağlamaktadır. Alt bölümlerde sıklıkla kullanılan klasik yöntemlere değinilecektir.

3.1.1 Spektral Çıkarma

Konuşma iyileştirmede çok yaygın kullanılan yöntemlerden birisi spektral çıkarma yöntemidir. Basit bir mantığa dayalı olmasına rağmen etkilidir. Gürültülü konuşma

spektrumu üzerinden, gürültü spektrumunu çıkarmak vasıtasıyla temiz konuşma spektrumunu elde etmeye çalışır.

$$Y(\omega) = S(\omega) + G(\omega) \tag{3.2}$$

Eşitlik 3.2'de verilen ifadeler Eşitlik 3.1'de verilen zaman ifadelerinin Fourier analizi karşılığı frekans spektrumudur.

$$\hat{S}(\omega) = Y(\omega) - \hat{G}(\omega) \tag{3.3}$$

Eşitlik 3.3'de gösterildiği gibi kestirilen gürültü spektrumu, gürültülü konuşma spektrumundan çıkarılıp temiz konuşma spektrumu elde edilir. Buradaki başlıca problem gürültü dağılımını kestirmektir. Gürültünün modeli ile ilgili bir ön bilgi mevcutsa ve gürültü spektrumu yüksek doğruluk ile kestirilebiliyorsa başarılı sonuçlar verebilen bir yöntemdir [13].

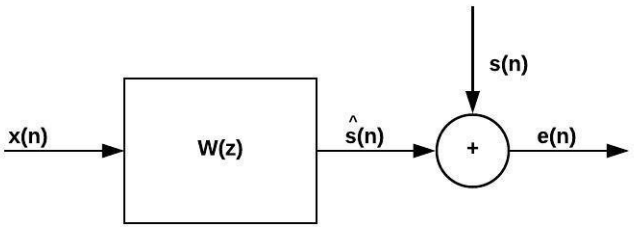
3.1.2 Wiener Süzgeci

Optimum süzgeç olarakta adlandırılan Wiener süzgeci bir çok alanda yaygın bir şekilde kullanılmaktadır. Tasarlanan süzgeç yardımıyla gürültülü sinyaden, istenen sinyali kestirmek amacıyla kullanılır. Eşitlik 3.1'de verilen denklemi düşünelim ve sinyallerin geniş anlamda durağan olduğunu varsayalım. İstenilen sinyalin, yani bizim problemimiz için temiz konuşma sinyalin, minimum ortalamalı karesel hatasını (minimum mean-square error) en küçük yapacak süzgeç katsayıları aranır [14].

$$e(n) = s(n) - \hat{s}(n) \tag{3.4}$$

$$\xi = E\{|e(n)|^2\} \tag{3.5}$$

Eşitlik 3.5'de *E{.}* şeklinde verilen ortalama (mean) operatörüdür. Şekil 3.1'de genel bir Wiener süzgeci modeli verilmiştir. Burada belirtilen *W(z)* Wiener süzgeci katsayılarıdır.



Şekil 3.1 Genel Wiener süzgeç modeli [14]

Wiener süzgeci, sonlu dürtü tepkili (finite impulse response) ve sonsuz dürtü tepkili (infinite impulse response-FIR) olmak üzere ikiye ayrılır. Kullanım açısından sonlu dürtü tepkili (SDT) süzgeç kararlı ve göreceli olarak değerlendirmesi daha kolay olduğu için yalnızca SDT Wiener süzgecine değinilecektir.

Sonlu dürtü tepkili bir Wiener süzgeci için Wiener-Hopf denklemi kullanılır. Burada Wiener-Hopf denklemini türetilmeyecektir. İlgili denklemin türetilmesi için [14] referansının 337-339 sayfaları arasına bakılabilir.

$$\mathbf{R}_x\mathbf{w} = \mathbf{r}_{sx} \tag{3.6}$$

Eşitlik 3.6'da Wiener-Hopf denklemi verilmiştir. $\boldsymbol{R}_x$ gürültülü konuşma sinyalinin özilişki matrisidir, $\boldsymbol{w}$ süzgeç katsayıları vektörü ve $\boldsymbol{r}_{sx}$ ise temiz ile gürültülü konuşma sinyalleri arasındaki çapraz ilişkinin vektörüdür. Ayrıca *s(n)* ile *x(n)* sinyalleri birleşik geniş anlamda durağan (jointly wide sense stationary) olmalıdır. Süzgecin derecesine göre bahsi gecen matris ve vektörlerin boyutları değişecektir. Belirtilmesi gereken başka bir husus ise $\boldsymbol{R}_x$ özilişki matrisi, hermitian toeplitz matrisi özelliklerini gösterir.

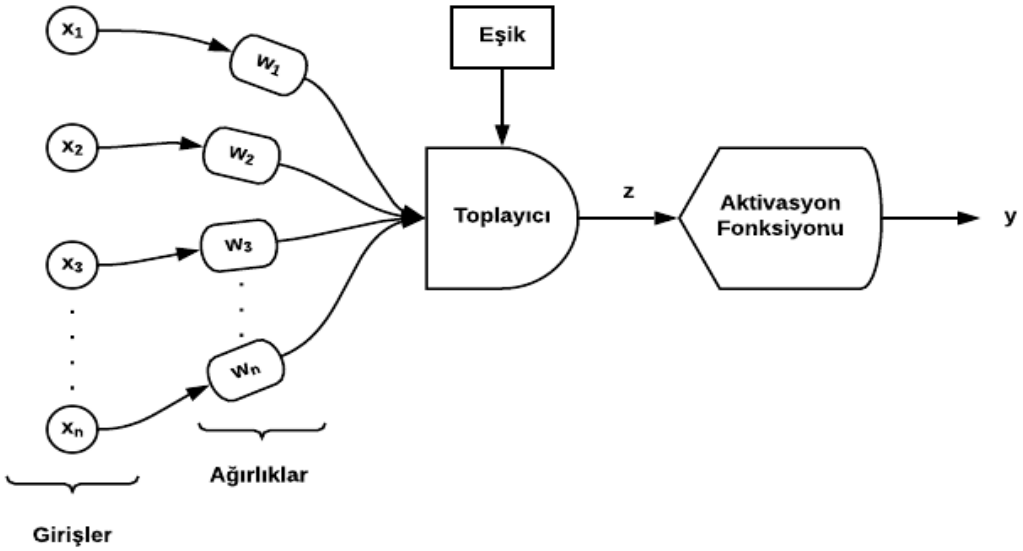
$$\mathbf{w} = \mathbf{R}_x^{-1}\mathbf{r}_{sx} \tag{3.7}$$

Eşitlik 3.7'de süzgeç katsayılarını elde etmek için Eşitlik 3.6'nın düzenlenmiş hali verilmiştir. Yaşanabilecek problemlerden birisi, $\boldsymbol{R}_x$ matrisinin tersini almakla ilgili olabilir. Örneğin ilgili matris tekil (singular) olabilir, böyle bir durumda analitik bir yöntemle tersini elde etmek mümkün değildir. Fakat sezgisel yöntemlerle yaklaşık bir çözüm bulunabilir. Ayrıca, Wiener süzgecinin uygun katsayılarını hesaplayabilmek için sinyalin ve gürültünün modelini doğru çıkarmak önemlidir.

3.2 Yapay Sinir Ağları

Bu bölümde yapay sinir ağlarına (YSA) değinilecektir. Böylece konuşma iyileştirmede kullanılan derin öğrenmeyle ilgili gerekli olan ön bilgi sunulmuş olacaktır. Yapay sinir ağlarının kullanım hissiyatının altında insan beyninin çalışma mantığı yatmaktadır. İnsan beyninde bulunan milyarlarca sinirin karmaşık bağlantıları, tarih boyunca ortaya çıkmış tüm teknolojinin, düşünecelerin ve bilimsel ilerlemenin tek kaynağıdır. Temel olarak tek bir sinir hücresinin tek başına yaptığı işlem çok basittir. Fakat yeterli

miktarda hücrenin bağlanmasıyla karmaşık problemler çözülebilir. Yarı iletken teknolojilerinin ilerlemesiyle ortaya çıkan bilgisayarlar sayesinde insan beynindeki ağ yapısına benzer yapay sinir ağları yapma imkanı doğmuştur. Araştırmacılar beyindeki sinir yapısının öğrenme kabiliyetini taklit ederek birçok probleme çözüm üretebilen modeller oluşturmayı başarmıştır. Yapay sinir ağları genel olarak, kuramı bilinmeyen problemlerde örneklerin gözlemlenmesi yoluylu geliştirilen bir model olarak özetlenebilir. İnsan beyninde bulunan sinir hücrelerine nöron denir. Temel görevleri uyarı sinyalleri almak, iletmek ve cevap vermektir. Bu işlemleri elektriksel ve kimyasal bir takım işlem sayesinde yapar. Nöron yapısı, hücre gövdesinden dallanan dendritler ve aksondan oluşmaktadır. Dendritler sayesinde farklı hücrelerden gelen sinyaller alınabilir ve hücre gövdesi üzerinden aksona aktarılmak suretiyle aksondan çıktı olarak bir başka nörona iletilebilir. Bu işleyiş yapay sinir ağında taklit edilmiştir.



