T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

HİBRİT KONUŞMA AKTİVİTE TESPİTİ KULLANILARAK D-VEKTÖR TABANLI BİR KONUŞMACI DİYARİZASYON SİSTEMİNİN TASARLANMASI

Yunus KORKMAZ

Doktora Tezi

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

Yazılım Bilim Dalı

ŞUBAT 2023

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Bilgisayar Mühendisliği Anabilim Dalı

Doktora Tezi

HİBRİT KONUŞMA AKTİVİTE TESPİTİ KULLANILARAK D-VEKTÖR TABANLI BİR KONUŞMACI DİYARİZASYON SİSTEMİNİN TASARLANMASI

Tez Yazarı
Yunus KORKMAZ

Danışman
Dr.Öğr.Üyesi Aytuğ BOYACI

ŞUBAT 2023
ELAZIĞ

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Bilgisayar Mühendisliği Anabilim Dalı

Doktora Tezi

Başlığı:	Hibrit Konuşma Aktivite Tespiti Kullanılarak D-Vektör Tabanlı Bir Konuşmacı Diyarizasyon Sisteminin Tasarlanması
Yazarı:	Yunus KORKMAZ
İlk Teslim Tarihi:	02.01.2023
Savunma Tarihi:	03.02.2023

TEZ ONAYI

Fırat Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına göre hazırlanan bu tez aşağıda imzaları bulunan jüri üyeleri tarafından değerlendirilmiş ve akademik dinleyicilere açık yapılan savunma sonucunda OYBİRLİĞİ ile kabul edilmiştir.

		İmza
Danışman:	Dr.Öğr.Üyesi Aytuğ BOYACI Milli Savunma Üniversitesi, Hava Harp Okulu	Onayladım
Başkan:	Prof.Dr. Galip AYDIN Fırat Üniversitesi, Mühendislik Fakültesi	Onayladım
Üye:	Dr.Öğr.Üyesi Mustafa KAYA Fırat Üniversitesi, Teknoloji Fakültesi	Onayladım
Üye:	Doç.Dr.Can EYÜPOĞLU Milli Savunma Üniversitesi, Hava Harp Okulu	Onayladım
Üye:	Dr.Öğr.Üyesi Oğuz ATA Altınbaş Üniversitesi, Mühendislik Fakültesi	Onayladım

Bu tez, Enstitü Yönetim Kurulunun ...../...../20..... tarihli toplantısında tescillenmiştir.

İmza

Prof. Dr. Burhan ERGEN
Enstitü Müdürü

BEYAN

Fırat Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım "Hibrit Konuşma Aktivite Tespiti Kullanılarak D-Vektör Tabanlı Bir Konuşmacı Diyarizasyon Sisteminin Tasarlanması" Başlıklı Doktora Tezimin içindeki bütün bilgilerin doğru olduğunu, bilgilerin üretilmesi ve sunulmasında bilimsel etik kurallarına uygun davrandığımı, kullandığım bütün kaynakları atıf yaparak belirttiğimi, maddi ve manevi desteği olan tüm kurum/kuruluş ve kişileri belirttiğimi, burada sunduğum veri ve bilgileri unvan almak amacıyla daha önce hiçbir şekilde kullanmadığımı beyan ederim.

03.02.2023

Yunus KORKMAZ

ÖNSÖZ

Konuşmacı Diyarizasyon sistemleri dijital bir ses kaydında "kim ne zaman konuşmuş?" sorusuna cevap verebilen yazılımsal sistemler olarak tanımlanmaktadır. Bu sistemler, dijital bir ses kaydındaki konuşmacı sayısını ve kişilerin konuştukları bölgeleri tespit edebildiğinden otomatik transkripsiyon işlemlerine de katkı sağlamaktadır. Konuşmacı Diyarizasyon sistemleri genel olarak konuşma aktivite tespiti, konuşma bölütleme, kümeleme gibi alt sistemlerden oluşmaktadır. Bahsi geçen alt sistemlerden konuşma aktivite dedektörlerinin yüksek doğrulukta gerçekleştirilmesi diyarizasyon sisteminin performansını doğrudan olumlu yönde etkilemektedir. Bu tez çalışmasında literatürde daha önce önerilmemiş yeni bir hibrit konuşma aktivite dedektörü üzerine kurulu d-vektör tabanlı bir konuşmacı diyarizasyon sistemi geliştirilmiştir.

Bu tez çalışması süresince maddi manevi katkılarını esirgemeyen ve tüm bilimsel çalışmalarda zaman gözetmeksizin her an destekçim olan değerli büyüğüm abim danışmanım Sayın Dr. Aytuğ BOYACI'ya teşekkürlerimi sunarım.

Çalışmamı, bugünlere ulaşmamı sağlayan, haklarını asla ödeyemeyeceğim ve bugüne kadar attığım her adımda destekçim olan kıymetli aileme ithaf ederim.

Yunus KORKMAZ
ELAZIĞ, 2023

İÇİNDEKİLER

Sayfa

ÖZET

Hibrit Konuşma Aktivite Tespiti Kullanılarak D-Vektör Tabanlı Bir Konuşmacı Diyarizasyon Sisteminin Tasarlanması

Yunus KORKMAZ

Doktora Tezi

FIRAT ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Şubat 2023, Sayfa: xi + 66

Teknolojik gelişmelerin hızla yaşandığı günümüzde insan yerine makinelerden ve yazılımlardan faydalanan sistemler gittikçe çoğalmaktadır. Bu sistemler, birçok alanda olduğu gibi Dijital Konuşma İşleme (DSP) alanında da geliştirilmeye ihtiyaç duymaktadır. DSP alanlarından biri olan Konuşmacı Diyarizasyonu, konuşma içeren bir ses kaydından, kimin ne zaman konuştuğu bilgisinin otomatik olarak çıkarılmasını gerektirmektedir. Yüksek performans ile çalışan bir konuşmacı diyarizasyon sistemi geliştirme günümüzde hala bu alanda çalışan arastırmacılar için güncel sorunların başında gelmektedir. Daha düşük hata oranına sahip konuşmacı diyarizasyon sistemlerini geliştirebilmek için, bu sistemleri oluşturan ses ön işleme, konuşma aktivite tespiti/detektörü, konuşmacı bölütleme ve konuşmacı kümeleme gibi alt sistemlerin iyileştirilmesi gerekmektedir.

Bu tez çalışmasında, konuşmacı diyarizasyon sistemlerinin aşamalarından biri olan Konuşma Aktivite Tespit sistemi için daha önce önerilmemiş hibrit bir model geliştirilerek düşük hata oranına sahip bir konuşmacı diyarizasyon sisteminin tasarlanması hedeflenmiştir. Denetimli ve denetimsiz öğrenmenin mantıksal operatörlerle birleştirildiği hibrit konuşma aktivite tespit sisteminde, denetimsiz öğrenme için özellik eşikleme, denetimli öğrenme için bir derin öğrenme mimarisi olan uzun-kısa süreli bellek (LSTM) kullanılmıştır. Konuşmacı diyarizasyon sisteminin devamında, önceden eğitilmiş bir yapay sinir ağından d-vektör'ler çıkarılmış, bu vektörler üzerinde Spektral Kümeleme uygulandıktan sonra ses kaydında kimin ne zaman konuştuğu tespit edilmiştir.

Geliştirilen konuşmacı diyarizasyon sisteminin değerlendirilmesinde, konuşmacı diyarizasyon sistemleri hata metriklerinden konuşma aktivite dedektörlerinden kaynaklanan Miss ve False Alarm (FA) hata metrikleri incelenmiş, literatüre göre nispeten düşük hata oranları elde edildiği görülmüştür.

Anahtar Kelimeler: Dijital konuşma işleme, Konuşma aktivite tespiti, Konuşmacı diyarizasyonu

ABSTRACT

Implementation Of A D-Vector Based Speaker Diarization System Using Hybrid Voice Activity Detection

Yunus KORKMAZ

Ph.D. Thesis

FIRAT UNIVERSITY
Graduate School of Natural and Applied Sciences

Department of Computer Engineering

February 2023, Pages: xi + 66

In today's world with rapid technological developments, systems that use machines and software instead of humans are increasing day after day. These systems need to be developed in the area of Digital Speech Processing (DSP), as in many other fields. Speaker Diarization, one of the DSP applications, requires automatic extraction of "who spoke when" from an audio recording containing speech. Developing a speaker diarization system working with high performance is still one of the challenging issues for researchers studying in this area. In order to develop speaker diarization systems with lower error rates, sub-systems such as Speech Pre-processing, Voice Activity Detection, Speaker Segmentation and Speaker Clustering, which constitute a speaker diarization system, need to be improved.

In this thesis, it is aimed to design a speaker diarization system with low error rate by developing a hybrid model that has not been proposed before for the voice activity detection system which is one of the stages of speaker dialization systems. In hybrid voice activity detection system where supervised and unsupervised learning is combined with logical operators, feature thresholding was used for unsupervised learning while long-short term memory (LSTM), a deep learning architecture, was utilized for supervised learning. In the continuation of the speaker dialization system, d-vectors were extracted from a pre-trained artificial neural network, and after Spectral Clustering was applied on these vectors, "who spoke when" was detected in the audio recording.

At the evalution phase of the proposed speaker diarization system, Miss and False Alarm (FA) metrics, which can be occurred due to used Voice Activity Detectors in Speaker Diarization Systems, were interpreted in detail. It was observed that using an Hybrid VAD in diarization systems has achieved low Miss and False Alarm (FA) error rate.

Keywords: Digital speech processing, Voice activity detection, Speaker diarization

ŞEKİLLER LİSTESİ

Sayfa

TABLOLAR LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR

Kısaltmalar

VAD	: Voice Activity Detection
DER	: Diarization Error Rate
DTW	: Dynamic Time Warping
SVM	: Support Vector Machines
KNN	: K Nearest Neighbors
LSTM	: Long-short Term Memory
DNN	: Deep Neural Network
ZCR	: Zero Crossing Rate
MFCC	: Mel-Frequency Ceptral Coefficients
SÇA	: Spektral Çıkarma Algoritması
VQ	: Vector Quantization

1. GİRİŞ

Konuşmacı Diyarizasyon sistemleri dijital bir ses kaydında "kim ne zaman konuşmuş?" sorusuna cevap verebilen yazılımsal sistemler olarak bilinmektedir. Bu sistemler, dijital bir ses kaydındaki konuşmacı sayısını ve kişilerin konuştukları bölgeleri tespit edebildiğinden otomatik transkripsiyon işlemlerine de katkı sağlamaktadır. Konuşmacı Diyarizasyon sistemlerinin yüksek doğrulukta çalışabilmesi, yani bir ses kaydında kimin ne zaman konuştuğunun otomatik tespit edilmesi, hala dijital konuşma işleme alanında çalışan araştırmacılar arasında zorlu bir problem olarak görülmektedir. Bahsi geçen zorluk, bu tür sistemler tasarlanırken özellikle konuşmacılar hakkında bir ön bilgi sahibi olunamamasından (no prior information) kaynaklanmaktadır. Bu tez çalışmasının nihai amacı bir konuşmacı diyarizasyon sistemini oluşturan alt sistemlerden Konuşma Aktivite Tespiti'nin daha yüksek doğrulukta gerçekleştirilerek konuşmacı diyarizasyon sistemlerinin dolaylı olarak iyileştirmesine katkı sunmaktır. Yüksek performansa sahip bir konuşmacı diyarizasyon sisteminin tasarlanması, bu sistemleri oluşturan alt sistemlerden; Konuşma Aktivite Tespiti (Voice Activity Detection-VAD), Konuşmacı Bölütleme, Konuşmacı Bazlı Özellik Çıkarma, Konuşmacı Kümeleme sistemlerinin iyileştirilmesi ile doğrudan ilişkilidir. Bir konuşmacı diyarizasyon sistemini oluşturan alt sistemlerden;

- *Konuşma Aktivite Tespiti*, dijital bir ses kaydında konuşma veya konuşma olmayan (speech/non-speech) bölgelerin temporal tespit edilmesi işlemi olarak bilinmektedir. Konuşmacı diyarizasyon sistemlerinde genellikle ilgi alanı (Region Of Interest-ROI) sadece konuşma bölgesi olduğundan, bu tür sistemler tasarlanır iken konuşma aktivite tespit sistemi büyük önem arz etmektedir. Konuşma olmayan (sessizlik, çevresel sesler, gürültü vs.) ses kısımlarının konuşma olarak yanlış etiketlenmesi diyarizasyon sonucuna doğrudan olumsuz yönde etki edecek, diyarizasyon hata oranını (Diarization Error Rate-DER) yükseltecektir. Konuşma aktivite tespit sistemleri denetimli veya denetimsiz olmak üzere iki farklı şekilde gerçekleştirilebilmektedir. Denetimli sistemlerde önceden eğitilmiş bir yapay sinir ağı modeli kullanılarak ses parçası (çerçeve) "konuşma veya konuşma değil" şeklinde etiketlenmektedir. Denetimsiz sistemlerde ise ses parçasından manuel (elle) çıkartılan sıfır geçiş oranı (ZCR), enerji, mel frekansı kepstrum katsayıları (MFCC) gibi özellikler belirli bir eşik değeri (thresholding) ile karşılaştırılarak ses parçasının konuşma olup olmadığı anlaşılmaya çalışılmaktadır.
- *Konuşmacı Bölütleme*, bir konuşmacı diyarizasyon sisteminde, konuşma aktivite tespitinden geçtikten ve boşluklar atıldıktan sonra elde edilen dijital ses sinyalinin makul sayılabilecek aralıklara bölünmesi işlemi olarak bilinmektedir. Bu aralığın genelde milisaniye cinsinden seçilmesi gerekmektedir. Konuşmacı segmentleme olarak ta bilinen bu aşamada amaç, her segmentte sadece bir konuşmacının bulunacağı şekilde bölütleme

yapmaktır. Bunun da sebebi bir sonraki aşama olan konuşmacı bazlı özellik çıkarma aşamasında çıkarılan özelliklerin tekil bir kullanıcıya ait olması. Aksi takdirde bir segmentte birden fazla konuşmacı bulunduğunda, o segmentten çıkarılacak özellikler tek bir konuşmacıya ait olmayacaktır. Konuşmacı bölütleme aynı zamanda konuşmacı diyarizasyon sisteminin çıkış çözünürlüğünü de belirlediğinden hata oranı çıkarmada önem arz etmektedir.

- *Konuşmacı Bazlı Özellik Çıkarma*, konuşmacı bölütleme işlemi sonucu oluşan her bölüt için konuşmacıları ayırt edebilecek nitelikteki özelliklerin (speaker discriminative features) çıkarılması aşaması olarak tanımlanmaktadır. Özellik çıkarma işlemi, her bölüt kendi içerisinde daha küçük ses parçacıklarına (çerçeve) ayrılarak, her ses parçasının özellikleri daha sonra ilgili bölüt için birleştirilip o bölüte ait tek bir özellik vektörü elde edileceği şeklinde tasarlanmaktadır. Konuşmacı diyarizasyon sisteminin bu aşamasında, daha önceden konuşmacı doğrulama veya konuşmacı tanıma için tasarlanmış hazır yapay sinir ağları (pre-trained network) kullanılmaktadır. Bu teknoloji, aktarım öğrenmesi (transfer learning) olarak ta bilinmektedir. Konuşmacı bölütleme aşamasında elde edilen bölütler tek tek önceden eğitilmiş ağa gönderilmekte ve bu ağdan belirli uzunlukta d-vektör (deep vektör) adındaki özellik vektörleri çekilmektedir. Konuşmacı bazlı özellik çıkarma aşaması sonrası her bölütü tek bir d-vektör temsil etmektedir. Bu aşama literatürde ayrıca konuşmacı gömülü özelliklerinin çıkarılması (speaker embedding extraction) olarak ta bilinmektedir.
- *Konuşmacı Kümeleme,* konuşmacı bazlı özellik çıkarma aşaması sonucu elde edilen d-vektörlerin (özellik vektörü) bir kümeleme algoritması kullanılarak gruplanmaya çalışıldığı aşamadır. Konuşmacı diyarizasyon işlemine girecek olan dijital ses kaydında, kaç kişinin konuştuğu (number of speakers) ve konuşan kişiler ile ilgili önceden bilgi (prior information) olmadığından, diyarizasyon sistemlerinde bir denetimsiz/gözetimsiz öğrenme türü olan kümeleme kullanılmaktadır. K-Means, Toplayıcı, Gaussian yöntemlerinin yanı sıra konuşmacı diyarizasyonunda en çok kullanılan kümeleme yöntemi Spektral Kümeleme olarak bilinmektedir. Spektral kümeleme işlemi öncesi elde edilmiş olan d-vektörler üzerinde konuşmacı sayısı belirleme önem taşımaktadır. Konuşmacı sayısı belirlemede Elbow ve Silhouette gibi yöntemler kullanılmaktadır. Spektral kümeleme işlemine tahmini konuşmacı sayısı belirlendikten sonra başlanması, kümeleme işleminin daha sağlıklı ve yüksek performans ile yapılabilmesi adına önem arz etmektedir.

Bu tez çalışmasında da yukarıda bahsedilen aşamalardan Konuşma Aktivite Tespiti sisteminin tamamen özgün bir şekilde tasarlanarak yeni ve başarılı bir konuşmacı diyarizasyon

sistemi tasarlanması hedeflenmiştir. Tasarlanan konuşma aktivite tespit sistemi, denetimli öğrenme ve denetimsiz öğrenme (eşikleme) yapılarının bir arada bulunduğu şekilde gerçekleştirilerek, bu sisteme "Hibrit VAD" ismi verilmiştir. Hibrit VAD sisteminin denetimli kısmında, Google AudioSet ses verisetinden manuel olarak üç adet "sessizlik", "gürültü" ve "konuşma" ses kaydı oluşturulmuş, bu ses kayıtlarından enerji, sıfır geçiş oranı (ZCR) ve 13.dereceden MFCC özellikleri çıkarıldıktan sonra bir uzun kısa süreli bellek ağı (Long Short Term Memory-LSTM) eğitilerek, konuşma ve konuşma olmayan bölgeleri ayırt edebilecek bir model geliştirilmiştir. Hibrit VAD sisteminin denetimsiz (metric-based) kısmında ise denetimli sistemdeki gibi enerji, sıfır geçiş oranı (ZCR) ve 13.dereceden MFCC özelliklerinden sadece birinci katsayıda olan (1^{st} *order MFCC*) değer, her bir özellik farklı bir eşik değerine tabi tutularak çerçevelerin konuşma olup olmadığı tespit edilmiştir. Bu hibrit yapıda, hem LSTM çıkışından hemde denetimsiz yapıdan elde edilen değerler mantıksal operatörler (AND/OR) ile birleştirilip nihai VAD kararı üretilmiştir. Sonuç olarak konuşma aktivite tespit sistemlerinin değerlendirme kriterleri olan düşük FEC (Front-End Clipping), MSC (Mid-Speech Clipping), OVER, NDS (Noise Detected as Speech) hata oranlarına sahip bir Hibrit Konuşma Aktivite Tespit sistemi tasarlanmıştır. Tasarlanan konuşmacı diyarizasyon sisteminde, konuşma aktivite detektöründen geçen ses kaydına daha sonra bölüt uzunluğu 400 ms (0,4 ms) olacak şekilde konuşmacı bölütleme işlemi uygulanmış. Bu değer geliştirilen diyarizasyon sisteminin aynı zamanda çıkış çözünürlüğü olarak belirlenmiştir. Bölütleme işlemi sonrası her bölüt konuşmacı tanıma/doğrulama için önceden eğitilmiş (transfer learning) bir LSTM ağına gönderilerek her bölüt'e ait d-vektör çıkarılmıştır. D-vektörler, önceden eğitilmiş ağ her bölüt için çalıştırıldıktan sonra, ağın son gizli katmanındaki değerlerin vektöre dönüştürülmesi sonucunda elde edilmektedir. Konuşmacı bazlı özellik (speaker embeddings) olarak ta bahsedeceğimiz d-vektörler üzerinde ilk önce konuşmacı sayısı belirleme ardından spektral kümeleme işlemleri uygulanarak nihai diyarizasyon sonucu elde edilmiştir. Böylece konuşmacı diyarizasyon sistemlerinin ilk aşaması olan konuşma aktivite tespiti için daha önce önerilmemiş hibrit bir yöntem geliştirilerek yeni bir konuşmacı diyarizasyon yaklaşımı ortaya çıkarılmıştır.

Konuşmacı diyarizasyon sistemleri günümüzde karşılıklı konuşma içeren bir ses kaydının yazıya dökülmesi (konuşmacı transkripsiyonu) ihtiyacını gerektiren her alanda kullanılabilmektedir. Başta adli makamlarda kullanılan dijital ses kayıt delillerinin yazıya dökülmesi olmak üzere çağrı merkezlerinde geçen konuşmaların analizi (hem çalışan performans ölçümü hem de müşteri memnuniyeti tespiti, duygu analizi), uzun ses kayıtlarının saniye saniye dinlenmesi yerine kimin ne zaman konuştuğunun yazılımsal olarak tespit edilmesi, medya sektöründe yapılan sesli röportajlar, ses izinden hastalık tespiti gibi bir çok alanda otomatik olarak gerçekleşen bir konuşmacı diyarizasyon sistemi toplumsal faydaya katkı sağlayacaktır.

Konuşmacı diyarizasyonu ile ilgili literatür incelendiğinde konunun özel sektörden akademik dünyaya kadar birçok araştırmacı tarafından çalışıldığı karşımıza çıkmaktadır. Bu araştırmaların genelde konuşmacı diyarizasyon sistemlerini oluşturan alt sistemleri geliştirerek konuşmacı diyarizasyon sistemlerini dolaylı bir şekilde iyileştirmeye odaklandığı görülmektedir. Diyarizasyon teknolojisinin tarihsel gelişimi 1990'lı yılların başında hava trafik kontrol konuşmalarında ve televizyon haberlerine ait ses kayıtlarında konuşmacıları ayırt ederek ve her konuşmacıya özgü model geliştirebilme çabalarıyla başlamıştır. 1992 yılında Man-Hung Siu, George Yu ve Herbert Gish çok konuşmacılı dalga formlarında konuşmacıları segmente edebilecek bir sıralı öğrenme algortiması önermiştir [1]. Günümüz konuşmacı diyarizasyon konusunu tamamen denetimsiz öğrenme yöntemi ile çözmeye çalışan araştırmada, ses sinyallerinden mel frekansı kepstrum katsayıları (MFCC) özellikleri çıkarıldıktan sonra, konuşmacı geçişlerini (speaker change point) tespit etmek için Gaussian Mixture Model (GMM) ve Expectation-Maximization (EM) algoritmaları kullanılmıştır. U.Jain ve arkadaşları 1996 yılında otomatik konuşma tanıma (Automatic Speech Recognition-ASR) uygulamalarında, kelimelerin sınırlarının bilinmesi ile daha başarılı sistemler tasaarlayacaklarını düşünerek, konuşmacı tanıma ve konuşmacı diyarizasyon konusuna odaklanmıştır [2]. Çalışmada, akustik ve dil modelleme şeklinde iki ayrı problem üzerinde duran ekip, farklı ortam akustiklerinde çalışabilen bir otomatik transkripsiyon sistemi tasarlamıştır. Bu süre zarfında konuşma bölütleri arasındaki mesafeyi hesaplayan ve kısa sürede konuşmacı diyarizasyonunda standart haline gelen Genelleştirilmiş Benzerlik Oranı (Generalized Likelihood Ratio-GLR) [3] ve Bayes Bilgi Kriteri (Bayesian Information Criterion-BIC) [4] algoritmaları geliştirilmiştir. Sırasıyla 1991 ve 1998 yıllarında geliştirilen bu sistemler bir ses kaydında geçen konuşmacılar arasında değişim (speaker change point) tespiti yapabilmektedir. Amerikan NIST (National Institute of Standards and Technology) kurumunun öncülüğü ile 2000'lerin başında düzenlenen yarışmalarda farklı algoritmalar önerilmiştir. J.Ajmera ve C.Wooters konuşmacılar ve konuşmacı sayısı hakkında hiçbir önbilgi sahibi olmadan (no prior knowledge) hem konuşmacı bölütleme hem de konuşmacı kümelemeyi aynı anda yapan bir yöntem önermiştir [5]. Önerilen yöntem, eşikleme yapma ihtiyacını ortadan kaldırmak, herhangi bir ön bilgiye sahip olma zorunluluğundan kurtulmak ve farklı akustik ortam şartlarına uyum sağlamak için halihazırda var olan gizli markov modeli (Hidden Markov Model-HMM), toplayıcı kümeleme (agglomerative clustering) ve BIC algoritmalarını birleştirmiştir. Aynı zamanda LPCC (Linear Prediction Cepstral Coefficient) ve MFCC olmak üzere iki farklı akustik ortam özelliği kullanılmıştır. 2004 yılında S.E. Tranter ve D.A. Reynolds televizyon haber ses kayıtlarını test ortamı olarak kullandıkları iki farklı konuşmacı diyarizasyon sistemi geliştirmiştir [6]. Araştırmacılar yeni bir kümeleme prosedürü ve BIC algortiması için durma kriteri üzerinde yoğunlaşmıştır. Bahsedilen çalışmada konuşmacı diyarizasyon sistemlerine konuşmacı cinsiyetinin de etki edebileceği vurgulanmıştır. 2005 yılında D.A. Reynolds ve P.T-Carrasquillo, o zaman kadar

yapılan tüm diyarizasyon çalışmalarını bir araya getirip bu çalışmaların performanslarını karşılaştırmıştır [7]. Diyarizasyon sisteminin genel çerçevesinin (framework) çıkarıldığı araştırmada DARPA (RT-04F) diyarizasyon yarışmasının ilkeleri uygulanmıştır. Bu çalışmada bir diyarizasyon sistemini oluşturan alt sistemlerden konuşma aktivite tespiti, konuşmacı değişim tespiti, cinsiyet sınıflandırma, kümeleme ve yeniden segmentasyon aşamaları detaylı şekilde açıklanmıştır. Değerlendirme kriteri olarak miss (gerçekte konuşmacı olan bölgenin konuşmacı yok şeklinde etiketlenmesi), false alarm (gerçekte konuşmacı olmayan bölgenin konuşmacı var şeklinde etiketlenmesi) ve speaker-error (konuşmacının tespit edilmesi lakin yanlış konuşmacının atanması) üçlüsünün toplandığı DER sistemi kullanılmıştır.

Konuşmacı diyarizasyon sistemleri geliştirilirken en önemli aşamalardan biri konuşmacıya özgü özelliklerin çıkarıldığı kısım olarak karşımıza çıkmaktadır. Bu aşamada N.Dehak ve arkadaşlarının 2011 yılında keşfettiği i-vektör [8] konuşmacıya özgü özelliklerin çıkarılması açısından büyük öneme sahiptir. Basitleştirilmiş faktör analizi (Joint Factor Analysis-JFA) olarak ta bilinen i-vektör özellikle konuşmacı tanıma alanında büyük başarı yakalamış, kısa konuşma bölütleri için özellik gösterimi olarak ta konuşmacı diyarizasyonuna uyarlanmıştır. I-vektör'ün literatürde kullanıldıktan sonra kendisinden önceki MFCC ve konuşmacı faktörü [9] özelliklerinin yerini aldığı görülmüştür. 2010 yıllarında derin öğrenmenin literatüre girmesiyle, konuşmacı diyarizasyon alanında derin öğrenmenin güçlü modelleme yeteneğini kullanan birçok çalışma gerçekleştirilmiştir. Derin öğrenme ile çıkarılan özellikler artık konuşmacı gömülü özellikleri (speaker embeddings) olarak tanımlanmaya başlanmıştır. Konuşmacı gömülü özellikleri d-vektör [10-12] ve x-vektör [13] olarak iki şekilde çalışılmıştır. D-vektör ve x-vektör, daha önce konuşmacı tanıma için eğitilmiş bir ağın son gizli katmanından konuşmacıya özgü özellik vektörü çıkarılması ile elde edilmektedir. Bu tez çalışması kapsamında geliştirilen konuşmacı diyarizasyon sistemi de d-vektör tabanlı inşa edilmiştir. I-vektör sistemlerinden bahsi geçen d-vektör ve x-vektör sistemlerine geçiş konuşmacı diyarizasyon sistemlerinde performans artışına, daha fazla veriyi daha kolay işleyebilmeye [14], konuşmacı ve akustik ortam çeşitliliğine karşı dayanıklılığa katkı sağlamıştır. Diğer yandan, geleneksel konuşmacı diyarizasyon sistemlerinde bulunan alt sistemleri tek parça nöral ağda birleştiren uçtan uca nöral diyarizasyon sistemleri de kullanılamaya başlanmıştır. Fujita Y. ve arkadaşları, 2019 yılında özellik çıkarma ve kümeleme gibi hiçbir şekilde ayrı modüllere sahip olmayan tek parça (end-to-end) konuşmacı diyarizasyon sistemi geliştirmişlerdir [15,16]. Bu sistemde konuşmacı diyarizasyonu problemine çok etiketli sınıflandırma bakış açısı ile yaklaşılmıştır. Konvansiyonel kümeleme tabanlı konuşmacı diyarizasyon sistemleri %28,77 diyarizasyon hata oranına (DER) sahipken, kendi sistemleri %12,28 DER ile çalışmıştır. Bu tür sistemlerde ağın eğitilmesi için büyük miktarda veriye ihtiyacın olması halen dezavantaj olarak görülmektedir.

