KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

AKUSTİK VE PROSODİK ÖZNİTELİKLERE DAYALI OLARAK KONUŞMACILARIN YAŞ VE CİNSİYET GRUBUNA GÖRE SINIFLANDIRILMASI

DOKTORATEZİ

Bil. Yük. Müh. Ergün YÜCESOY

HAZİRAN 2017
TRABZON

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

AKUSTİK VE PROSODİK ÖZNİTELİKLERE DAYALI OLARAK KONUŞMACILARIN YAŞ VE CİNSİYET GRUBUNA GÖRE SINIFLANDIRILMASI

Bil. Yük. Müh. Ergün YÜCESOY

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsünce
"DOKTOR (BİLGİSAYAR MÜHENDİSLİĞİ)"
Unvanı Verilmesi İçin Kabul Edilen Tezdir.

Tezin Enstitüye Verildiği Tarih : 30 / 05 /2017
Tezin Savunma Tarihi : 30 / 06 /2017

Tez Danışmanı : Prof. Dr. Vasif V. NABİYEV

Trabzon 2017

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Bilgisayar Mühendisliği Anabilim Dalında
Ergün YÜCESOY Tarafından Hazırlanan

AKUSTİK VE PROSODİK ÖZNİTELİKLERE DAYALI OLARAK KONUŞMACILARIN YAŞ VE CİNSİYET GRUBUNA GÖRE SINIFLANDIRILMASI

başlıklı bu çalışma, Enstitü Yönetim Kurulunun 06 /06/2017 gün ve 1705 sayılı kararıyla oluşturulan jüri tarafından yapılan sınavda DOKTORA TEZİ olarak kabul edilmiştir.

Jüri Üyeleri

Başkan : Prof. Dr. Rıfat YAZICI .....

Üye : Prof. Dr. Vasif V. NABİYEV .....

Üye : Prof. Dr. İsmail Hakkı ÇAVDAR .....

Üye : Doç. Dr. İbrahim Yücel ÖZBEK .....

Üye : Doç. Dr. Mustafa ULUTAŞ .....

Prof. Dr. Sadettin KORKMAZ
Enstitü Müdürü

ÖNSÖZ

Konuşmaya dayalı yaş ve cinsiyet tanıma sistemleri adli, güvenlik, reklamcılık, pazarlama, insan-robot etkileşimi, çocuk haklarının korunması gibi birçok uygulama alanına sahiptir. Ancak kullanılan mikrofon, kayıt ve iletişim ortamı, seslendirilen metnin içeriği, kişinin ruhsal durumu, hasta veya yorgun olması gibi birçok faktör konuşma sinyalini değiştirir. Hatta bu faktörlerden hiç biri değişmese bile farklı zamanlarda seslendirilen iki konuşma sinyali birebir aynı olmaz. Ancak sağladığı kolay kullanım, düşük maliyet, yüksek güvenirlik ve uzaktan kullanılabilme gibi avantajlar konuşma dayalı tanıma sistemlerine olan ilgiyi artmıştır. Bu tez çalışmasında, yeni yaklaşımlar ortaya konularak konuşmacıları yaş ve cinsiyet grubuna göre otomatik olarak sınıflandıran yeni sistem tasarlanmıştır.

Bilgi ve deneyimleri ile beni destekleyen, yardımını esirgemeyen, danışmanım Sayın Prof. Dr. Vasif V. NABİYEV'e, değerli görüş ve önerileriyle çalışmalarıma katkıda bulunan değerli jüri üyelerine, başta Öğr. Gör. Mustafa KARA olmak üzere beni destekleyen tüm mesai arkadaşlarıma, beni bugünlere getiren ve her zaman desteğini hissettiğim annem ve babama, beni motive eden, her an yanımda olduğunu hissettiğim eşime, varlığı ve sevgisi ile bana güç veren, yaşam kaynağım biricik oğluma,

Sonsuz teşekkür ederim.

Ergün YÜCESOY

Trabzon 2017

TEZ ETİK BEYANNAMESİ

Doktora Tezi olarak sunduğum "Akustik ve Prosodik Özniteliklere Dayalı Olarak Konuşmacıların Yaş ve Cinsiyet Grubuna Göre Sınıflandırılması" başlıklı bu çalışmayı baştan sona kadar danışmanım Prof. Dr. Vasif V. NABİYEV'ın sorumluluğunda tamamladığımı, verileri/örnekleri kendim topladığımı, deneyleri/analizleri ilgili laboratuvarlarda yaptığımı/yaptırdığımı, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma sürecinde bilimsel araştırma ve etik kurallara uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim. 30/06/2017

Ergün YÜCESOY

İÇİNDEKİLER

Sayfa No

Doktora Tezi

ÖZET

AKUSTİK VE PROSODİK ÖZNİTELİKLERE DAYALI OLARAK KONUŞMACILARIN YAŞ VE CİNSİYET GRUBUNA GÖRE SINIFLANDIRILMASI

Ergün YÜCESOY

Karadeniz Teknik Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Prof. Dr. Vasif V. NABİYEV
2017, 165 Sayfa

Bu çalışmada konuşmacının yaş ve cinsiyet grubunun otomatik olarak belirlenmesi konusu ele alınmıştır. Başta ticari, medikal ve adli olmak üzere geniş bir uygulama alanına sahip olan otomatik yaş ve cinsiyet tanıma sistemleri doğrudan bir servisin seçiminde kullanılabileceği gibi farklı tanıma sistemlerinde ön işlem olarak da kullanılır. Ancak konuşma sinyali oldukça değişkendir ve başarılı bir sistemin gerçekleştirilmesi için konuşmayı etkileyen tüm faktörlerin değerlendirilmesi gerekir. Bu çalışmada ses işleme alanında kullanılan çeşitli öznitelik çıkarma ve sınıflandırma yöntemleri incelenerek bu yöntemlerle geliştirilen yaş ve cinsiyet sınıflandırma sistemlerinin performans değerlendirmeleri yapılmıştır. Her bir sistemin avantaj ve dezavantajları ortaya koyularak bu sistemler için en uygun model büyüklüğü, konuşma süresi, öznitelik boyutu gibi parametreler belirlenmiştir. Çalışmada, yaygın olarak kullanılan akustik ve prosodik özniteliklerin yanı sıra ses kaynağından çıkarılan parametrelerde incelenmiştir. Sınıflandırma yöntemi olarak dinamik zaman bükme, vektör nicemleme, Gauss karışım modeli (GMM), Destek Vektör Makineleri ve GMM süpervektörler kullanılmıştır. Çalışmada ayrıca 7 farklı alt sistemin skor seviyeli birleşimine dayanan yeni bir sistem önerilerek %5 civarında başarı artışı sağlanmıştır. Sıkıntı öznitelik projeksiyonu (NAP) yöntemi ile gerçekleştirilen kanal dengelemenin başarı üzerindeki etkisi ise %1.5 olmuştur.