Şekil 3.2 Yapay sinir ağının genel şeması

Bir yapay sinir ağının yapısı Şekil 3.2'de gösterilmiştir. Girişler (x) farklı katsayılar (w) ile ağırlandırılmak suretiyle toplayıcıya aktarılır, toplanan sinyallere eşik değeri (θ) eklenir. Ardından oluşturulan ara değer (z) aktivasyon fonksiyonundan geçerek çıkış sinyali (y) oluşturulur.

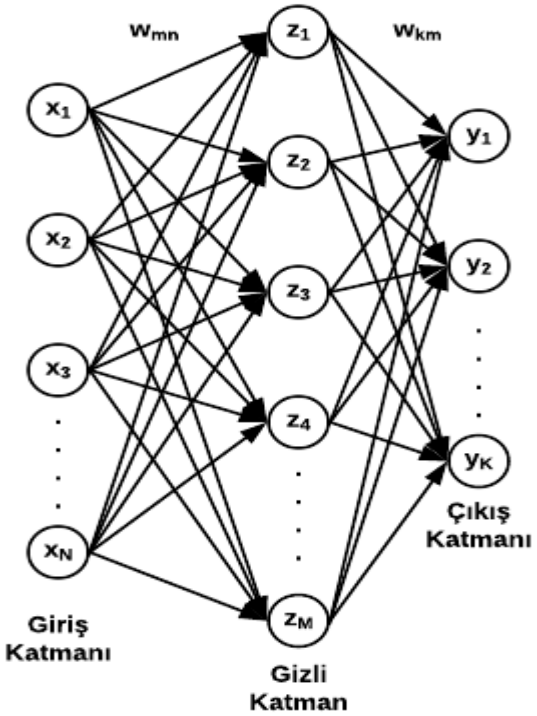
$$z = \left(\sum_{i=1}^{n} w_i x_i \right) + \theta \tag{3.8}$$

$$y = f(z) \tag{3.9}$$

Eşitlik 3.8 ve Eşitlik 3.9'da giriş sinyalleri ile çıkış sinyali arasındaki ilişki verilmiştir. Burada bahsi geçen aktivasyon fonksiyonu *f(.)* olarak kullanılan birçok fonksiyon vardır. Bunlardan en bilinenleri parçalı doğrusal fonksiyon, sigmoid fonksiyonu ve hiperbolik tanjant fonksiyonudur. Aktivasyon fonksiyonun görevi, gelen değerleri belirli iki sayı arasında haritalamak (mapping) ve çıkış değerini oluşturmaktır. Örneğin sigmoid fonksiyonu 0 ile 1 arasında değerler üretirken hiperbolik tanjant fonksiyonu -1 ile 1 arasında üretmektedir. Eşitlik 3.10'da sigmoid fonksiyonu verilmiştir.

$$f(z) = \frac{1}{1+e^{-z}} \tag{3.10}$$

Nöronların birbirleriyle bağlanmaları sonucunda çeşitli sinir ağları yapılarını oluştururlar. Bu yapılar ileri beslemeli sinir ağı (feedforward neural network), tekrarlamalı sinir ağı (recurrent neural network) ve evrişimsel sinir ağı (convolutional neural network) gibi mimarilere sahiptir. YSA genel olarak döngüsel (cyclic), döngüsüz (acyclic), tamamen bağlı (fully connected) yada kısmen bağlı (partially connected) formlarda olabilir. YSA için tamamen bağlı özelliği, tüm katmanlardaki nöronların kendilerinden sonra gelen katmandaki tüm nöronlara bağlı olması anlamına gelir. En basit ve sık kullanılan örneği ileri beslemeli sinir ağıdır (İSA). Şekil 3.3'de İSA mimarisi verilmiştr [20].



Şekil 3.3 İSA mimarisi [20]

Giriş katmanı verinin boyutuyla ilgilidir yani giriş sayısı kadar giriş katmanında nöron bulunmalıdır. Gizli katmandaki nöron sayısının nasıl belirleneceğine dair bir kural yoktur fakat genellikle giriş katmanında fazla seçilir. Çıkış katmanı ise sistemde kaç çıkış varsa o sayıda seçilmektedir.

$$y_k = f_2\left(\sum_{m=1}^{M} w_{km} f_1\left(\sum_{n=1}^{N} w_{mn} x_n\right)\right) \tag{3.11}$$

Yukarıda verilen İSA mimarisinin matematiksel gösterimi Eşitlik 3.11'de yapılmıştır. Katmanlar arasındaki katsayılar w_{km} ve w_{mn} şeklinde verilmiştir, sırasıyla giriş katmanından gizli katmana ve gizli katmandan çıkış katmanına geçişin katsayılarıdır. Ayrıca f_1 ve f_2 sırasıyla birinci ve ikinci katman arasındaki aktivasyon fonksiyonlarıdır. İnsan beynindeki nöronların sinyali ileri yönde iletme, yani dendritten aksona doğru, gerçeğinden yola çıkılarak İSA mimarileri geliştirilmiştir [17]. Bunun yanı sıra, sinir hücrelerindeki geri yönde oluşan hareketin varlığından da esinlenerek tekrarlamalı sinir ağları (TSA) geliştirilmiştir [18]. Evrişimsel sinir ağları (ESA) ise görsel kortekstekı bölgelerden geliştirilmiştir [19].

Bir başka önemli konu ise oluşturulan YSA mimarileri için uygun katsayıların hesaplanmasıdır. Daha öncede bahsedildiği gibi kuramı bilinmeyen bir problem için kullanılan YSA'ların katsayılarını bulabilmek adına en akılcı çözüm şekilde adaptif bir yöntem ile sağlanabilir. Geri yayılım algoritması (backpropagation algorithm) YSA katsayılarının hesaplanması için yaygın kullanılan bir yöntemdir. Temelinde hatayı geri yaymak suretiyle en uygun YSA katsayılarını bulmaya çalışır. Bu aşamada problemin bir optimizasyon problemi gibi düşünülmesinde bir sakınca yoktur. Ağın uygun katsayılarının hesaplanması sürecine ağın eğitilmesi denir. Gözetimli öğrenme (supervised learning) ve gözetimsiz öğrenme (unsupervised learning) olmak üzere ikiye ayrılır. Gözetimsiz öğrenmede çıkışlara karşılık gelen çıkış bilgisi mevcut değildir, daha çok aynı özelliklere sahip girişleri bir bölgede sınıflandırmak amacıyla kullanılır. Gözetimli öğrenmede ise girişlere karşılık gelen çıkış bilgileri mevcuttur ve bunlar etiketli veri olarak da dile getirilebilir. Geri yayılım algoritması (GYA) bir hata fonksiyonu üzerinden elde ettiği hatayı en aza indirecek olan katsayıları seçmeye çalışır.

$$E = \frac{1}{2}\sum_{k=1}^{K}(d_k - y_k)^2 \tag{3.12}$$

Eşitlik 3.12'de verilen hata fonksiyonu karesel toplam hatadır. Eşitlikte d_k ile gösterilen hedef çıkışlardır.

$$\Delta\mathbf{w}(t) = -\alpha\frac{\partial E}{\partial\mathbf{w}(t)} + \beta\Delta\mathbf{w}(t-1) \tag{3.13}$$

Eşitlik 3.13'de verilenler **w**=[w_{k1}, w_{k2}, w_{k3} . . . w_{kM}] vektörü, α öğrenme katsayısı, β ise momentum katsayısıdır. Hatanın katsayılar vektörüne göre türevi ile eğim yönü bulunur ve eğim yönünün tersine doğru α adım büyüklüğünde ilerlenir ayrıca önceki adımın değişim miktarı da β katsayısı oranında eklenir böylece Δ**w** değişim miktarı hesaplanmış olur. Unutulmaması gereken bir nokta da *E'*nin **w** katsayılarının bir fonksiyonu olduğudur.

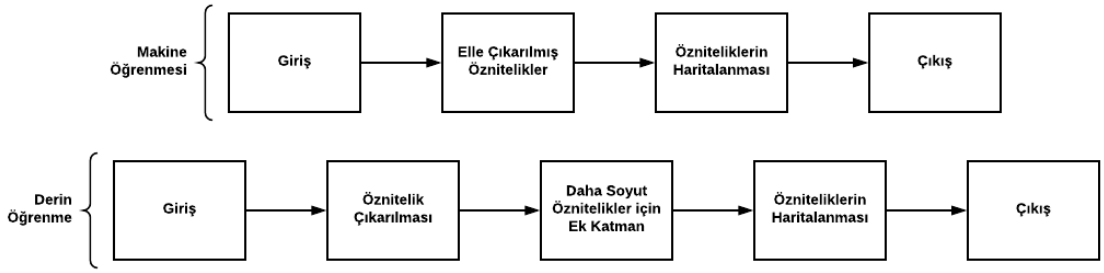
$$\mathbf{w}(t) = \Delta\mathbf{w}(t) + \mathbf{w}(t-1) \tag{3.14}$$

Eşitlik 3.14'de **w** katsayılarının nasıl güncelleneceği verilmiştir. İstenilen hata eşiğinin altına düşene ya da belirlenen iterasyon sayısına ulaşılana kadar güncellenen katsayılar

ile tekrar hata fonksiyonu hesaplanır, öğrenme adımı boyunca ilerleme ve güncelleme işlemleri tekrar edilir. İstenilen hata değerinin altına inildiğinde ya da belirlenmiş iterasyon sayısına ulaşıldığında ise eğitim tamamlanmış olur. Matematiksel ifadesi verilmiş olan **w** katsayıları gizli katman ile çıkış katmanı arasındaki katsayılardır. Bu işlemlere benzer olarak giriş katmanı ile gizli katman arasındaki katsayılarda hesaplanır. Buradaki maksat algoritmanın mantığından bahsetmek olduğu için hesaplamaların tamamına yer verilmemiştir. İlgili denklerim [20] referansının 210-216 sayfaları arasında bulunabilir.