Konuşmacı diyarizasyonu ile ilgili aynı zamanda literatürde yayınlanmış bir çok yüksek lisans ve doktora tezi bulunmaktadır. 2017 yılında Beatriz Martinez Gonzalez toplantı ses kayıtları için güçlü bir konuşmacı diyarizasyon sistemi gerçekleştirmeye çalışmıştır [17]. Çalışmada mikrofon dizisi (micophone arrays) şeklinde toplanan ses kayıtlarında mikrofonlar arası gecikmeden (delay) yola çıkarak konuşmacı değişim noktalarını tespit etmiştir. Diyarizasyon hata oranını düşürmede, farklı mikrofonlara gelen farklı sinyaller arasındaki çapraz korelasyon seçiminin büyük katkısının olduğu tespit edilmiştir. Bunun yanısıra Silhouette (silüet) katsayısı tabanlı K-Means kümeleme ve Temel Bileşen Analizi (Principal Component Analysis-PCA) kullanımının diyarizasyon sistemlerini iyileştirmede büyük başarı sağladığı belirtilmiştir. Tezin başka bölümünde toplantılarda tek kanallı ses sinyali olması durumunda da glotal ses özelliklerinin kullanılması gerektiği vurgulanmıştır. Konuşmacı bölütlemeye farklı bir modifikasyon yapılarak diyarizasyon sistemlerinin geliştirilebileceği öngörülmüştür. 2020 yılında Amerika Birleşik Devletlerinde bulunan MIT (Massachusetts Teknoloji Enstitüsü) üniversitesinde yayınlanan bir tez çalışmasında Adedotun J.O. toplantı senaryosu çerçevesinde bir konuşmacı diyarizasyon sistemi geliştirmiştir [18]. Tamamen çevrimiçi veya tamamen çevrimdışı çalıştırılabilecek bu yapının, ses kaydetme veya ses transfer etme ihtiyacı olmadan diyarizasyon işlemini gerçekleştirilebiliyor olması vaad edilmiştir. Konuşmacıların cinsiyetinin ve aynı anda birden fazla kişinin konuştuğu anda ortaya çıkan "örtüşen konuşma" bölütlerinin de diyarizasyon başarı oranına etki edebileceği göz önünde bulundurulmuştur. Kümeleme algoritması olarak K-Means, konuşma aktivite tespiti (VAD) sistemi için de ResNet50 (Residual Network 50) mimarisi kullanılmıştır. Çevrimdışı senaryoda önerilen sistem %27,8 diyarizasyon hata oranı ile çalışmıştır. Ses kaydında örtüşen ses olmaması, %44,3 küçük örtüşen ses olması ve %50 büyük ölçüde örtüşen ses olması şeklinde ihtimallerde de aynı hata oranı elde edilmiştir. Sistem çevrimiçi çalıştırıldığında ise yukarıdaki her üç ihtimal için sırasıyla %16,9, %37,2 ve %45,6 hata oranları gözlenmiştir. Yine 2020 yılında Tayland'ta bulunan Thammasat Üniversitesinde, Pantid Chantangphol, lisansüstü tezinde televizyon haber ses kayıtlarında konuşmacı diyarizasyon sistemini ele almıştır [19]. Geliştirilen sistemde Python programlama dili kullanılarak düşün diyarizasyon hata oranına (DER) sahip adım adım bir konuşmacı diyarizasyon sistemi tasarlanması hedeflenmiştir. Önerilen sistemde farklı arka plan gürültüleri de hesaba katılarak, DenseNet (Dense Convolutional Network) derin öğrenme mimarisi tabanlı özellik çıkarma konuşmacı ayırt etme işlemleri gerçekleştirilmiştir. Konuşmacıya özgü özellik çıkarmada Mel Frekansı Kepstrum Katsayıları (MFCC) ve log-mel spektrogramın kullanıldığı çalışmada, konuşmacı bölütleri arasındaki mesafeyi bulmak için kosinüs benzerliği metriğinden faydalanılmıştır. Çalışmada konuşma aktivite tespiti (VAD) aşaması süresince özellik çıkarma ve derin öğrenme kullanımının başarıyı arttırdığı vurgulanmıştır. Ayrıca log-mel spektrogram ve MFCC'den özellik çıkarma aşamasında, ResNet96, DenseNet121 ve modifiye edilmiş DenseNet'ten derin öğrenme mimarisi kısmında faydalanarak, bu teknolojilerin farklı

kombinasyonda kullanımı sonucunda VAD'ın ne kadar geliştirilebileceği ölçülmüştür. 2020 yılının Eylül ayında İsveç'in başkenti Stokholm'de bulunan KTH Royal Teknoloji Enstitüsünde, Yi Li, lisansüstü tezinde çağrı merkezleri ses kayıtları için bir konuşmacı diyarizasyon sistemi önermiştir [20]. Çalışmasının temelini konuşmacı doğrulama sistemleri üzerine inşa eden araştırmacı, MFCC-vektör adını verdiği konuşmacı özellikleri kullanmıştır. Tez çalışmasında faydalandığı konuşmacı doğrulama sistemi mevcutta telefon ses kaydında konuşmacının kimliğini doğrulayan bir çağrı merkezi uygulaması olarak seçilmiştir. Özellik çıkarma da 13.dereceden MFCC kullanan bir konuşma aktivite tespit (VAD) sistemi geliştirilmiştir. Daha sonra, bölütleme ve Gaussian Mixture Model (GMM) ve BIC skoru tabanlı lineer ve hiyerarşik kümeleme uygulamıştır. Özellikle VAD algoritmasında değişken çerçeve boyutlarına sahip agresiv VAD gerçekleştirilmiş, böylece ses kaydındaki konuşma olmayan bölgeleın atılması daha sık şekilde uygulanmıştır. Önerilen sistemin değerlendirme metriği olarak, konuşma içeren sesin uzunluğunu belirten Scored Speech Time (SST), algoritmanın konuşma bölgelerini yanlış bir şekilde konuşma yok şeklinde etiketlenmesi sonucu oluşan Missed Speech Time (MST), MST'nin tam tersine konuşma olmayan bölgelerin konuşma olarak işaretlendiği süreyi belirten False-alarm Speech Time (FST), algoritmanın yanlış konuşmacı atadığı zamanı belirten Speaker Error Rate (SER) ve MST, FST, SET toplamlarının SST'ye bölünmesi ile elde edilen Overal Speaker Diarization Error (OSDE) kullanılmıştır. 2006 yılında İspanya'nın Barselona şehrinde bulunan Katalonya Politeknik Üniversitesi'nde (BarcelonaTech) yapılan bir doktora tezinde, Xavier Anguera Miro, toplantı ses kayıtları üzerine güçlü bir konuşmacı diyarizasyon sistemi önermiştir [21]. Derin öğenme mimarilerinin henüz popüler olmadığı süre zarfında, araştırmacı, tek kanallı ses kayıtlarında hiyerarşik yukarıdan aşağıya yaklaşımına sahip konvansiyonel/geleneksel bir konuşmacı diyarizasyon sistemi tasarlamıştır. Tek kanallı (mono) konuşmacı diyarizasyon sistemlerini geliştirmeye odaklanan çalışmada, NIST'in zengin transkripsiyon yarışmalarında kullandığı RT05s ve RT06s veri setleri kullanılmıştır. Toplantılarda birden fazla mikrofon bulunması halinde Sinyalin Varış Zamanının Gecikmesi (Time Delay of Arrival-TDOA) yaklaşımının kullanılabileceği vurgulanmıştır. 2015 yılında Singapur'da bulunan Nanyang Teknoloji Üniversitesinde, Nguyen Trung Hieu, toplantı ses kayıtları için konuşmacı diyarizasyon yöntemi öneren bir doktora tezi yayınlamıştır [22]. Çalışmada özellikle konuşmacı diyarizasyonunun alt aşaması olan konuşmacı kümeleme iyileştirilmeye çalışılmıştır. Konuşmacı kümeleme için en uygun metriğin tespit edilebilmesi için farklı uzaklık metrikleri değerlendirilmiştir. Bayesian Information Criterion (BIC) gibi Generalized Likelihood Ratio (GLR) tabanlı metriklerin güçlü olmamasından ötürü konuşmacı diyarizasyon sistemlerinde kullanılmaması gerektiği üzerinde durulmuştur. Çalışmada, Information Change Rate (ICR) adında yeni bir uzaklık ölçüm yöntemi önerilmiştir. Önerilen bu uzaklık metriğinin farklı uzunluktaki konuşmacı bölütlerinden olumsuz anlamda çok fazla etkilenmediği, hatta GLR tabanlı metriklere oranla diyarizasyon hata oranını %10'a yakın iyileştirdiği vurgulanmıştır. Önerilen

yöntemin değerlendirilmesinde NIST zengin transkripsiyon yarışmalarında kullanılan RT07s kullanılmıştır. 2020 yılında Fransa'nın başkenti Paris'te bulunan Sorbonne Üniversitesinde gerçekleştirilen bir doktora tezinde, José María PATINO VILLAR, konuşmacı değişiminde düşük gecikmeye sahip bir konuşmacı diyarizasyon sistemi önermiştir [23]. Metrik ve model tabanlı yaklaşımların kombinasyonu ile inşa edilen yöntemde konuşmacıları ayırt edebilmek için spektral kümeleme yönteminden faydalanılmıştır. Konuşmacı modellenmesi için konuşmacı doğrulama yöntemleri kullanılmıştır. Modelleme için özellikle ikili anahtar (binary keys-BK) yaklaşımına başvurulmuştur. Bu yöntemin büyük veri eğitme ihtiyacını ortadan kaldırdığı vurgulanmıştır. Yine BK'ların kullanımı ile konuşmacı değişim noktalarının daha etkili tespit edilebileceği önerilmiştir. Tez çalışması, çoğunlukla EURECOM Üniversitesinin bir konuşmacı tespit projesi olan ANR ODESSA projesine odaklanmıştır. Asıl amacın konuşmacı tespitinin düşük gecikmeyle tahmin edilmesi olduğu belirtilmiştir. 2019 yılında Finlandiya'da bulunan Aalto Üniversitesinde, Tuomas Kaseva, lisansüstü tezinde SphereDiar adında toplantı ses kayıtlarında konuşmacı diyarizasyonu gerçekleştiren bir sistem geliştirmiştir [24]. Önerilen yöntem kendileri tarafından daha önce geliştirilmiş; konuşmacı modelleme SphereSpeaker sinir ağı, homojenlik tabanlı bölütleme ve en yukarıdaki iki Silhouette adındaki üç sistemin birleşimi olarak tasarlanmıştır. Derin öğrenme yöntemlerine de başvurulan çalışmada, özellik çıkarma aşamasında MFCC kullanılmıştır. Önerilen konuşmacı diyarizasyon yönteminde konuşmacı kümeleme için Spherical K-means (K-means'in modifiye edilmiş hali) uygulanmıştır. Geliştirilen diyarizasyon sitemi araştırmacının kendi oluşturduğu ve 200'den fazla toplantı ses kaydının bulunduğu bir veri seti ile değerlendirilmiş, mevcutta bulunan iki farklı diyarizasyon sisteminden daha başarılı sonuç aldığı vurgulanmıştır. 2015 yılında İskoçya'da bulunan Edinburgh Üniversitesinde, Mark Sinclair doktora tezi kapsamında, transkripsiyon ve çeviri işlemlerinde kullanılabilecek bir konuşmacı diyarizasyon sistemi tasarlamıştır [25]. Tez çalışmasında mevcut GMM-HMM tabanlı yöntemlerin yetersiz kaldığı ispatlanmaya çalışılmış, konuşma bölütlemede daha güçlü olabilecek yeni bir derin öğrenme ağı (Deep Neural Network-DNN) tabanlı yaklaşım önerilmiştir. Değerlendirme kriteri olarak, konuşma bölütlemede VAD hata oranı, konuşmacı bölütlemede ise DER hata oranı dikkate alınmıştır. Özellik çıkarma aşamasında MFCC ve PLP'den faydalanılmıştır. Yine 2020 yılında Amerika Birleşik Devletleri Georgia Eyaletinde bulunan Georgia Teknoloji Enstitüsünde, Yufeng Yang lisansüstü tezi kapsamında Voxceleb veri seti kullanarak otomatik bir konuşmacı doğrulama ve konuşmacı diyarizasyon sistemi tasarlamıştır [26]. Bu çalışma ses ile görüntünün (videonun) birleştirilerek (audio visual) diyarizasyonun gerçekleştirildiği literatürdeki tek lisansüstü tezi olmuştur. EACeleb adı verilen ve araştırmacının ürettiği görüntü veri setinde doğu asya dillerini konuşan ünlülere yer verilmiştir. Önceden eğitilmiş x-vektör ağına nazaran önerilen yöntemin %25'e kadar daha başarılı sonuç ürettiği belirtilmiştir. Ayrıca konuşmacı doğrulama görevinin EACeleb veri seti ile eğitilmesi sonucunda, Voxceleb'e göre yaklaşık %36'ya yakın daha başarılı

çalıştığı gözlenmiştir. Önerilen yöntemde çerçeve seviyesinde (frame-level) x-vektör'lerden faydalanılmıştır.

Konuşmacı diyarizasyonu alanında 2022 yılında da güncel olarak bir çok çalışmanın yapılmış olduğu görülmektedir. Ahmed Isam ve arkadaşları tarafından "*Channel and channel subband selection for speaker diarization*" başlıklı araştırmada çok mikrofonlu ortamda, farklı mikrofonlardan gelen sinyallerde ayrı ayrı özellik çıkarma işlemi uygulanarak kanal tabanlı bir konuşmacı diyarizasyon sistemi önerilmiştir [27]. Çalışmada özellik seçimi olarak Hızlı Fourier Dönüşümü kutucukları (Fast Fourier Transform-FFT Bins) kullanılmıştır. GMM-BIC ve Toplayıcı Hiyerarşik Kümeleme (Agglomerative Hierarchical Clustering-AHC) yöntemlerinden faydalanılan konuşmacı diyarizasyon sisteminde %52,83 diyarizasyon hata oranı elde edilmiştir. 2023 yılı tarihli başka bir çalışmada, Meysam Shamsi ve arkadaşları tarafından insan kontrollü konuşmacı diyarizasyon modeli geliştirmeye çalışılmıştır [28]. Çalışmada x-vektör ve i-vektör farklılıklarına da değinilmiştir. Hayatboyu öğrenme (lifelong learning) prensibi ile yaptıkları araştırmada, önerilen konuşmacı diyarizasyon sistemi canlıda kullanıldıkça yeni gözlemleri öğrenmiş olduğu veri setine entegre etmektedir. Böylece sistem karşılaştığı durumlardan da öğrenme gerçekleştirebilmektedir. 2022 yılında Vijay K. ve Rajeswara R.R. derin öğrenme kullanarak konuşmacı diyarizasyonunun alt aşaması olan konuşmacı bölütlemedeki konuşmacı değişim noktalarını optimize etmiştir [29]. Kümeleme algoritması olarak Derin Gömülü Kümeleme (Deep Embedded Clustering-DEC) yönteminden, optimizasyon algoritması olarak ta COVID-19'un önlem yaklaşımı ve balinaların avlanma yaklaşımlarını birleştiren Kesirli Antikorona Balina Optimizasyonundan (Fractional Anticorona Whale Optimisation-FAWO) faydalanılmıştır. Test doğruluğunda yaklaşık %90, diyarizasyon hata oranında ise %60 değerleri elde edilmiştir. Özellik çıkarma aşamasında MFCC, LPCC, spektral yayılım, spektral rolloff ve spektral skewness gibi yöntemlere başvurulmuştur. Başka bir 2023 tarihli çalışmada, Miquel India ve arkadaşları tarafından telefon görüşmeleri ses kayıtları kullanılarak konuşmacı diyarizasyonu için dil ve akustik modelleme gerçekleştirmiştir [30]. Önerilen sistem, bir LSTM ağının konuşmacı sınıflandırıcı olarak kullanıldığı yinelemeli bir algoritmaya dayanmaktadır. Akustik özelliklerin ve dilsel içeriğin birleşimi, HMM/VB temel sistemiyle karşılaştırıldığında kelime düzeyinde DER açısından %84,29'luk bir gelişme gösterdiği anlaşılmıştır. Or Haim Anidjar ve arkadaşları tarafından 2023 tarihli yayınlanan bir çalışmada konuşmacı değişimi ve diyarizasyonu sistemi için konuşma ve çokdilli doğal dil modeli önerilmiştir [31]. Çalışmada konuşmacı sırası değişiminin tespitinden sonra konuşmacı diyarizasyonu için de kümeleme yapılmıştır. Çoğunlukla İngilizce, Fransızca ve İbranice dilleri üzerinde durulan çalışmada, dilden bağımsız yapı önerilmiştir. Kapsamlı değerlendirme sonucunda çok dilli Konuşmacı Değişimi Tespiti çatısının (Speaker Change Detection-SCD Framework), tek dilli veri kümeleri üzerindeki çatılar ile karşılaştırıldığında yeterince başarılı olduğu gözlenmiştir. Konuşmacı diyarizasyonu konusunda

günümüze kadar yapılmış en geniş derleme çalışmasını 2021 yılında Tac Jin Park ve arkadaşları "A Review of Speaker Diarization: Recent Advances with Deep Learning" başlıklı araştırma ile gerçekleştirmiştir [32]. Çalışmalarında, konuşmacı diyarizasyonunun tarihsel gelişiminden, geleneksel ve modern (derin öğrenme) diyarizasyon sistemlerine ve kullanılan metotlara kadar birçok alana değinilmiştir. Değerledirme kriterleri olarak, diyarizasyon hata oranı (DER), Jakart Hata Oranı (Jaccard Error Rate-JER), Kelime Seviyesi Diyarizasyon Hata Oranı (Word-level Diarization Error Rate-WDER) detaylı bir şekilde açıklanmıştır. Çalışmada özellikle son zamanlarda derin öğrenme yöntemlerinin diyarizasyona olan katkıları vurgulanmış, sonuç olarak mevcut birkaç yöntemin karşılaştırılması yapılarak gelecekte gerçekleştirilebilecek çalışmalar anlatılmıştır.

Bu tez çalışmasının tamamen özgün olan konuşma aktivite tespiti aşaması dışındaki diğer adımlarına ilham kaynağı olan araştırma 2017 yılında Google firmasına bağlı çalışan Quan Wang ve arkadaşları tarafından "*Speaker Diarization with LSTM*" başlıklı çalışma ile gerçekleştirilmiştir [33]. Araştırmacılar bu çalışmada, daha önce gerçekleştirmiş oldukları konuşmacı doğrulama [34] sistemlerini kullanarak yeni bir LSTM tabanlı konuşmacı diyarizasyon sistemi önermiştir. PLP özelliklerinin i-vektör şeklinde kullanılarak geliştirilen GMM tabanlı bir konuşma aktivite dedektörü (VAD) kullanılmıştır. Yazarların önerdiği diyarizasyon sistemi genel olarak konuşma aktivite tespiti, konuşma bölütleme, konuşmacı özelliği çıkarma ve konuşmacı kümeleme olmak üzere dört aşamadan oluşmaktadır. Konuşmacı gömülü özellikleri daha önceden konuşmacı doğrulama için kendi geliştirdikleri 3 katmanlı LSTM içeren bir nöral ağdan çıkarılmış ve buna d-vektör adı verilmiştir. Kümeleme aşamasında ise spektral kümeleme üzerinde bir takım geliştirme (refinement) işlemi uygulanmıştır. Çalışmanın sonunda i-vektör ve d-vektör konuşmacı özelliklerinin Naive, Links, K-Means ve Spekral kümeleme algoritmaları ile birlikte performansları ölçülmüştür. En düşük hata oranının d-vektör'ün spektral kümeleme ile kullanımı sonucunda elde edildiği gözlenmiştir.

Bu tez çalışması toplamda altı ana bölümden oluşmaktadır. Tezin Giriş bölümünü takip eden ikinci kısmında insanda ve doğada ses ve konuşmanın nasıl oluştuğu, ses bilimi (fonetik), Türkçe dilindeki sesler ve özellikleri anlatıldıktan sonra sesin insanda algılanması yani işitmenin yapısına değinilmiştir. Üçüncü bölümde geçmişten günümüze dijital (sayısal) konuşma işleme teknikleri; ön işlemler, konuşma/konuşmacı tanıma, ses iyileştirme ve konuşma sentezleme kategorilerine ayrılarak detaylı bir şekilde açıklanmıştır. Aynı kısımda bu kategorilerin her birinde kullanılan yöntemlere de değinilmiştir. Tezin dördüncü yani materyal ve metot başlıklı kısmında önerilen konuşmacı diyarizasyon sisteminin tüm aşamaları derinlemesine incelenmiştir. Bu bölümde sırasıyla denetimli VAD sistemi için oluşturulan veri seti, LSTM ve eşikleme tabanlı gerçekleştirilen Hibrit VAD yapısı, d-vektörlerin (konuşmacı özelliklerinin) çıkarımı, spektral kümeleme işlemi ve oluşan nihai diyarizasyon sistemi anlatılmıştır. Oluşan konuşmacı

diyarizasyon sisteminin dalga formu grafiğinde orjinal ses ile karşılaştırması yapılmıştır. Tezin beşinci kısımda özgün olan VAD sistemi ve sonucunda oluşan diyarizasyon yapısı ile ilgili bulgulara yer verilmiş, altıncı ve son kısmında elde edilen sonuçlar yorumlanmıştır.

2. SES VE KONUŞMA FİZYOLOJİSİ

2.1. Ses ve Konuşma

Ses, belirli frekans değeriyle titreyen bir cisim tarafından oluşturulan ve işitme hissi uyandıran basınç dalgası olarak tanımlanmaktadır. Ses dalgalarının yayılabilmesi için bulunduğu ortamın iletici bir özelliğe sahip olması gerekmektedir. Bulunduğu ortamın (katı, sıvı, gaz) özelliklerine göre farklı yayılma hızına sahiptirler. Ses, yayılma hızı, frekans, basınç, dalga boyu, desibel, genlik, şiddet, renk ve tını gibi fiziksel özelliklere sahip olabilmektedir.

İnsanlarda ses oluşturma sisteminin 3 ana bölümden oluşur:

- Akciğer
- Gırtlak (ses telleri, ses kıvrımları)
- Ses Bölgesi (yutak,ağız)

Bu sistemde, akciğerler hava kaynağı, ses telleri titreşim elemanı ve ağız rezonatör bölge olarak görev yapmaktadır. Akciğerlerden gelen hava,ses tellerinden geçerek bu tellerinin titreşimini sağlayarak sesi oluşturmaktadır.

2.2. Ses ve Konuşma Fizyolojisi

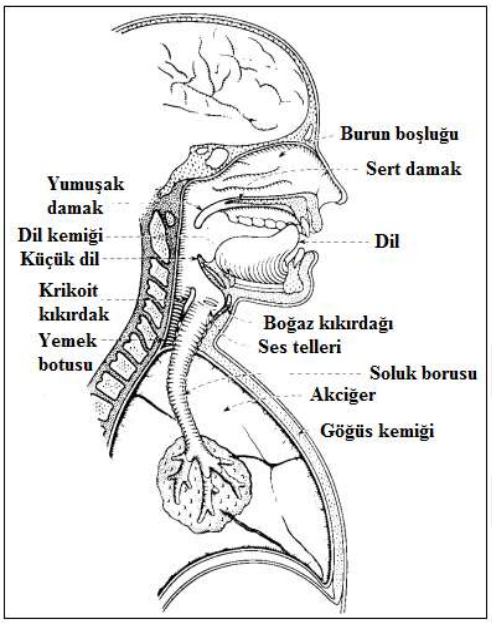
Ses dalgası, insanlarda ses üretim mekanizması tarafından üretilen akustik bir hava basıncı dalgası olarak ortaya çıkmaktadır. Akciğerler, trake (soluk borusu), ses telleri, gırtlak, boğaz, ağız ve burun bu mekanizmanın başlıca organlarıdır. Akciğerler ve akciğerde bulunan kaslar ses üretim mekanizmasında hava kaynağı olarak görev yapmaktadırlar. Akciğerde bulunan kaslar, havayı bronşlar ve trake boyunca akciğer dışına çıkarırlar. Konuşma sesleri genel olarak sesli kısım (voiced) ve sessiz kısım olmak (unvoiced) üzere ikiye ayrılmaktadır. Sesli kısım akciğerlerden gelen havanın gerilmiş ses tellerini titrettiği zaman oluşmaktadır. Sessiz kısım ise üretilen havanın ağızda ani ve düzensiz patlaması sonucu oluşur.

2.2.1. İnsanda Ses ve Konuşma Üretimi

İnsan sesinde ayırt edici özellikler fiziksel ve öğrenilmiş olmak üzere iki genel başlığa ayrılmaktadır. Ses telleri ve yukarısında kalan organlar olarak tanımlanan vokal yolu, yapı itibari ile insan sesini farklı kılan fiziksel özelliktir. Şekil 2.1'de görselleştirildiği gibi ses üretim organları şu şekilde sıralanabilir:

- Alt yutak (laryngeal pharynx) (gırtlak kapağının altında)
- Orta yutak (oral pharynx) (dilin arkasında bulunan yumuşak damak ile gırtlak kapağı arası)

- Ağız boşluğu (oral cavity) (yumuşak damak ile başlayıp dil ve dudak ile sonlanan kısım)
- Üst yutak (nasal pharynx) (yumuşak damağın üstü, burun boşluğunun arka ucu)
- Burun boşluğu (nasal cavity) (sert damağın üstü ve yutaktan başlayıp burun deliklerinde sonlanan kısım)



Şekil 2.1. İnsanda ses oluşturma/çıkarma organları

Yetişkin bir erkeğin vokal yolu yaklaşık 17 cm uzunluğundadır. Gırtlak, ses telleri, aritenoid kıkırdak, gırtlak kıkırdağı ve adem elması olarak bilinen gırtlak çıkıntısından oluşmaktadır. Ses telleri, gırtlak çıkıntısı ile aritenoid kıkırdak arasında gerilmiş halde bulunmaktadır. Ses telleri araasında bulunan alan ya da küçük dilin bulunduğu yer nefes borusunun ağzıdır. Akustik dalga vokal yolundan geçtiğinde frekans değeri vokal yolunun yankı ve çınlama gibi özelliklerinden dolayı değişime uğramaktadır. Vokal yolunun bu şekilde oluşturduğu rezonans etkisine "*formant*" (biçimlendirici) denmektedir. Vokal yolu yapısı, böylece ses sinyalinin spektral şeklinden (formantların konumu ve spektral eğim yardımıyla) belirlenebilmektedir.

2.2.2. Ses Oluşumu

Günümüzde ses doğrulama sistemleri sadece vokal yolundan çıkartılan özelliklerle gerçekleştirilmektedir. Şekil 2.1'de de görüldüğü gibi insan ses üretme sisteminde uyarım kaynağı bulunmaktadır. Akciğerlerin uyarım kaynağı olarak ürettiği hava akışının nefes borusu (trake) yardımıyla ses tellerine geçmesiyle fiziki olarak ilk ses dalgası üretim işlemi başlamaktadır.