Anahtar Kelimeler: Yaş ve cinsiyet tanıma, Akustik ve Prosodik öznitelik, Gauss karışım modeli, Ses kaynağı, Skor seviyeli birleşim, Süpervektör, Destek vektör makinesi

PhD. Thesis

SUMMARY

CLASSIFICATION OF SPEAKERS BASED ON ACOUSTIC AND PROSODIC FEATURES ACCORDING TO AGE AND GENDER GROUPS

Ergün YÜCESOY

Karadeniz Technical University
The Graduate School of Natural and Applied Sciences
Computer Engineering Graduate Program
Supervisor: Prof. Dr. Vasif V. NABİYEV
2017, 165 Pages

In this study, age and gender determination of a speaker is investigated. Automatic age and gender recognition systems having applications mainly in trade, medicine and forensic can directly be used for selection of a service or as an initial operation for different recognition systems as well. However, speech signal is quite variable. Therefore all factors affecting speech are required to realize a successful system. In this study by examining feature extraction and classification methods used in speech processing, performance evaluations of age and gender classification systems developed by these methods are carried out, pros and cons of each system are presented and the most suitable parameters such as model size, speech duration and feature size for these systems are determined. Beside, commonly used acoustic and prosodic features and parameters obtained from the voice source are also examined. Dynamic time warping, vector quantization, Gaussian mixture model (GMM), support vector machine, and GMM supervectors are used as classification methods. In the study, moreover, a new system based on score-level fusion of 7 subsystems is proposed and %5 success rate increase is achieved. The effect of channel compensation developed with nuisance attribute projection method on success rate became as 1.5%

Key Words: Age and gender recognition, Acoustic and prosodic features, Gaussian mixture model, Voice source, Score-level fusion, GMM supervector, Support vector machine

ŞEKİLLER DİZİNİ

Sayfa No

TABLOLAR DİZİNİ

Sayfa No

SEMBOLLER DİZİNİ

ACF : Otokorelasyon fonksiyonu (Autocorrelation Function)
ANN : Yapay sinir ağları (Artificial Neural Network)
ASR : Otomatik konuşma tanıma (Automatic Speech Recognition)
DCT : Ayrık kosinüs dönüşümü (Discrete Cosine Transform)
DTW : Dinamik zaman bükme (Dynamic Time Warping)
EM : Beklenti maksimumlaştırılması (Expectation Maximization)
FFT : Hızlı Fourier dönüşümü (Fast Fourier Transform)
GLDS : Genelleştirilmiş dorusal ayırtaç dizisi (Generalized Linear Discriminant Sequence Kernel)
GMM : Gauss karışım modeli (Gaussian Mixture Model)
HMM : Saklı Markov modeli (Hidden Markov Model)
kNN : *k*-en yakın komşuluk (k-Nearest Neighbour)
LDA : Doğrusal ayrıştırma analizi (Linear Discriminant Analysis)
LPC : Doğrusal öngörülü kodlama (Linear Predictive Coding)
LPCC : Doğrusal öngörü kepstrum katsayıları (Linear Prediction Cepstral Coefficients)
LVQ : Vektör nicemleme öğrenme (Learning Vector Quantization)
MAP : En büyük ardıl olasılık (Maximum A Posteriori)
MFCC : Mel Frekanslı Kepstral Katsayılar (Mel Frequency Cepstral Coefficients)
ML : Maksimum olabilirlik (Maximum Likelihood)
MLLR : En büyük olabilirlik doğrusal regresyon (Maximum Likelihood Linear Regression)
MLP : Çok katmanlı algılayıcı (Multilayer Perceptron)
NAP : Sıkıntı öznitelik projeksiyonu (Nuisance Attribute Projection)
PCA : Temel bileşenler analizi (Principal Component Analysis)
PLP : Algısal doğrusal öngörü (Perceptual Linear Prediction)
SV : Süper vektör (Super Vector)
SVM : Destek vektör makineleri (Support Vector Machines)
UBM : Genel arka plan modeli (Universal Background Model)
VAD : Ses etkinliği algılama (Voice Activity Detection)

VQ : Vektör nicemleme (Vector Quantization)

ZCR : Sıfır geçiş oranı (Zero Crossing Rate)

1. GENEL BİLGİLER

1.1. Giriş

Bilgisayarların günlük hayatımızdaki önemi arttıkça insan-makine iletişiminin önemi de her geçen gün artmaktadır. İnsan-makine iletişiminde genellikle klavye, mouse, dokunmatik ekran gibi araçlar kullanılır. Ancak bu araçlarla kurulan iletişim hem yavaş hem de taşınan bilgi bakımından oldukça sınırlıdır. Gelişen teknoloji ile birlikte gelecekte bu araçların yerini insanlar için en kolay ve etkin iletişim şekli olan konuşmanın alması muhtemeldir. Günümüzde ses teknolojilerinin kullanıldığı birçok ürün vardır. Ancak bu ürünlerin hiçbiri mükemmel değildir ve geliştirilmesi için çalışmalar devam etmektedir.

Konuşma sahip olduğu zengin boyutlu karakterinden dolayı en güçlü iletişim biçimi olarak değerlendirilir. Konuşma sinyali, iletilmek istenen mesaj bilgisi dışında konuşmacı hakkında yaş, cinsiyet, dil, aksan, psikolojik durum gibi kişisel bilgiler de içerir. Bu bilgiler insanlar ve çeşitli teknolojik uygulamalar tarafından değişik amaçlar için kullanır. Örneğin; bir telefon konuşmasında sadece sesini duyduğumuz bir kişinin yaşını, cinsiyetini, aksanını veya psikolojik durumunu tahmin ederek hitap şeklimizi ona göre belirleyebiliriz. Bu tahmin sürecinde konuşmacının doğum yeri, eğitim ve sosyo-ekonomik durum gibi özelliklerini ortaya çıkaran semantik, diksiyon, şive ve telaffuz bilgileri en üst seviyesinde kullanılır. Orta seviyede konuşmacının kişisel ve ana-baba etkilerini ortaya çıkaran prosodi, ritim, hız, tonlama ve modülasyon hacmi bilgileri, en düşük seviyede ise genizden okuma (nasality), nefeslilik (breathness) veya pürüzlülük (roughness) gibi sesin akustik yönleri kullanılır [1]. İnsanlar tarafından başarılı bir şekilde gerçekleştirilen bu tahmin süreci aslında oldukça zor bir iştir. Çünkü aynı şeyi söyleseler bile iki kişinin seslendirdiği konuşmalar farklı olacaktır. Hatta bu durum aynı konuşmacının farklı telaffuzları için bile geçerlidir. Bunun nedeni konuşma sürecinde kişinin hem fiziksel hem de zihinsel sistemlerinin birlikte kullanılmasıdır. Bu sistemler insanlar arasında faklıdır ve sürekli değişmektedir. Bu nedenle de mesaj aynı olsa bile konuşmalar farklı olacaktır.