3.3 Derin Öğrenme

Derin öğrenme, YSA ve farklı işlevli yapıları içeren çok katmanlı mimariye sahip makine öğrenmesinin bir alt dalıdır. Son zamanlarda popüler olmasına rağmen tarihi nazaran eskiye dayanmaktadır. Derin öğrenme terimi 1986 yılında Rina Dechter tarafından tanıtılmıştır [21]. Özellikle grafik işlemci ünitelerinin (graphics processing unit) gelişmesiyle derin öğrenmeye dair mimarileri gerçeklenmesi mümkün kılınmıştır. Derin öğrenmede geçen "derin" ifadesi verilerin dönüştürüldüğü katmanları ifade eder [22]. DÖ, öğreticili öğrenme için çok güçlü bir yapı sağlar. Bir girdi vektörünü bir çıkış vektörüne eşleştirmekten ibaret olan ve bir insanın hızlı bir şekilde yapabileceği çoğu görev derin öğrenme vasıtasıyla yapılabilir. Fakat bunun için uygun modeller ve yeterince büyük etiketli eğitim verilerine ihtiyaç duyulur. Bir vektörü diğerine ilişkilendirmek olarak tanımlanamayan veya bir insanın görevi yerine getirmek için düşünmek zorunda olduğu görevler, DÖ konusundaki mevcut durum göz önüne alındığında şimdilik derin öğrenmenin kapsamı dışında kalır [15].



Şekil 3.4 Makine öğrenmesi ile derin öğrenmenin akış şemaları [15]

Şekil 3.4'de klasik makine öğrenmesi mimarisiyle derin öğrenme mimarisi karşılaştırılmıştır. En belirgin fark özniteliklerin çıkarılmasıyla ilgili olan katmandır. Klasik makine öğrenmesi sırasında öznitelikler elle çıkarıldıktan sonra oluşturulan mimari eğitilmektedir. Fakat derin öğrenmede ise öznitelikler ağın içerisindeki katmanlarda çıkarılır ve ek katmanlar vasıtasıyla bu öznitelikler farklı boyutlara dönüştürülebilir. Böylece özniteliklerin çıkarıldığı katmanlarında eğitilmesiyle en uygun öznitelikler seçilebilir. Bu sayede klasik özniteliklerin arasında bulunmayan farklı uzaylardaki öznitelikler ile çalışma imkanı olur. Bu yüzden derin öğrenme mimarisi daha karmaşıktır ve daha zorlu problemler için giriş ve çıkış arasında karmaşık ilişkiler kurmayı başarır. Bu büyük karmaşık mimarinin eğitilmesi için klasik makine öğrenmesi yöntemlerinde ihtiyaç duyulandan çok daha fazla örnek gereklidir. Derin öğrenme mimarisi için gerekli olan büyük veri setinin, klasik makine öğrenmesinde olduğu gibi tek bir yığın (batch) şeklinde eğitime katılması fazlasıyla maliyetlidir. Bu yüzden küçük yığınlar (mini batch) şeklinde parça parça eğitilir. Böylece büyük miktardaki verinin eğitime dahil edilmesi problemi aşılmış olur. Fakat küçük yığınların boyutlarıyla ilişkili olarak doğru öğrenme oranını seçmek konusuna dikkat edilmelidir. DÖ için hatanın geri yayılımı esnasında stokastik dereceli alçalma algoritması (stochastic gradient descent) kullanılması yaygındır. Stokastik dereceli alçalma (SDA) dereceli alçalma (DA) algoritmasının bir versiyonudur. Hata geri yayılımı için kullanılan bu algoritmalar bir maliyet fonksiyonu marifetiyle toplam hatayı en aza indirmeye çalışır.

$$J(\theta) = \mathrm{E}_{x,y \sim \hat{p}_{veri}} L(\mathbf{x}, y, \theta) = \frac{1}{m} \sum_{i=1}^{m} L(\mathbf{x}^{(i)}, y^{(i)}, \theta) \tag{3.15}$$

Eşitlik 3.15'de verilen koşullu log-olabilirlikdir. Belirtilen θ ağın değişkenlerinin vektörünü temsil eder, L ise örnek başına kayıptır ve $L(\mathbf{x}, y, \theta) = -\log p(y \mid \mathbf{x}; \theta)$ şeklinde gösterilir.

$$\nabla_{\theta} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\theta} L(\mathbf{x}^{(i)}, y^{(i)}, \theta) = \mathbf{g} \tag{3.16}$$

$$\theta_{güncel} = \theta - \varepsilon \mathbf{g} \tag{3.17}$$

Eşitlik 3.16'da θ'e göre hesaplanan gradyan gösterilmiştir. Eşitlik 3.17'de ise ε öğrenme katsayısıdır ve değişkenlerin güncellenmesi gösterilmiştir [23].

Üzerinde durulması gereken bir başka fark ise makine öğrenmesi ile derin öğrenmenin farklı problem çözme biçimleridir. Derin öğrenme problemi bir bütün şeklinde çözmeye çalışırken makine öğrenmesinde problem alt bölümlere ayrılır. Yani, örneğin görüntü tanıma için bir problem ele alındığında, derin öğrenme görüntüyü girdi olarak alır ve çıkışta sınıfını söyler. Fakat makine öğrenmesi sırasında ise önce görüntüyü tanımlamak için bir öğrenme algoritması ardından ise bir sınıflandırma algoritmasına ihtiyaç duyulur. Bu özellikler sayesinde derin öğrenme farklı alanlara (domain) ve uygulamara kolaylıkla uyum sağlayabilir. Bu üstünlüklerin yanı sıra derin öğrenmenin de sorunları vardır. Bir problemi çok başarılı bir şekilde çözse bile nedenini açıklamaz. Matematiksel olarak tüm mimarinin modelini çıkarılsa bile nöronların neyi modellediği ve birbirleriyle olan ilişkileri bilinmemektedir. Bu nedenle problemin çözümünü yorumlamak mümkün olmaz [24].

3.3.1 Evrişimsel Sinir Ağları

Evrişim sinir ağları, YSA'nın özel bir halidir. Izgara benzeri topoloji (grid like topology) sayesinde özellikle görüntü sinyalleri üzerinde etkili çalışır. İsminde geçen "evrişim" iyi bilinen matematiksel doğrusal bir operatördür. İki fonksiyonun birbirlerini nasıl değiştirdiğini ifade etmek için üçüncü bir fonksiyon üreten işlemdir. Yani bir sinyal sisteme uygulandığında oluşturulacak çıkışı oluşturan operatördür. Örneğin, gürültülü bir sinyalin filtreden geçirilmesi olayı matematiksel olarak evrişim işlemiyle temsil edilir.

$$y(t) = \int x(a)h(t-a)da \tag{3.18}$$

Evrişim integrali Eşitlik 3.18'de verilmiştir.

$$y(t) = x(t) * h(t) \tag{3.19}$$

Eşitlik 3.19'da ise yaygın kullanılan evrişim operatörü gösterilmiştir. Burada verilenler, *y(t)* çıkış sinyal, *x(t)* giriş sinyali ve *h(t)* sistemin dürtü tepkisidir. Evrişim operatörünün bilgisayar ortamında kullanılabilmesi için ayrık versiyonu Eşitlik 3.20'de verilmiştir.