Uyarımın fonasyon, fısıldama, friksiyon (sürtünme), basınçlandırma, titreşim ve ya bunların kombinasyonu şeklinde çeşitleri bulunmaktadır.

Fonasyon, akciğerlerden gelen havanın ses telleriyle şekillenmesi sonucu ortaya çıkmaktadır. Ses telleri kapalı durumdayken alttan gelen basınç ile patlayıp ayrılırlar. Basınçı hava sonlandığında ses telleri gerilimin ve sahip oldukları esnekliğin sayesinde tekrar eski hallerini alırlar. Bu durum hava akışının salınım yapan ses tellerinden başlayarak oluşan vuruşlar olarak şekillenmesini sağlar. Ses tellerindeki bu salınımın frekansı temel frekans (fundamental frequency) diye adlandırılmaktadır ve bu frekans ses tellerinin uzunluğuna, gerilimine ve ağırlığına bağlı değişmektedir. Böylece sesin fiziksel olarak ayırt edici karakteristik özelliklerinden birinin de temel frekans olduğu söylenebilmektedir.

Fısıldama, akciğerden gelen havanın, kapalı sayılabilecek açıklıkta olan ses tellerinin arka kısmında yer alan aritenoid kıkırdakları arasındaki üçgen küçük açıklıktan geçmesiyle oluşmaktadır. Bu işlem, akciğerden gelen havanın yüksek gürültü oranı ile türbülanslı hava akışına dönüşmesine neden olur.

Friksiyon (sürtünme), vokal yolundaki büzülmeler ve daralmalar sonucu oluşan sestir. Vokal yolundaki bu daralmanın yeri, şekli ve açısı oluşacak geniş bant gürültünün şeklini belirler. Friksiyon ile sürtünmeli sessiz (fricatives) ve ıslıklı sessiz (sibilants) sesleri oluşur.

Basınçlandırma, tamamen kapalı olan vokal yolunun basınca maruz kalıp açılmasıyla oluşan sestir. Bu durumda kısa bir sessizliğin ardından (vokal yolunda basınç birikir) ani ve kısa bir ses patlaması olur. Basınç ani bir şekilde sonlandırılması ile patlayıcı (plosive) bir ses çıkarken basıncın kademeli kaldırılması ile yarı kapalı sessiz (affricate) ses oluşur.

Titreşim, ses telleri yerine vokal yolundaki kapalılığa (özellikle dil şekilleriyle) maruz kalan havanın meydana getirdiği ses olarak tanımlanmaktadır (örneğin: "r" sesini çıkarma).

Fonasyon ile üretilen sese sesli (voiced), fonasyon ve friksiyon karışımı ile üretilen sese karışım (mixed) ve diğer şekiller ile üretilen sese sessiz (unvoiced) denilmektedir. Fonasyon ve fısıldama'nın aksine friksiyon, basınçlandırma ve titreşim'de sesin üretim yeri vokal yoludur. Göğüs bölgesi, vokal sistemin rezonans özellikleri açısından önemli rol oynar. Burada bulunan nefes borusu genelde 12 cm uzunluğunda, 2 cm çapında birbirlerine bağ dokusuyla bağlanmış kıkırdak halkaların iç içe geçmesiyle oluşan bir organdır. Ses telleri titreştiği anda altında ve üstünde yankılanmalar (sublottal resonance) oluşturur. Nefes borusunun yapısına göre değişebilen bu yankılanmalar konuşmacıya bağlı fiziksel bir özelliktir.

2.3. Ses Bilimi (Fonoloji)

Ses Bilimi ve başka bir deyişle Fonoloji, dillerin veya lehçelerin seslerini veya işaret dilleri için işaretleri oluşturan parçaların sistematik olarak nasıl organize edildiğini inceleyen dilbilim dalı olarak tanımalanmaktadır. Terim ayrıca belirli bir dil çeşidinin ses veya işaret sistemine de atıfta

bulunmaktadır. İlk zamanlarda, fonoloji çalışması sadece konuşma dillerindeki fonem sistemlerinin incelenmesiyle ilgilenmekte iken, şimdi herhangi bir dilbilimsel analizle ilgili fonoloji çalışmaları kapsamında yapılabilmektedir. Fonoloji kavramsal olarak, dilin seslerinin veya işaretlerinin fiziksel üretimi, akustik iletimi ve algılanması ile ilgili olarak fonetikten ayrı bir bilim dalı olarak görülmektedir. Fonoloji, anlamı kodlamak için belirli bir dil içinde veya diller arasında nasıl işlev gördüklerini açıklamaktadır. Pek çok dilbilimci için fonetik betimleyici, dilbilime ve fonoloji teorik dilbilime aittir, ancak bir dilin fonolojik sistemini oluşturmak zorunlu olarak bazı teorilerdeki fonetik kanıtların analizine teorik ilkelerin uygulanmasını gerektirmektedir. Özellikle 20. yüzyılın ortalarında modern fonem kavramının geliştirilmesinden önce, ayrımın her zaman yapılmadığı gözlenmiştir. Modern fonolojinin bazı alt alanları psikolinguistik ve konuşma algısı karşımıza çıkmaktadır. Ayrıca eklemleyici (bağlayıcı) fonoloji veya laboratuvar fonolojisi gibi belirli alanlar da bulunmaktadır.

2.3.1. Türkçe Dilindeki Sesler

Konuşma organlarının bir arada uyumla ve düzenli çalışmasıyla anlam ihtiva eden sözcükler ve tümceler oluşturmak için ağızdan çıkarılan birimlere ses (phon) denmektedir. Bir dilin sesli ifadelerinde birim eleman olarak ta bilinen seslerden yazıya geçişte her sese bir alfabetik simge verilmektedir. Ses sayısı çok olan dolayısıyla yazıya geçişi karışık dillerde seslerin kümelenmesiyle anlam ayırıcı özelliği bulunan fonemler meydana gelmektedir. Ses diller üstü bir kavram iken fonem anlam ayırıcı özelliği nedeniyle dillere özgüdür. IPA (*International Phonetic Association*), tüm diller için geçerli olan ve her sesin bir karakter ile temsil edilmesini amaçlayan *Internation Phonetic Alphabet* fonetik alfabesini tanımlamıştır. Türkçe dilinde, her foneme tek bir alfabetik simge (harf) atandığından, ifadeden yazıya ve yazıdan ifadeye geçişin oldukça yalın olduğu söylenebilmektedir. Türkçe'de diğer dillerde olduğu gibi fonemler parçalı (segmental) ve parçalarüstü (bürün) (suprasegmental) olmak üzere ikiye ayırmak mümkündür.

Parçalı Fonemler

Türkçe'de parçalı fonemler

- Ünlü
- Ünsüz
- Kayan ünlü

olmak üzere 3 başlıkta incelenebilir.

Ünlü sesler akciğerlerden gelen havanın hiçbir sürtünme veya engele maruz kalmadan dilin üstünden geçerken çıkardığı seslerdir. Bu sesler çıkartılırken konuşma organlarında herhangi bir kapanma ya da daralma olmadığı için gürültüsüz ses olarak bilinmektedirler. Türkçe'de 8 adet ünlü

harf bulunmaktadır. Her ünlünün bir açık bir de kapalı formu vardır. Fakat Türkçe'de ünlülerin açık ya da kapalı olması içinde yer aldığı sözcüğün anlamını değiştirmediğinden tüm ünlüler 8 adet fonem ile temsil edilmektedir. Bu fonemler: /a/, /e/, /o/, /ö/, /u/, /ü/, /ı/, /i/ şeklindedir. Ünlü sesler, dilin ağızdaki konumuna, dudakların biçimine, genizin açık/kapalı durumuna göre sınıflandırılmaktadır.

- Çene açıklığı (dar, geniş)
- Dil konumu (ön, arka)
- Dudakların biçimi (yuvarlak, düz)
- Geniz (açık, kapalı)

Türkçe'de geniz dışındaki durumlar önem arz etmektedir. Tablo 2.1'de ünlü fonemler çene açıklığı, dil konumu ve dudakların biçimine göre tablo şeklinde gösterilmiştir.

Tablo 2.1. Türkçe dilinde bulunan sesli fonemlerinin bazı parametrelere göre kategorize edilmesi

	Dilin Konumu				**Dudak Biçimi**		**Çene Açıklığı**	
Ünlü	**Düz Ön**	**Yuvarlak Ön**	**Orta**	**Arka**	**Yuvarlak**	**Düz**	**Dar**	**Geniş**
a				+		+		+
e	+					+		+
o				+	+			+
ö		+			+			+
ı			+			+	+	
i	+					+	+	
u				+	+		+	
ü		+			+		+	

Bir ünlü fonemin diğer bir ünlü fonem ile yer değiştirmesi sonucu anlam değişikliğinin meydana gelip gelmediğinin kontrol edildiği sözcük çiftlerine yalın sözcük çiftleri denmektedir. Türkçe'deki ünlü fonemleri dilin konumu, dudak biçimi ve çene açıklığı esas alınarak tespit etmede yararlanılan yalın sözcük çiftleri Tablo 2.2, Tablo 2.3 ve Tablo 2.4'te verilmiştir.

Tablo 2.2. Çene durumu açısından yalın çiftler

Geniş/Dar	**ı**	**i**	**u**	**ü**
a	kar/kır	kar/kir	kar/kur	sar/sür
e	kes/kıs	tez/tiz	bez/buz	ses/süs
o	koş/kış	sos/sis	koş/kuş	son/sün
ö	söz/sız	söz/siz	son/sun	söz/süz

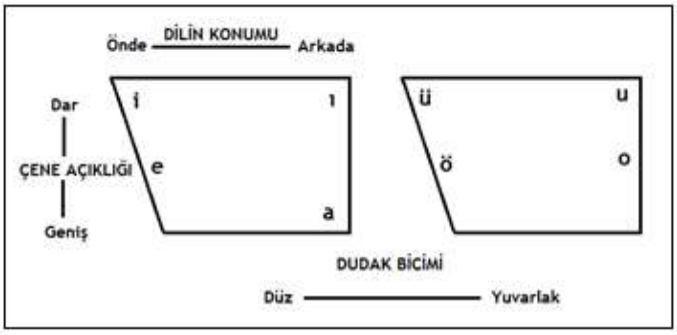
Tablo 2.3. Dudakların şekli açısından yalın çiftler

Düz/Yuvarlak	o	ö	u	ü
a	kar/kor	kar/kör	kar/kur	sar/sür
e	kel/kol	gel/göl	ser/sur	kel/kül
ı	kış/koş	sır/sör	tır/tur	tır/tür
i	kir/kor	kir/kör	kir/kur	kir/kür

Tablo 2.4. Dil pozisyonu açısından yalın çiftler

Arka/Ön	e	ö	i	ü
a	kas/kes	kar/kör	kar/kir	sar/sür
o	kol/kel	kor/kör	kor/kir	kor/kür
u	sur/ser	sun/sön	kur/kir	kul/kül

Ünlü fonemlerini sınıflandırmada kullanılan bir başka yöntem ise Şekil 2.2'de gösterilen ünlü dörtgenidir (vowel diagrams). İki boyutlu olarak çizilen ünlü dörtgeni çene açıklığı ve dilin konumuna göre oluşturulmaktadır. Yatay eksende dilin arka ve ön konumda bulunması, dikey eksende çene açıklığının dar ya da geniş olması dikkate alınır. Dudak biçimi olan düz ve yuvarlak parametreleri için iki ayrı dörtgen belirtilmektedir.



Şekil 2.2. Türkçe dili için ünlü dörtgeni diyagramı

Türkçe'de bir sözcüğün ilk hecesindeki ünlü kendisinden sonra gelen ünlüleri etkiliyorsa ünlü uyumundan bahsedilir. İlk hecede bulunan ünlü söylenilirken dil konumu önde ise, yani başka bir deyişle /e/, /ö/, /i/, /ü/ fonemlerinden biri ise, kendisinden sonra gelen ünlü de öndil ünlüsü olmak zorundadır. Sözcüğün ilk hecesindeki ünlü arkadil ünlüsü ise, yani /a/, /o/, /u/ fonemlerinden

biri ise, kendisinden sonra gelen ünlü ya ortadil ünlüsü /ı/, ya da arkadil ünlüsü olur. Ünlü uyumunda, ilk hecedeki ünlü düz dudak ünlüsü ise (/ı/,/i/,/a/,/e/) diğer hecelerdeki ünlülerde düz dudak ünlüsü olmak zorunda. Eğer ilk hecede yer alan ünlü yuvarlak dudak ünlüsü ise, devamında gelen ünlüler düz ya da yuvarlak dudak ünlüsü olabilmektedir. Ayrıca, geniş çene ünlüsü ile biten bir köke ünlü ile başlayan bir ek gelirse, araya geçiş sesi /y/ eklenerek, bu geniş çene ünlüsü dar çene ünlüsüne dönüştürülmektedir.

Ünlü fonemlerin çıkışı sırasında akciğerden gelen hava akımının hiçbir engele rastlamaksızın özgür çıkışına karşın, ünsüz fonemlerin çıkışında akciğerden gelen hava konuşma organlarının herhangi bir yerinde engele rastlar. Bu engeller dudaklar, dişler, alt ve üst damak ve dil ile oluşmaktadır. Ünsüz fonemler çıkış biçimlerine, çıkış yerlerine ve ses tellerinin titreşim durumuna göre Tablo 2.5.'te gösterildiği gibi sınıflandırılmaktadır.

Tablo 2.5. Sessiz harflerin kategorize edilmesi

		b	c	ç	d	f	g	h	j	k	l	m	n	p	r	s	ş	t	v	y	z
Çıkış Biçimi	**Patlamalı**	+			+		+			+				+				+			
	Genizden											+	+								
	Çarpmalı														+						
	Daralmalı										+										
	Sızmalı		+	+		+		+	+							+	+		+	+	+
Çıkış Yeri	**Çift Dudak**	+										+		+							
	Alt dudak-Üst diş					+													+		
	Dil ucu-Diş ardı				+													+			
	Dil ucu-Diş eti												+		+	+					+
	Dil ucu-Ön damak										+										
	Dil-Ön damak		+	+					+								+			+	
	Dil-damak sonu						+			+											
	Ses telleri							+													
Titreşim	**Ötümlü**	+	+		+		+		+		+	+	+		+				+	+	+
	Ötümsüz			+		+		+		+				+		+	+	+			

Parçalarüstü Fonemler (Bürünler)

Fonemlerin en önemli özelliği anlam ayırıcı olmalarıdır. Anlam ayırıcı olma süre, ton, kavşak, durak, vurgu ve ezgi ile sağlanabilmektedir. Bunlar, parçalarüstü sesbirimler veya bürünler olarak tanımlanmaktadır. Türkçe için genelde bürünlerin anlam ayırıcı özelliği bulunmadığı kabul edilmektedir. Fakat Japonca ve Çince gibi vurguya dayalı yapıda olan dillerdebu kavramlar anlam açısından oldukça önemlidir. Bürünler sesli ifade tanıma kapsamında, sesli ifade örüntüsünün yakalanmasına yardımcı olmaktan çok sesli ifadeden anlam çıkarma aşamasında önem taşımaktadır.

Süre, bir sesin söylenirken kısa ya da uzun zaman alması ile ilgilidir. Sesler her dilde farklı sürelerde kullanılmaktadır. Bir dilde uzun söylenen bir fonem başka bir dilde kısa söylenebilir. Türkçe'de süre anlamı değiştirecek bir öğe değildir. Ancak yabancı dillerden Türkçe'ye geçen sözcüklerden söyleniş süresini uzatma anlam değişikliğine yol açabilmektedir. Ayrıca iç ve son seslerde kullanılan /ğ/ ve /h/ fonemlerinin yutulmasıyla uzayan sözcüklerde de anlam değişikliği söz konusudur. Örneğin "düğün" sözcüğü söylenirken süre kısa tutulursa "dün" sözcüğü ile anlam karışıklığı oluşturabilir. Türkçe'de süre seslenme ve buyurma biçimlerinde de ortaya çıkabilmektedir.

Bir sesteki sıklığın yüksek ya da düşük olması *ton* olarak tanımlanmaktadır. Ses sıklığının düşük veya yüksekliğine göre tiz ve pes kavramları ortaya çıkmaktadır. Kişiden kişiye (aynı yaşta ve cinsiyette kişiler arasında bile) farklılık gösterir. Bir hecenin tiz ya da bas söylenmesi ile de alakalıdır. Çince ve Nijerya dilleri gibi bazı dillerde sözcük anlamını ayırmada kullanılan ton, Türkçe'de genellikle tek sözcüklü bildirilerde anlam ayırıcı özelliği bulunmaktadır. Örneğin; "aferin" sözcüğü ile ya beğeni ya da yanlış yapılan bir işe serzeniş dile getirme ton (perde değişimi) ile sağlanır.

Ünsüz ile biten bir kök ünlü ile başlayan ek aldığında hece düzenin değişmesi *kavşak* olarak tanımlanmaktadır. Örneğin "ki-lim", "ki-li-min" sözcüklerinde olduğu gibi. Türkçe'de ünlüyle başlayan sözcüklerde de kavşak bulunur. Örneğin; ulak sözcüğü "ul-ak" yerine "u-lak" şeklinde okunur. Türkçe'de sözcük başında var olan bu durum, Arapça sözcüklerde iç seste ve son seste bulunabilmektedir. *Durak*, bir tümcede anlamda farklılık oluşturabilecek şekilde duraklama yapılmasıdır. Örneğin "Kara, deniz, hava yolları" ve "Karadeniz havayolları" tümceleri arasındaki fark gibi.

Bir sözcükteki herhangi bir hecenin diğer hecelere göre daha baskılı bir şekilde söylenmesine *vurgu* denmektedir. Bazen cümle içinde sözcük vurgusu da söz konusu olmaktadır. Bazı dillerde vurgunun yeri her zaman aynıdır. Örneğin Fransızca'da sözcüklerin son, Fince'de sözcüğün sondan ikinci ve Çek dilinde sözcüğün ilk hecesinde vurgu gerçekleştirilmektedir. Ancak Türkçe, İspanyolca, İtalyanca ve Rusça gibi dillerde vurgunun sözcük içindeki yeri değişebilmektedir. Vurgunun anlam ayırt edici özelliği vardır. Örneğin "varmış" sözcüğü için yapılan vurgu "Epeyce parası varmış" ve "eve varmış" cümlelerinde farklı yerlerde bulunmakta olmasından anlam ayırt edici özellik sağlamaktadır.

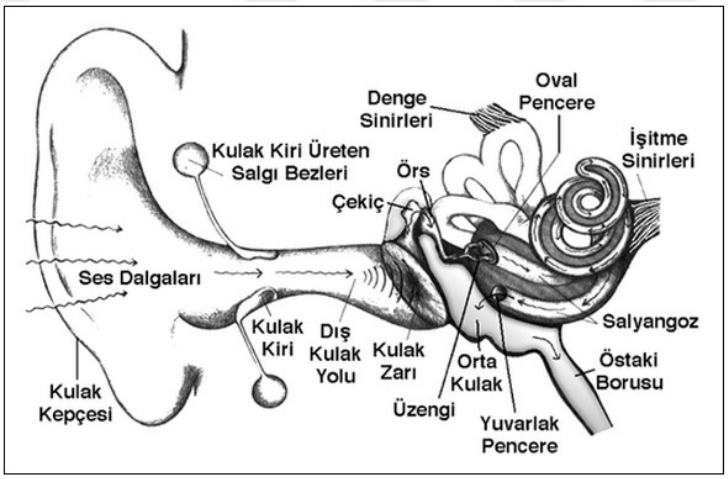
Sesli ifadedeki hece, durak ve vurguya bağlı ortaya çıkan ton değişimlerine *ezgi* denmektedir. Ezgi tamamen konuşmacıya bağlı olduğundan, kesin kurallar koymak zordur. Ezginin söz konusu parametreleri hece, durak ve vurgu değişimleridir. Ruhsal etkenlerle, tümce içinde kimi sözcüklerdeki ton yükselmesi ya da alçaltılması, anlatılmak istenende belirli anlam ayrımı sağlar.

2.4. İnsan'da İşitme

Canlılar arasında iletişim, konuşma ve işitme olarak iki yönlü gerçekleşmektedir. İletişimde önemli bir role sahip olan işitme sisteminde, kulak işitmenin ilk adımı olan organdır. Doğuştan gelişmiş olan insan kulağı en zayıf veya en yüksek sesleri algılayabilmektedir.

2.4.1. Kulak Yapısı ve Görevi

Kulak denildiğinde genel olarak kafanın yan taraflarında bulunan belirgin, kıvrımsal, kulak kanalının en başında gelen organ akla gelmektedir. Bu organ, Şekil 2.3'te gösterildiği gibi hem harici kulak kanalını koruma görevini üstlenir hem de duyulabilir yüksek frekanslarda çok yönlü ses alabilen yapısıyla gelen sesin konumunu belirler.



Şekil 2.3. İnsanda kulak organı.

Dış Kulak

İnsanlarda dış kulak kanalı yaklaşık 2,7 cm uzunluğunda, 0,7 cm çapındadır ve ortalama 1 cm^3 hacmindedir. Bir ucu açık ve diğer ucu kapalı boru şeklinde olan dış kulak kanalı kulak kepçesiyle başlayıp kulak zarı ile sonlanır. Kulak zarı, yüzey alan 0,8 cm^2 olan, nispeten sert içeriye doğru yönelmiş yaklaşık 135 derecelik koni şeklindedir. Dış kulak kanalının bazı frekanslarda oluşan normal titreşim modları da vardır. Dış kulak kanalına giren bir ses, kulak zarına varıncaya kadar ses basın şiddetini 5-10 dB arası arttırmaktadır.

Orta Kulak

Kulak zarının hemen sonrasında ortak kulağın başlangıcı olan ve içinde insan vücudunun en küçük kemikçikleri çekiç, örs ve üzengi'yi bulunduran orta kulak boşluğu yer almaktadır. Bu kemikçiklerin asıl işlevi hava ortamından kulak zarına gelen ses dalgalarını mekanik hale

dönüştürüp iç kulağın sıvı ortamına iletmektir (empedans dönüşümü). Çekiç kemikçiğinin bir ucu kulak zarına yapışık halde dururken diğer ucu örs kemikçiğine bağlanmaktadır. Örs kemikçiği, çekiç kemikçiği ve üzengi kemikçiği arasında bulunmaktadır. Üzengi kemiğinin iç kulakla birleştiği noktaya oval pencere denmektedir ve titreşimler bu noktadan iç kulağa iletilmektedir.

İç Kulak

İç kulakta üç ana bölüm bulunmaktadır. Bunlar; koklea (salyangoz), yarım daire kanalları (labirent) ve giriş kanalı şeklinde tanımlanmaktadır. Koklea insan kulağının işitme kabiliyetini destekler iken giriş ve yarım daire kanalları dengeyi sağlamaktadır. Salyangoz sıvı ile doludur ve taban adı verilen geniş bir uçtan tepe adı verilen dar bir başlığa doğru sivrilen gerçek bir salyangoz şeklindedir. Taban en çok yüksek perdeli seslere (kuş cıvıltıları gibi) tepki verirken, tepe noktası en düşük perdeli seslere (bas davul gibi) tepki vermektedir. Salyangoz iki ince zarla üç tüpe ayrılmaktadır. Bu zarlardan biri üzerinde Korti organının oturduğu elastik bir duvar görevi görmektedir. Korti organında tüylü hücreler adı verilen çok küçük hücreler bulunmaktadır. Bu hücreler o kadar küçüktür ki, salyangozdaki yaklaşık 18.000 hücre bir toplu iğnenin başı boyutunda ölçülmektedir.

Ses Dalgalarının Kulak İçerisinde İlerlemesi

Ses dalgaları iç kulağa ve ardından salyangoz şeklindeki bir organ olan kokleaya girmektedir. Salyangoz'un iç kısmı, oval pencereden gelen titreşimlere tepki olarak hareket eden (dalgalanan) bir sıvı ile doludur. Sıvı hareket ettikçe 25.000 sinir ucu harekete geçmektedir. Bu sinir uçları, titreşimleri, daha sonra sekizinci kraniyal sinir (işitme siniri) boyunca beyne giden elektriksel uyarılara dönüştürmektedir. Beyin organı ise iletilegelen bu sinyali yorumlamakla görevlidir. İç kulakta ayrıca insan denge sistemini yöneten vestibüler organlarda bulunmaktadır.

3. DİJİTAL KONUŞMA İŞLEME TEKNİKLERİ

Dijital konuşma işlemenin çoğu uygulamasındaki ilk adım, akustik dalga biçimini bir sayı dizisine dönüştürmektir. Modern Analog'tan Dijital'e dönüştürücülerin çoğu, çok yüksek bir hızda örnekleme yaparak, önceden belirlenmiş bir bant genişliğini korumak için kesme ayarlı bir dijital düşük geçiş filtresi uygulayarak ve ardından örnekleme hızını istenen örnekleme hızına (iki katına kadar düşebilen) düşürerek çalışır. Dijital filtrenin kesme frekansı keskin kesme olarak tanımlanmaktadır. Bu ayrık zamanlı temsil, çoğu uygulama için başlangıç noktasıdır. Bu noktadan itibaren, diğer temsiller dijital işleme ile elde edilir. Çoğunlukla, bu alternatif temsiller konuşma zincirinin işleyişi hakkındaki bilgileri birleştirmeye dayanmaktadır. Hem konuşma üretimi hem de konuşma algılama sürecinin yönlerini dijital temsil ve işlemeye dahil etmek mümkün olmaktadır. Bu tez çalışmasında dijital konuşma işleme teknikleri olarak ön işlemler, konuşma/konuşmacı tanıma, ses iyileştirme ve konuşma sentezleme konuları açıklanmaktadır.

3.1. Ön İşlemler

Konuşma işleme alanında yapılan çalışmalarda genelde ses sinyalleri üzerinde hiçbir işlem uygulanmadan önce ön işlemler adı altında bir takım iyileştirmeler gerçekleştirilmektedir. Bu işlemlerden en çok kullanılanlar çerçeveleme/çerçeve bloklama (frame blocking), pencereleme (windowing) ve filtreleme işlemleri olarak karşımıza çıkmaktadır.

3.1.1. Çerçeveleme

Durağan olmayan sinyaller söz konusu olduğunda, kısa parçalardaki/dizilerdeki spektral özellikler çok faydalıdır. Bu nedenle, sinyali birden çok aralığa ayrıştırmak, bu tür öznitelik çıkarımına gitmenin yoludur. Bu teknik, çerçeve engelleme veya çerçeveleme olarak bilinir. Çerçeve bloklama veya çerçeveleme, orijinal sinyalin, genellikle çerçeveler olarak adlandırılan X adet bloğa bölünmesinden oluşan temel bir sinyal işleme tekniğidir. Örtüşen çerçeveleme sıklıkla kullanılan bir yöntem olarak karşımıza çıkmaktadır. Çerçevelerin üst üste binmesi, bitişik çerçeveler arasında bilgi kaybının önlenmesine yardımcı olur. Konuşma durağan olmayan bir sinyaldir, sonuç olarak istatistiksel özellikleri zaman içinde sabit değildir. Bu nedenle, spektral özellikleri ve diğer karakteristik özellikleri (örneğin: kısa süreli enerji, MFCC vb.) sinyalin küçük bloklarından çıkarılmalıdır. Her şeyden öte, çerçeve bloklama, verimliliği en üst düzeye çıkardığı için gerçek zamanlı ve sabit işlem yükünü birçok örneğe dağıtmak isteyen sistemlerde sıklıkla kullanılmaktadır.

3.1.2. Pencereleme

Pencereleme (windowing), sinyal işlemede klasik bir yöntem olarak bilinmekte ve giriş sinyalinin geçici parçalara bölünmesini ifade etmektedir. Segmentlerin sınırları daha sonra gerçek dünya sinyaliyle uyumsuz olan süreksizlikler olarak görünmektedir. Segmentlemenin sinyalin istatistiksel özellikleri üzerindeki etkisini azaltmak için zamansal segmentlere pencereleme uygulanmaktadır. Pencereleme fonksiyonları, sınırlarda sıfıra giden yumuşak fonksiyonlar olarak bilinmektedir. Giriş sinyalini bir pencere fonksiyonu ile çarparak, pencereleme fonksiyonu da sınırda sıfıra gider, böylece sınırdaki süreksizlik görünmez olur. Pencereleme bu nedenle sinyali değiştirmekte, ancak bu değişiklik, sinyal istatistikleri üzerindeki etkisi en aza indirilecek şekilde tasarlanmıştır.