Konuşma sinyalindeki konuşmacıya özel karakteristiklerin otomatik belirlenmesi başta ticari, medikal ve adli olmak üzere geniş bir uygulama alanına sahiptir [2-7]. Örneğin çok dilli bir çağrı merkezinde müşterinin dili tahmin edilerek çağrı o dil ile eşleşen temsilciye yönlendirilebilir [3]. En uygun temsilcinin belirlenmesinde otomatik

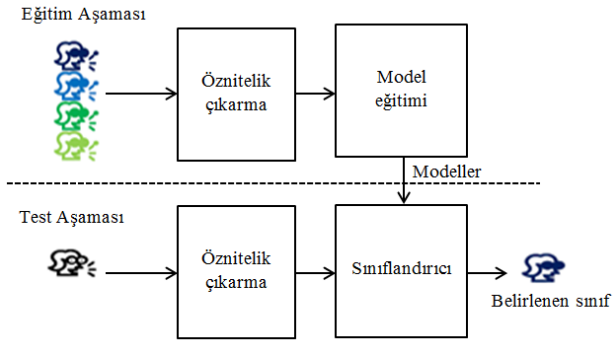
diyalekt/aksan tanıma sistemi de düşünülebilir. Böylece müşteri temsilci görüşmesinde yanlış anlaşılmalar önlenebilir. Ayrıca otomatik yaş tahmini de uygulanarak yaşlı müşterilerin yavaş konuşan temsilci tercihi sağlanabilir [7]. Son yıllarda sesli etkileşiminin oldukça arttığı internet üzerinden reklamcılık diğer bir uygulama alanıdır. Bu durum için kullanıcının sesinden elde edilecek dil/aksan, yaş ve cinsiyet bilgileri kullanıcıya uygun ürün ve servislerin sunulmasında kullanılabilir [7]. Video oyunlarında kullanıcı hakkındaki bilgiler oyunu kullanıcıya göre adapte etmeye yardımcı olabilir. Örneğin genç bir erkeğin müzik tercihi yetişkin bir kadından oldukça farklı olabilir. Konuşmacı karakteristikleri ayrıca otizm ve parkinson gibi farklı hastalıkların tanı, analiz ve takibi için kullanılabilir. Medikal alanında konuşma teknolojisine dayalı farklı uygulamalar vardır [8-15]. Konuşmacı karakteristiklerinin otomatik belirlenmesi otomatik konuşma tanıma sistemlerinin (ASR) performansını da arttırabilir. ASR sistemlerinin gerçek dünya uygulamalarında kullanımlarındaki temel zorluk yerli olmayan konuşmacılar için önemli performans düşüşünün olmasıdır [16, 17]. Bu nedenle aksan/diyalekt tanıma sistemleri bu sorunu ortadan kaldırmak için kullanılabilir. Birçok adli senaryo için de ses teknolojileri kullanılabilir. Bir bireyin kimliğini doğrulamak için değişik biyometrik teknikler kullanılır. Parmak izi, yüz karakteristikleri, el geometrisi, imza dinamiği ve ses örüntüsü bu tekniklerden bazılarıdır. Bir yöntemin seçimi onun belirli bir uygulama için güvenirliğine ve mevcut veriye bağlıdır. Bazı kriminal durumlarda mevcut deliller kayıtlı telefon görüşmeleri olabilir ve bu kayıtlar önemli bilgiler içerebilir [4]. Örneğin bir kişinin konuşma örüntüsü onun yaşı, cinsiyeti, diyalekti, duygusal veya psikolojik durumu ve belirli bir sosyal veya bölgesel grup üyesi olup olmadığı hakkında bilgi sağlayabilir. Bu nedenle kaçırma, tehdit çağrıları ve yanlış alarm gibi durumlarda konuşma verileri kullanılabilir.

1.1.1. Konuşmacı Sınıflandırma Sistemleri

Öncelikle konuşmacı sınıflandırma ile konuşmacı belirleme ve doğrulama olmak üzere ikiye ayrılan konuşmacı tanıma arasındaki farklar bilinmelidir. Konuşmacı belirleme konuşmacının sesinden elde edilen bilgilere dayalı olarak kimin konuştuğunun belirlenmesi işlemidir. Konuşmacı doğrulama bir konuşmacının kimlik iddiasının kabulü veya reddi sürecidir. Konuşmacı sınıflandırma ise bir konuşma örneğinin yaş, cinsiyet, aksan veya duygusal durum gibi belirli bir sınıfa atanması işidir. Konuşmacı sınıflandırma

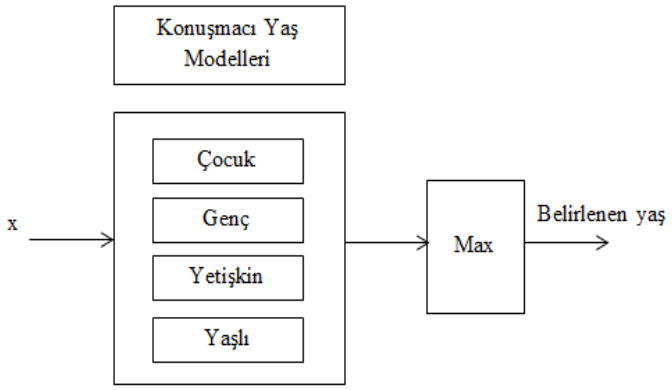
her sınıfın bir konuşmacı olduğu durumda konuşmacı belirleme olarak düşünülebilir. Örneğin cinsiyet sınıflandırma bir test konuşmasının bir erkek konuşmacıdan mı yoksa bir kadın konuşmacından mı olduğunun belirlenmesi olarak düşünülebilir.

Bir otomatik konuşmacı sınıflandırma sistemi iki aşamadan oluşur; eğitim aşaması ve test aşaması. Genel yapısı Şekil 1.1'de verilen otomatik konuşmacı sınıflandırma sitemlerinin eğitim aşamasında kayıtlı konuşmacılara ait ses sinyalleri işlenerek öznitelik olarak isimlendirilen vektörlere dönüştürülür.