$$y(n) = \sum_{a=-\infty}^{\infty} x(a)h(n-a) \tag{3.20}$$

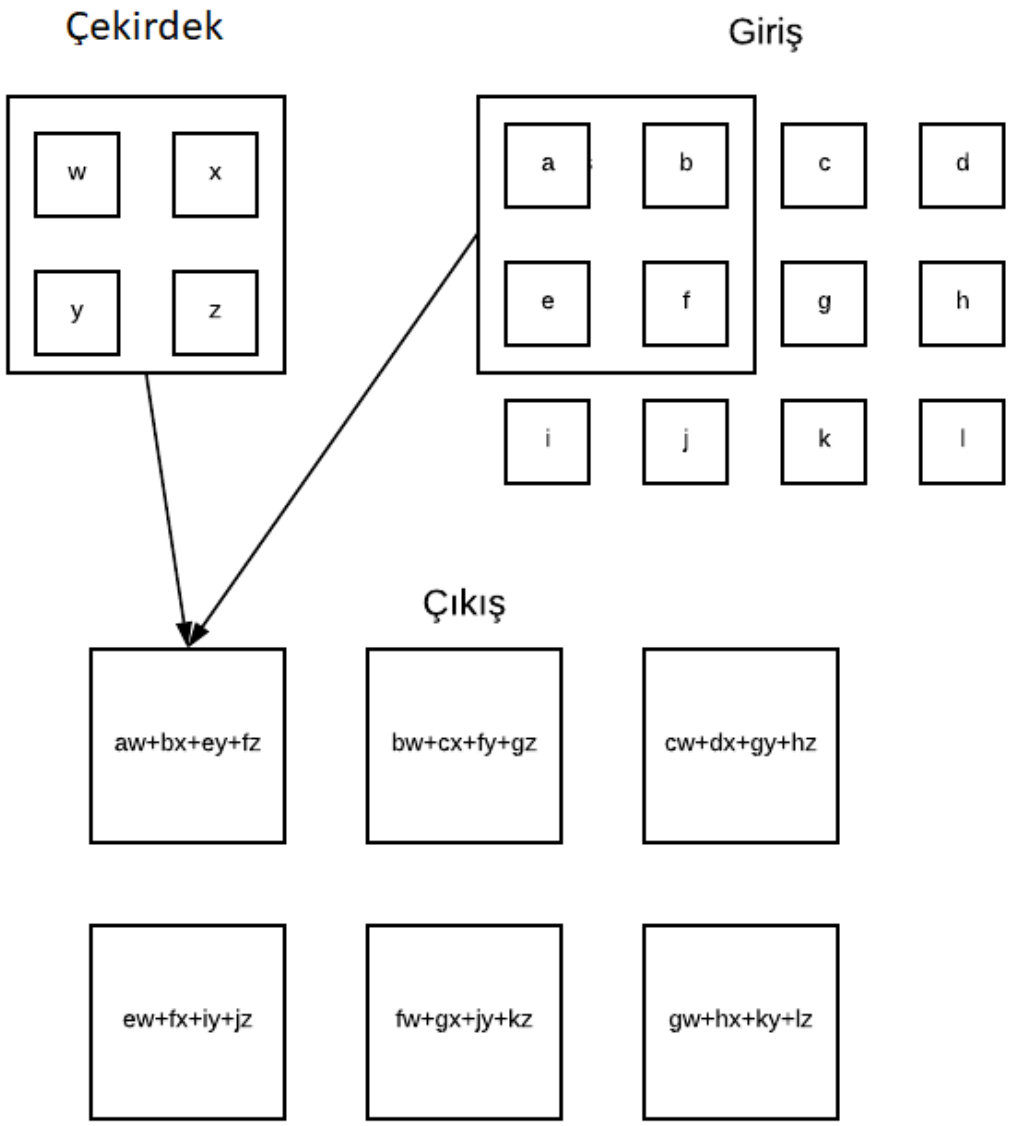
Toplam işaretinin aralığı sonsuz olsada, uygulamalarda sonsuz uzunluğunda sinyal ya da sistemler mevcut değildir. O yüzden yalnızca örnekleri bulunduğu noktalarda işlem yapılabilecek şekilde aralık daraltılır, örneklerin bulunmadığı tüm indisler sıfırdır. ESA terminolojisinde giriş fonksiyonu *x(n)* "giriş", *h(n)* fonksiyonuna "çekirdek" (kernel) ve *y(n)* çıkış fonksiyonuna öznitelik haritası (feature map) denir [15]. Görüntü gibi 2 boyutlu girişler için Eşitlik 3.21'de verilen 2 boyutlu bir evrişim operatörü kullanılır.

$$y(i,j) = \sum_m \sum_n x(m,n)h(i-m, j-n) \tag{3.21}$$

Görüntü *x(m,n)* şeklinde iki boyutlu bir fonksiyon, *h(m,n)* ise iki boyutlu çekirdek olarak verilmiştir. Şekil 3.5'de iki boyutlu evrişim örneği verilmiştir. ESA'nın evrişim katmanında gerçekleşen işlemleri anlamak için 2 boyutlu evrişim işleminin içselleştirilmesi önemlidir. Herhangi bir ESA mimarisinde filtre olarak adlandırılan çekirdekler vasıtasıyla evrişim işlemleri yapılmaktadır. Bu filtrelerin sayısı ve boyutları ilgili problemdeki giriş verisi boyutlarına göre tasarlanabilir. Tabiki, evrişim katmanının şekillenmesi yalnızca veriyle ilgili değildir. Arka arkaya kaç katman ekleneceği ve kurulan ağ ile nasıl bir problem çözülmek istendiğine de bağlı olarak değişir. Evrişim katmanı ardından gelen aktivasyon fonksiyonun tıpkı klasik YSA da olduğu gibi bir haritalama görevi vardır. Klasik aktivasyon fonksiyonlarından farklı olarak ESA için yaygın kullanılan doğrultucu doğrusal birimdir (rectified linear unit-ReLU).

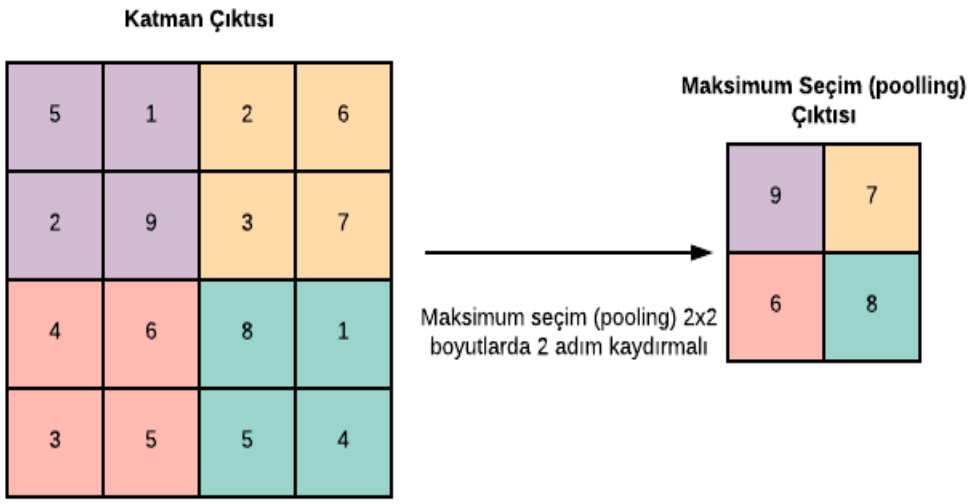
$$f(x) = x^+ = \max(0, x) \tag{3.22}$$

Eşitlik 3.22'da matematiksel gösterimi verilmiştir. Pozitif kısımda sınırsız olması, gradyanın sıfıra gitmesi problemini azaltmıştır. Fakat ReLU'nun negatif tarafının sıfır olması ve tüm negatif değerleri sıfıra götürmesi, modelin verileri uygun şekilde uydurma ve eğitme yeteneğini azaltır. Bu yüzden farklı aktivasyon fonksiyonları da geliştirilmiştir. Örneğin, sızdıran ReLU (leakly ReLU) negatif tarafın da bir doğruyla modellendiği bir fonksiyondur.



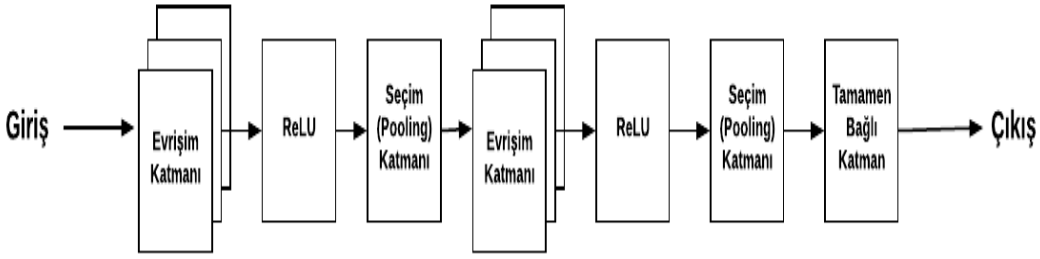
Şekil 3.5 İki boyutlu evrişim örneği [15]

Yaygın kullanılan bir başka katman ise seçim (pooling) katmanıdır. Genelde evrişim katmanları arasında sistematik olarak katman çıkışlarının boyutunu azaltmaya yönelik kullanılır. Böylece değişken sayısı azaltılır ve ağın karmaşıklığının önüne geçilir. Katman çıktılarının üzerinde dolaştırılan ve altörnekleme yapan matris şeklinde bir operatör olarak düşünülebilir. Örneğin, 2×2 boyutunda bir seçim matrisi olsun ve 16×16 büyüklüğünde ara katman çıktısı üzerinde 2 örnek uzunluğunda adımlansın (stride) böylece 4 adımda tüm yüzeyde dolaşır ve üzerinde işlem yaptığı her dört örneği bir örneğe düşürür. Pooling işlemi farklı şekillerde yapılabilir. Örneğin, Şekil 3.6'da gösterildiği gibi maksimum değer alınarak yapılabilir.



Şekil 3.6 Maksimum seçim (pooling) örneği

Birbiri ardına bağlı olan evrişim ve pooling katmanları ardından kullanılan bir başka katman ise tamamen bağlı katmandır (fully connected layer). Klasik YSA'dan farklı yoktur.