3.1.3. Filtreleme

Filtreleme işlemi dijital sinyallerde istenmeyen aralıktaki frekansları elemek için kullanılmaktadır. Bu tez çalışmasında pre-emphasis adı verilen ön vurgu filtresi tüm seslere ön işlemler aşamasında uygulanmıştır. Ön vurgu filtresi ortalama spektral şekli telafi etmek için kullanılan yaygın bir ön işleme aracıdır. Daha yüksek frekansları vurgulayan ön vurgudur. Tipik olarak ön vurgu, bir serbest parametreli bir zaman alanlı FIR filtresi olarak uygulanır; örneğin, 8 kHz veya 12.8 kHz örnekleme hızında konuşma kodlamasında ön vurgu filtresi kullanılır. Ön vurguyu ayarlamanın çok sayıda farklı yolu vardır. İlk olarak, ortalama spektrum azalıyor olsa da, sessiz sürtünmeliler tipik olarak yüksek frekanslarda daha fazla enerjiye sahiptir. Aşırı ön vurgu bu nedenle sürtünmeliler için sorunlara neden olur. Ön vurgu aynı zamanda hem algısal hem de istatistiksel modelleme üzerinde ve doğrusal tahmin modellerinin tahmininde de etkiye sahiptir. Bu nedenle, ön vurgunun en iyi miktarı büyük ölçüde uygulamaya ve uygulama ayrıntılarına bağlıdır.

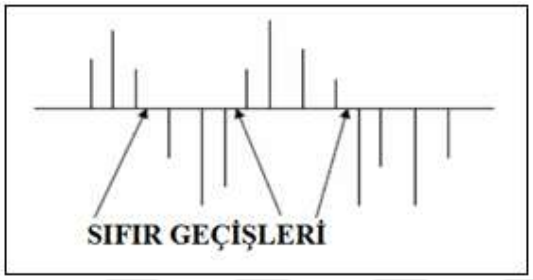
3.2. Özellik Çıkarma

Dijital ses sinyallerini tanımak (recognition) veya sınıflandırmak (classification) için kullanılan ilk aşamalardan biri de özellik çıkarma aşamasıdır. Bu kısımda sıfır geçiş oranı (ZCR), enerji, Fourier dönüşümü ve MFCC özellik çıkarma yöntemleri açıklanmıştır. Tez çalışmasında ise enerji, sıfır geçiş oranı (ZCR) ve MFCC özelliklerinden faydalanılmıştır.

3.2.1. Sıfır Geçiş Oranı

Sıfır geçiş oranı (ZCR), bir sinyalin pozitiften sıfıra, negatife veya negatiften sıfıra, pozitife değişme hızıdır. Değeri, hem konuşma tanımada hem de müzik bilgisi almada yaygın olarak kullanılmaktadır ve vurmalı sesleri sınıflandırmak için önemli bir özelliktir olarak görülmektedir. ZCR, bir sinyalin gürültülülüğünün (ne kadar gürültülü olup olmadığının) bir ölçüsü olarak

yorumlanabilmektedir. Örneğin, genellikle gürültülü sinyaller durumunda daha yüksek değerler sergiler. Ayrıca, bir sinyalin spektral özelliklerini oldukça kaba bir şekilde yansıttığı da bilinmektedir. ZCR'nin bu tür özellikleri, hesaplanmasının kolay olması gerçeğiyle birlikte, konuşma-müzik ayrımı, konuşma algılama ve müzik türü sınıflandırması dahil olmak üzere çok sayıda uygulama tarafından benimsenmesine yol açmıştır. Şekil 3.1'de örnek ayrık bir ses sinyalinin sıfır geçişleri gösterilmiştir.



Şekil 3.1. Ayrık zamanlı bir sinyaldeki sıfır geçişleri

3.2.2. Enerji

Bir sinyalin enerjisi, sinyalin toplam büyüklüğüne karşılık gelmektedir. Ses sinyalleri için bu, kabaca sinyalin ne kadar yüksek sesli olduğu anlamına gelmektedir. Sessiz bölümlerin genliği genellikle sesli bölümlerin genliğinden çok daha düşüktür. Konuşma sinyalinin kısa süreli enerjisi, bu genlik değişimlerini yansıtan uygun gösterim sağlamaktadır. Bir konuşma çerçevesinin enerjisi, çerçevenin etkinliği için bir parça bilgi vermekte ve karar vermede enerji eşik değeri kullanılmaktadır. Zaman domeninde geliştirilen VAD algoritmaları için, bir çerçevedeki bir konuşmanın genliği, çerçeveleri konuşma veya konuşma olmayan olarak sınıflandırmak için önemli bir parametre olarak görülmektedir.

3.2.3. Fourier Dönüşümü

Bir ses sinyali, ortamda bir rahatsızlık (basınç değişimi) olarak birlikte hareket eden çoklu 'tek frekanslı ses dalgalarından' oluşan karmaşık bir sinyal olarak görülmektedir. Ses kaydedildiğinde, yalnızca bu çoklu dalgaların sonuçtaki genlikleri yakalnmış olmaktadır. Fourier Dönüşümü, bir sinyali bileşen frekanslarına ayrıştırabilen matematiksel bir kavram olarak tanımlanmaktadır. Fourier dönüşümü sadece sinyalde bulunan frekansları vermez, aynı zamanda sinyalde bulunan her bir frekansın büyüklüğünü de verir. Ters Fourier Dönüşümü, Fourier Dönüşümünün tam tersidir. Belirli bir sinyalin frekans alanı temsilini girdi olarak alır ve orijinal sinyali matematiksel olarak sentezler. N örnekli bir sette Hızlı Fourier Dönüşümü Denklem 3.1'deki gibi tanımlanmaktadır.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N} \qquad k = 0,1,\dots,N-1 \tag{3.1}$$

3.2.4. Mel Frekansı Kepstrum Katsayıları

Mel Frekansı Kepstrum Katsayıları (MFCC), tanıma görevlerinde kullanılmak üzere konuşma sinyallerinden çıkarılan en yaygın özelliklerden biridir. Konuşmanın kaynak filtre modelinde, MFCC'nin filtreyi (ses yolu) temsil ettiği anlaşılmaktadır. Ses yolunun frekans yanıtı nispeten yumuşaktır, oysa sesli konuşmanın kaynağı bir dürtü dizisi olarak modellenebilmektedir. Ses yolunun bir konuşma bölümünün spektral zarfı ile tahmin edilebilmektedir. MFCC'nin motive edici fikri, kokleanın anlaşılmasına dayalı olarak ses yolu (düzleştirilmiş spektrum) hakkındaki bilgileri az sayıda katsayıya sıkıştırmaktır. MFCC hesaplamak için gerekli adımlar şu şekilde özetlenebilir;

- İlk aşamada, ses sinyalimizi 8kHz veya 16kHz örnekleme frekansı ile analogdan dijital formata dönüştürülmektedir.
- Daha sonra Bölüm 3.1.3'te de bahsedilen ön vurgu filtresi uygulanmaktadır.
- MFCC tekniği, konuşmadaki fonemlerin algılanması için kullanılabilecek ses sinyalinin özelliklerini geliştirmeyi amaçladığından filtrelemeden sonra pencereleme uygulanmaktadır.
- Ses sinyallerini frekans alanında analiz etmek, zaman alanında analiz etmekten daha kolay olduğundan Fourier dönüşümü uygulanarak sinyal zaman alanından frekans alanına dönüştürülmektedir.
- Kulağımızın sesi algılama şekli, makinelerin sesi algılama şeklinden farklıdır. Kulaklarımız, daha yüksek bir frekanstan daha düşük bir frekansa göe daha yüksek çözünürlüğe sahiptir. Yani 200 Hz ve 300 Hz'de ses duyuyorsak, aralarında 100 Hz fark olmasına rağmen, 1500 Hz ve 1600 Hz'deki sesleri karşılaştırdığımızda kolayca ayırt edebiliriz. Oysa makine için çözünürlük tüm frekanslarda aynıdır. Özellik çıkarma aşamasında insan işitme özelliğinin modellenmesinin modelin performansını artıracağı fark edilmiştir.

3.3. Konuşma-Konuşmacı Tanıma

Konuşma/konuşmacı tanıma, konuşma sinyali üzerinde yapılabilecek işlemlerden biridir. Konuşma/konuşmacı tanıma, Bölüm 3.1'de bahsedilen ön işlemler, Bölüm 3.2'de bahsedilen özellik çıkarma ve Bölüm 3.2'de değinilecek olan tanıma aşaması olarak 3 ana kısımdan oluşmaktadır.

3.3.1. Dinamik Zaman Eşleştirme

Dinamik zaman eşleştirme (Dynamic Time Warping-DTW), sesten gelen iki veri arasındaki en uygun çarpıtma yolunu hesaplayan bir algortima olarak tanımlanmaktadır. Böylece çıktı, yol çarpıtma değerleri ve iki veri arasındaki mesafe olur. Eşleştirme yolu, iki modelin karşılaştırması arasındaki mesafedir, üretilen çözgü yolu ne kadar küçük olursa, iki modelin aynı olduğu söylenebilmektedir. Aynı kullanıcı tarafından aynı kelimeden iki kelime farklı sürelere sahip olabilir. Örneğin, iki, iki veya iki ile telaffuz edilebilir. DTW, kelimeleri doğru bir şekilde hizalayarak ve iki kelime arasındaki minimum mesafeyi hesaplayarak bu sorunu çözmektedir. Konuşma hizalamasının farklı zamanlaması, konuşma tanımada mesafe ölçümü için temel bir problemdir. Küçük vardiyalar yanlış tanımlamaya neden olur. DTW, zaman hizalama problemlerini çözmek için etkili bir yöntemdir. Bu nedenle, bu algoritmanın bir örüntünün benzerliğini ölçmede (kalıp/şablon eşleştirme) kullanılması daha gerçekçidir. İşlenen veriler her zaman zaman dilimindedir, bu nedenle elimizdeki veri dizisinin zamana göre değiştiği kabul edilir.

3.3.2. Örüntü Eşleştirme

Örüntü tanıma, bir makine öğrenimi algoritması kullanarak örüntüleri tanıma işlemidir. Örüntü tanıma, halihazırda elde edilmiş bilgilere veya örüntülerden ve/veya bunların temsilinden çıkarılan istatistiksel bilgilere dayalı olarak verilerin sınıflandırılması olarak tanımlanabilir. Örüntü tanımanın önemli yönlerinden biri uygulama potansiyelidir. Konuşma işleme alanında örüntü tanıma işlemi gerçekleştirebilmek için büyük miktarda verinin sisteme (modele) kaydedilmesi gerekmektedir. Dijital konuşma işleme'de örüntü tanıma denetimli veya denetimsiz öğrenme yoluyla gerçekleştirilebilmektedir. Denetimsiz öğrenmede Destek Vektör Makinaları (Support Vector Machines-SVM), K En Yakın Komşuluk (K Nearest Neighbours-KNN) gibi klasik makine öğrenme algoritmaları kullanılıyor iken, denetimli öğrenmede ise günümüzde yapay sinir ağlarının yaygın kullanımı mevcuttur.

3.3.3. Vektör Niceleme

Vektör Niceleme (Vector Quantization-VQ), konuşma kodlama, görüntü kodlama, konuşma tanıma ve konuşma sentezi ve konuşmacı tanımada kullanılan temel ve en başarılı tekniktir. Bu teknikler ilk olarak, büyük vektör uzayının o uzayda sınırlı sayıda bölgeye eşlendiği konuşmanın analizinde uygulanır. VQ teknikleri, ayrık veya yarı sürekli HMM tabanlı konuşma tanıma sistemi geliştirmek için yaygın olarak kullanılmaktadır. VQ'de, sıralı bir sinyal örnekleri veya parametre seti, giriş vektörünü önceden tanımlanmış bir kod çizelgesindeki benzer bir model veya kod vektörü (kod sözcüğü) ile eşleştirerek verimli bir şekilde kodlanabilir. VQ teknikleri, çeşitli disiplinlerde veri kümeleme yöntemleri olarak da bilinir. Birçok uygulamada yaygın olarak kullanılan

denetimsiz bir öğrenme prosedürüdür. Veri kümeleme yöntemleri, sert ve yumuşak kümeleme yöntemleri olarak sınıflandırılır. Bunlar, Bregman sapmaları olarak bilinen geniş bir distorsiyon fonksiyonları sınıfına dayanan merkez tabanlı parametrik kümeleme teknikleri olarak bilinmektedir.

3.3.4. K En Yakın Komşular

K en yakın komşular (K Nearest Neighbor-KNN veya k-NN) olarak da bilinen k-en yakın komşu algoritması, bireysel bir veri noktasının gruplandırılması hakkında sınıflandırmalar veya tahminler yapmak için yakınlığı kullanan, parametrik olmayan, denetimli bir öğrenme sınırlandırıcısı olarak tanımlanmaktadır. Hem regresyon hem de sınıflandırma problemleri için kullanılabilse de, tipik olarak benzer noktaların birbirine yakın bulunabileceği varsayımıyla çalışan bir sınıflandırma algoritması olarak faydalanılır. Sınıflandırma problemlerinde, çoğunluk oyu temel alınarak bir sınıf etiketi atanır. Örneğin belirli bir veri noktası etrafında en sık temsil edilen etiket kullanılır. Bu teknik olarak "çoğunluk oyu" olarak kabul edilirken, literatürde "çoğunluk oyu" terimi daha yaygın olarak kullanılmaktadır. Bu terminolojiler arasındaki fark, "çoğunluk oylaması"nın teknik olarak %50'den fazla çoğunluk gerektirmesidir ve bu, yalnızca iki kategori olduğunda işe yarar. Birden fazla dersiniz olduğunda örneğin dört kategori, bir sınıf hakkında bir sonuca varmak için oyların %50'sine ihtiyacınız olması gerekmez; %25'in üzerinde oyla bir sınıf etiketi atanabilmektedir.

3.3.5. Gizli Markov Modeli

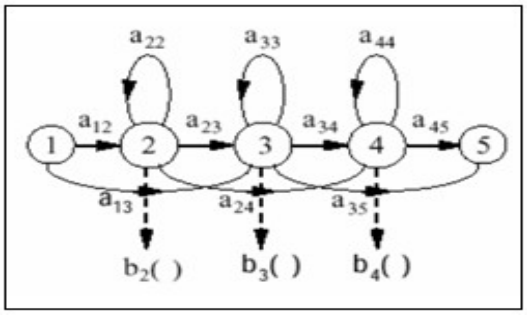
Gizli bir Markov modeli (HMM), modellenen sistemin gözlemlenemeyen ("gizli") durumlara sahip bir Markov süreci olduğunun varsayıldığı istatistiksel bir Markov modelidir. Markov ve Gizli Markov modelleri, zaman içindeki gözlemlerin "dizisi" olarak temsil edilebilecek verileri işlemek için tasarlanmıştır. Gizli Markov modelleri, gözlemlenen verilerin birkaç (gizli) dahili durumdan biri tarafından üretilen bir dizi çıktı olarak modellendiği olasılık çerçeveleridir. İlk olarak konuşma tanımada kullanıldıktan sonra biyolojik dizilerin analizine de başarıyla uygulanabilmektedir. Gizli markov modelindeki durum geçiş matrisi Denklem 3.2'deki gibidir.

$$A = [a_{ij}],$$
$$a_{ij} = P(q = j / q_{t-1} = i), \quad i, j = 1, \ldots, X \tag{3.2}$$

Sistemin başlangıç aşaması q_0 olarak belirlenmektedir. İlk durumdan sonra ortaya çıkan herhangi bir aşama dizisi $q = (q_0, q_1, \ldots, q_r)$'nin bir markov süreci ile üretilebilme ihtimali Denklem 3.3'te gösterilmiştir.

$$P(q \mid A) = \pi_{q_0} a_{q_0 q_1}, a_{q_1 q_2}, \dots, a_{q_{T-1} q_r} \quad (3.3)$$

Durum dizisi q hızlı bir şekilde tespit edilemiyorsa markov süreci gizli olarak tanımlanmaktadır. Bu durumda durumların dolaylı olarak gözlemlenebilmesi söz konusudur. Bir gözlem ve bir durum arasında her ne kadar birebir bir gereklilik olmazsa bile her bir durumun Şekil 3.2'deki gibi belirli bir olasılıkla gözlenmesi gerekmektedir.

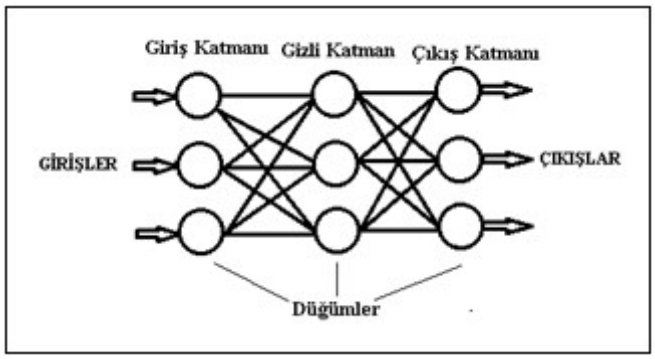


Şekil 3.2. 5 şarta dayanan (sağa doğru yönlü) saklı markov modeli

Gizli Markov modelleri (HMM'ler) sıralı modellerdir. Yani, sözcükler gibi bir girdi dizisi verildiğinde, bir HMM aynı uzunluktaki bir dizi çıktıyı hesaplayacaktır. Bir HMM modeli, düğümlerin etiketler üzerindeki olasılık dağılımları olduğu ve kenarların bir düğümden diğerine geçiş olasılığını verdiği bir grafiktir. Bunlar birlikte, giriş dizisi verilen bir etiket dizisinin olasılığını hesaplamak için kullanılmaktadır.

3.3.6. Yapay Sinir Ağları

Yapay sinir ağları, bir insanın doğduğu andan itibaren öğrenme şeklini çalışma prensibi olarak benimsemiş sistemler olarak bilinmektedir. Nasıl ki bir insan doğduğundan itibaren nesneleri olayları görüp öğreniyorsa, yapay sinir ağı sistemleri de yazılımsal olarak sisteme yüklenen gözlemlerden (imaj, ses dosyası, metin verisi vs.) eğitim gerçekleştirmektedir. Şekil 3.3'te YSA'nın genel modeli verilmiştir.



Şekil 3.3. Yapay Sinir Ağı'nın en genel yapısı

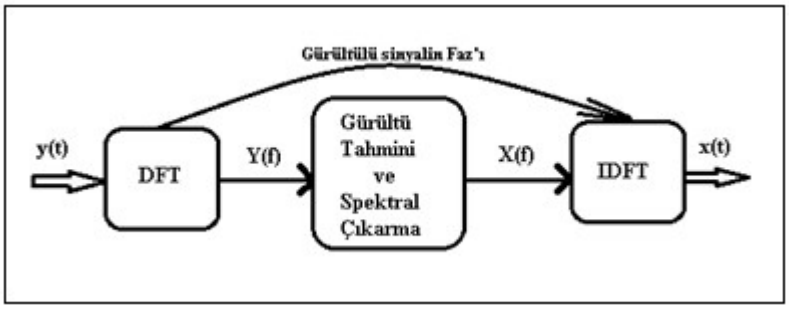
Geleneksel olarak sinir ağları, nöronlar veya devre olarak adlandırılır. Sinir ağları terimi, yapay nöronlardan veya düğümlerden oluşan Yapay Sinir Ağı olarak anılmaktadır. Bir girdi alan, bu girdiye göre durumu değiştiren ve bir çıktı üreten yapay nöronlar olarak bilinen temel elementlerden oluşan bir ağdır. Birbirine bağlı bir doğal veya yapay nöron grubu, iletişime bağlantıcı yaklaşıma dayalı bilgi işleme için matematiksel bir model kullanır. "Yapay sinir ağı" terimi, her bir sistemin farklı katmanlarındaki nöronlar arasındaki ara bağlantıları ifade eder. Matematiksel olarak bir nöronun ağ işlevi f(x), başka işlevlerin bir bileşimi olarak tanımlanabilecek olan diğer gi(x) işlevlerinin bir bileşimi olarak tanımlanır. Genellikle ileri besleme olarak adlandırılmaktadır. Döngüleri olan ağlara tekrarlayan sinir ağları denir.

3.4. Ses İyileştirme

3.4.1. Spektral Çıkarma Algoritması

Spektral çıkarma, tarihsel olarak, tek kanallı konuşmanın geliştirilmesi için önerilen ilk algoritmalardan biridir. Bu yöntemde, gürültü spektrumu konuşma duraklamaları sırasında tahmin edilir ve net konuşmayı tahmin etmek için gürültülü konuşma spektrumundan çıkarılır. Bu aynı zamanda gürültülü konuşma spektrumunun bir kazanç fonksiyonu ile çarpılması ve daha sonra gürültülü konuşmanın fazı ile birleştirilmesiyle elde edilir. Bu yöntemin dezavantajı, artık gürültü adı verilen işleme bozulmalarının varlığıdır. Son yıllarda dezavantajı gidermek için yöntemin bir dizi varyasyonu geliştirilmiştir. Bu değişkenler, bir spektral çıkarma tipi algoritma ailesi oluşturmaktadır. SÇA algoritmasında Şekil 3.4'te gösterildiği gibi sadece normal ve ters ayrık Fourier dönüşümleri kullanılmaktadır. Tek kanallı konuşma geliştirmenin spektral çıkarma yöntemi, ek gürültüyü azaltmak için en yaygın kullanılan geleneksel yöntemdir. Kalan geniş bant gürültüsü ve müzikal gürültü olarak adlandırılan dar bant ton gürültüsü gibi tipik olarak spektral çıkarma ile ilişkili problemlerin üstesinden gelmek için birçok iyileştirme önerilmiştir. Diğer spektral çıkarma varyantları arasında spektral aşırı çıkarma, çok bantlı spektral çıkarma, Wiener

filtreleme, yinelemeli spektral çıkarma ve algısal özelliklere dayalı spektral çıkarma yer almaktadır. Spektral çıkarma tipi algoritma, spektral aşırı çıkarma, çok bantlı spektral çıkarma, Wiener filtreleme, yinelemeli spektral çıkarma ve algısal özelliklere dayalı spektral çıkarma gibi spektral çıkarma yönteminin farklı varyantlarının ailesi olarak görülmektedir. Bu nedenle, spektral çıkarma tipi algoritmaların ilkesi, tahmini gürültüyü gürültülü konuşma spektrumundan çıkararak veya gürültülü spektrumu kazanç fonksiyonlarıyla çarparak konuşmanın kısa süreli spektral büyüklüğünü tahmin etme olarak belirlenmiştir.



Şekil 3.4. Spektral Çıkarma Algoritmasının Yapısı.

3.4.2. Wiener Filtreleme

Sinyal işlemede, Wiener filtresi, bilinen durağan sinyal ve gürültü spektrumları ve ek gürültü varsayılarak, gözlemlenen gürültülü bir işlemin doğrusal zamanla değişmeyen (Linear Time Invariant-LTI) filtrelenmesi yoluyla istenen veya hedef rasgele işlemin bir tahminini üretmek için kullanılan bir filtredir. Wiener filtresi, tahmini rasgele işlem ile istenen işlem arasındaki ortalama kare hatasını (Mean Square Erro-MSE) en aza indirmektedir. Wiener filtresinin amacı, ilgili bir sinyali girdi olarak kullanarak bilinmeyen bir sinyalin istatistiksel bir tahminini hesaplamak ve tahmini bir çıktı olarak üretmek için bilinen sinyali filtrelemektir. Örneğin, bilinen sinyal, ilave gürültü ile bozulmuş, bilinmeyen bir ilgilenilen sinyalden oluşabilir. Wiener filtresi, ilgilenilen sinyalin bir tahminini sağlamak için bozuk sinyalden gelen gürültüyü filtrelemek için kullanılabilmektedir. Wiener filtresi, istatistiksel bir yaklaşıma dayanmaktadır. Denklem 3.4'te Wiener filtresinin input-output formülizasyonu verilmiştir.

$$x(m) = \sum_{k=0}^{P-1} w_k y(m-k) \tag{3.4}$$

Verilen eşitlikte;

m : süre endeksi,

$y^T = [y(m), \ldots, y(m\text{-}P\text{-}1)]$: analiz edilecek sinyali,

$x(m)$: Wiener filtre sonucunu,

$w^T = [w_0, w_1, \ldots, w_{P\text{-}1}]$: filtrenin katsayı değerlerinin vektör gösterimini ifade etmektedir.

3.4.3. İstatistiksel Model Tabanlı Yöntemler

Ses iyileştirme konusunda istatistiksel model tabanlı yöntemler kendi içerisinde kısa dönemli (short-term) modeller, uzun dönemli (long-term) modeller ve model-parametre tahmini şeklinde ayrılmaktadır. Konuşmanın istatistiksel modellemesi, kodlama, geliştirme ve tanıma dahil olmak üzere neredeyse tüm konuşma işleme uygulamaları için ilgili bir konudur. İstatistiksel bir model, parametreleştirilmiş bir olasılık yoğunluk fonksiyonları (Probability Density Function-PDF) seti olarak görülebilir. Parametreler belirlendikten sonra model, konuşma sinyalinin, sinyalin stokastik davranışını açıklayan bir PDF'sini sağlamaktadır. Bir olasılık yoğunluk fonksiyonu, istatistiksel bir çerçevede en uygun niceleyicileri, tahmin edicileri ve tanıyıcılarını tasarlamak için kullanışlı görülmektedir. Konuşma sinyalini 20-30 ms çerçeveler içinde yerel olarak durağan olan stokastik bir süreç olarak modellemek daha yaygındır. Konuşma sinyali daha sonra çerçeve bazında işlenir. Konuşma işlemedeki istatistiksel modeller, çalışma kapsamlarına bağlı olarak kabaca kısa vadeli ve uzun vadeli sinyal modelleri olmak üzere iki sınıfa ayrılabilir. Kısa vadeli bir model, belirli bir çerçeve için vektörün istatistiklerini tanımlamaktadır. İstatistikler, çerçeveler üzerinde değişirken, her çerçeve için kısa vadeli bir modelin parametreleri belirlenmektedir. Öte yandan, çoklu sinyal çerçeveleri üzerinden istatistikleri açıklayan bir model, uzun vadeli bir model olarak adlandırılmaktadır.

3.4.4. Altuzay Algoritmaları

Sinyal işlemede, sinyal alt uzay yöntemleri, boyut azaltma ve gürültü azaltma için ampirik doğrusal yöntemlerdir. Bu yaklaşımlar son zamanlarda konuşma işleme, konuşma modelleme ve konuşma sınıflandırma araştırması bağlamında büyük önem kazanmıştır. Sinyal alt uzayı, MUSIC algoritması kullanılarak radyo yön bulmada da kullanılmaktadır. Alt uzay tabanlı yöntemler, gürültülü verilerin iki veya daha fazla bileşene dağıtılabileceği varsayımına dayanmaktadır. ürme algoritmaları (blok algoritmalar, özyinelemeli algoritmalar) gibi algoritmalardan oluşmaktadır. Yüksek çözünürlüklü yöntemler veya süper çözünürlüklü yöntemler olarak da bilinen alt uzay yöntemleri, korelasyon matrisinin bir öz analizine veya öz ayrışmasına dayalı bir sinyal için frekans bileşeni tahminleri üretir. Örnekler, çoklu sinyal sınıflandırması yöntemi veya özvektör (eigenvector) yöntemidir. Bu yöntemler en çok çizgi spektrumları, yani sinüzoidal sinyallerin spektrumları, için uygundur ve özellikle sinyal gürültü oranları düşük olduğunda, gürültüye

gömülmüş sinüsoidallerin tespitinde etkili olmaktadır. Sinyal alt uzayı gürültü azaltma yöntemi, Wiener filtre yöntemleriyle karşılaştırıldığında iki temel fark ortaya çıkmaktadır;

- Wiener filtrelemede kullanılan temel sinyaller genellikle, bir sinyalin Fourier dönüşümü ile ayrıştırılabileceği harmonik sinüs dalgalarıdır. Buna karşılık, sinyal alt uzayını oluşturmak için kullanılan temel sinyaller ampirik olarak tanımlanır ve örneğin cıvıltılar veya saf sinüsoidler yerine belirli tetikleyici olaylardan sonra geçici olayların belirli karakteristik şekilleri olabilmektedir.
- Wiener filtresi, sinyalin hakim olduğu lineer bileşenler ile gürültünün hakim olduğu lineer bileşenler arasında sorunsuz bir şekilde derecelendirme yapmaktadır. Gürültü bileşenleri filtrelenir, ancak tam olarak değil; sinyal bileşenleri korunur, ancak tam olarak değil; kısmen kabul edilmiş bir geçiş alanı (zone) vardır. Buna karşılık, sinyal alt uzayı yaklaşımı keskin bir kesmeyi temsil etmektedir: ortogonal bir bileşen ya sinyal alt uzayında yer alır, bu durumda %100 kabul edilir ya da buna dik olursa, bu durumda %100 reddedilir. Sinyali çok daha kısa bir vektöre soyutlayarak, boyutsallıktaki bu azalma, yöntemin özellikle çok sık kullanılan bir özellik hale gelmesini sağlamaktadır.