Şekil 1.1. Otomatik konuşmacı sınıflandırma sistemlerinin yapısı

Daha sonra tüm sınıfların öznitelik vektörleri kullanılarak bir sınıflandırıcının konuşmacı sınıf modelleri eğitilir. Test aşamasında ise giriş ses sinyali eğitim aşamasındaki gibi öznitelik vektörlerine dönüştürüldükten sonra eğitilen sınıflandırıcıya giriş olarak uygulanır. Sınıflandırıcı tarafından her modelin skoru hesaplanır ve giriş en yüksek skoru veren modelle eşleştirilir (Şekil 1.2).



Şekil 1.2. Yaş sınıflandırma sisteminin yapısı

1.1.2. Cinsiyet ve Yaş Sınıflandırma Sistemleri

Konuşmaya dayalı sistemlerin artmasıyla birlikte konuşmacıların yaş ve cinsiyet gibi kişisel özelliklerine göre sınıflandırılması konusuna olan ilgide artmış durumdadır. Konuşmacının cinsiyet ve yaş özellikleri birçok uygulamada doğrudan ve dolaylı olarak kullanılmaktadır. Örneğin etkileşimli sesli yanıt sistemlerinde konuşmacının yaş ve cinsiyeti otomatik olarak tespit edilerek bu bilgiler istatistiksel amaçlarla veya bekleme kuyruğunda konuşmacıya uygun reklam veya müzik seçiminde kullanılabilir. Yaş ve cinsiyet bilgisinin kullanıldığı diğer bir alanda adli uygulamalardır. Adli birçok olayda şüphelilerin telefon konuşmaları delil olarak kullanılır. Bu durumda konuşmalar üzerinde yaş ve cinsiyet analizi yapılarak şüpheli sayısı azaltılabilir. Ayrıca altyazılı TV yayınlarında yazı rengi konuşmacının cinsiyetine göre değiştirilerek, işitme güçlüğü çeken insanların yazının hangi konuşmacıya ait olduğunu anlamalarına yardımcı olunabilir. Bu uygulamalar dışında cinsiyet bilgisi konuşmacı ve konuşma tanıma sistemlerinde ön bilgi olarak da kullanılır. Cinsiyet ön bilgisi konuşmacı tanıma sistemlerinde araştırma uzayını aynı cinsiyetli konuşmacılarla sınırlandırarak; konuşma tanıma sistemlerinde ise cinsiyet bağımlı modellerin tanımlanmasına imkan sunarak performans artışı sağlar. [18] çalışmasında cinsiyet bağımlı konuşma modellerinin cinsiyet bağımsız modellere kıyasla daha başarılı olduğu gösterilmiştir. Android cihazlarda kullanılan Google'ın en son konuşma tanıma sisteminde ve "Glass" ürününde konuşmacının cinsiyeti bulunduktan sonra konuşma tanıma işlemi yapılmaktadır. Daha önceden tek cinsiyetli modellerin kullanıldığı bu ürünlerde cinsiyet bağımlı modellerin kullanılmasıyla birlikte başarı oldukça artmıştır. Yakın zamanda piyasaya sürülen "Kinect" isimli bir online giyinme odası uygulamasında da kişinin konuşması kullanarak cinsiyeti belirlenmekte ve ona göre kıyafet sunulmaktadır.

Cinsiyet sınıflandırma sistemlerinin çoğunda yalnızca yetişkin konuşmacılar erkek ve kadın olarak sınıflandırılırken özellikle çocuk istismarı videolarının otomatik tespiti gibi uygulamalar için çocuk seslerinin de sınıflandırılması oldukça önemlidir. Ancak ergenlik öncesinde sesten cinsiyet ayrımı yapılamadığından çocuk konuşmacılar genellikle tek bir sınıfta değerlendirilmektedir. Günümüzde konuşmacıların yaş ve cinsiyet özelliklerine dayalı olarak geliştirilen sistemlerde farklı sınıf tanımları kullanılmaktadır. Bunlardan konuşmacıların cinsiyetlerine göre erkek, kadın ve çocuk olarak üç, yaşlarına göre çocuk, genç, yetişkin ve yaşlı olarak dört ve her iki özelliğine göre (yaş ve cinsiyet) ise 7 sınıfta

tanımlandığı üç kategori en yaygın kullanılan yaş-cinsiyet kategorileridir [19]. Ancak birçok çalışmada farklı sayıda sınıftan oluşan yaş ve cinsiyet kategorileri de kullanılmaktadır. Örneğin konuşmacıların cinsiyetlerine göre erkek ve kadın olarak iki, yaşlarına göre ise genç, yetişkin ve yaşlı olarak üç yada yetişkin ve yaşlı olarak iki sınıfta tanımlandığı çalışmalar da vardır.

1.1.3. Literatür Özeti

Konuşmacıların yaş ve cinsiyet özelliklerinin ses karakteristiği üzerindeki etkileri ile ilgili ilk çalışmalar 1950'li yıllarda başlamıştır [8]. Bu tarihten itibaren konuya olan ilgi devam etmesine rağmen konuşmacıları yaş ve cinsiyetlerine göre otomatik olarak sınıflandıran ilk gerçek sistemler yeni yeni geliştirilmeye başlanmıştır. Bu çalışmaların bazılarında yaş ve cinsiyet bilgileri birlikte kullanılırken bazılarında ise bu bilgiler ayrık bir şekilde ele alınmıştır. Bu tür sistemlerin performansını etkileyen birçok faktör vardır. Bu faktörlerin başında ses materyalinin metne bağımlı olup olmaması gelir. Metne bağımlı sistemlerde konuşmacı önceden belirlenmiş bir kelime veya ifadeyi seslendirirken, metinden bağımsız sistemlerde böyle bir kısıtlama yoktur. Metne bağımlı sistemlerde konuşmacı görevin farkındadır ve kişinin işbirliği ve tutarlılığı sonucunda metinden bağımsız sistemlere kıyasla daha yüksek başarı elde edilir. Performansı etkileyen bir diğer faktör de şablonların temsilinde kullanılan öznitelik kümesinin seçimidir. Seçilen öznitelik kümesi hem kolay hesaplanabilir hem de gürültüye karşı güçlü olmalıdır. Özellikle metinden bağımsız sistemlerde öznitelik kümesinin arka plan gürültüsünden etkilenmemesi gerekir. Çünkü bu sistemlerde konuşmacı çoğunlukla görevin farkında değildir ve çalışma ortamı gürültü ve dış müdahalelere açık olabilir.