Şekil 3.7 ESA mimarisi örneği

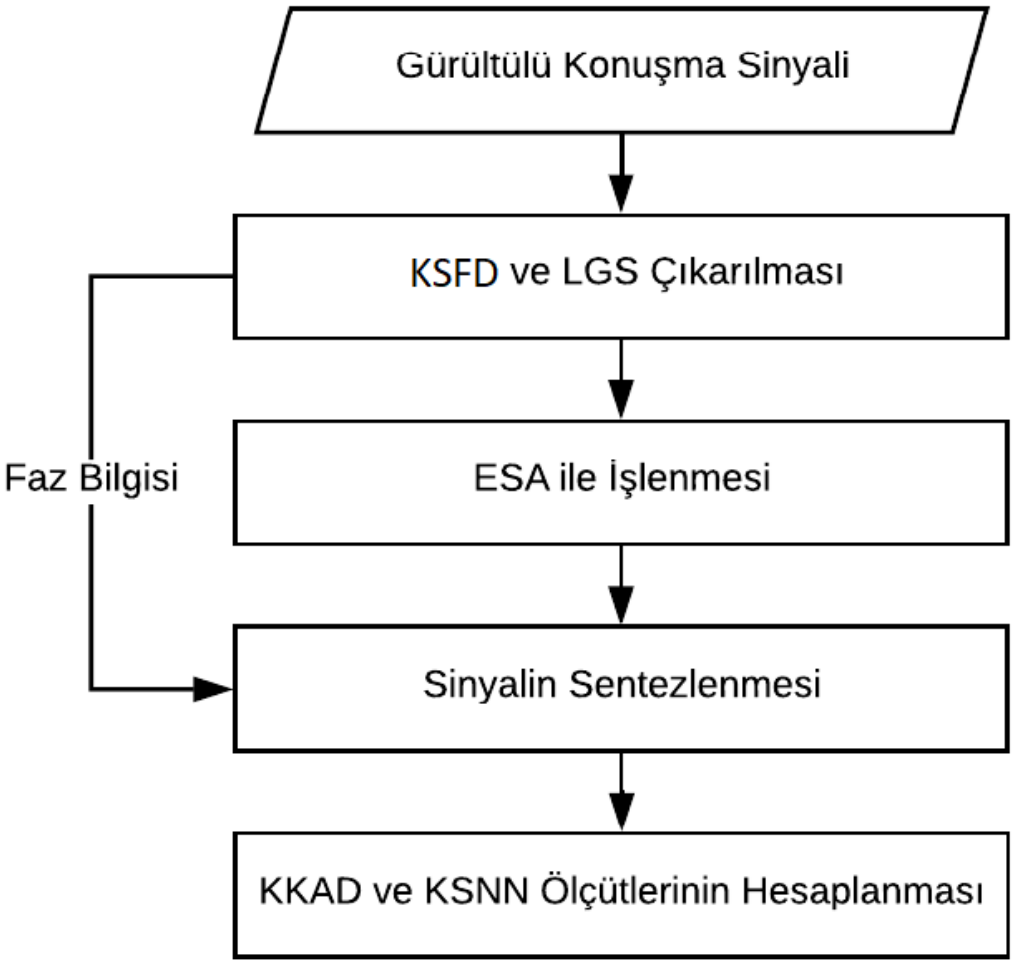
Şekil 3.7'de ESA mimarisi örneği verilmiştir. Probleme bağlı olarak ağın derinliği değiştirilir. Seçim katmanı bazı problemler için kullanılmayabilir.

BÖLÜM 4

EVRİŞİM SİNİR AĞI İLE KONUŞMA İYİLEŞTİRME

Bu bölümde çalışmanın uygulama detayları verilecektir. Önerilen mimarinin detayları ve nasıl gerçeklendiği detaylı bir şekilde aktarılacaktır. Kullanılan veri setinin hazırlanması, algoritmaların hangi ortamda geliştirildiği ve hangi kütüphaneler kullanıldığı detaylandırılacaktır.

Çalışmanın akış şeması Şekil 4.1'de verilmiştir. Farklı SGO'ya sahip veri setleri oluşturulur. Böylece önerilen mimarinin yüksek ve düşük SGO'da ayrı ayrı eğitilmesinin performansı gözlenebilecektir. Gürültülü konuşma sinyalinin LGS katsayılarını elde etmek için KSFD genlik katsayıları üzerinden işlem yapılır. Bu aşamada gürültülü sinyalin fazı saklanır. Böylece konuşma sinyalinin tekrar oluşturulacağı basamakta bu faz bilgisi kullanılacaktır. ESA mimarisi gürültülü LGS katsayılarıyla temiz LGS katsayıları arasında doğrusal olmayan bir haritalama yapacak şekilde tasarlanmıştır. Ayrıca ESA mimarisinin içerisine tamamen bağlı katmandan sonra bir regresyon katmanı eklenmiştir. Ağın çıktısındaki iyileştirilmiş LGS katsayıları önce Fourier genlik katsayılarını dönüştürülür bu aşamada faz bilgisiyle genlik bilgisi birleştirilerek tekrar KSFD katsayıları elde edilir. Ardından ise ters KSFD ile iyileştirilmiş konuşma sinyali elde edilir. Son olarak ise KKAD ve KSNN ölçütlerinin hesaplanması ile ağın konuşma iyileştirme üzerindeki performansı gözlenir.



Şekil 4.1 Çalışmanın akış şeması

4.1 Veri Seti

Çalışmada yer alan veri seti Edinburg Üniversitesi, Konuşma Teknolojisi Araştırma Merkezi (The University of Edinburg, The Centre Speech Technology Research) veri tabanından alınmıştır [25]. Veri setinde biri kadın biri erkek olmak üzere iki konuşmacıya ait 460 örnek cümle vardır. Cümlelerin ortalama uzunluğu 3 saniye uzunluğunda 16 kHz ile örneklenmiştir. Gürültülü konuşmayı oluşturmak için, veri setindeki örneklerin üzerine MATLAB ortamında gevezelik gürültüsü eklenmiştir. Farklı SGO'ya sahip veri setleri oluşturulmuştur. Yani her biri örnek 460 cümleyi de içeren -10 dB, -5 dB, 0 dB, 5dB, 10 dB SGO değerlerinde 5 farklı veri seti oluşturulmuştur. Durağan olmayan konuşma sinyalinin LGS katsayılarını elde etmek için öncelikle KSFD ile Fourier katsayıları elde edilmiştir. KSFD katsayıları hesaplanırken pencere uzunlukları 16 ms olarak ayarlanmıştır. Konuşma örneklerinin 16 kHz ile örneklendiği düşünülürse 256

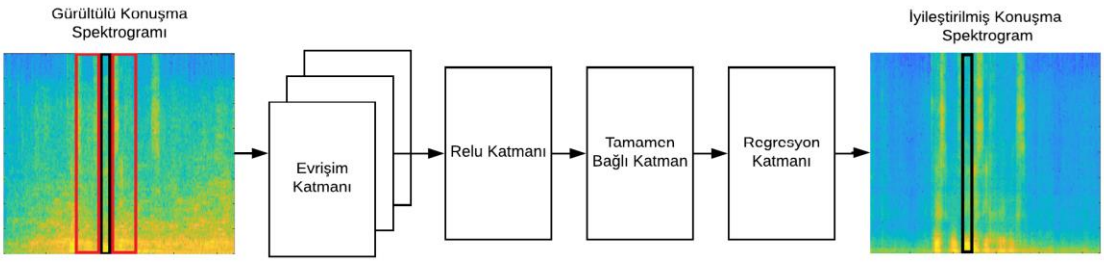
noktalı örnek 16 ms'ye karşılık gelir. Böylece konuşma sinyalinin durağan gibi davrandığı uzunluklar üzerinde çalışılır. KSFD 256 noktalı Hamming pencere (Hamming window) uzunluğu, 128 noktalı örtüşme ve 256 noktalı dönüşüm ile yapılmıştır. 460 konuşma cümlesinin dönüşümü yapıldığında toplam 226852 tane 256 Fourier katsayısına sahip vektör oluşturmuştur. Konuşma sinyali gerçek olduğu için Fourier katsayıları çift simetri özelliği gösterir, bu yüzden katsayıların yarısını almak (129 katsayı) yeterli olmuştur. Yani veri seti 129×226852 boyutlarında bir matris oluşturmuş olur. KSFD katsayılarının genlik ve faz bilgileri hesaplanmıştır. Gürültülü konuşma sinyallerinden elde edilen faz bilgisi iyileştirme işlemi sonrasında kullanılmak için saklanmıştır. Genlik değerlerinden Eşitlik 4.1'de verildiği gibi LGS katsayıları hesaplanmıştır.

$$L(n) = \log_{10} S(n)^2 \qquad n = 1, 2 \ldots 129 \tag{4.1}$$

$L(n)$ LGS katsayılarını, $S(n)$ ise KSFD genlik katsayılarını temsil etmektedir. Kısaca, temiz konuşma sinyal seti yanı sıra oluşturulan farklı SGO değerlerine sahip gürültülü konuşma sinyaller setlerinin de ayrı ayrı LGS katsayıları hesaplanır ve her SGO değerinden gürültülü sinyallerin faz bilgisi saklanır. Çünkü iyileştirilmiş ses tekrar oluşturulurken saklanan gürültülü faz bilgisi kullanılacaktır. Önerilen ESA mimarisi daha öncede belirtildiği gibi gürültülü LGS katsayıları ile temiz LGS katsayıları arasında doğrusal olmayan haritalama yapar. Her SGO için önerilen ağ ayrı ayrı eğitilir.

4.2 Evrişimsel Sinir Ağı Mimarisi

Çalışmada kullanılan ESA mimarisi Şekil 4.2'de verilmiştir. Ağın girişine yedi çerçeve (frame) çıkışına ise bir çerçeve verilerek eğitilmiştir. Yani, iyileştirilmek istenilen gürültülü konuşma çerçevesi ile onun 3 öncesindeki ve sonrasındaki çerçeveleri girişe dahil ederek, iyileştirilmek istenen indise denk gelen temiz konuşma çerçevesine doğrusal olmayan haritalama yapılmıştır. Şekil 4.2'de giriş kısmındaki siyah çerçeveli vektör iyileştirilmek istenilen gürültülü LGS katsayılarıdır, çevresindeki kırmızı çerçeveli üçer vektörden oluşanlar ise öncesindeki ve sonrasındaki gürültülü LGS katsayılarından oluşan vektörlerdir. Böylece zaman frekans ilişkisini içeren bir giriş oluşturulur. Ağdan geçen bu 7 çerçeve hedefdeki çerçevenin iyileştirilmiş halini çıkışa iletir.