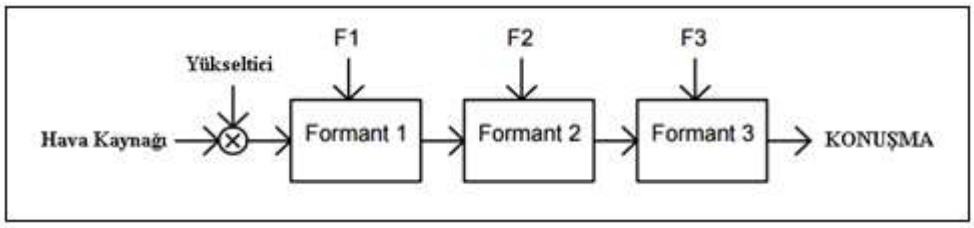
3.4.5. Gürültü Tahmin Algoritmaları

Gürültü tahmin algoritmaları, konuşma işleme alanında önemli rol oynamaktadır. Otomatik konuşmacı tanıma sistemi için konuşma işleme, insan-makine iletişimi, ses tanıma sistemleri, konuşma kodlayıcılar, işitme cihazları, video konferans ve birçok uygulama gürültü tahmin algoritmaları kullanmaktadır. Tüm bu sistemler gerçek dünya sistemleri olarak görüldüğünden bu sistemler için mevcut girdi yalnızca gürültülü konuşma sinyali olarak karşımıza çıkmaktadır. Gürültü tahmin algoritmaları ile ses iyileştirme yapabilmek için gürültülü konuşma sinyalinden gürültü bileşeninin çıkartılması gereklidir. Çoğu ses iyileştirme algoritmasında, bir gürültü spektrumu tahmininin mevcut olduğu varsayılır. Gürültü tahmini kritik bir kısımdır ve ses iyileştirme algoritmaları için önem arz etmektedir. Gürültü tahmininin düşük gerçekleştirilmesi ses iyileştirmenin başarız yapılmasına, yüksek gerçekleştirilmesi ise atılmak istenmeyen örneğin konuşma gibi elemanların da duyulabilirlik kalitesinin düşmesine yol açmaktadır. Üç sınıf gürültü tahmin algoritması vardır. Bunlar; Minimal İzleme Algoritmaları, Zaman Özyinelemeli Algoritmalar ve Histogram Tabanlı Algoritmalardır. İlk önce sinyal, tipik olarak 20-30 milisaniye olan kısa üst üste binen (örtüşmeli) çerçevelerden hesaplanan kısa süreli spektrumlar kullanılarak analiz edilmektedir. Ardından, gürültü spektrumunun hesaplanmasında analiz segmenti adı verilen birkaç ardışık çerçeve kullanılmaktadır. Bu bölümün tipik zaman aralığı 400 milisaniye ile 1 saniye arasında değişebilmektedir. Gürültü tahmin algoritmaları, analiz bölümünün konuşma duraklamalarını ve düşük enerjili sinyal bölümlerini içermek için yeterince uzun olduğu ve analiz bölümünde bulunan gürültünün konuşmadan daha durağan olduğu varsayımına dayanmaktadır.

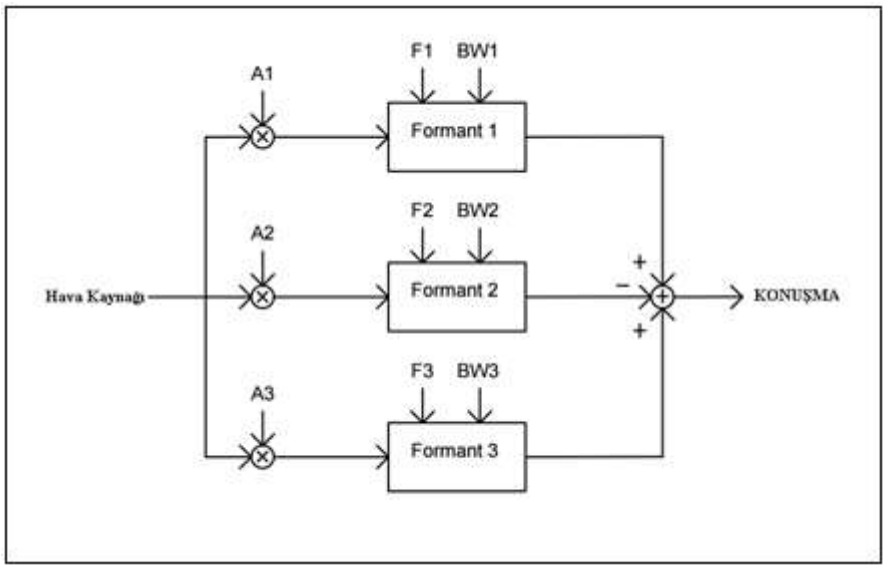
3.5. Konuşma Sentezleme

3.5.1. Formant Sentezleme

Konuşma Formantlar, spektrumda yüksek derecede enerjiye sahip frekans tepe noktaları olarak tanımlanmaktadır. Özellikle sesli harflerde belirgin görülmektedir. Her biçim, ses yolundaki bir rezonansa karşılık gelir (kabaca konuşursak, spektrumun her 1000 Hz'de bir biçimi vardır). Formantlar filtreler olarak kabul edilebilmektedir. Ses boşluklarının ortak rezonans etkisi nedeniyle, kaynak seste sesin filtrelenmesi gerçekleşir, bazı frekans bileşenleri (bazı harmonikler) güçlendirilir ve diğerleri zayıflatılmaktadır. Konuşma bilimi ve fonetikte, bir formant, insan ses yolunun akustik rezonansından kaynaklanan geniş spektral maksimum olarak geçmektedir. Akustikte, bir formant genellikle spektrumda geniş bir tepe veya yerel maksimum olarak tanımlanır. Harmonik sesler için, bu tanımla, biçimlendirme frekansı bazen bir rezonansla en çok artırılan harmoniğin frekansı olarak alınır. Bu iki tanım arasındaki fark, "formantların" bir sesin üretim mekanizmalarını mı yoksa üretilen sesin kendisini mi karakterize ettiği ile ilişkilidir. Uygulamada, bir spektral tepenin frekansı, şans eseri harmoniklerin rezonans frekansıyla hizalandığı durumlar dışında, ilişkili rezonans frekansından biraz farklıdır. Vokal yollar rezonatörlerin Şekil 3.5'teki gibi ardı sıra halinde (cascade) veya Şekil 3.6'daki gibi alt alta olarak bağlanmasıyla modellenebilmektedir. Ses üretici, formantları modelleyen rezonatörlerin yanında ses yolunun yapısı, dudak salınımlarını modelleyen filtreleri ve burundan çıkan sesleri modelleyebilmek için anti-rezonatörleri de barındırmaktadır.



Şekil 3.5. Kaskat formant sentezleme yapısı



Şekil 3.6. Alt alta formant sentezleme yapısı

3.5.2. İfadesel Sentezleme

İfadesel konuşma sentezlemede metinle birlikte istenen ifade de metin işleme aşamasına ek bir girdi olarak kullanılmaktadır. Girilen metin, Tarafsız Konuşma Sentezinde (Neutral Speech Synthesis-NSS) olduğu gibi soyut dilsel temsile dönüştürülmektedir. Ek olarak, nötr konuşmanın sentezinden önce veya sonra ifade edici bilgi de dahil edilmektedir. İlk durumda, ifade bilgisi dil bilgisi ile birlikte kodlanır ve konuşma, dil ve ifade bilgisi kullanılarak metinden sentezlenmektedir. Daha sonraki durumda, konuşma başlangıçta herhangi bir ifade olmadan yani nötr konuşma sentezlenip uygun bir ses dönüştürme tekniği kullanılarak istenen ifade eklenmektedir. Duygu, insan-bilgisayar etkileşimlerinin performansında temel bir unsur olarak kabul edilmektedir. Konuşma sentezleme işlemi sonucunda oluşan konuşmanın etkileyici ve gerçekçi olması isteniyorsa, ince ve karmaşık duygusal durumları yansıtan duygusal konuşma oluşturmak önem arz etmektedir. Sentezlenmiş konuşmanın ifade gücünü geliştirmek için, anlamlı konuşma sentez modeli, düşük seviyeli konuşma özelliklerinden bir sözcenin yüksek seviyeli duygu durumlarını üretmeyi amaçlamaktadır. Duygu üreten görevler için perde, enerji ve süre ile ilgili özelliklerin yanı sıra konuşma hızı da dahil olmak üzere birçok akustik özellik araştırılmıştır. Bu özellikler çeşitli duygu durumlarında oldukça farklıdır. Örneğin, öfkeli konuşma buna bağlı olarak yüksek ve hızlıdır ve yüksek bir ortalama perde, geniş perde aralığı, güçlü yüksek frekans enerjisi ve yüksek konuşma hızı ile ifade edilmektedir. Buna karşılık, hüzünlü konuşma yumuşak ve yavaştır ve daha düşük bir ortalama perde, daha dar perde aralığı, az yüksek frekans enerjisi ve düşük konuşma hızı ile ifade edilmektedir.

3.5.3. Bitiştirerek Sentezleme

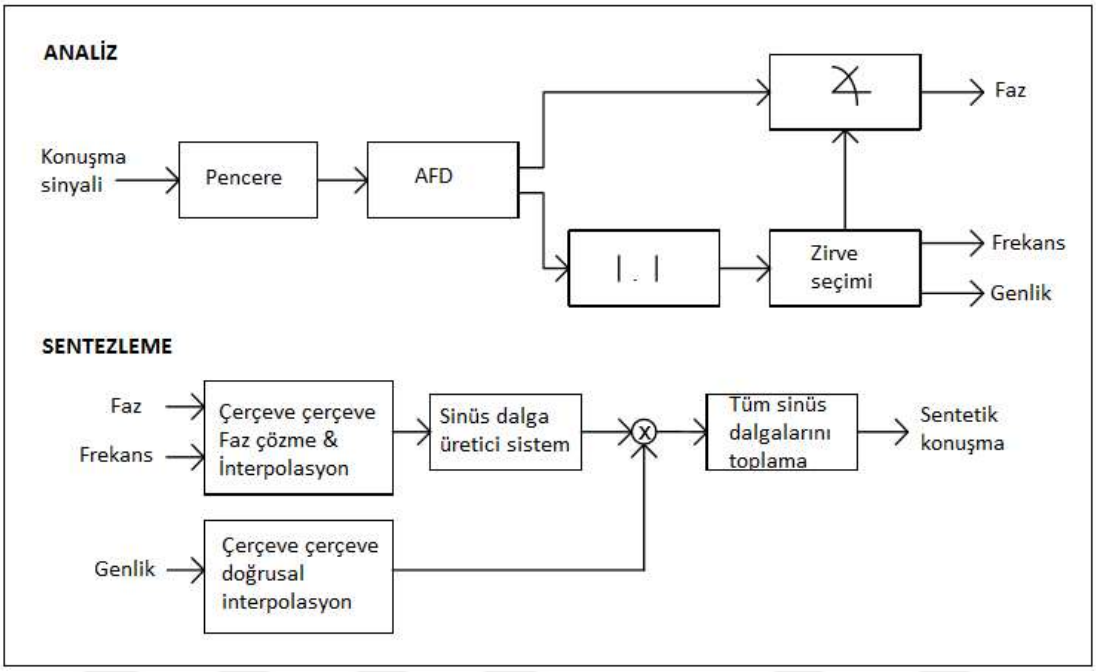
Bitiştirerek sentezleme, kaydedilen sesin kısa örneklerini (birimler olarak adlandırılan) bitiştirerek sesleri sentezlemek için kullanılan bir tekniktir. Birimlerin süresi kesin olarak tanımlanmamıştır ve uygulamaya göre kabaca 10 milisaniye ile 1 saniye arasında değişebilmektedir. Diğer dizilerin kayıtlarından oluşturulan bir veri tabanından (genellikle korpus olarak adlandırılır) kullanıcı tarafından belirlenen ses dizilerini oluşturmak için konuşma sentezinde ve müzik sesi sentezinde kullanılmaktadır. Granüler sentezin aksine, birleştirmeli sentez, belirtilen kritere en iyi uyan birimleri belirlemek için kaynak sesin analiziyle yürütülmektedir. Birim seçmeli konuşma sentezi olarak da bilinen bitiştirerek konuşma sentezi (Concetenative Speech Synthesis-CSS), istatistiksel parametrik konuşma sentezi ile birlikte iki temel modern konuşma sentezi tekniğinden biridir. Adından da anlaşılacağı gibi CSS, anlaşılır yüksek kaliteli konuşma oluşturmak için önceden kaydedilmiş konuşma bölümlerinin birleştirilmesine dayanmaktadır. Bu yaklaşımın avantajı, sistem iyi tasarlanmış olduğu ve geliştirilmesi için uygun konuşma verileri mevcut olduğu sürece üretilen konuşmanın son derece yüksek doğal olmasıdır. Dezavantajı ise, kullanılan tüm konuşma bölümlerinin önceden kaydedilmesi gerektiğinden, depolama ihtiyacının fazla olması olarak görülmektedir. Bir birim veri tabanı mevcut olduğunda, Bitiştirerek sentezleme (CSS)'nin temel adımları şu şekilde sıralanmaktadır: Önce giriş metninin, ses perdesi, süre ve güç gibi ek prosodik özelliklerle birlikte sentezlenecek fonem dizisini içeren bir hedef belirtime dönüştürülmesi gerekmektedir. Daha sonra spesifikasyona göre her fonem segmenti için birim seçimi gerçekleştirilmektedir. Son olarak olası birleştirme kusurlarının etkisini azaltmak için son işleme aşaması uygulanmaktadır.

3.5.4. Sinüsoidal Sentezleme

Konuşma dalga formunun olası temsillerinden biri, konuşmanın, ses yolunun rezonans özelliklerini modelleyen ve zamanla değişen doğrusal bir filtreden gırtlaksı uyarılma dalga formunun geçirilmesinin sonucu oluştuğu varsayımıdır. Çoğu durumda uyarma sinyalinin, sesli veya sessiz konuşmaya karşılık gelen iki olası durumdan birinde olduğunu varsaymak uygun görülmektedir. Gırtlaksı uyarımı modellemek için birçok yaklaşım bulunmaktadır. Sinüsoidal modelde gırtlaksı uyarım, zamanla değişen bir filtreye uygulandığında konuşma dalga biçimlerinin istenen sinüsoidal temsiline yol açan sinüs dalgalarının toplamı cinsinden temsil edilmektedir. Denklem 3.5'te olduğu gibi, bu tür sentezlemede *s(n)* orjinal ses sinyali *L* adet sinüsoid'in toplamı şeklinde modellenmektedir.

$$s(n) = \sum_{i=1}^{L} A_i \cos(w_i n + f_i) \tag{3.5}$$

Verilen eşitlikte, A_i sinüsoidalin genlik miktarını, f_i faz miktarını ve w_i frekans miktarını belirtmektedir. A_i ve f_i parametrelerini bulmak için pencereleme uygulanmış sinyal çerçevelerinin ayrık fourier dönüşümü (Discrete Fourier Transform-DFT) değeri bulunup Şekil 3.7'deki gibi çerçevelerden spektral büyüklüğün en üst noktası ayıklanmaktadır.



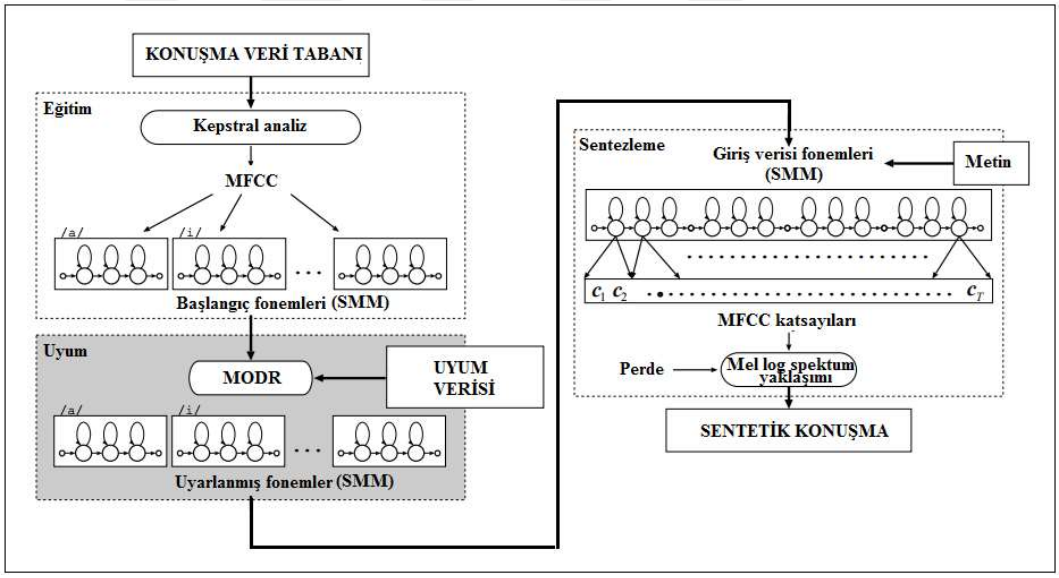
Şekil 3.7. Sinüsoidal sentezleme genel yapısı

Sinüzoidal modelde gırtlaksı uyarım, zamanla değişen ses yolu filtresine uygulandığında konuşma dalga biçimleri için istenen sinüzoidal gösterime yol açan sinüs dalgalarının toplamı cinsinden temsil edilmektedir. Konuşma parametreleri, konuşma dalga formunun yarı durağan bir kısmına kısa süreli Fourier dönüşümü (Short Time Fourier Transform-STFT) uygulanarak tahmin edilmektedir. Konuşmanın STFT'si, tüm perde harmoniklerinde meydana gelen zirvelere sahip olmaktadır. Bu nedenle, altta yatan sinüs dalgalarının frekansları, STFT'nin tepe noktalarına karşılık gelmektedir. Bileşen dalgalarının genlikleri ve fazları, basit bir tepe noktası seçme algoritması kullanılarak yüksek çözünürlüklü STFT'den tepe noktalarının çıkarılması ile tahmin edilebilmektedir.

3.5.5. Gizli Markov Modeli Tabanlı Sentezleme

Kendi alanındaki istatistiksel teorilerin çoğunu geliştiren Rus matematikçi Andrey Andreyevich Markov'un adını taşıyan Gizli Markov modelleri (Hidden Markov Model-HMM), 1970'lerin başlarında keşfedilmiştir. İlk olarak konuşma tanımada kullanılmış ve 1980'lerin

sonlarından beri biyolojik dizilerin analizine başarıyla uygulanmaktadır. Günümüzde, Bayes teorisine dayanan dinamik Bayes ağlarının belirli bir biçimi olarak kabul edilmektedir. HMM'ler, gözlemlenebilir sıralı sembollerden (örneğin, bir nükleotid dizisi) gizli bilgileri yakalamak için istatistiksel modeller olarak görülmektedir. Dizi analizinde, özellikle genomik DNA'daki ekzonları ve intronları tahmin etmek, proteinlerdeki fonksiyonel motifleri (alanları) belirlemek (HMM profili), iki diziyi hizalamak (çift HMM) olmak üzere pek çok uygulamaya sahiptirler. Bir HMM'de, modellenen sistemin bilinmeyen parametrelere sahip bir Markov süreci olduğu varsayılır ve buradaki zorluk, gözlemlenebilir parametrelerden gizli parametreleri belirlemektir. İyi bir HMM, gözlemlenen gerçek verilerin gerçek dünya kaynağını doğru bir şekilde modeller ve kaynağı simüle etme yeteneğine sahiptir. HMM'lere dayalı birçok makine öğrenmesi tekniği, konuşma tanıma, optik karakter tanıma, hesaplamalı biyoloji dahil problemlere başarıyla uygulanmıştır. Özellikle biyoinformatik'te temel bir araç haline gelmiştir: sağlam istatistiksel temeller, kavramsal basitlik ve işlenebilme açısından çeşitli sınıflandırma problemlerine uydurulabilmektedir.

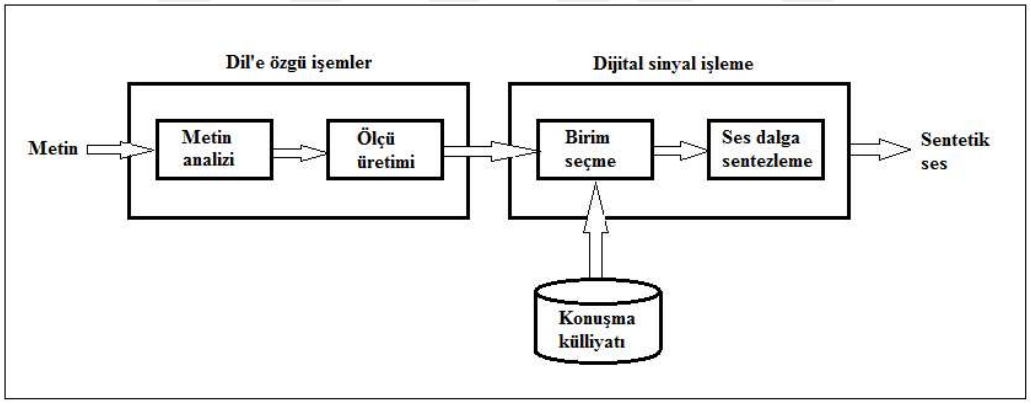


Şekil 3.8. Gizli Markov Modeli tabanlı konuşma sentezleme yapısı

Gizli markov modeli tabanlı konuşma sentezlemede Şekil 3.8'de de gösterildiği gibi eğitim, uyum ve sentezleme olmak üzere üç aşama bulunmaktadır. Ses sinyallerine MFCC analizi uygulanarak başlangıç fonemleri elde edilmektedir. Başlangıç fonemleri uyarlanmış fonemlere dönüştürüldükten sonra sentez modülüne aktarılmaktadır. Burada sentezlenmek istenen metin ve önceki aşamadan elde edilen uyarlanmış fonemler giriş verisi olarak kullanılmaktadır. Oluşan sentez ses sinyaline perde özelliğide (etkisi) eklendikten sonra sentetik konuşma üretilmektedir.

3.5.6. Birim Seçerek Sentezleme

Korpus tabanlı ses sentezleme adıylada bilinen birim seçerek ses sentezlemede daha önceden veri tabanına kaydedilmiş birimler içerisinden sentezlenecek sese göre birim seçimi gerçekleştirilmektedir. Oluşturulacak veritabanı en küçük ses parçasından (fonlar, fonemler) en büyük ses parçasına (hece, kelime, cümle vs) kadar çeşitli uzunlukta ses birimleri içerebilmektedir. Şekil 3.9'da gösterildiği üzere mevcut veri tabanından birim seçme ihtiyacı ve sonrasında bunları birleştirme işlemleri yapılması gerektiğinden bitiştirerek sentezleme ile benzerlik göstermektedir. Sentezlenen sese temel frekans ve formant frekansları ile doğallık eklenmektedir. Bu tür konuşma sentezleme de veri tabanlarında kayıtlı birim uzunlukları kısa olduğunda, hem etiketleme hemde bu birimleri toplama işlem süresi uzayacağından, küçük birimler depolama dezavantajlı olarak görülmektedir. Ayrıca küçük birimler kullanılarak yapılan sentezleme de birimler arasındaki geçişlerin insan kulağıyla algılanma ihtimali yükselmektedir. Bu da sesin doğallığını bozmaktadır. Uzun birim kullanma depolamada tek parçanın çok yer kaplaması dezavantajı oluşturuken, sentez konuşmada birimler arasında çok fazla geçiş olmadığından ses doğallığının iyi olması avantajı barındırmaktadır.



Şekil 3.9. Birim (unit) seçerek konuşma sentezlemenin genel yapısı

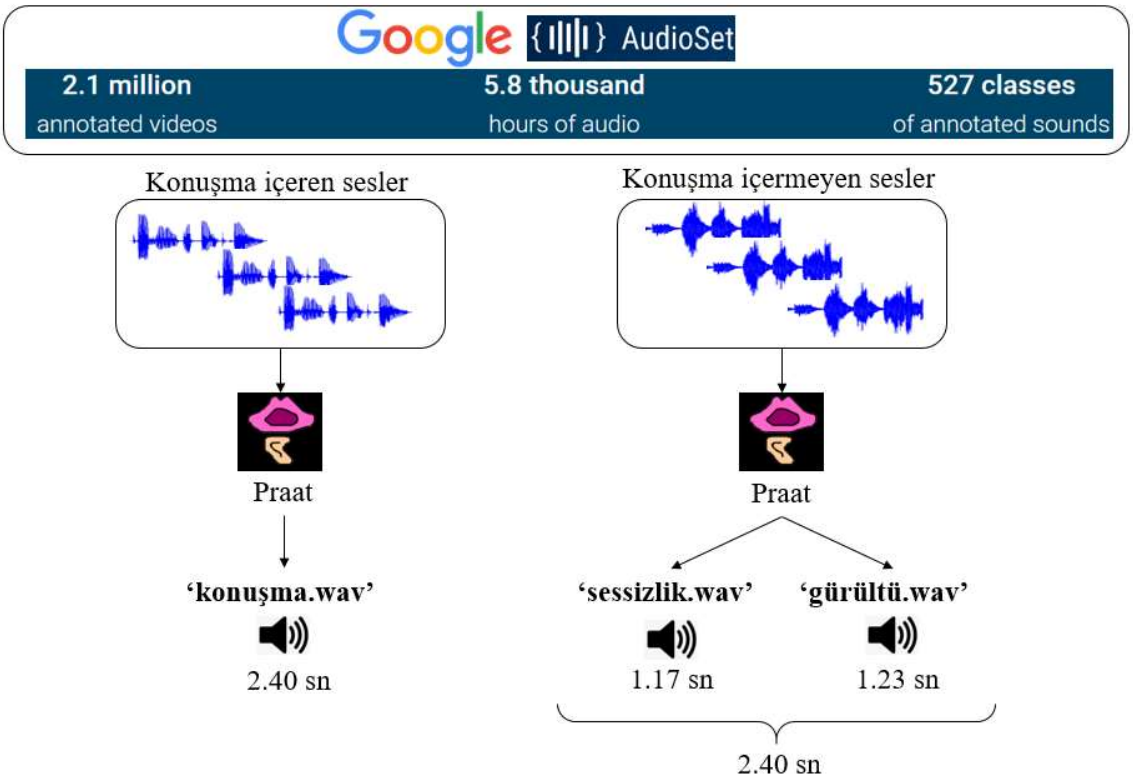
Veri tabanı oluşturma sırasında, kaydedilen her söz bazı bireysel fonemlere, hecelere, morfemlere, kelimelere, deyimlere ve cümlelere göre kaydedilmektedir. Daha sonra temel frekans, perde, süre, hecenin durumu, önceki ve sonraki fonemler gibi segmentasyon ve akustik parametrelere dayalı olarak konuşma veritabanındaki birimlerin bir ekdeksi yapılmaktadır. Bu yöntem çıkış konuşmasında diğer tekniklere göre doğallık sağlamaktadır.