Otomatik konuşma tanıma sistemleri başta olmak üzere birçok tanıma sisteminde ön işlem olarak uygulanan cinsiyet tanıma, cinsiyet bağımlı modellerin kullanılmasına imkan sunarak bu sistemlerin performansında önemli artışlar sağlamaktadır. Vergin ve arkadaşları tarafından yapılan çalışmada ilk iki formant frekansının konumuna dayalı bir cinsiyet sınıflandırıcı önerilmiştir [20]. ATIS (Air Travel Information System) veri kümesi üzerinde %85 sınıflandırma başarısı sağlayan bu sınıflandırıcının cinsiyet bağımlı konuşma modellerinin seçiminde kullanılmasıyla konuşma tanıma başarısında %14 artış sağlanmıştır.

Otomatik cinsiyet belirleme konusunda Parris ve Carey tarafından yapılan çalışmada ise her cinsiyet için HMM sınıflarını eşleştiren akustik analiz ile ortalama perde frekansı öznitelikleri doğrusal bir sınıflandırıcı kullanılarak birleştirilmiştir [21]. Geliştirilen sistemin Britanya İngilizcesi konuşan konuşmacıların ses kayıtlarıyla yapılan testlerinde %99'un üzerinde bir tanıma başarısı elde edilirken OGI veritabanındaki diğer 11 dile ait ses kayıtlarıyla yapılan testlerde ise en yüksek hata oranı %5.2, ortalama hata oranı ise %2 olarak tespit edilmiştir.

Slomka ve Sridharan tarafından yapılan çalışmada çeşitli ses parametreleri (Mel-tabanlı kepstral katsayılar, yansıma katsayıları, otokorelasyon katsayıları ve logaritmik alan oranı) kullanılarak eğitilen GMM çiftleri ile iki farklı yöntemle (AEP ve AEPSP) tahmin edilen ortalama perde frekansı öznitelikleri doğrusal bir sınıflandırıcı kullanılarak birleştirilmiştir [22]. Çalışmada bu 6 bilgi kaynağının tüm olası kombinasyonları ile oluşturulan toplam 63 farklı otomatik cinsiyet belirleme (AGI) sistemi OGI ve Switchboard veri tabanlarından seçilen 6 farklı veri kümesi ile test edilmiştir. Bu testlerde seçilen veri kümesine bağlı olarak, geliştirilen 63 AGI sisteminin ortalama doğruluğu %91.19 ile %95.42 arasında hesaplanmıştır.

Zeng ve arkadaşları tarafından yapılan çalışmada erkek ve kadın ses karakteristiklerinin modellenmesi için perde frekansı ve RASTA-PLP katsayılarının birleşimi ile oluşturulan GMM'ler kullanılmıştır [23]. Çalışmada farklı kovaryans matrisli ve bileşen sayılı GMM'ler eğitilmiş ve geliştirilen sistem temiz, gürültülü ve çok dilli ortamlarda test edilmiştir. Bu testler sonucunda önerilen yöntem için 4-8 bileşenli GMM'lerin yeterli olduğu ve geliştirilen cinsiyet sınıflandırıcının dilden bağımsız ve gürültüye karşı güçlü olduğu görülmüştür. Çalışmada temiz konuşmalar %98.1 doğrulukta sınıflandırılırken, 0 dB gürültülü konuşmalar ise %95 doğrulukta sınıflandırılmıştır.

Hu ve arkadaşları tarafından yapılan çalışmada perde frekansına dayalı olarak geliştirilen iki seviyeli bir cinsiyet belirleme sistemi önerilmiştir [24]. Önerilen sistemin ilk aşamasında eşik seviyeye dayalı basit bir kurala göre perde frekansından cinsiyet ayrımı yapılabilen konuşmacılar belirlenmiştir. İkinci aşamada ise şüpheli (ilk aşamada cinsiyetine karar verilemeyen konuşmacılar) konuşmacılar için GMM'ye dayalı ikinci bir inceleme yapılmış ve bu sayede hem yüksek doğruluk hem de düşük karmaşıklık sağlanmıştır. Önerilen sistem TIDIGITS veri kümesi kullanılarak test edilmiş, bu testler sonucunda konuşmacıların %98.65 doğrulukta sınıflandırdığı ve yaklaşık 5 saniye

uzunluğundaki konuşmaların öznitelik çıkarma ve sınıflandırma için yeterli olduğu görülmüştür.

Harb ve Chen tarafından yapılan çalışmada perde frekansı ve spektral özniteliklerin perceptron sınıflandırıcı ile birlikte kullanıldığı bir cinsiyet belirleme sistemi önerilmiştir [25]. Önerilen sistemin 4 Fransız ve 1 İngiliz radyo istasyonun kayıtlarıyla yapılan testlerinde ortalama sınıflandırma başarısı %93, switchboard veritabanı ile yapılan testlerinde ise %98.5 olarak verilmiştir.

Shue ve Iseli tarafından yapılan çalışmada ses kaynağı ile ilişkili akustik ölçümlerinin otomatik cinsiyet sınıflandırmada üzerindeki etkisi incelenmiştir [26]. Bu amaçla SVM'ler kullanılarak geliştirilen sistem, yaşları 8 ile 39 arasında değişen 205 erkek ve 160 kadın konuşmacının 3880 konuşmasından çıkarılan format frekansları, format bant genişlikleri, açıklık oranı (open quotient) ve spektral eğim (spectral tilt) ölçümleri ile test edilmiştir. Yaşı 16 ile 39 arasında değişen konuşmacıların ses yolu parametreleri ile yapılan testlerinde sınıflandırma başarısı %95.15 olarak verilmiştir.

Ting ve arkadaşları tarafından yapılan çalışmada cinsiyet tanıma performansını arttırmak için yeni bir yöntem önerilmiştir [27]. MFCC kullanılarak eğitilen iki GMM modelinin (erkek ve kadın) olabilirlik skorlarıyla ses dizisinin tüm çerçevelerinden hesaplanan ortalama perde frekansının doğrusal olarak birleştirilmesi fikrine dayanan bu yöntem SRMC veritabanı ile test edilmiş ve bu testler sonucunda en yüksek hata oranı %3.3 olarak bulunmuştur.

Djemili ve arkadaşları tarafında yapılan çalışmada dört faklı sınıflandırıcı (GMM, MLP, VQ ve LVQ) ve MFCC katsayıları kullanılarak geliştirilen cinsiyet belirleme sistemleri IViE veri kümesi kullanılarak test edilmiştir [28]. Her konuşmacının yalnızca 1 saniyelik konuşmaları kullanılarak yapılan bu testlerde en yüksek tanıma başarısı %96,4 olarak 128 noktalı VQ yöntemi ile sağlanmıştır.