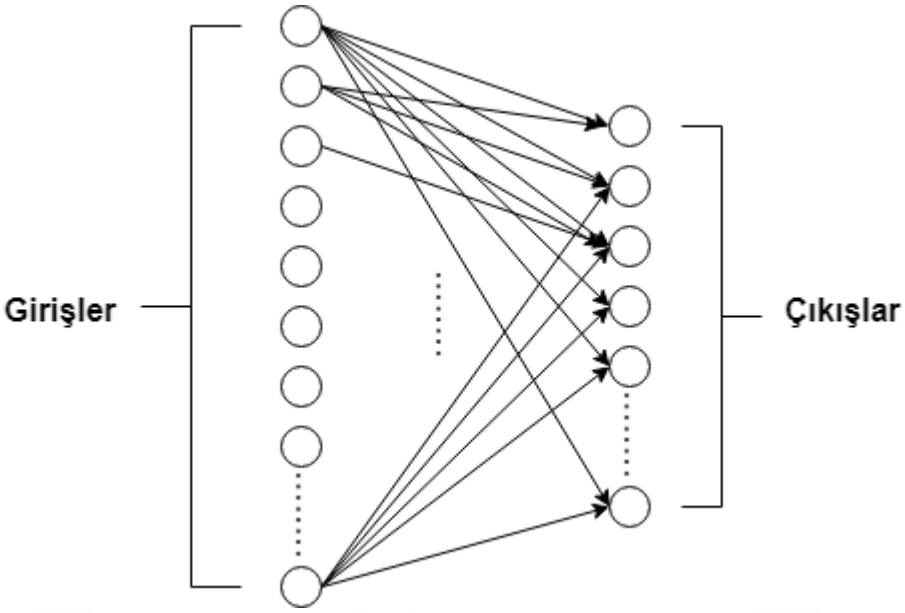


Şekil 4.2 Çalışmada kullanılan ESA mimarisi

Ağın girişi 7×129 boyutlarında bir matristir. Girişe uygulanan matrisi oluşturan her vektör gürültülü konuşmanın LGS katsayılardır ve 7 adet vektör uygulanır, her vektörde 129 adet LGS katsayısı vardır. Ortadaki vektöre N. Vektör denirse solundaki (önceki) vektörler N-1, N-2, ve N-3 sağındaki (sonraki) vektörler ise N+1, N+2 ve N+3 şeklinde indislenir.

Evrişim katmanındaki filtreler marifetiyle girişte verilen zaman frekans bilgisi ilişkilendirilerek alt uzayda yeni vektörler oluşturur. Bu yüzden, evrişim katmanındaki filtreler 7×10 boyutlarında 20 adet seçilir. Giriş matrisi üzerinde birer örnek kaydırılarak uygulanır. Böylece zaman bilgisi belirli frekans aralıkları boyunca birleştirilmiş olur. Hedef vektördeki gürültü bileşenlerinin yanı sıra çevresindeki gürültü bileşenlerininde hesaba katılması gevezelik gürültüsünün frekansta ve zamandaki karakterisliğini çözümlemekte faydalı olur. Aktivasyon fonksiyonu olarak ReLU kullanılmıştır. Ardından ise tameman bağlı katman vardır. Bu basamaklı (cascade) yapı içerisinde ReLU önemli bir rol oynamaktadır ve katman üzerinde doğrultulmuş ilişki (rectified correlations on a sphere) modeline dayanmaktadır. Böylece evrişim katmanının çıktısı olan özniteliklerin hedef ile olan ilişkisini göz önüne alarak, benzer olan kısımları seçer [26].

Ardından tamamen bağlı katman gelmektedir. Evrişim katmanından çıkıp ReLU'dan geçen veriler yüksek seviyedeki özellikleri temsil eder. Bunların çıkışa iletilmesi ve aralarında doğrusal olmayan kombinasyonlarının öğrenilmesini sağlar. Böylece düşük boyutlu bir öznitelik alanı sağlar ve doğrusal olmayan bir işlev öğrenir. Şekil 4.3'de kullanılan tamamen bağlı katman gösterilmiştir. Evrişim katmanından gelen 2400 giriş ve 129 çıkışı vardır. Ağırlık matrisi 129×2400 ve eşik vektörü ise 129×1 boyutundadır.



Şekil 4.3 Tamamen bağlı katman

Ağın eğitilmesi MATLAB 2017a ortamında gerçekleştirilmiştir. Stokastik dereceli alçalma algoritması kullanılmıştır, başlangıç eğitim hızı 3×10^{-4}, momentum katsayısı 0.9, küçük yığın (mini batch) 128 ve en fazla yaklaşım sayısı ise 40 olarak seçilmiştir. SDA algoritması Eşitlik 4.2'de gösterilmiştir.

$$\theta_{l+1} = \theta_l - a\nabla E(\theta_l) + \gamma(\theta_l - \theta_{l-1}) \tag{4.2}$$

Buradaki θ değişkenleri, a öğrenme katsayısını, γ momentum katsayısını ve E kayıp fonksiyonunu temsil etmektedir. Kayıp fonksiyonu Eşitlik 4.3'de verilen L2 normudur [27].

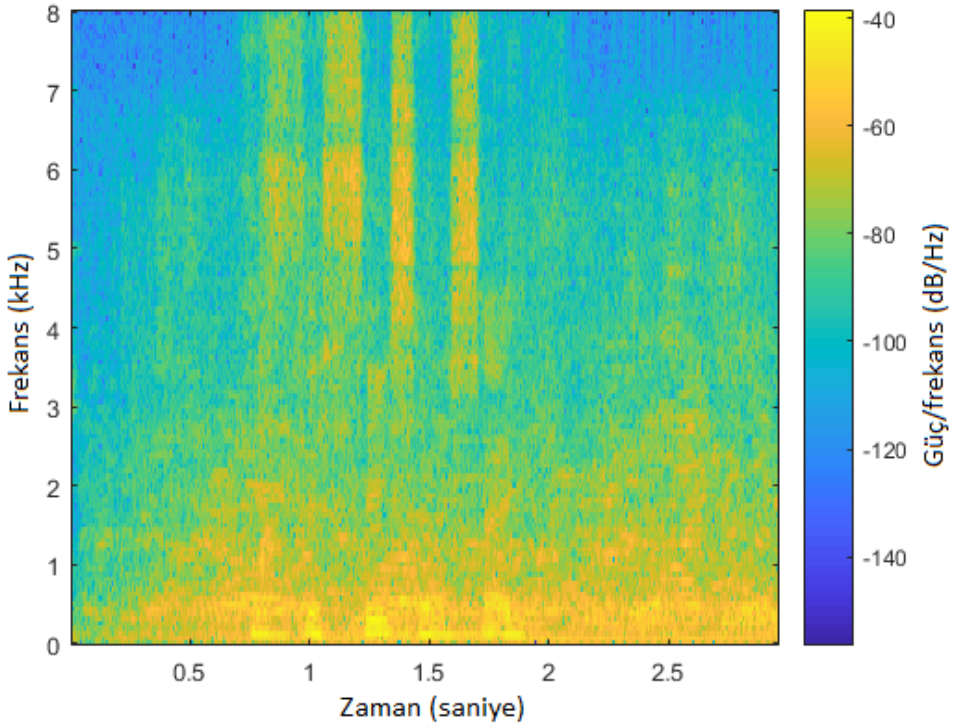
$$E = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{4.3}$$

Bu değişkenler çok sayıda denemenin sonuçunda en iyi sonuçları verenler arasından deneme yanılma yöntemiyle seçilmiştir. Veri setinden rasgele seçilmiş 25 konuşma sinyali eğitim için kullanılmıştır. Yani, 129×10546 boyutlarında bir matristir. Geri kalanlar ise test için ayrılmıştır.