4. MATERYAL VE METOT

Bu tez çalışmasında önerilen konuşmacı diyarizasyon sistemi hibrit konuşma aktivite dedektörü (hibrit VAD), konuşmacı bölütleme ve konuşmacı özelliklerinin çıkarımı (d-vektör), konuşmacı kümeleme (spektral kümeleme) olmak üzere 3 genel aşamadan oluşmaktadır. Tezin literatüre katkısı olan kısım tamamen özgün geliştirilen hibrit konuşma aktivite tespit sistemidir. Hibrit konuşma aktivite tespit sistemi denetimli ve denetimsiz (eşikleme) iki farklı konuşma aktivite dedektörünün mantıksal operatörler (OR/AND) aracılığıyla birleştirilmesi sonucu oluşmaktadır. Denetimli VAD ve Denetimsiz VAD için ses sinyallerinden aynı özellikler (zaman domeninde enerji, sıfır geçiş oranı-ZCR, MFCC) çıkarılmıştır. Denetimli VAD sistemi için Google AudioSet ses verisetinden elle anote edilmiş ses parçacıkları ile eğitilen bir uzun-kısa dönemli hafıza ağı (Long-Short Term Memory-LSTM) kullanılmıştır. Denetimsiz VAD sistemi için ise ses işaretlerinden çıkarılan yukarıdaki özelliklerin herbiri anlamlı bir eşik değerine tabi tutulmuştur. Konuşmacı diyarizasyonunun ilk aşaması olarak hibrit bir konuşma aktivite tespit sistemi geliştirilmiştir. Hibrit VAD sisteminde ise ilk olarak denetimli yapı tasarlanmıştır.

4.1. Google AudioSet: Denetimli VAD için Kullanılan Ses Veri Seti

2017 yılında Google araştırmacıları Jort F. Gemmeke ve arkadaşları tarafından bir Google platformu olan Youtube'tan çekilen seslerle toplanan bir veriseti "*Audio Set: An ontology and human-labeled dataset for audio events*" başlıklı çalışma ile açık kaynak kodlu olarak yayınlanmıştır [35]. Bu çalışma IEEE'nin International Conference on Acoustics, Speech, and Signal Processing (ICASSP) adlı konferansında sunulmuştur. 10'ar saniyelik Youtube videolarından toplanan veriseti içerisinde toplamda 632 farklı ses sınıfı/türü bulunmaktadır. Veriseti özellikle ses olay tespiti/tanıma (sound/audio event detection/recognition) alanlarında çalışan araştırmacılar için hazırlanmıştır. Veri seti içerisindeki ses kayıtlarının toplam uzunluğu 5800 saattir. Yayınlanan ses veriseti ayrıca yaklaşık 2.1 milyon ses kaydı barındırmaktadır. Bu tez çalışmasında geliştirilen konuşmacı diyarizasyon sisteminin ilk aşaması olan hibrit konuşma aktivite tespiti iki kısımdan oluşmaktadır. Bu iki kısımdan ilki denetimli konuşma aktivite tespit (VAD) sistemdir. Denetimli VAD sistemi için yukarıda bahsedilen Google AudioSet ses verisetinden Şekil 4.1'de gösterildiği gibi "konuşma.wav", "sessizlik.wav" ve "gürültü.wav" olmak üzere .wav formatında üç farklı ses dosyası oluşturulmuştur. "konuşma var" ve "konuşma yok" şeklinde iki farklı ses kategorisiyle eğitilecek olan LSTM ağı için, "konuşma.wav" ses doyası "konuşma var", "sessizlik.wav" ve "gürültü.wav" ses dosyaları "konuşma yok" kategorisine dahil edilmiştir.



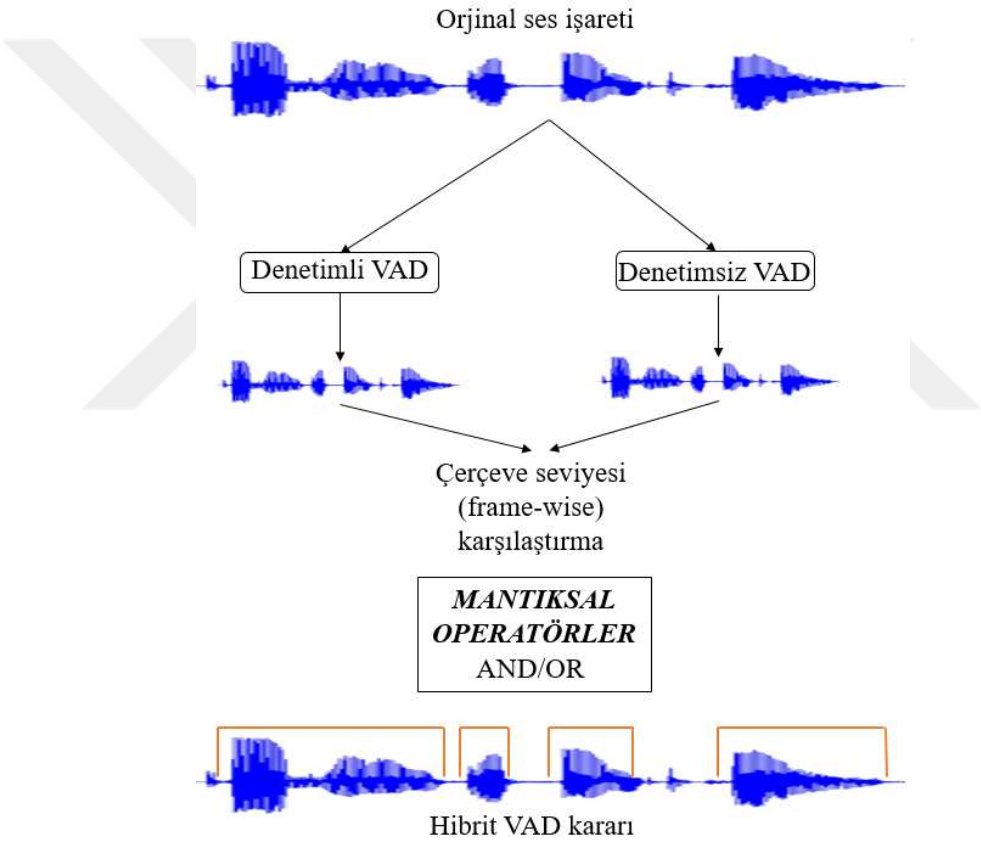
Şekil 4.1. Google AudioSet ses verisetinden LSTM ağında eğitilecek seslerin çıkarımı.

Bu ses kayıtlarının uzunluğu "konuşma.wav" için 2.40 saniye, "sessizlik.wav" için 1.17 saniye ve "gürültü.wav" için 1.23 saniye olarak belirlenmiştir. Öyle ki hem "konuşma var" hem de "konuşma yok" kategorilerinde bulunan ses kayıtlarının toplam ses uzunluğu 2.40 saniye de eşitlenerek eğitilecek olan LSTM ağı için dengeli (balanced) bir veri seti oluşturulması amaçlanmıştır. Denetimli VAD sisteminde kullanılacak üç adet ses kaydını oluşturmak için açık kaynak kodlu Praat [36] yazılımdan faydalanılmıştır. Google AudioSet ses verisetinde bulunan "konuşma var" ve "konuşma yok" kategorisindeki sesler Praat yazılımında açılıp, hem dalga formu hem de ses sinyallerinin spektrogramı dikkate alınmak suretiyle ses parçaları çıkarılmıştır. Çıkarılan bu ses parçaları birleştirilerek yukarıda bahsedilen ve Şekil 4.1'de görselleştirilen "konuşma.wav", "sessizlik.wav" ve "gürültü.wav" ses kayıtları oluşturulmuştur.

Bu tez çalışması süresince gereken tüm yazılımlar ve modüller Pyhton programlama dili ve bu dil için hazırlanmış kütüphaneler kullanılarak kodlanmıştır. Geliştirme ortamı olarak Google Colab Online Editöründen [39] faydalanılmış, depolama ihtiyacıda Google'ın Drive uygulamasından karşılanmıştır. Denetimli VAD sistemindeki LSTM ağı için tensorflow [40] ve keras [41] kütüphanelerinden, önceden eğitilmiş LSTM ağından konuşmacı gömülü özelliklerinin (d-vektör) çıkarılması için de pytorch [42] kütüphanesinden faydalanılmıştır.

4.2. Hibrit VAD Sisteminin Tasarlanması

Tez çalışması kapsamında geliştirilen konuşmacı diyarizasyon sistemi hibrit konuşma aktivite dedektörü (hibrit VAD), konuşmacı bölütleme ve konuşmacı özelliklerinin çıkarımı (d-vektör), konuşmacı kümeleme (spektral kümeleme) olmak üzere 3 genel aşamadan oluşmaktadır. Tezin literatüre katkısı olan kısım tamamen özgün geliştirilen hibrit konuşma aktivite tespit sistemidir. Hibrit konuşma aktivite tespit sistemi denetimli ve denetimsiz (eşikleme) iki farklı konuşma aktivite dedektörünün mantıksal operatörler (OR/AND) aracılığıyla birleştirilmesi sonucu oluşmaktadır. Hibrit konuşma aktivite tespit (hibrit VAD) sisteminin genel yapısı Şekil 4.2'de verilmiştir.



Şekil 4.2. Özgün Hibrit Konuşma Aktivite Dedektörü'nün genel yapısı.

Bu yapıdan da anlaşılacağı üzere hem denetimli hem de denetimsiz VAD sonucu oluşan kararlar çerçeve bazında mantıksal operatörlere tabi tutularak nihai Hibrit VAD kararı elde edilmiştir. Denetimli ve denetimsiz konuşma aktivite dedektörlerinin her ikisinde de ses işareti çerçevelere bölünüp analiz edilmiştir. Çerçeve boyutu her iki yaklaşım için de 0,02 saniye (20 milisaniye) olarak belirlenmiştir. Denetimli VAD yaklaşımında, ses sinyali çerçevelere bölündükten sonra çıkarılan enerji, sıfır geçiş oranı (ZCR) ve MFCC özellikleri ile LSTM ağı

beslenerek bir LSTM ağı modeli oluşturulmuştur. Daha sonra orjinal ses işaretinin her 20 milisaniyelik çerçevesi bu önceden eğitilen ağa sorularak her çerçeve için 1 (konuşma) veya 0 (konuşma değil) kararı alınmıştır. Denetimsiz VAD yaklaşımında ise ses orjinal ses işareti çerçevelere bölünmüş, her çerçeve için enerji, sıfır geçiş oranı (ZCR) ve 13.dereceden MFCC'nin sadece 1.derecesindeki (1^{st} MFCC) katsayı çıkarılıp özellik olarak kullanılmıştır. Daha sonra her bir özellik için mantıklı eşik değerleri belirlenmiştir. Zaman domeninde enerji için 0,2, sıfır geçiş oranı (ZCR) için 0,08 ve MFCC katsayılarının ilki için -700 değerleri kullanılmıştır. Her üç özellik için de bu değerler üzerinde kalan çerçevelere 1 (konuşma), altında kalan çerçevelere ise 0 (konuşma değil) kararı atanmıştır. Tüm çerçevelere; LSTM, zaman domeninde enerji, sıfır geçiş oranı (ZCR) ve ilk MFCC katsayısı açısından 1 veya 0 atandıktan sonra her çerçevenin bu dört kararı Tablo 4.1'deki mantıksal işleme tabi tutulmuştur.

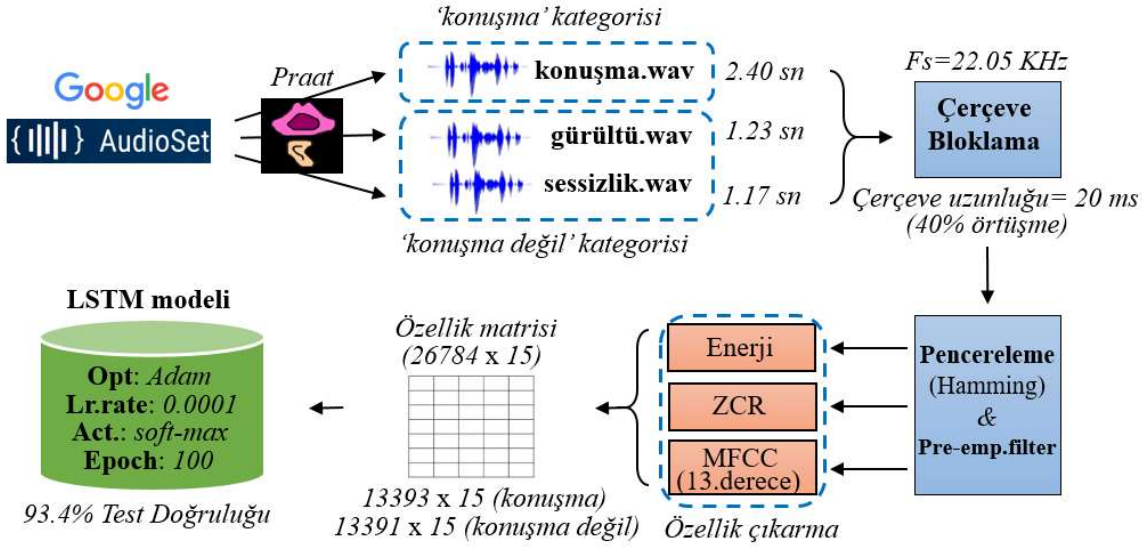
Tablo 4.1. Denetimli ve denetimsiz VAD kararlarının birleşim operatörü

Hibrit VAD Mantıksal Operatörü
Karar = (LSTM ∨ ENERJİ ∨ ZCR ∨ 1.MFCC) ∧ (ZCR ∨ 1.MFCC)

Bu tez çalışması kapsamında geliştirilen özgün hibrit konuşma aktivite dedektörü "*Hybrid voice activity detection system based on LSTM and auditory speech features*" başlığı altında [37] uluslararası hakemli ve Science Citation Index (SCI) kapsamında taranan Biomedical Signal Processing & Control dergisinde yayınlanmıştır.

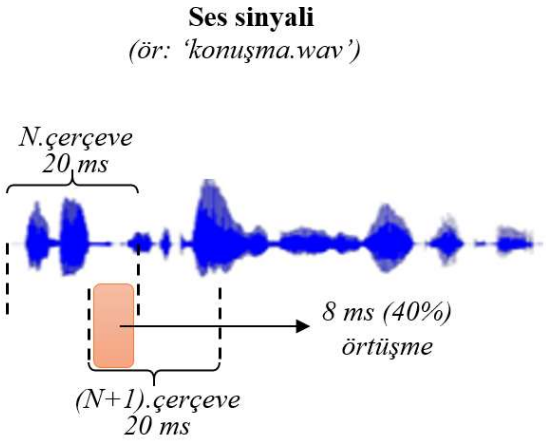
4.2.1. LSTM Tabanlı Denetimli VAD

Hibrit konuşma aktivite dedektörünün (VAD) denetimli safhasında Google AudioSet ses verisetinden elde edilen "konuşma.wav", "sessizlik.wav" ve "gürültü.wav" olmak üzere üç farklı ses dosyası LSTM ağını eğitmek için kullanılmıştır. Hibrit VAD sistemi tasarlanırken denetimli ve denetimsiz her iki alt sistemde de aynı ses özelliklerine bavurulmuştur. Denetimli VAD sistemi için Şekil 4.3'teki gibi bir LSTM modeli oluşturulmuştur.



Şekil 4.3. Denetimli VAD sistemi için LSTM modeli tasarlanması.

LSTM modeli oluşturmak için Google AudioSet ses verisetinden oluşturulan her üç ses dosyası için Şekil 4.3'teki Çerçeve Bloklama, Pencereleme, Filtreleme (pre-emphasis filter) ve özellik çıkarma işlemleri aynı olacak şekilde uygulanmıştır. Her bir ses sinyali ilk önce Çerçeveleme veya Çerçeve Bloklama (frame blocking) olarak bilinen işleme tabi tutulmuştur. Burada Google'dan toplanan seslerin örnekleme frekansı (sampling frequency-Fs) 22,05 KHz olarak belirlenmiştir. Çerçeve Bloklama işleminde Şekil 4.4'ki gibi her çerçeveye 0,02 saniye (20 milisaniye) değeri atanmış, çerçeveler üzerinde gezerken örtüşme seviyesi %40 (0,008 saniye yani 8 milisaniye) olarak ayarlanmıştır.



Şekil 4.4. Ses sinyali üzerinde çerçeve bloklama işlemi.

Çerçeveleme işleminde örtüşen çerçeveleme kullanılmasının sebebi ses sinyalinde ilgilendiğimiz özellikle ilgili hiçbir detayı kaçırmamaktır. Google AudioSet ses verisetinden elde

edilen her üç ses dosyası için de yukarıdaki çerçeve bloklama işlemi uygulanmıştır. Ses sinyalinin çerçeve bloklanmasından sonra konuşma sinyalini başlangıç ve bitiş noktalarında yumuşatmak için "Hamming" adı verilen bir pencereleme işlevi gerçekleştirilmiştir. Bu aşamada kullanılan pencereleme (windowing) fonksiyonunun matematiksel gösterimi Denklem 4.1'deki gibidir [38]. Ses sinyali üzerinden pencereleme işlemi uygulandıktan sonra, herhangi bir gürültü etkisini azaltmak için (minimum seviyede tutmak için) Denklem 4.2'de matematiksel gösterimi verilen ön vurgu filtresi kullanılmıştır.

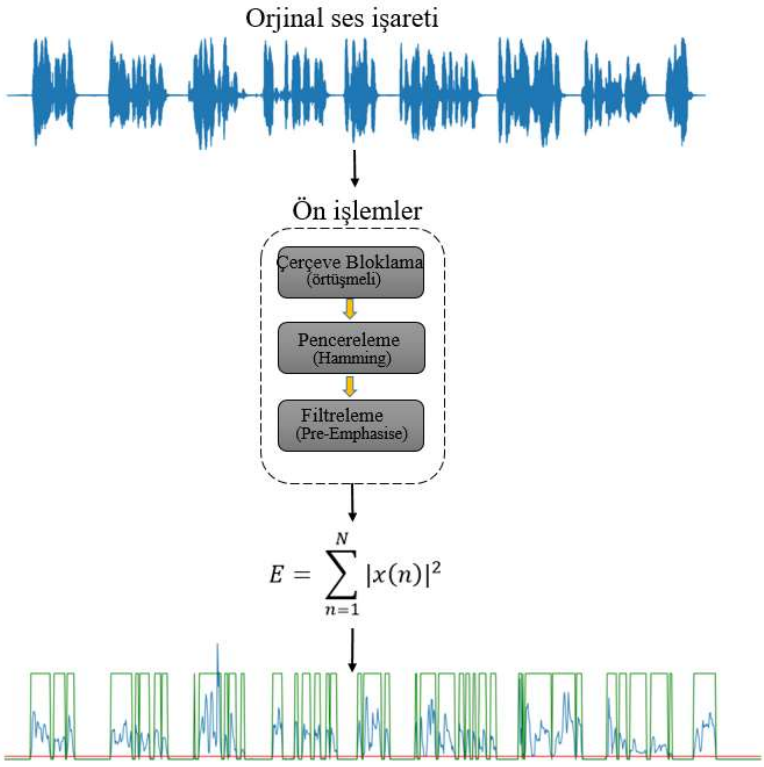
$$w_{hamming}(n) = 0{,}54 - 0{,}46\ \cos\frac{2\pi n}{N-1} n = 0,1, \dots N-1 \tag{4.1}$$

$$y(n) = x(n) - a.x(n-1) \tag{4.2}$$

Bu çalışması kapsamında "konuşma.wav", "gürültü.wav" ve "sessizlik.wav" ses sinyallerine uygulanan ön işlemler çerçeve bloklama, pencereleme ve filtreleme işlemlerinden oluşmaktadır. Ses sinyallerinden konuşma veya konuşma olmayan bölgeleri ayırt edebilmeye yönelik özellik çıkarma aşamasında enerji, sıfır geçiş oranı (ZCR) ve MFCC özelliklerinden faydalanılmıştır. Ses sinyalinin zaman domenindeki enerji değeri sinyal genliğindeki değişimleri yasıtmaktadır. Genel olarak ses sinyallerine ait özellikler zaman domeninde önemli ölçüde çeşitlilik göstermektedir. Ses sinyalinin enerji değeri sesli-sessiz bölge ayrımı yapmaktakta da kullanılmaktadır. Örneğin bir konuşma sinyalinde, sesli bölgelerde temel frekansın (fundamental frequency) ölçülebilir boyutta değiştiği gözlemlenebilmektedir. Bu yüzden zaman domeninde yapılan işlemler sonucunda sinyalde konuşma veya konuşma olmayan bölgeleri ayırt edebilecek yoğunluk, perde ve formant frekansı gibi vokal yol özellikleri elde edilebilmektedir. İncelenen ses sinyali çerçevesinin zaman domeninde enerji değeri Denklem 4.3'teki eşitlik yardımıyla hesaplanmıştır.

$$E = \sum_{n=1}^{N} |x(n)|^2 \tag{4.3}$$

Yukarıdaki formülden de anlaşılacağı üzere, zaman domeninde bir ses sinyali çerçevesinin enerji değeri, o çerçeve içerisindeki tüm örneklerin (*N* adet sample) mutlak değerli karelerinin toplamına şeklinde bulunmaktadır. Eşitlik uygulandıktan sonra bir çerçevenin enerji değeri bulunmaktadır. Ses sinyalinde konuşma olan bölgelerin/çerçevelerin enerji değeri konuşma olmayan bölgelerin/çerçevelerin enerji değerinden her zaman yüksek ölçülmektedir. Şekil 4.5'te bir ses sinyaline ön işlemler ardından enerji formülizasyonu uygulandıktan sonra ses sinyalinde konuşma olan bölgelerin bariz bir şekilde görünebildiği gösterilmiştir.



Şekil 4.5. Enerji bulma işlemi sonucu konuşma olan bölgeler.

Şekil 4.5'te orjinal ses işareti ön işlemlerden geçirildikten sonra çerçeve seviyesinde enerji değerleri hesaplanmıştır. Alttaki ses dalgası incelendiğinde, yeşil ile gösterilen (yani konuşma olan) alanlarda enerji değerlerinin yüksek olduğu görülmektedir. Ses sinyallerinden enerji değerleri çıkarıldıktan sonra benzer şekilde sıfır geçiş oranı ve 13.dereceden MFCC katsayıları elde edilmiştir. Denetimli VAD sistemi için LSTM modeli oluştururken, orjinal ses sinyalinin her çerçevesi için 1 adet enerji değeri, 1 adet sıfır geçiş oranı (ZCR) değeri ve 13 adet MFCC katsayıları olmak üzere 15 sütunlu bir özellik vektörü elde edilmiştir. Şekil 4.3'te de gösterildiği üzere, uzunluğu 2,40 saniye olan "konuşma.wav" ses dosyasından, çerçeve uzunluğu 0,02 saniye ve örtüşüm oranı %40 olmak üzere, çerçeveleme işlemi sonucunda toplamda 13393 çerçeve çıkarılmıştır. Her çerçevenin *1x15* boyutunda bir özellik vektörüne sahip olduğu göz önünde bulundurulduğunda, LSTM ağını eğitmek için "konuşma" kategorisinde *13393x15* boyutunda bir özellik matrisi elde edilmiştir. Aynı ön işlemler ve özellik çıkarma adımları, uzunluğu 1,23 saniye olan "gürültü.wav" ve 1,17 saniye olan "sessizlik.wav" ses dosyalarına uygulandığında, "konuşma değil" kategorisi için toplamda *13391x15* boyutunda bir özellik matrisi ortaya çıkmıştır. Hem "konuşma" hem de "konuşma değil" kategorisi için hazırlanan özellik matrisleri yatay olarak birleştirildiğinde LSTM ağına giriş (input) özellik matrisi olarak *26784x15* bir matris elde

edilmiştir. Bu matris LSTM ağı için eğitim seti olarak kullanılmıştır. Tablo 4.2'de denetimli VAD sistemi için geliştirilen LSTM ağı modelinin tüm hiper parametreleri detaylı bir şekilde verilmiştir.

Tablo 4.2. Denetimli VAD için geliştirilen LSTM modeline ait hiper parametreler.

Hiper parametre	Değer
Optimizasyon	Adam
Aktivasyon (dense)	ReLU
Aktivasyon (çıkış)	Soft-max
Giriş şekli	(15,1)
Dropout (dense)	0,3
Kayıp fonksiyonu	Sparse Cat.Cros.Entropy
Öğrenme oranı	0,0001
Epoch (iterasyon)	100
Test doğruluk oranı	%94,93

Oluşturulan basit LSTM modeli sadece 100 iterasyon çalıştırılmış ve %94,4 oranında test doğruluk oranı elde etmiştir. Bu model, kararı daha sonra denetimsiz VAD sisteminde enerji, ZCR ve 1.MFCC katsayısı ile karşılaştırılmak üzere saklanmıştır.

4.2.2. Eşikleme Tabanlı Denetimsiz VAD

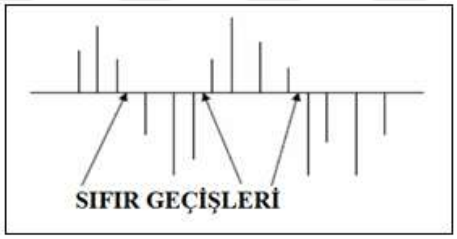
Geliştirilen özgün hibrit konuşma aktivite dedektörünün denetimsiz VAD bölümünde, denetimli sistemdekine benzer şekilde önce ses sinyaline ön işlemler uygulanmıştır. Çerçeveleme işleminde çerçeve boyutu yine 0,02 saniye (20 milisaniye) olarak belirlenmiş, denetimli VAD sisteminin aksine herhangi bir örtüşüm (overlapping) senaryosu uygulanmamıştır. Bunun sebebi denetimsiz VAD sisteminde, denetimli VAD sistemindeki gibi eğitilecek herhangi bir sinir ağının olmamasıdır. Doğrudan çerçeveler çıkarılmış her bir çerçevenin enerji, sıfır geçiş oranı (ZCR) ve 1.MFCC katsayısı elde edilmiştir. Enerji değeri bir önceki bölümde de anlatıldığı üzere konuşma olan yerlerde yüksek çıkmıştır.

Doğadaki ses sinyalleri dijital (sayısal) ortama aktarıldıktan sonra, sinyaller genlik (amplitude), frekans (veya periyot) ve şiddet parametrelerine sahip olmaktadır. İçerisinde konuşma bulunan bir ses dosyasının dijital ortamda dalga formu incelendiğinde, konuşma olan bölgelerde ses sinyalinin frekans değerinin gürültü veya sessizlik bölgelerine göre daha düşük olduğu gözlenmektedir. Genel olarak konuşma olmayan ses sinyallerinde konuşma ses sinyallerine göre daha yükses frekans değerleri ölçülmektedir. Sıfır geçiş oranı (ZCR), ses sinyalinin belli zaman aralığında sıfırı kaç defa geçtiğini belirten bir değer olarak tanımlanmaktadır. Konuşma olan yerlerde düşük frekanslı sinyaller bulunduğundan ZCR oranı düşük ölçülmekte iken konuşma olmayan yerlerde (gürültü, sessizlik vs.) yüksek frekans gözlendiğinden ZCR oranı yüksek

çıkmaktadır. Bu da ZCR oranının sinyalin frekans değeriyle doğru orantılı olduğunu göstermektedir. İncelenen ses sinyali çerçevesinin zaman domeninde enerji değeri Denklem 4.4'teki eşitlik yardımıyla hesaplanmıştır.