Gaikwad ve arkadaşları tarafından yapılan çalışmada perde frekansı, azalma (roll off) ve enerji gibi cinsiyet bağımlı özniteliklerle MFCC'nin kombinasyonuna dayanan yeni bir cinsiyet sınıflandırma sistemi SVM ile geliştirilmiştir [29]. Geliştirilen sistem erkek konuşmacıları %93.22 doğrulukta sınıflandırılırken, kadın konuşmacıların sınıflandırma başarısı ise %86.90 olarak elde edilmiştir. Farklı yaş grupları üzerinde yapılan testlerde ise en yüksek cinsiyet sınıflandırma başarısı %95 ile 25-30 yaş grubunda, en düşük başarı ise %89 ile 20-23 yaş grubunda sağlanmıştır.

Phoophuangpairoj ve arkadaşları tarafından yapılan çalışmada [30] farklı ses tonlarına sahip hecelerden konuşmacının cinsiyetini tanıyabilen iki seviyeli bir yöntem önerilmiştir. Ortalama perde frekansı, MFCC öznitelikleri ve HMM kullanılarak gerçekleştirilen bu yöntemin ilk aşamasında cinsiyet tanıma için konuşmacıların ortalama perde frekansı kullanılmıştır. İlk aşamada tam olarak sınıflandırılamayan konuşmacılar ise ikinci aşamaya iletilmiş ve burada MFCC öznitelikleri, erkek ve kadın konuşmacılar için tanımlanan akustik hece modelleri ve gramerler kullanılarak sınıflandırma işlemi tamamlanmıştır. Çalışmada elde edilen deneysel sonuçlara göre önerilen cinsiyet sınıflandırma sistemi ile konuşmacılar %98,92 doğrulukta cinsiyetlerine göre sınıflandırılmıştır. Ayrıca önerilen yöntem perde frekansı ve ANN kullanılarak gerçekleştirilen geleneksel iki yöntemle karşılaştırılmış ve bu karşılaştırma sonucunda önerilen yöntemin daha başarılı olduğu görülmüştür.

Bakir tarafından yapılan çalışmada Almanca için geliştirilen bir cinsiyet tanıma sistemi tanıtılmıştır [31]. 50 erkek ve 50 kadın konuşmacı tarafından seslendirilen yaklaşık 2658 ses kaydından oluşan bir veri kümesinin kullanıldığı çalışmada farklı sayıda (1, 3, 5 ve 9) MFCC katsayıları ile temsil edilen konuşma sinyalleri üç farklı yöntemle (ANN, HMM ve DTW) sınıflandırılmıştır. Yapılan testlerde kadın konuşmacıların erkek konuşmacılara kıyasla daha başarılı sınıflandırıldığı ve kullanılan öznitelik vektörünün boyutu arttıkça sınıflandırma sistemin başarı oranının arttığı görülmüştür. Çalışmada %98.34 kadın ve %97.02 erkek tanıma oranı ile 9 MFCC katsayılı HMM sistemi en başarılı cinsiyet tanıma sistemi olurken onu %87.37 kadın ve %86.33 erkek tanıma oranı ile DTW sistemi takip etmiştir.

Konuşmaya dayalı tanıma sistemlerde yaygın olarak kullanılan bir diğer bilgide konuşmacının yaşıdır. Bu bilginin otomatik olarak belirlenmesi konusunda yapılan çalışmalarda genellikle iki farklı yaklaşım kullanılır; yaş-grubu sınıflandırma ve yaş regresyonu. Yaş-grubu sınıflandırmada konuşmacılar önceden belirlenmiş yaş gruplarından birine atanırken, yaş regresyonunda ise konuşmacıların yaşı yıl olarak tahmin edilir. Başta çağrı merkezleri olmak üzere birçok uygulamada doğrudan veya dolaylı olarak kullanılan konuşmacı yaşının otomatik olarak belirlenebilmesi konusunda birçok çalışma yapılmıştır. Bu çalışmalarda genellikle konuşmacılar çocuk, genç, yetişkin ve yaşlı olmak üzere dört yaş grubuna sınıflandırılırken cinsiyet ve yaş gruplarının birlikte kullanıldığı çalışmalarda yapılmıştır.

Minematsu ve arkadaşları tarafından yapılan çalışmada konuşmacıları yaşlı ve yaşlı olmayan şeklinde iki gruba ayıran yeni bir yöntem önerilmiştir [10]. MFCC ve delta regresyon katsayılarının GMM ile birlikte kullanıldığı çalışmada 43 yaşlı ve aynı sayıda yaşlı olmayan konuşmacı seçilmiş ve her birisi 5 saniye olan cümleler kullanılarak testler yapılmıştır. Bu testler sonucunda önerilen sınıflandırma sisteminin konuşmacıları %90 başarı ile sınıflandırdığı görülmüştür. Çalışmada ayrıca mevcut özniteliklere konuşma oranı ve yerel güç dalgalanma katsayıları eklenmiş ve bu parametrelerin eklenmesiyle başarı oranın %95.3'e çıkarılmıştır.

Shafran ve arkadaşları tarafından yapılan çalışmada MFCC ve temel frekans öznitelikleri ile HMM tabanlı sınıflandırıcılar kullanılarak konuşmacıları 5 farklı yaş grubuna ($< 25, \approx 25, 26-50, \approx 50\ ve\ > 50$) sınıflandıran yeni bir sistem önerilmiştir [13]. Bu sistemin eğitim için bir müşteri hizmetleri sisteminin 1854 telefon görüşmesine ait toplam 5147 cümleden oluşan bir veritabanı kullanılmıştır. %65'i erkek, %35'i ise kadın konuşmacılara ait olan bu veritabanı ile yapılan testlerde tüm test örneklerinin en olası sınıfa atanmasıyla oluşturulan sınıflandırıcının başarısı (%33.3) taban seviye olarak kabul edilmiştir. Yapılan testler sonucunda önerilen sistemin başarısı yalnız kepstral katsayıların kullanımı ile %68.4, temel frekans ve kepstral katsayıların birlikte kullanımı ile %70.2 olarak elde edilmiştir.