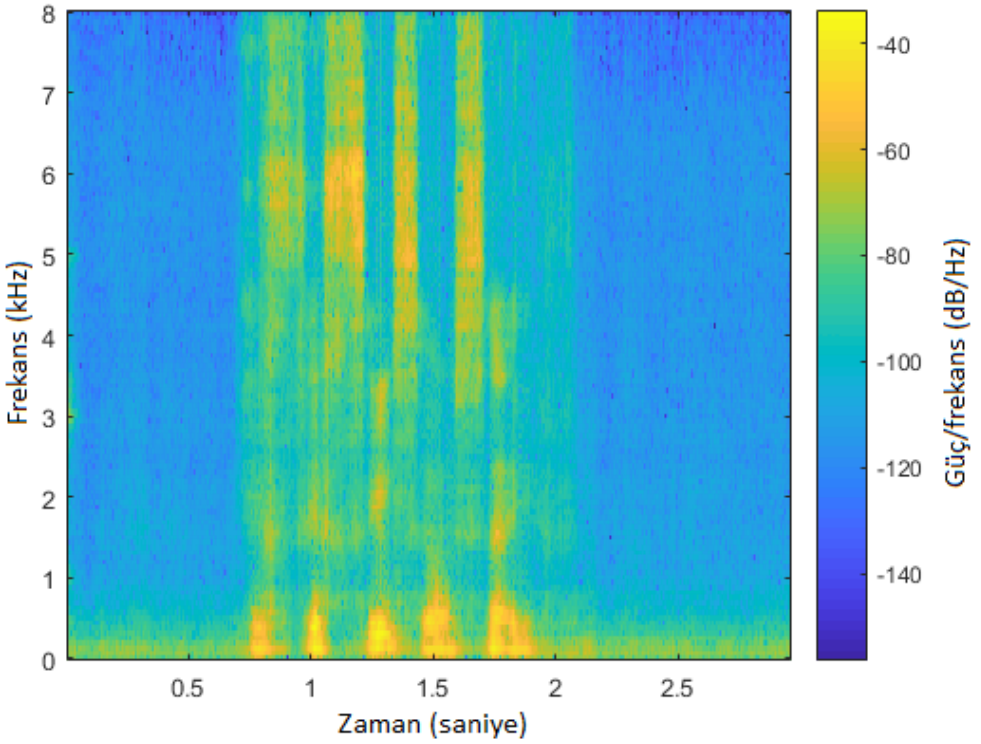
Tamamen bağlı katmanın çıkışları bir regresyon katmanına bağlıdır. Genelde ESA mimarileri, çıkış olarak tamamen bağlı katmanı kullanınır. Fakat eklenen 129 giriş ve 129 çıkışı olan regresyon katmanı ile LGS katsayılarının haritalanmasındaki başarımın artılması hedeflenmiş ve ortalama karesel hata ile kullanılmıştır. Ardından iyileştirilmiş

olan LGS katsayılarından sinyalin genlik spektrumuna geçilip ve saklanan gürültülü faz ile konuşma tekrar elde edilmiştir.

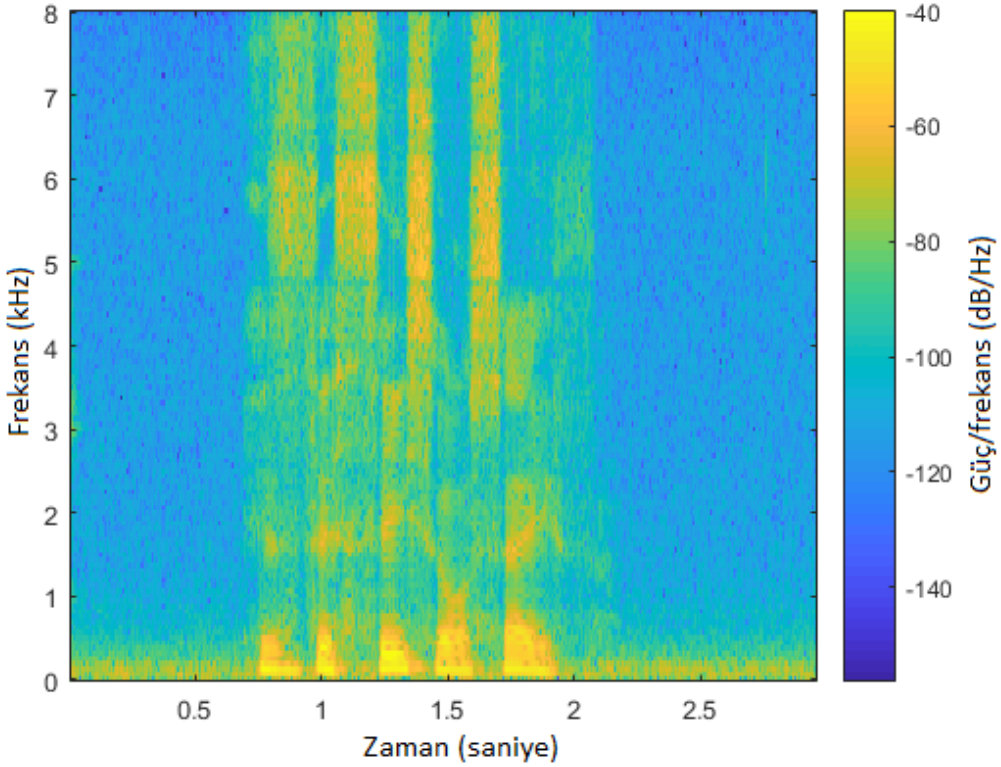
Şekil 4.4'de ağın girişine uygulanan gürültülü konuşma sinyalinin spektrogramı verilmiştir. Gevezelik gürültüsü ile bozulmuş 0 dB SGO sahip bir örnek seçilmiştir. Şekil 4.5'de ağın çıkışından elde edilmiş iyileştirilmiş konuşma sinyalinin spektrogramı ve Şekil 4.6'da temiz konuşma sinyalinin spektrogramı verilmiştir. Böylece iyileştirilmiş sinyalin temiz haline ne kadar benzediği konusunda bir karşılaştırma yapılabilir. Temiz spektrogram ile gürültülü spektrograma bakıldığında gevezelik gürültüsünün hangi frekans bileşenlerinde ne kadar var olduğunu kestirilebilir. Alçak frekans bileşenlerinde gücünün yüksek olduğu fakat birkaç dalga şeklinde yüksek frekans bileşenlerininde de olduğu aşikardır. İyileştirilmiş spektrogram ile temiz spektrogram birbirine gayet yakın görünmektedir. Fakat temiz sinyale ait özellikle yüksek frekans bileşenleri cevresindeki keskin karakteristlik geçişler, iyileştirilmiş spektrogramda daha yumuşak geçişler haline gelmiştir.



Şekil 4.4 Gürültülü konuşma sinyalinin spektrogramı



Şekil 4.5 İyileştirilmiş konuşma sinyalinin spektrogramı



Şekil 4.6 Temiz konuşma sinyalinin spektrogramı

İyileştirilmiş sinyalin KKAD ve KSNN değerlerine bakarak daha objektif bir değerlendirme yapılmıştır. KKAD için Hu vd. [28] tarafından 2008 yılında, KSNN için ise Taal vd. [16] tarafından 2010 yılında yapılan çalışmalardan paylaşılan toolboxlar kullanılmıştır. Çizelge 4.1'de farklı SGO'lardaki veri setleri için iyileştirilen sinyallerin KKAD ve KSNN sonuçları verilmiştir.

Çizelge 4.1 Farklı SGO değerlerine karşı KKAD ve KSNN ölçütleri

SGO Değerleri	KKAD	KSNN
10 dB	2.76	0.711
5 dB	2.50	0.674
0 dB	2.24	0.601
-5 dB	1.96	0.517
-10 dB	1.70	0.408
Ortalama	2.23	0.582

KKAD'nin değer aralığı -0.5 ile 4.5 ve KSNN'nin değer aralığı ise -1 ile 1 arasındadır. Bu değerlerin artması kalitenin arttığı anlamına gelmektedir. Çalışmanın, literatürdeki derin öğrenme mimarilerini kullanan güncel çalışmalar ile karşılaştırılması sonuçları daha anlaşılır hale getirecektir.

Çizelge 4.2 Farklı çalışmaların karşılaştırılması

Mimariler	KKAD
Regresyon Temelli	2.24
Ham Dalga Formlu [8]	1.99
Kodlayıcı-Kod Çözücü [4]	2.34
İlerici Öğrenme [6]	2.01

Çizelge 4.2 Farklı çalışmaların karşılaştırılması (Devamı)

LSTM-RNN [10]	2.13
DSA [3]	2.24
İkincil Öznitelikli DSA [9]	2.26

Xu vd. [3] tarafından 2015 yılında regresyon tabanlı DSA modeli, Park vd. [4] tarafından 2016 yılında yapılan çalışmada bir kodlayıcı-kod çözücü derin öğrenme mimarisi, Gao vd. [6] tarafından 2016 yılında DSA dayalı konuşma iyileştirme için ilerici öğrenme (progressive learning) modeli, Sun vd. [10] tarafından 2017 yılında uzun kısa süreli hafızalı yineleyen sinir ağı (long short term memory recurrent neural network-LSTM-RNN) modeli, Xu vd. [9] tarafından 2017 yılında doğrudan kullanılan LGS katsayıları yanı sıra mel frekansı kepstral katsayıları ve ideal ikili maske (ideal binary mask) gibi ikincil öznitelikleri kullanan DSA mimarisi ve Fu vd. [8] tarafından 2017 yılında yapılan ESA mimarisi kullanılmıştır. Bahsi geçen çalışmaların tamamında KKAD ölçütü kullanılmasına rağmen bir kısmında KSNN kullanılmamış onun yerine farklı ölçütler kullanılmıştır. Bu yüzden ortak bir taban oluşturmak için 0 dB'de gevezelik gürültüsüyle (geneli gevezelik gürültüsünü kullanmış olmasına rağmen az sayıda çalışma farklı gürültü türleri tercih etmiş) elde edilmiş KKAD sonuçları karşılaştırılmıştır. Bu çalışmalarda kullanılan veri setlerinin ve bazı çalışmaların gürültü türünün de farklı olmasından dolayı çok sağlıklı bir sonuca varılması güçtür. Fakat fikir vermesi bağlamında yararlı olmuştur. Farklı SGO değerlerinin ortalama sonuçlarının karşılaştırılması daha sağlıklı olurdu fakat çalışmalar farklı değerlerde SGO değerleri kullanılmıştır, bu nedenle tüm çalışmalarda ortak olan 0 dB SGO seviyesinde karşılaştırma uygun bulunmuştur. Çizelge 4.2'de bahsi geçen çalışmalar, bu tezde önerilen regresyon temelli mimari ile karşılaştırılmıştır. İlgili tablo incelendiğinde, önerilen mimari litaretürdeki çalışmalar ile rekabet edebilecek bir iyileştirme sağladığı görülmüştür. Bir sonraki bölümde bu sonuçlar detaylı bir şekilde tartışılacaktır.