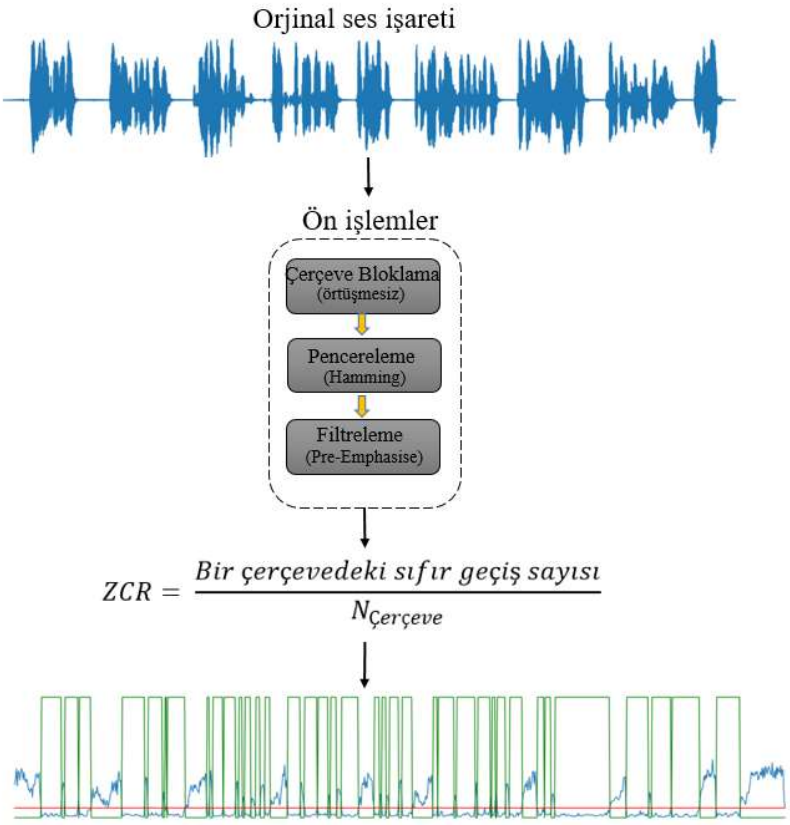
$$ZCR = \frac{Bir\ çerçevedeki\ sıfır\ geçiş\ sayısı}{N_{Çerçeve}} \tag{4.4}$$

Bu eşitlikte *N* değeri, incelenen çerçevedeki örnek (sample) sayısıdır. Örneğin Şekil 4.6'da bulunan çerçevedeki sıfır geçiş sayısı 3 iken, çerçevede bulunan örnek sayısı görseldeki dikey mumların (çizgilerin) sayısı kadar yani 14'tür. Örnek şekildeki çerçevenin ZCR değeri 3/14 yani yaklaşık 0,21 olarak hesaplanmaktadır. Tez çalışmasında kullanılan ses kayıtlarının örnekleme frekansı 22050 Hz (22,05 Khz), hibrit VAD sisteminin denetimsiz VAD alt modülünde çerçeve boyutu 0,02 ms olduğundan, her bir çerçeve 441 adet örnek içermektedir. Yani tez çalışması kapsamındaki tüm ZCR oranı hesaplamalarında Denklem 4.4'teki formülün paydası 441 olarak alınmıştır.



Şekil 4.6. Örnek bir çerçevede sıfır geçişleri

Şekil 4.7'de tez çalışması kapsamında bir ses sinyaline ön işlemler ardından ZCR formülizasyonu uygulandıktan sonra ses sinyalinde konuşma olan bölgelerin bariz bir şekilde anlaşılabildiği görülmüştür.



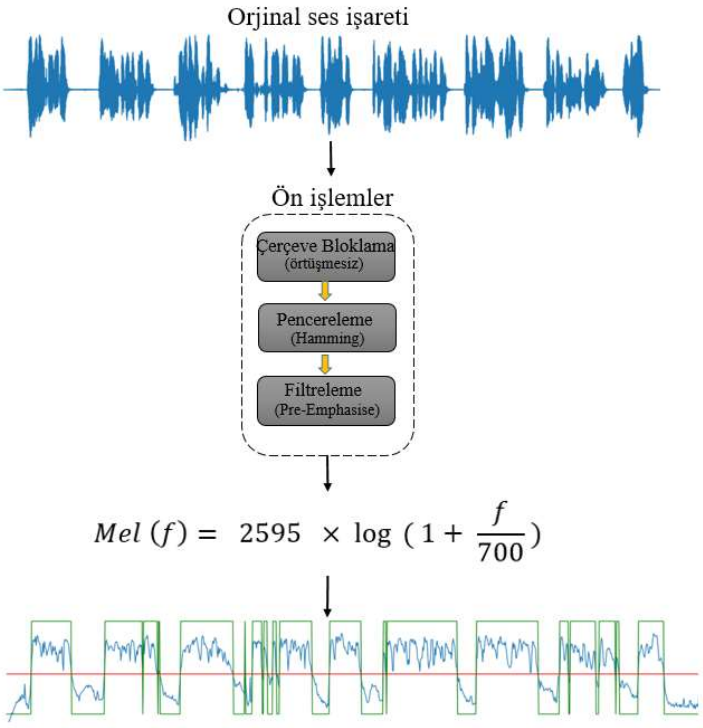
Şekil 4.7. ZCR hesaplaması sonucu konuşma olan bölgeler

Şekil 4.7'de altta bulunan sinyal grafiğinde yeşil olan bölgeler konuşma olan bölgeler olarak belirlenmiştir. Grafikten de anlaşılacağı üzere Şekil 4.5'te enerji için anlatılan durumun tam tersine, konuşma olan yerlerde ZCR oranının konuşma olmayan bölgelere göre daha düşük çıktığı görülmektedir. Bu da yukarıda bahsedilen konuşma sinyallerinde frekans değerinin konuşma olmayan sinyallerde frekans değerinden daha düşük olmasından kaynaklanmaktadır. Alttaki sinyal grafiğinde yer alan kırmızı yatay çizgi ise 0,08'lik eşik değerini temsil etmektedir.

Ses sinyalinden zaman domeninde enerji değeri ve ZCR oranı çıkarıldıktan sonra MFCC hesaplanmıştır. MFCC katsayıları gerçek hayattaki insan kulağını modelleyebildiğinden konuşma işleme uygulamalarındaki özellik çıkarma işlemlerinin kalbi sayılabilecek niteliktedir [43]. Günümüzde konuşma/konuşmacı tanıma, ses olay tespiti, konuşmacı doğrulama, konuşma aktivite dedektörleri ve konuşmacı diyarizasyonu gibi özellik çıkarma aşaması ihtiyacı olan konuşma işleme alanlarında merkezi bir özellik çıkarma yöntemi olarak bilinmektedir. MFCC doğrusal frekans skalasının Denklem 4.5'te verilen formül yardımıyla Mel skalasına dönüştürülmesini amaçlamaktadır.

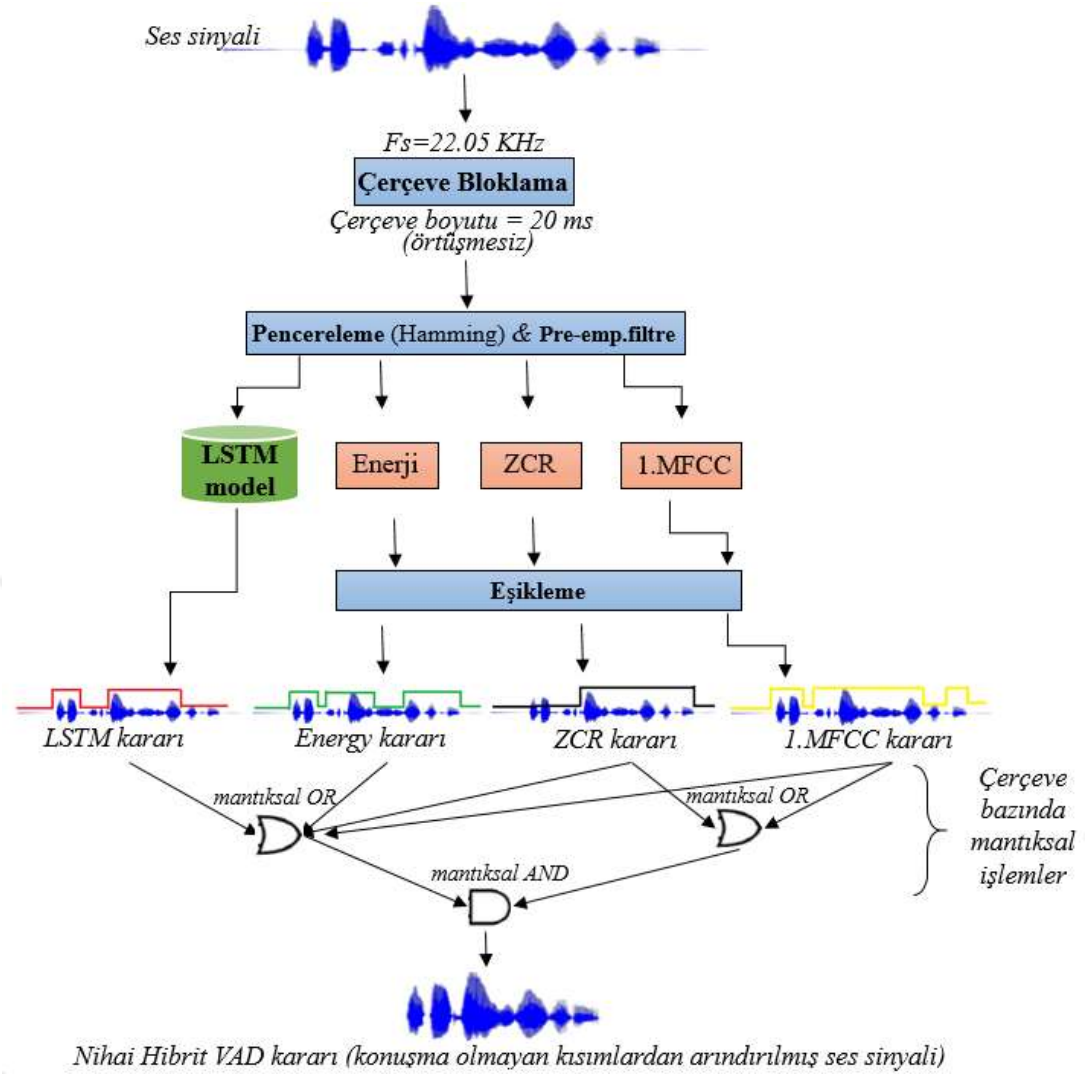
$$Mel\,(f) = \ 2595\ \times \log\,(1 + \frac{f}{700}) \tag{4.5}$$

Şekil 4.8'de tez çalışması kapsamında bir ses sinyaline ön işlemler ardından MFCC özellik çıkarma işlemi uygulandıktan sonra ses sinyalinde konuşma olan bölgelerin net bir şekilde anlaşılabildiği görülmektedir.



Şekil 4.8. 1.MFCC katsayısı hesaplandıktan sonra konuşma olan bölgeler

Şekil 4.8'de altta bulunan sinyal grafiğinde yeşil olan bölgeler konuşma olan bölgeler olarak tespit edilmiştir. Grafikte konuşma olan (yeşil) bölgelerde 1.MFCC katsayısının yüksek olduğu, konuşma olmayan bölgelerde bariz şekilde düşük olduğu görülmüştür. Dolayısıyla 1.MFCC katsayısının bir ses çerçevesinin konuşma içerip içermediği konusunda bariz ipuçlarına sahip olduğu kanıtlanmıştır. Şekil 4.8'de en altta bulunan sinyal grafiğinde yer alan kırmızı yatay çizgi ise -700 olan eşik değerini temsil etmektedir. Denetimli VAD sistemi (sinir ağı modeli tabanlı) ve denetimsiz VAD (eşikleme tabanlı) sisteminin birleşiminden oluşan hibrit VAD yapısı Şekil 4.9'da görselleştirilmiştir.



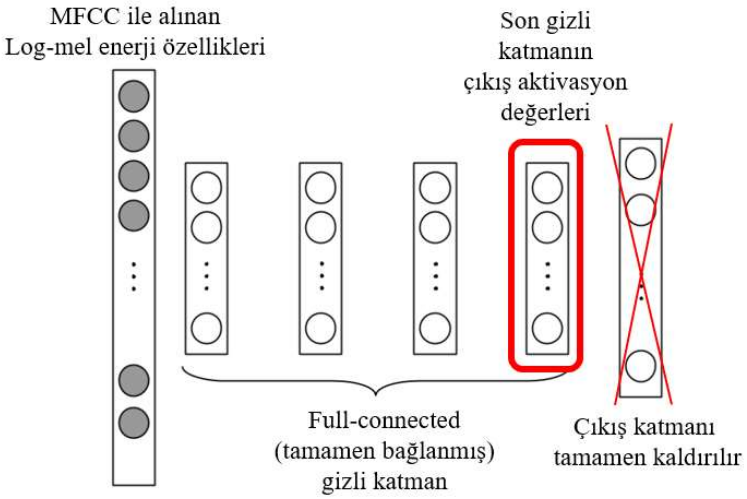
Şekil 4.9. Hibrit VAD sisteminin genel yapısı

Bu tez çalışması kapsamında geliştirilen konuşmacı diyarizasyon sisteminin en önemli (özgün) kısmı olan hibrit konuşma aktivite dedektörüne genel olarak bakıldığında, 22.05 KHz örnekleme frekansına (sampling frequency) orjinal ses sinyali öncelikle örtüşümsüz 0,02 saniyelik çerçevelere ayrılmaktadır. Ses çerçevesine daha sonra ön işlem olarak kullanılan pencereleme (hamming) ve ön vurgu filtresi (pre-emphasis filter) uygulanmaktadır. Ses sinyalinden elde edilen tüm çerçeveler öncelikli olarak Bölüm 4.2.1'de elde edilen LSTM modeline gönderilip, tüm çerçevelerin 1 veya 0 karşılıklarından oluşan 1x*N* (*N*;orjinal ses sinyalindeki toplam çerçeve sayısı) boyutunda LSTM kararı çıkarılmaktadır. Daha sonra önişlemlerden geçmiş tüm çerçevelere sırasıyla enerji, ZCR ve 1.MFCC katsayısı bulma işlemleri uygulanmaktadır. Enerji, ZCR ve 1.MFCC katsayısı bulma işlemleri sonucunda oluşan karar vektörlerinin boyutu, LSTM kararındaki gibi 1x*N*'dir. Şekil 4.9'da farklı renklerde gösterilen bu kararlara daha sonra çerçeve seviyesinde (frame-level/frame-wise) Tablo 4.1'deki hibrit VAD mantıksal operatörü uygulandıktan sonra,

nihai hibrit VAD kararına ulaşılmıştır. Sonuç olarak orjinal ses sinyalinden konuşma olmayan bölgeler atılmış ortaya daha kısa ve sadece konuşma olan bölgeleri içeren bir ses sinyali çıkmıştır. Böylece bu tez çalışması kapsamında önerilen konuşmacı diyarizasyon sisteminin özgün olan konuşma aktivite dedektörü (VAD) hibrit bir şekilde gerçekleştirilmiştir. Buradaki hibrit ifadesi denetimli ve denetimsiz yaklaşımların bir arada kullanılması anlamına gelmektedir.

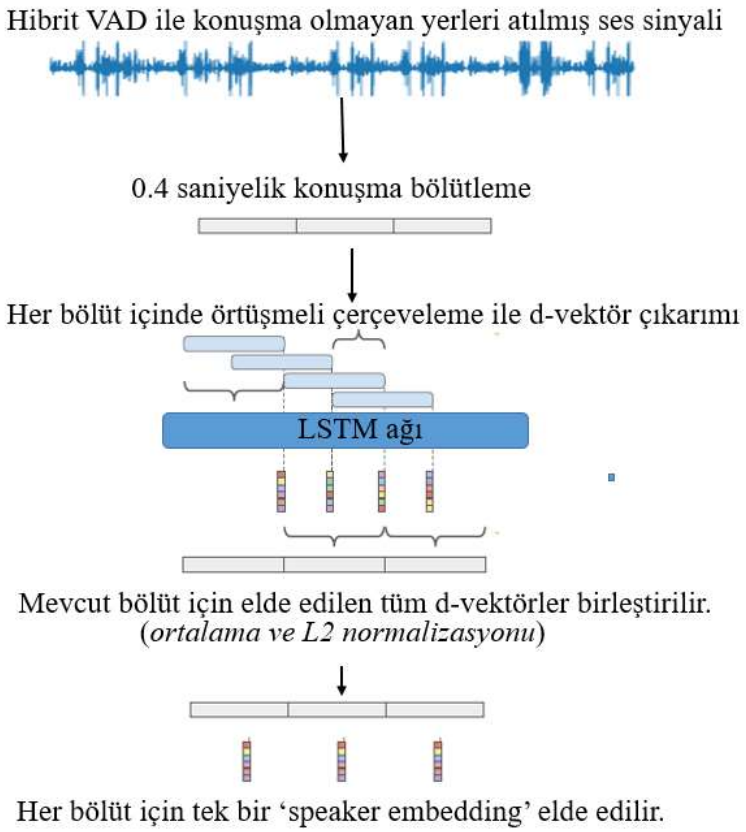
4.3. D-Vektörlerin Çıkarımı

Konuşmacı diyarizasyonu konusunda son zamanlarda yapılan çalışmaların çoğunda derin öğrenme yöntemlerine başvurulmaktadır. Derin öğrenme yöntemlerinin yaygınlaşması ile konuşmacı doğrulama için büyük miktarda veri kullanılarak eğitilen sinir ağlarından faydalanılmaya başlanmıştır. Konuşmacı doğrulama için eğitilen ağlar, çok sınırlı sayıda gözlem ile aynı sesi veren kullanıcıyı tanımaya çalışmaktadır. Son zamanlarda bu alanda %90'ın üzerine çıkan konuşmacı doğrulama doğruluk oranları elde edilmektedir. Konuşmacı diyarizasyonu probleminin konuşmacıya özgü özellik çıkarma aşamasında da konuşmacı doğrulama için önceden büyük miktarda veri ile eğitilmiş hazır ağlar kullanılabilmektedir. Bu tez çalışmasında daha önce Quan Wang ve arkadaşlarının [34] konuşmacı doğrulama için yapmış olduğu bir çalışmada oluşturduğu ve açık kaynak olarak paylaştığı bir konuşmacı doğrulama LSTM ağından faydalanılmıştır. Araştırmacılar bahsi geçen çalışmada aynı zamanda d-vektör kullanımının i-vektör kullanımdan daha başarılı sonuçlar ürettiğini vurgulamışlardır. D-vektör ilk defa 2014 yılında Google çalışanlar ve Johns Hopkins Üniversitesi araştırmacıları tarafından "*Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification*" başlıklı çalışma ile keşfedilmiştir [44].



Şekil 4.10. D-vektör'ün çıkarılacağı konuşmacı doğrulama sinir ağı'nın yapısı

Deep-vektör (derin vektör) ifadesinin kısaltılmış hali olarak kullanılan d-vektör, Şekil 4.10'da görselleştirildiği gibi daha önce konuşmacı doğrulama çalışmaları için eğitilmiş bir derin öğrenme ağının son gizli katmanından alınan aktivasyonların ortalaması olarak hesaplanmakta ve konuşmacı özellik çıkarıcısı (speaker feature extractor) amacıyla kullanılmaktadır. Bu tez çalışmasında da yapıldığı gibi özellik çıkarıcı ağın giriş değerleri ses bölütünün log-mel enerji değerleridir. Konuşmacı doğrulamada her konuşmacı için çıkarılan d-vektör arasında mesafe ölçümü yapılarak karar verilmektedir. Enerji değerleri çıkarıldıktan sonra her çerçeveye ait bu değerler önceden konuşmacı doğrulama için eğitilmiş hazır LSTM ağına gönderilerek, bu ağın son gizli katmanının aktivasyon değerleri d-vektör olarak kabul edilmiştir. Bu tez çalışmasında ise d-vektörler 0,4 saniyelik (400 milisaniye) (yalnızca bir konuşmacının yer aldığından emin olunan bölüt uzunluğu) her bölütü temsil eden ve daha sonra üzerinde spektral kümeleme işlemi uygulanacak özellik vektörleri olarak kullanılmıştır. Hibrit konuşma aktivite tespit sistemi uygulanan ses sinyaline bölütleme gerçekleştirildikten sonra Şekil 4.11'daki gibi her bölüt için tek bir d-vektör (speaker embedding) elde edilmiştir. Bu d-vektörler bir sonraki aşamada spektral kümeleme işlemine tabi tutulmuştur.



Şekil 4.11. Konuşma bölütleme ve d-vektör çıkarma işlemi

4.4. Konuşmacı Kümeleme

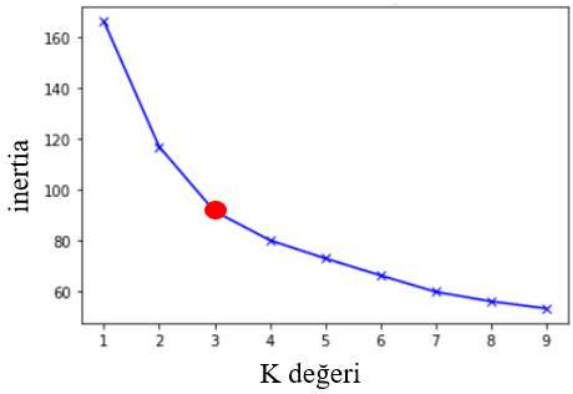
Konuşmacı diyarizasyonunda en etkili kümeleme yöntemlerinden biri de K-means kümeleme algoritması üzerinde bir takım iyileştirme işlemleri yapıldıktan sonra ortaya çıkan Spektral Kümeleme yöntemidir. Bu tez çalışmasında bir önceki bölümde elde edilen d-vektör'ler spektral kümeleme algortimasına tabi tutulmuştur. Spektral kümeleme algoritması genel olarak 4 aşamadan oluşmaktadır. Bunlar;

- Benzerlik (affinity) matrisi A_{ij}'nin oluşturulması. Bu benzerlik matrisinde A_{ij}, *i.* ve *j.* konuşma bölütlerine ait d-vektör'lerin kendi aralarındaki kosinüs benzerliği değerini ifade etmektedir. Bu matrisin diyagonal elemanlarındaki değer her segmente ait d-vektörün kendisi ile olan uzaklığını belirttiğinden benzerlik matrisinin o satırındaki maksimum değeridir.
- A matrisi üzerinde; Gauss bulanıklığı (Gaussian Blur), satır bazlı eşikleme, simetrizasyon, difüzyon ve satır bazlı maksimum normalizasyonu gibi bir takım ince ayar/iyileştirme (refinement) işlemlerinin yapılması. Bu işlemler yakınlık matrisi üzerinde bir nevi yumuşatma ve gürültü azaltma görevi görmektedir. Burada önemli olan benzer konuşmacı özelliklerine (d-vektör'lere) sahip bölütlerin, yakınlık matrisinde benzer değerlere sahip olmasıdır. Gauss bulanıklığı, verileri yumuşatmak ve aykırı değerlerin etkisini azaltmak için hareket eder. Satır bazında eşikleme, iki farklı konuşmacıya ait yerleştirmeler arasındaki benzerlikleri sıfırlamaya hizmet eder. Simetrizasyon, spektral kümeleme algoritması için çok önemli olan matris simetrisini geri yükler. Difüzyon, Difüzyon Haritaları algoritmasından [45] ilham alır ve farklı konuşmacılara ait yakınlık matrisinin bölümleri arasında net sınırlar ile sonuçlanan görüntüyü keskinleştirmeye hizmet eder. Son olarak, satır bazlı maksimum normalizasyon, sonraki spektral kümeleme adımı sırasında istenmeyen ölçek etkilerinin oluşmamasını sağlamak için matrisin spektrumunu yeniden ölçeklendirmeye yarar.
- A matrisi üzerinde ikinci maddedeki tüm iyileştirmeler yapıldıktan sonra, A matrisine Eigen Dekompozisyonu (ayrıştırma) uygulanması.
- Son olarak konuşmacı özelliklerinin (d-vektör'lerin) eigen ayrıştırması sonucu oluşan eigen vektörlerle yer değiştirilmesi

Tüm bu işlemlerden sonra elde edilen yeni konuşmacı özellikleri olarak dikkate alınmıştır. İyileştirme işlemleri sonucu oluşan bu konuşmacı özelliklerine K-Means kümeleme algoritması [46] uygulanmıştır. K-Means algortiması literatürde konuşmacı diyarizasyonu alanında bir çok çalışma tarafından da kullanıldığından [47-49], merkezi bir yöntem olarak kabul edilmektedir.

D-vektör çıkarma aşamasından sonra spektral kümeleme aşamasına geçmeden başlangıç (initial) kümeleme sayısını belirlemek gerekmektedir. K-means kümeleme için başlangıç küme

sayısını belirlemede elbow (dirsek) ve Silhouette (silüet) yöntemleri kullanılmaktadır. Bu tez çalışmasında elbow yönteminden faydalanılmıştır. Şekil 4.12'de tez çalışmasında kodlanan elbow yönteminin görseli verilmiştir.



Şekil 4.12. Inertia yöntemi kullanılarak Elbow yönteminin gerçekleştirilmesi

Şekil 4.12'de gösterilen yıldızların herbiri inertia noktası olarak tanımlanmaktadır. Inertia değeri küme sayısı 1'den 10'a kadar olmak üzere her başlangıç küme sayısı için hesaplanmıştır. Her başlangıç küme sayısı için bulunan inertia değerleri yukarıdaki şekilde bir grafik oluşturmuştur. Elbow yöntemi bu grafikteki en bariz kırılma noktasını (dirseği) tespit etmeyi gerektirmektedir. Bu tez çalışmasında K-means kümeleme ve inertia değerlerinin hesaplanmasında Python dilinde Scikit-learn [50] kütüphanesinden faydalanılmıştır. Çıkarılan inertia değerleri arasındaki dirsek (elbow) noktasının tespitinde ise kneed [51] kütüphanesi kullanılmıştır. Yukarıdaki grafikte başlangıç küme sayısı 3 olarak belirlenmiştir. Başlangıç küme sayısı belirlendikten sonra spektral kümeleme (refinement adımları dahil) işlemi için Quan Wang ve arkadaşlarının [52, 33]'deki çalışmalarında da faydalandığı ve açık kaynak olarak yayınlanan "spectralcluster" kütüphanesi kullanılmıştır. Şekil 4.13'te 48 saniyelik 3 konuşmacı içeren 22.05 KHz örnekleme frekansına sahip .wav uzantılı bir ses sinyalinin şimdiye kadar anlatılan tüm işlemlerden geçerek spektral kümeleme sonrasında bölüt etiketleme sonucu gösterilmiştir.

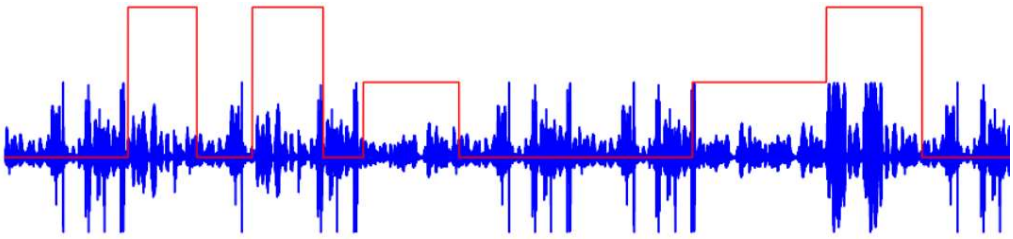
```
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1,
 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2,
 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0]
```

Şekil 4.13. 3 konuşmacılı bir ses sinyaline spektral kümele uygulanması sonucu oluşan bölüt etiketleri

Şekil 4.13'te görüldüğü gibi 0 değeri ilk konuşmacıya, 2 değeri ikinci konuşmacıya, 1 değeri ise üçüncü konuşmacıya atanarak, iyileştirme (refinement) işlemi sonucu oluşan d-vektör (veya eigen vektör) matrisinin kümelenmesi gerçekleştirilmiştir. Referans ses sinyalinde de yukarıdaki grafikte olduğu gibi sırasıyla 0, 2, 0, 2, 0, 1, 0, 1, 2, 0 konuşmacıları konuşmaktadır.