Müller ve arkadaşları tarafından yapılan ve konuşmacıların yaş (yaşlı, yaşlı değil) ve cinsiyet (erkek, kadın) gruplarına göre sınıflandırıldığı çalışmada [12] beş farklı sınıflandırma yönteminin performans karşılaştırması yapılmıştır; Karar ağacı (DT), Yapay sinir ağları (ANN), k-en yakın komşuluk (kNN), Naive Bayes (NB) ve SVM . Çalışmada temel frekans mikro dalgalanmaları (jitter ve shimmer) otomatik olarak çıkarılmış ve akustik öznitelik olarak kullanılmıştır. 347'si 60 yaş üstü, 46'sı ise 60 yaş altı olmak üzere toplam 393 konuşmacıdan oluşan veritabanı ile yapılan testlerde 6 yöntemin de sınıflandırma başarısı her zaman en sık rastlanan sınıf seçilerek belirlenen taban seviyeden (Yaşlı:%88, erkek:%59) yüksek olmuştur. Çalışmada ANN yöntemi %96.6 yaş ve %81.1 cinsiyet tanıma oranı ile en başarılı yöntem olarak verilmiştir. Bu yöntem Müller tarafından daha da geliştirilmiş ve konuşmacıların cinsiyetine göre 2, yaşına göre 4 sınıfa (çocuk, genç, yetişkin ve yaşlı) ayrıldığı [32] çalışması yapılmıştır. Bu çalışmada önceki çalışmada kullanılan 5 makine öğrenme yöntemine ilave olarak GMM modeli ve jitter ve shimmer özniteliklerine ilave olarak da F0, HNR, konuşma oranı, duraklama süresi ve frekans öznitelikleri kullanılmıştır. Çoğunluğu çocuk ve yaşlı olmak üzere 507 kadın ve

657 erkek konuşmacı ile yapılan testlerde 4 yaş sınıfı için en yüksek başarı (63.5%) ANN yöntemi ile elde edilmiştir.

Metze ve arkadaşları tarafından yapılan çalışmada konuşmacıları yaş ve cinsiyetlerine göre 7 sınıfa ayıran dört faklı yaklaşım insan performansıyla karşılaştırılmıştır [33]. Bu yaklaşımların ilkinde otomatik dil tanıma sisteminden türetilen bir paralel fonem tanıyıcı kullanılırken ikincisinde dinamik Bayes ağı ile farklı prosodik özniteliklerin birleştirildiği bir sistem kullanılmıştır. Doğrusal öngörü analizine dayalı bir sistem üçüncü, MFCC'ye dayalı GMM modelleri ise dördüncü yaklaşım olarak kullanılmıştır. 4000 Alman konuşmacının bir kayıt sistemini arayarak seslendirdiği bir dizi sayı, kelime ve cümleden oluşan German SpeechDat II veritabanı ile yapılan testlerde bu yaklaşımların başarısı sırasıyla %54, %40, %27 ve %42 olarak verilmiştir. Çalışmada paralel fonem tanıyıcıya dayalı sistemin insana yakın bir performans sunduğu ancak bu yaklaşımın performansının cümle uzunluğuna bağımlı olduğu, prosodik özniteliklere dayalı sistemin performansının ise cümle uzunluğundan çok az etkilendiği belirtilmiştir.

Bocklet ve arkadaşları tarafından yapılan çalışmada okul öncesi ve ilkokul çağındaki çocukların yaşlarının otomatik olarak belirlenmesi konusu ele alınmıştır [34]. Bu amaçla her çocuk için bir GMM modeli eğitilmiş (UBM'den MAP yöntemiyle uyarlanarak) ve bu modele karşı gelen ortalama süpervektörler SVM'e uygulanarak sınıflandırma ve regresyon işlemi yapılmıştır. Çalışmada eğitim için üç farklı veri kümesinden toplam 212 konuşmacının ses kayıtları kullanılırken test için 100 konuşmacıdan oluşan bir veri kümesi kullanılmıştır. Yapılan testler sonucunda önerilen sınıflandırma sistemi ile konuşmacıların yaş gruplarına (<7, 7, 8, 9+10 ve >10) göre %83 doğrulukta sınıflandırıldığı, yaş regresyonunun ise ortalama 0.8, maksimum 3 yıl hata ile yapıldığı görülmüştür.

Bocklet ve arkadaşları tarafından yapılan diğer bir çalışmada ise otomatik yaş ve cinsiyet tanıma işi için 5 farklı sistem karşılaştırmalı olarak incelenmiştir [35]. Bu sistemlerin üçünde MFCC, PLP ve TRAPS özniteliklerinden elde edilen GMM süpervektörleri kullanılırken diğer ikisinde ise prosodik ve glottal uyarım öznitelikleri kullanılmıştır. SVM sınıflandırıcısı kullanılarak geliştirilen bu 5 sistem aGender veri kümesi ile test edilmiş ve bu testler sonucunda bireysel başarısı en yüksek sistem %42.4 başarı oranı ile GMM-MFCC sistemi olmuştur. Çalışmada ayrıca iki farklı birleştirme yaklaşımı (skor ve öznitelik seviyeli) kullanılarak bu sistemlerin farklı kombinasyondaki birleşimleri de incelenmiştir. Yapılan testlerde en yüksek sınıflandırma başarısı 5 sistemin skor seviyeli birleştirmesi sonucunda %47.8 olarak bulunmuştur.

Heerden (2010) ve arkadaşları tarafından yapılan çalışmada yaş/cinsiyet sınıflandırma işi için yeni bir yöntem önerilmiştir [16]. Önerilen bu yönteme göre önce konuşmacının yaşı SVM regresyonu ile tahmin edilmiş, daha sonra ise iyi eğitilmiş cinsiyet sınıflandırıcılarının önsel olasılıkları ile tahmin edilen yaş bilgisi birleştirilerek 7 sınıflı yaş/cinsiyet sınıflandırma işi tamamlanmıştır. Çalışmada sınıflandırıcı ve regresörlerin eğitimi için iki farklı öznitelik türü kullanılmıştır; uzun ve kısa süreli öznitelikler. Uzun süreli öznitelikler perde frekansı, jitter, shimmer ve güç gibi öznitelikleri içeren 22 elemanlı bir kümeden seçilirken, kısa süreli öznitelikler ise MFCC katsayıları ile eğitilen GMM süpervektörlerinden seçilmiştir. Yaklaşık 700 Alman konuşmacının telefon konuşmalarını içeren bir veri kümesi ile yapılan testlerde standart 7-sınıflı sınıflandırıcının başarı oranı uzun süreli öznitelikler için %45.7, süpervektör öznitelikleri için %45 olarak elde edilirken önerilen yöntemle başarı oranı %50.7'ye çıkarılmıştır.