BÖLÜM 5

SONUÇ VE ÖNERİLER

Çalışmanın amacı ve hipotez kısmında belirtilen yaklaşımın ne kadar uygun olup olmadığı üzerinde durulmalıdır. Kısaca, tüm problemi büyük bir optimizasyon problemi gibi düşünüp eldeki ön bilgileri kullanarak en az değişkenle tatmin edici bir sonuç bulmak amaçlanmıştı. Bu amaç doğrultusunda önerilen ağın LGS katsayılarıyla beslenmesi, bir ön bilgi kullanılması ve optimizasyon problemine bir kısıt getirmesi olarak düşünülebilir. Bu sayede SDA algoritması ağın parametrelerini araştırırken hem daha az değişkenle çalışmak durumunda kalmıştır hem de araştırma alanı daraltılmıştır. Doğruca ham veri ile beslenme fikrine yatkın olan ESA'lardan beklenen arka arkaya sıralanmış evrişim katmanları sayesinde uygun özniteliklerin oluşturulması sıklıkla başvurulan bir yoldur. Fakat litaretürde de yaklaşımımıza benzer şekilde bazı önerilen ağlar LGS katsayıları ile beslenmiştir. Böylece evrişim katmanının derinliği azaltılabilir. LGS katsayılarının tercih edilmesinin nedeni frekans ile zaman arasında ilişki kurarak sinyalin faydalı özelliklerini ön plana çıkarmasıdır. Böylece optimizasyon problemi için iyi bir başlangıç noktası oluşturur. Bu yaklaşımdan yola çıkarak, ağın sinyalin dalgaçık dönüşümü katsayılarıyla beslenmesi durumunda da iyi bir sonuç verme ihtimali ortaya çıkar. Bu önerinin altında dalgaçık dönüşümünün KSFD'den farklı olarak her zaman aralığı için farklı frekans çözünürlüklerini sunuyor olması yatmaktadır. Böylece ağın girişine verilen katsayılarda aynı zaman aralındaki sinyalin farklı frekans çözünürlüğü bilgileri mevcut olacaktır. Böylece farklı frekans çözünürlüklerinde mevcut olan bilgierin faydalı öznitelikler çıkarma ihtimali doğacaktır. Ayrıca eklenen regresyon katmanı maliyeti düşük olmasına rağmen katsayıların haritalanması işlemine önemli katkı sağlamıştır. Regresyon katmanı yerine bir tamamen bağlı katman daha

bağlanabilirdi fakat içerdiği yüksek sayıda değişkenden dolayı tercih edilmedi. Böylece ağın karmaşıklığı azaltılmış oldu.

Çizelge 4.1 ve 4.2'deki sonuçlara bakıldığında tatmin edici bir sonuç elde edilebildiği söylebilir. Karşılaştırılan mimarilerin ağ derinliği göz önüne alındığında (örneğin, Park'ın çalışması [4] 15 katman, Fu'nun çalışması [8] 4 evrişim katmanı ve 2 tamamen bağlı katman içerir), bu çalışmadaki ağ nazaran sığ olmasına rağmen başarısı rekabet edebilecek düzeydedir. Akademik bağlamda katman sayısını artırarak başarımı artırmak makul olsada ilgili ağların harici bir uygulama üzerine gömülmesi işlemi güçleşir. Önerdiğimiz mimaride başarıdan en az tavizi vermek suretiyle tatmin edici derecede başarılı ve sığ bir ağ mimarisi sunulmuştur.

Ayrıca değinilmesi gereken ve başarıyı etkileyen bir başka konu ise sinyalin tekrar oluşturulması kısmında kullanılan gürültülü fazdır. Genlik spektrumu iyileştirilmişken faz spektrumunun gürültülü halinin sinyali oluşturmak için kullanılması KKAD ve KSNN değerlerinde düşüşe neden olmuştur. Ayrıca bir fazın da iyileştirilmesi için bir kestirimci kullanılması uygun olur.

KAYNAKLAR

[1] 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.

[2] Gerkmann, T., Krawczyk, M., Rehr, R., (2012). "Phase estimation in speech enhancement—unimportant, important, or impossible?", Electrical & Electronics Engineers in Israel, 14-17 November, Eilat.

[3] Park, S. R., Lee, J., (2016). "A fully convolutional neural network for speech enhancement", arXiv preprint arXiv:1609.07132.

[4] Xu, Y., Du, J., Dai, L. R., Lee, C. H., (2015). "A regression approach to speech enhancement based on deep neural networks", IEEE/ACM Transactions on Audio, Speech and Language Processing, 23(1): 7-19.

[5] Fu, S. W., Tsao, Y., Lu, X., (2016). "SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement", Interspeech, 3768-3772.

[6] Gao, T., Du, J., Dai, L. R., Lee, C. H., (2016). "SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement" Interspeech, 3713-3717.

[7] Wang, H. M., (2017). "Audio-Visual Speech Enhancement based on Multimodal Deep Convolutional Neural Network", arXiv preprint arXiv:1703.10893.

[8] Fu, S. W., Tsao, Y., Lu, X., Kawai, H., (2017). "Raw waveform-based speech enhancement by fully convolutional networks", arXiv preprint arXiv:1703.02205.

[9] Xu, Y., Du, J., Huang, Z., Dai, L. R., Lee, C. H., (2017). "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement", arXiv preprint arXiv:1703.07172.

[10] Sun, L., Du, J., Dai, L. R., Lee, C. H., (2017). "Multiple-target deep learning for LSTM-RNN based speech enhancement", Hands-free Speech Communications and Microphone Arrays, 1-3 March 2017, San Francisco.

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

[12] Loizou, P. C., (2007). Speech enhancement: theory and practice, Second Edition, CRC Press, Boca Raton.

[13] Deller, J. R., Hansen, J. H., Proakis, J. G., (2000). Discrete-time processing of speech signals, First Edition, Wiley-IEEE Press, Lansing.

[14] Hayes, M. H., (1996). Statistical digital signal processing and modeling, First Edition, John Wiley & Sons Inc., New York.

[15] Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., (2016). Deep learning (Vol. 1), First Edition, MIT Press, Boston.

[16] Taal, C. H., Hendriks, R. C., Heusdens, R., Jensen, J., (2010). "A short-time objective intelligibility measure for time-frequency weighted noisy speech", IEEE International Conference on Acoustics, Speech and Signal Processing, 14-19 March 2010, Dallas.

[17] Hall, J. E., (2011). Textbook of medical physiology, 12th Edition, Saunders, Mississippi.

[18] Freeman, W. J., (1975). Mass action in the nervous system, First Edition, Academic Press, San Diego.

[19] LISA Lab, University of Montreal., (2015). Deep learning tutorials, Montreal.

[20] Alpaydın, E., (2011). Yapay öğrenme, Birinci Baskı, Boğaziçi Üniversitesi Yayınevi, İstanbul.

[21] Dechter, R. (1986). Learning while searching in constraint-satisfaction problems, First Edition, University of California Press, Los Angeles.

[22] Schmidhuber, J., (2015). "Deep learning in neural networks: An overview", Neural networks, 61: 85-117.

[23] YAZAN, E., Talu, M. F., (2017). "Comparison of the stochastic gradient descent based optimization techniques", International Artificial Intelligence and Data Processing Symposium, 16-17 September 2017, Malatya.

[24] LeCun, Y., Bengio, Y., Hinton, G., (2015). "Deep learning", Nature International Journal of Science, *521*:436-444.

[25] The Centre for Speech Technology Research, MOCHA-TIMIT, www.cstr.ed.ac.uk/research/projects/artic/mocha.html, 20 Ekim 2017.

[26] Kuo, C. C. J., (2016). "Understanding convolutional neural networks with a mathematical model", Journal of Visual Communication and Image Representation, 41:406-413.

[27] Robert, C. (2014)., "Machine learning, a probabilistic perspective", Change, 27(2):62-63.

[28] 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.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı :MUSTAFA ERSEVEN

Doğum Tarihi ve Yeri :20.08.1992, Çorum-Alaca

Yabancı Dili :İngilizce

E-posta :mustafaerseven@gmail.com

ÖĞRENİM DURUMU

Derece	**Alan**	**Okul/Üniversite**	**Mezuniyet Yılı**
Lisans	Elektrik ve Elektronik Mühendisliği	Ankara Üniversitesi	2016
Lise	Fen Bilimleri	İzmir Anadolu Öğretmen Lisesi	2010

YAYINLARI

Bildiri

1. Erseven, M., Bolat, B., (2018). “Regression-based speech enhancement by convolutional neural network”, IEEE 26th Signal Processing and Communications Applications Conference, 2-5 May 2018, İzmir.

2. Erseven, M., Bilgin, G., (2017). “Statistical-Spatial Approach for Cell Classification in Histopathological Imagery”, IEEE 21st National Biomedical Engineering Meeting, 24-26 November, İstanbul.