4.5. Konuşmacı Diyarizasyonu

Bir ses sinyalinde geçen konuşmacılardan kimin ne zaman konuştuğunu tespit edebilen konuşmacı diyarizasyon sistemlerinin alt aşamalarından hibrit konuşma aktivite dedektörü (hybrid VAD), konuşma bölütleme, d-vektör çıkarma ve spektral kümeleme adımları gerçekleştirildikten sonra, kümeleme sonucu oluşan etiketlerin zaman ekseninde orjinal ses dalga formuyla eşleştirilmesi gerekmektedir. Bu tez çalışmasının son kısmında kümeleme sonucu Şekil 4.13'teki gibi oluşan küme tahminlerinin (etiketlerinin) zaman ekseninde genişletilmesi yapılarak, orjinal ses sinyalinin dalga formuyla örtüşmesi işlemi yapılmıştır. Tez çalışmasının sonucunda Şekil 4.14'te gösterildiği gibi test edilen ses sinyallerinin çıkışı gözlenmiştir. Grafikten referans (groud-truth) ve hipotez karşılaştırması yapıldığında, orjinal seste 0, 2, 0, 2, 0, 1, 0, 1, 2, 0 sırasıyla konuşan konuşmacıların diyarizasyon sonucunda başarılı bir şekilde tespit edildiği görülmektedir.



Şekil 4.14. Önerilen özgün hibrit VAD tabanlı diyarizasyon sonucunun orjinal ses üzerinde gösterimi

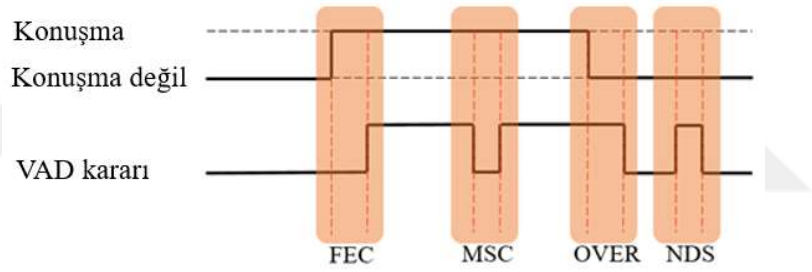
Şekil 4.14'te mavi ile gösterilen orjinal ses sinyalinin hibrit VAD işlemi sonucundaki dalga formu (referans/ground-truth) iken, kırmızı ile gösterilen (y ekseninde 0, 1, 2 değerlerine karşılık gelen) değerler ise önerilen diyarizasyon sisteminin çıkışı (hipotez) olarak belirlenmiştir. Bu tez çalışmasında geliştirilen diyarizasyon sistemi kaydedilmiş bir ses kaydını analiz edebilen çevrimdışı (offline) sistem olarak tasarlanmış, çevrimiçi (online) veya gerçek zamanlı (real-time) analiz yapmamaktadır.

5. BULGULAR VE TARTIŞMA

Bu tez çalışmasında tamamen özgün ve daha önce literatürde önerilmemiş bir hibrit konuşma aktivite dedektörü (VAD) üzerine inşa edilmiş d-vektör tabanlı bir konuşmacı diyarizasyon sisteminin tasarlanmıştır. Önerilen hibrit konuşma aktivite tespit sistemi için literatürde olan bazı değerlendirme kriterleri göz önünde bulundurulmuştur.

5.1. Hibrit VAD sisteminin değerlendirilmesi

Önerilen konuşma aktivite tespit sisteminin değerlendirilmesi için literatürde konuşma aktivite tespit sistemlerinde faydalanılan Front-End Clipping (FEC), Mid-Speech Clipping (MSC), Over Hang (OVER), Noise Detected as Speech (NDS) hata oranları (değerlendirme metrikleri) kullanılmaktadır [53]. Şekil 5.1'de bu dört değerlendirme metriğinin ses sinyalinin hangi bölgesinde hesaplandığı gösterilmiştir.



Şekil 5.1. VAD sistemleri için değerlendirme kriterleri

Geliştirilen konuşma aktivite tespit sistemi, FEC hatası konuşma sinyalinin başında, MSC hatası konuşma süresince, OVER hatası konuşma bittiğinde ve NDS hatası hiç konuşma yokken ortaya çıkabilecek hatalar olarak tanımlanmıştır. Bu hata türlerinden;

- Front-End Clipping (FEC), orjinal ses sinyalinde (referans) konuşma başladığı halde, hipotez VAD kararının konuşmanın başlangıcını geç algılaması sırasında oluşmaktadır. Geç algılanma çerçeve adedi kadar FEC hatası oluşmaktadır.
- Mid-Speech Clipping (MSC), orjinal ses sinyalinde konuşma aralıksız devam ediyor iken, hipotez VAD kararının konuşma yokmuş gibi görmesi sonucu oluşmaktadır. Gerçekte konuşma olduğu halde konuşmanın algılanmadığı çerçeve adedi kadar MSC hatası ortaya çıkmaktadır.
- Over Hang (OVER), orjinal ses sinyalinde konuşma bitmesine rağmen hipotez VAD kararının hala konuşma devam ediyor gibi görmesiyle oluşan hata türüdür. Konuşmanın

bittiği süreden, hipotez sistemin konuşmanın bittiğini algıladığı zamana kadar geçen çerçeve adedi ile OVER hatası hesaplanmaktadır.

- Noise Detected as Speech (NDS), adından da anlaşılacağı üzere, orjinal ses sinyalinde tüm konuşmalar bittikten sonra herhangi bir konuşma yokken, hipotez VAD kararının konuşma varmış gibi görmesi sonucunda oluşmaktadır. Konuşma olmadığı halde konuşma olarak tespit edilen süre içerisindeki çerçeve adedi ile ölçülmektedir.

Bu tez kapsamında geliştirilen özgün hibrit VAD sisteminin Front-End Clipping (FEC), Mid-Speech Clipping (MSC), Over Hang (OVER) ve Noise Detected as Speech (NDS) hata oranları Denklem 5.1, 5.2, 5.3 ve 5.4'te verilen formüller ile hesaplanmıştır.

$$FEC = \frac{FEC\ oluşan\ çerçeve\ adedi}{Konuşma\ içeren\ toplam\ çerçeve\ adedi} \tag{5.1}$$

$$MSC = \frac{MSC\ oluşan\ çerçeve\ adedi}{Konuşma\ içeren\ toplam\ çerçeve\ adedi} \tag{5.2}$$

$$OVER = \frac{OVER\ oluşan\ çerçeve\ adedi}{Konuşma\ içermeyen\ toplam\ çerçeve\ adedi} \tag{5.3}$$

$$NDS = \frac{NDS\ oluşan\ çerçeve\ adedi}{Konuşma\ içermeyen\ toplam\ çerçeve\ adedi} \tag{5.4}$$

Önerilen hibrit VAD sisteminin değerlendirilmesinde kullanılan ses kayıtları gürültü etkisini en aza indirmek ve ayrıca bir gürültü azaltma algoritmasına başvurma ihtiyacını ortadan kaldırmak için gürültü seviyesi düşük ses kayıtları arasından seçilmiştir. Bu çalışmanın literatür özetinin bulunduğu Giriş kısmında da açıklandığı üzere, hem konuşmacı diyarizasyonu hemde konuşma aktivite tespit sistemlerine konuşmacının cinsiyetinin etkisi olabileceğinden, bu çalışmanın eğitim (training) aşamasında ve hibrit konuşma dedektörünün test (değerlendirme) aşamasında erkek cinsiyetine ait ses sinyalleri tercih edilmiştir. Tüm algoritma ve kodlar Google Colab online geliştirme editöründe yazılımıştır. Hibrit konuşma aktivite tespit sisteminin FEC, MSC, OVER ve NDS kapsamında değerlendirilebilmesi için kullanılan ses kayıtlarında, "böcek boncuk", "cumartesi yüreklilik", "çekilişten araba kazandım", "çabuk kaçalım, yoksa bizi öldürecek" gibi cümleler yer almaktadır. Ses sinyalleri çerçeve seviyesinde değerlendirilmiştir. Test gerçekleştirmek için kullanılan ses kayıtlarının "konuşma", "konuşma olmayan" ve toplam çerçeve sayıları Tablo 5.1'de verilmiştir.

Tablo 5.1. Her cümle (C) için her kayıttaki (K) toplam (T), konuşma içeren (K), konuşma içermeyen (KD) çerçeve sayısı.

			Kayıtlar									
			K1	**K 2**	**K3**	**K4**	**K5**	**K6**	**K7**	**K8**	**K9**	**K10**
Cümleler	**C1**	***T***	200	234	92	109	134	130	143	114	175	124
		K	36	41	35	40	38	41	39	39	38	35
		KD	164	176	57	69	96	89	104	75	137	89
	C2	***T***	217	233	113	115	130	141	138	132	166	132
		K	57	54	54	49	51	52	54	57	53	55
		KD	160	179	59	66	79	89	84	75	113	77
	C3	***T***	91	102	94	105	72	94	97	99	105	89
		K	49	47	40	48	41	45	45	47	43	42
		KD	42	55	54	57	31	49	52	52	62	47
	C4	***T***	128	104	118	107	91	104	102	121	122	134
		K	64	64	56	65	63	55	52	60	59	61
		KD	64	50	62	42	28	49	50	61	63	73

Tablo 5.1'deki sayılar çerçeve 44,1 KHz'lik, yani başka deyişle bir saniyesinde 44100 adet örnek içeren, ses sinyallerinin 0,03 saniye (30 milisaniye) uzunluğundaki çerçevelerinde bulunan örnek sayısıdır. FEC, MSC, OVER ve NDS değerleri 0,03, 0,02 ve 0,01 çerçeve uzunlukları için Tablo 5.2'deki gibi hesaplanmış olup Tablo 5.1 uzun tutulmaması adına 0,02 ve 0,01 saniye çerçeve uzunluğundaki örnek sayıları yazılmamıştır. Mevcut durumda Tablo 5.1'deki değerlere göre zaten 44,1 KHz örnekleme frekansı ve 0,03 saniye çerçeve boyutu göz önünde bulundurularak 0,02 ve 0,01 saniyelik çerçeve boyutundaki örnek sayıları kolayca türetilebilecektir.

Tablo 5.2. Hibrit VAD sisteminin farklı çerçeve boyutlarında sonuçları

		Değerlendirme metriği			
Çerçeve boyutu	**Gösterim**	**FEC**	**MSC**	**OVER**	**NDS**
0,01 saniye	Değer	0,0121	0,0723	1e-4	8e-6
(441 örnek)	Oran	%1,21	%7,23	%0,01	%0,0008
0,02 saniye	Değer	0,0245	0,0361	1.7e-4	5e-6
(882 örnek)	Oran	%2,45	%3,61	%0,017	%0,0005
0,03 saniye	Değer	0,0524	0,0085	2.9e-4	5e-6
(1323 örnek)	Oran	%5,24	%0,85	%0,029	%0,0005

6. SONUÇLAR

Bu tez çalışmasında konuşmacı diyarizasyon sistemleri için denetimli ve denetimsiz yaklaşımların birleştirildiği bir hibrit konuşma aktivite dedektörü (VAD) tabanlı konuşmacı diyarizasyon sistemi önerilmiştir. Geliştirilen özgün hibrit VAD yapısının denetimli kısmında zaman serileri analizlerinde kullanılan LSTM sinir ağından, denetimsiz kısmında ise eşikleme yaklaşımından faydalanılmıştır. Tablo 5.2'deki değerlerden çerçeve boyutu arttıkça FEC ve OVER hatalarının neredeyse iki kat yükseldiği kolay bir şekilde çıkarılabilmektedir. Bunun temel sebebi, hibrit VAD sisteminin denetimli kısmında kullanılan LSTM modelinin çerçeve uzunluğu arttıkça yanlış VAD kararı ürettiği olarak gözlenmiştir. Denetimli VAD kısmında kullanılacak eğitim veri setinin büyütülmesi ve LSTM ağının daha komplex (gelişmiş) hale getirilmesi ile bu dezavantajın önüne geçilebilmektedir. Tablo 5.2'den çıkarılabilecek bir başka sonuç ise ses sinyali çerçeve uzunluğu düştükçe MSC hata oranının yükseldiği görülmüştür. Bunun sebebi ise bazı Türkçe kelimeler içerisinde (ortalarında) geçen "s", "ş", "ç", "v", "f" ve "z" gibi fonasyonların (harf seslendirmelerinin) gürültü olarak algılanmasıdır. Sistemin çerçeve uzunluğu azaltıldıkça VAD sistemi daha agresif ve daha detaylı (kelime hatta hece seviyesinde) çalışmaktadır. Bu da bahsi geçen fonasyonların gürültü gibi algılanmasına sebep olmaktadır. Türkçe dili için /p/, /b/, /t/, /d/, /k/ ve /g/ patlamalı ünsüz sesleri ile başlayan kelimeler ses başlangıç süresi (Voice Onset Time-VOT) özelliğine sahip olmaktadır. VOT, patlamalı ünsüz üretimi esnasında tam oral daralmanın serbest bırakılması ile gırtlaksı titreşimlerin başlangıcı arasındaki süre olarak karakterize edilen anlık bir akustik parametredir [54]. Çerçeve boyutu arttırılması FEC hata oranını düşürürken diğer yandan gerçekte konuşma bölgesi olmasına rağmen VOT bölgesinin gürültü olarak yorumlanmasına sebep olmaktadır. Bu durum VAD sistemlerinin değerlendirilmesinde FEC hatası için kelime/cümle seçiminin ne denli önemli olduğunu kanıtlamaktadır. VAD sonucu ile çerçeve uzunluğu arasındaki ilişki şu şekilde özetlenebilir; çerçeve uzunluğu azaltıldıkça VAD sistemi daha agresif ve detaylı inceleme yapmaktadır. Diğer bir deyişle, VAD sisteminin kelimeler arasındaki hatta heceler arasındaki boşlukları elemesi isteniyorsa, çerçeve boyutu azaltılabilir. Tablo 6.1'de bu tez çalışması kapsamında önerilen özgün hibrit VAD sisteminin literatürde var olan bazı VAD sistemleri ile karşılaştırılması yapılmıştır. Daha sağlıklı bir karşılaştırma yapabilmek için karşılaştırma için kullanılan çalışmaların değerlendirme kriteri olarak FEC, MSC, OVER ve NDS'yi içermesine dikkat edilmiştir. Önerilen hibrit VAD yönteminin genel olarak Tablo 6.1'deki çalışmalardan daha başarılı olduğu gözlenmiştir. Ayrıca hibrit VAD sistemi içerisinde eşikleme ihtiyacı olan bir denetimsiz VAD yapısının kullanılması farklı akustik ortamlardaki ses sinyalleri için bir dezavantaj teşkil edebilir. Fakat önceki kısımlarda da açıklanadığı gibi LSTM derin öğrenme mimarisinin kullanılması bu dezavantajı en az seviyeye indirmektedir. LSTM tabanlı denetimli VAD sistemi sayesinde hibrit VAD sistemi farklı akustik ortamlardan alınan ses

sinyallerinde de sadece eşikleme tabanlı VAD sistemlerine nazaran çok daha başarılı bir şekilde çalışabilmektedir.

Tablo 6.1. Hibrit VAD ile literatürdeki bazı yöntemlerin karşılaştırılması

Yazar	Akustik Özellik	Değerlendirme Metriği	Sonuçlar
Javier R. [55]	Spektral zarf, Gürültü spektral ortalaması	FEC, MSC, OVER, NDS	FEC: 0,000 MSC: 0,073 OVER: 59,636 NDS: 10,629
Bachu R.G. [56]	Enerji, ZCR	FEC, MSC, OVER, NDS	FEC: 1,760 MSC: 4,817 OVER: 0,531 NDS: 1,949
Beritelli F. [57]	Diferansiyel ZCR, Spektral bozulma	FEC, MSC, OVER, NDS	FEC: 0,499 MSC: 0,029 OVER: 14,374 NDS: 0,000
Kirill S. [58]	Enerji	FEC, MSC, OVER, NDS	FEC: 19,239 MSC: 48,605 OVER: 0,000 NDS: 0,000
Hibrit VAD	Enerji, ZCR, MFCC	FEC, MSC, OVER, NDS	FEC: 0,0121 MSC: 0,0085 OVER: 0,01 NDS: 0,0005

ÖNERİLER

Bu tez çalışmasında hibrit konuşma aktivite dedektörü (VAD) kullanılarak d-vektör tabanlı bir konuşmacı diyarizasyon sistemi tasarlanmıştır. Çalışmanın literatüre en önemli katkısı özgün şekilde gerçekleştirilen hibrit VAD sistemidir. Bu sistem denetimli VAD içererek farklı akustik ortamlarda güçlü çalışabilme, denetimsiz VAD içererek tamamen eğitim veri setine bağlı kalmadan çalışabilen bir sistem olarak inşa edilmiştir. Tez çalışması süresince çıkarılan en önemli sonuç bu alanda geliştirecek herhangi bir sistemin dil ile alakalı bazı avantaj ve dezavantajları taşıyabilmesidir. Tamamen dilden bağımsız sistemler geliştirmek, yapılacak çalışmaların daha geniş kitlelerce kullanılabilmesine olanak sağlayacaktır. Araştırmacılar tarafından farklı verisetleri ve farklı değerlendirme metriklerinin kullanımı, bu alanda yeni çalışmaya başlayan araştırmacılar için bir dezavantaj olarak görülmektedir.

KAYNAKLAR

[1] M.-H. Siu, Y. George, H. Gish (1992). "*An unsupervised, sequential learning algorithm for segmentation for speech waveforms with multiple speakers*", in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 189–192.

[2] U. Jain, M. A. Siegler, S.-J. Doh, E. Gouvea, J. Huerta, P. J. Moreno, B. Raj, R. M. Stern (1996). "*Recognition of continuous broadcast news with multiple unknown speakers and environments*", in: Proceedings of ARPA Spoken Language Technology Workshop, pp. 61–66.

[3] H. Gish, M. . Siu, R. Rohlicek (1991). "*Segregation of speakers for speech recognition and speaker identification*", in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 873–876.

[4] S. S. Chen, P. S. Gopalakrishnan (1998). "*Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion*", in: Tech. Rep., IBM T. J. Watson Research Center, pp. 127–132.

[5] J. Ajmera, C. Wooters (2003). "*A robust speaker clustering algorithm*", in: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 411–416.

[6] S. E. Tranter, D. A. Reynolds (2004). "*Speaker diarisation for broadcast news*", in: Odyssey, 2004, pp. 337–344.

[7] D. A. Reynolds, P. Torres-Carrasquillo (2005). "*Approaches and applications of audio diarization*", in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 953–956.

[8] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet (2011). "*Front-end factor analysis for speaker verification*", IEEE Transactions on Audio, Speech, and Language Processing 19.

[9] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, C. Vair (2008). "*Stream-based speaker segmentation using speaker factors and eigenvoices*", in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4133–4136

[10] E. Variani, X. Lei, E. McDermott, I. L. Moreno, J. G-Dominguez (2014). "*Deep neural networks for small footprint text-dependent speaker verification*", in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4052–4056.

[11] G. Heigold, I. Moreno, S. Bengio, N. Shazeer (2016). "*End-to-end textdependent speaker verification*", in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5115–5119.

[12] Q.Wang, C. Downey, L.Wan, P. A. Mansfield, I. L. Moreno (2018). "*Speaker diarization with LSTM*", in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5239–5243.

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur (2018). "*Xvectors: Robust DNN embeddings for speaker recognition*", in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5329–5333.

[14] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, C. Wang (2019). "*Fully supervised speaker diarization*", in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6301–6305.

[15] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, S. Watanabe (2019). "*Endto- end neural speaker diarization with permutation-free objectives*", in: Proceedings of the Annual Conference of the International Speech Communication Association, pp. 4300–4304.

[16] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, S. Watanabe (2019). "*End-to-end neural speaker diarization with self-attention*", in: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, IEEE, pp. 296–303.

[17] Beatriz Martinez Gonzalez (2017). "*Analysis and Development of Robust Speaker Diarization for Meetings*", Doktora Tezi, Madrid Politeknik Üniversitesi, İspanya.

[18] Adedotun J Oseni-Adegbite (2020). "*Speaker Diarization in a Meeting Scenario*", Yüksek Lisans Tezi, Massachusetts Teknoloji Enstitüsü, A.B.D.

[19] Pantid Chantangphol (2020). "*Speaker Diarization in Broadcast News*", Yüksek Lisans Tezi, Thammasat Üniversitesi, Tayland.

[20] Yi Li (2020). "*Speaker Diarization System for Call-center data*", Yüksek Lisans Tezi, KTH Royal Teknoloji Enstitüsü, İsveç.

[21] Xavier Anguera Miro (2006). "*Robust Speaker Diarization for Meetings*", Doktora Tezi, Katalonya Politeknik Üniversitesi, İspanya.

[22] Nguyen Trung Hieu (2015). "*Speaker Diarization in Meetings Domain*", Doktora Tezi, Nanyang Teknoloji Üniversitesi, Singapur.

[23] Jose Maria Patino Villar (2020). "*Efficient Speaker Diarization and Low-latency Speaker Spotting*", Sorbonne Üniversitesi, Fransa.

[24] Tuomas Kaseva (2019). "*SphereDiar-an efficient speaker diarization system for meeting data*", Yüksek Lisans Tezi, Aalto Üniversitesi, Finlandiya.

[25] Mark Sinclair (2015). "*Speech Segmentation and Speaker Diarisation for Transcription and Translation*", Doktora Tezi, Edinburg Üniversitesi, İskoçya.

[26] Yufeng Yang (2020). "*Automatic Speaker Verification and Diarization on Voxceleb Data Collection*", Yüksek Lisans Tezi, Georgia Teknoloji Enstitüsü, A.B.D.

[27] Ahmed I.A., John P.C., David L.N., Mahmoud M.A. (2022). "*Channel and channel subband selection for speaker diarization*", Computer Speech & Language, 75 (101367).

[28] Meysam S., Anthony L., Loic B., Sylvain M., Yevheni P., Marie T., Ambuj M., Simon P., Olivier G., Samuel G., Andre A., Sebastien M., Marta R.C. (2023). "*Towards lifelong human assisted speaker diarization*", Computer Speech & Language, 77 (101437).

[29] Vijay K.K., Rajeswara R.R. (2022). "*Optimized speaker change detection approach for speaker segmentation towards speaker diarization based on deep learning*", Data & Knowledge Engineering, https://doi.org/10.1016/j.datak.2022.102121.

[30] Miquel I., Javier H., Jose A.R.F. (2023). "*Language modelling for speaker diarization in telephonic interviews*", Computer Speech & Language, 78 (101441).

[31] Or H.A., Yannick E., Chen H., Amit D., Itshak L. (2023). "*Speech and multilingual natural language framework for speaker change detecion and diarization*", Expert Systems with Applications, 213 (119238).

[32] Tae J.P., Naoyuki K., Dimitrios D., Kyu J.H., Shinji W., Shrikanth N. (2021). "*A Review of Speaker Diarization: Recent Advances with Deep Learning*", arXiv preprint arXiv:2101.09624v4.

[33] Quan W., Carlton D., Li W., Philip A.M., Ignaico L.M. (2017). "*Speaker Diarization with LSTM*", arXiv preprint arXiv:1710.10468v7.

[34] Li W., Quan W., Alan P., Ignacio LM. (2017). "*Generalized end-to-end loss for speaker verification*" arXiv preprint arXiv:1710.10467.

[35] Jort F.G., Daniel P.W.E., Dylan F., Aren J., Wade L., Moore C.R., Manoj P., Marvin R. (2017). "*Audio Set: An ontology and human-labeled dataset for audio events*", International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[36] Boersma, Paul (2001). "*Praat, a system for doing phonetics by computer*", Glot International 5:9/10, 341-345.

[37] Korkmaz Y., Boyacı A. (2023). "*Hybrid voice activity detection system based on LSTM and auditory speech features*", Biomedical Signal Processing & Control, 80-2 (104408).

[38] Antoniou A. (2005). "*Digital Signal Processing: Signals, Systems, and Filters*", McGraw-Hill, A.B.D.

[39] Ekaba B. (2019). "*Google Colaboratory*", Building Machine Learning and Deep Learning Models on Google Cloud Platform, pp. 59-64.

[40] Abadi M., Barham P., Chen J., Chen Z., Davis A. (2016). "*TensorFlow: A system for large-scale machine learning*", Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16), Google Brain.

[41] Chollet F. & others. (2015). *Keras*. https://keras.io.

[42] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G. (2019). "*PyTorch: An Imperative Style, High-Performance Deep Learning Library*", In: Advances in Neural Information Processing Systems 32.

[43] Lawrence R. Rabiner ve Ronald W. Schafer. (2007). "*Introduction to Digital Speech Processing*", Foundations and Trends® in Signal Processing: Vol. 1: No. 1–2, pp. 1-194, http://dx.doi.org/10.1561/2000000001.

[44] Ehsan V., Xin L., Erik M., Ignacio L.M., Javier G.D. (2014). "*Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification*", 2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP).

[45] Ronald R Coifman ve St´ephane Lafon. (2006). "*Diffusion maps*", Applied and computational harmonic analysis, vol. 21, no. 1, pp. 5–30.

[46] Lloyd, Stuart P. (1982). "*Least squares quantization in PCM*", Information Theory, IEEE Transactions on 28.2: 129-137.

[47] Oshry Ben-Harush, Ortal Ben-Harush, Itshak Lapidot ve Hugo Guterman. (2012). "*Initialization of iterative-based speaker diarization systems for telephone conversations*", IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no.2, pp. 414–425.

[48] Stephen H Shum, Najim Dehak, R´eda Dehak ve James R Glass. (2013). "*Unsupervised methods for speaker diarization: An integrated and iterative approach*", IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 10, pp. 2015–2028.

[49] Dimitrios Dimitriadis ve Petr Fousek. (2017). "*Developing on-line speaker diarization system*", in INTERSPEECH.

[50] Fabian Pedregosa, Alexandre Gramfort. "*Scikit-learn: Machine Learning in Python*".

[51] Ville S., Jeannie A., David I., Barath R. "*Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior*", University of Massachusetts, International Computer Science Institute, Berkeley, CA, Williams College, Williamstown.

[52] Wei X., Han L., Quan W., Anshuman T., Yiling H., Ignacio L.M., Hasim S. (2021). "*Turn-to-Diarize: Online Speaker Diarization Constrained by Transformer Transducer Speaker Turn Detection*", arXiv preprint arXiv:2109.11641v3.

[53] Freeman D.K., Cosier G., Southcott C.B., Boyd I. (1989). "*The voice activity detector for the PAN-European digital cellular mobile telephone service*", In: Internat. Conf. on Acoust. Speech Signal Process., 1, pp. 369–372.

[54] Lisker L., Abramson A.S. (1964). "*A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements*", Word 20, Taylor & Francis (Routledge), pp. 384-422.

[55] Ramirez J., Segura J.C., Benitez C., Torre A.D.L., Antonio R. (2004). "*Efficient voice activity detection algorithms using long-term speech information*", Speech Communication (4), pp. 271-287.

[56] Bachu R.G., Kopparthi S., Adapa B., Barkana B.D. (2009). "*Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy*", Advanced Techniques in Computing Sciences and Software Engineering, pp 279–282.

[57] Beritelli F., Casale S., Ruggeri G., Serrano S. (2002). "*Performance Evaluation and Comparison of G.729/AMR/Fuzzy Voice Activity Detectors*", IEEE Signal Processing Letters, 9(3).

[58] Kirill S. (2009). "*Dynamical Energy-Based Speech/Silence Detector for Speech Enhancement Applications*", Proceedings of the World Congress on Engineering, London (UK).

ÖZGEÇMİŞ

Yunus KORKMAZ

[redacted]

AKADEMİK FAALİYETLER

Makaleler (SCI, SCI-E):

1. Korkmaz Y., Boyacı A. (2023). "*Hybrid voice activity detection system based on LSTM and auditory speech features*", Biomedical Signal Processing & Control, 80-2 (104408).
2. Korkmaz Y., Boyacı A. (2022). "*milVAD: a bag-level MNIST modelling of voice activity detection using deep multiple instance learning*", Biomedical Signal Processing & Control, 74 (103520).
3. Korkmaz Y., Boyacı A. (2022). "*A Comprehensive Turkish Accent/Dialect Recognition System using Acoustic Perceptual Formants*", Applied Acoustics, 193 (108761).
4. Korkmaz Y., Boyacı A. (2020). "*Unsupervised and Supervised VAD systems using cobination of time and frequency domain features*", Bio. Sig. Proces. & Cont., 61 (102044).
5. Korkmaz Y., Boyacı A., Tuncer T. (2019). "*Turkish Vowel Classification based on acoustical and decompositional features optimized by Genetic Alg.*", Applied Acoustics, 154 (28-35).