Porat ve arkadaşları tarafından yapılan çalışmada GMM ve SVM sınıflandırıcılarının eşsiz bir kombinasyonuna dayanan yeni bir yaş tanıma sistemi önerilmiştir [36]. Önerilen bu sistemde MAP uyarlamasıyla hesaplanan ortalama süpervektörleri yerine her konuşmacıyı temsil eden Gaussian ağırlık süpervektörleri hesaplanmış ve bu vektörler SVM sınıflandırıcısı için öznitelik olarak kullanılmıştır. Çalışmada iki veri kümesi (Hebrew ve aGender) kullanılarak testler yapılmış ve bu testler sonucunda önerilen yaklaşımın 4 gruplu (çocuk, genç, yetişkin ve yaşlı) yaş sınıflandırma başarısı sırasıyla %53.75 ve %56.18 olarak bulunmuştur. Düzenleyiciler tarafından sağlanan ve oldukça kısa ses bölümlerinden oluşan veri kümesiyle (test dosyaları) yapılan test sonucunda ise önerilen yaş tanıma sisteminin başarısı %43.08 olarak bulunmuştur.

Meinede ve Isabel tarafından yapılan çalışmada konuşmacıları yaş ve cinsiyete göre sınıflandıran iki ayrı sistem sunulmuştur [37]. Bu sistemlerden yaş sınıflandırma sistemi 4 alt sistemin, cinsiyet sınıflandırma sistemi ise 6 alt sistemin birleşiminden oluşmaktadır. Uzun ve kısa süreli akustik ve prosodik öznitelikler, farklı sınıflandırma stratejileri (GMM-UBM, MLP ve SVM) ve dört farklı konuşma veritabanı kullanılarak geliştirilen bu alt sistemlerin birleştirilmesi için çok sınıflı doğrusal lojistik regresyon yöntemi kullanılmıştır. aGender veritabanının geliştirme bölümü ile yapılan testlerde önerilen yaş sınıflandırma sistemin başarı oranı %51.2, cinsiyet sınıflandırma sisteminin başarı oranı ise %83.1 olarak bulunmuştur. Temel seviye olarak verilen sistemle [19] karşılaştırıldığında önerilen yaş sınıflandırma sisteminin %4.1, cinsiyet sınıflandırma sisteminin ise %5.8 daha başarılı

olduğu görülmektedir. Önerilen yaş ve sınıflandırma sisteminin test kümesi ile başarısı ise sırasıyla %48.7 ve %84.3 olarak verilmiştir.

SVM yaş ve cinsiyet sınıflandırma problemlerinin çözümü için kullanılan güçlü bir araçtır. Ancak bu yöntem gürültü ve aşırılıklara karşı oldukça hassastır. Nguyen ve arkadaşları tarafından yapılan çalışmada bu sorunun ortadan kaldırılması için eğitim noktalarının eşit olarak değerlendirilmemesi gerektiği varsayımına dayanan yeni bir fuzzy SVM yaklaşımı önerilmiştir [38]. Çalışmada eğitim noktalarının eşit değerlendirilmemesi için her bir eğitim noktasına atanan bir fuzzy üyelik ağırlığı kullanılmıştır. Önerilen fuzzy SVM yaklaşımı aGender veritabanı ile test edilmiş ve bu testlerde hem yaş hem de cinsiyet sınıflandırmada başarı artışı sağlandığı görülmüştür. Çalışmada önerilen yaş sınıflandırma sisteminin başarısı %49.06, cinsiyet sınıflandırma sisteminin başarısı ise %81.65 olarak verilmiştir.

Kockmann ve arkadaşları tarafından yapılan çalışmada 6 alt sistemin çok sınıflı lojistik regresyon ile birleştirilmesine dayalı bir yaş ve cinsiyet sınıflandırma sistemi tanıtılmıştır [39]. Bu alt sistemlerin birinde resmi openSMILE öznitelikleri ve SVM sınıflandırıcısı kullanılırken dört alt sistemde ise MFCC ve delta öznitelikleri kullanılmıştır. MFCC'ye dayalı alt sistemlerin ilkinde UBM'den uyarlanarak elde edilen sınıfa özgü GMM'ler kullanılmıştır. Bu sistemde UBM'nin eğitimi için EM algoritması, uyarlama için MAP yöntemi kullanılmıştır. MFCC'ye dayalı ikinci sistemde ise sınıfa özgü GMM'lerin eğitiminde doğrudan ML (Maximum Likelihood) yöntemi kullanılmıştır. Üçüncü alt sistemde JFA temelli bir konuşmacı tanıma sisteminden [40] tahmin edilen hiper-parametreler, dördüncü alt sistemde ise JFA temelli konuşmacı tanıma sisteminin [40] skorları kullanılmıştır. Son alt sistem ise PLP özniteliklerine dayalı bir GMM-SVM sistemidir. Önerilen birleşik sistem aGender veri kümesinin değerlendirme bölümü ile test edilmiş ve bu testler sonucunda önerilen yaş sınıflandırma sisteminin başarısı %52.35, cinsiyet sınıflandırma sisteminin başarısı ise %83.14 olarak verilmiştir.

Li ve arkadaşları tarafından yapılan çalışmada akustik ve prosodik seviyede 7 farklı yöntemin birleştirildiği yeni bir yaş ve cinsiyet belirleme yaklaşımı sunulmuştur [4]. (1) MFCC özniteliklerine dayalı GMM sistemi, (2) GMM ortalama süpervektörlerine dayalı SVM sistemi ve (3) 450 boyutlu cümle seviyeli özniteliklere (OpenSMILE) dayalı SVM sistemi çalışmada kullanılan üç temel alt sistemdir. Çalışmada bu temel alt sistemlere ilave olarak dört yeni alt sistem daha önerilmiştir. Bu alt sistemlerden ilki GMM-UWPP-SVM sistemi, ikincisi GMM–UWPP–Sparse representation sistemi, üçüncüsü GMM-MLLR-

SVM sistemi ve dördüncüsü ise SVM-Prosody sistemidir. Önerilen dört alt sistemle farklı yaş ve cinsiyet gruplarının sınıflandırmasında etkin ve rekabetçi sonuçlar elde edilmiştir. Çalışmada genel sınıflandırma performansının daha da arttırılması için 7 alt sistemin skor seviyesinde birleştirilmesine dayalı bir sistem önerilmiştir. aGender veritabanının geliştirme bölümü ile yapılan testlerde önerilen birleşik sistemin SVM tabanlı temel sisteme (3) kıyasla cinsiyet sınıflandırmada %4.2, yaş sınıflandırmada ise %5.6 daha başarılı olduğu görülmüştür. Önerilen birleşik sistemin resmi test kümesi üzerindeki sınıflandırma başarısı ise yaş için %52.0, cinsiyet için %85.0 olarak verilmiştir.

Safavi ve arkadaşları tarafından yapılan çalışmada OGI çocuk veritabanı ve GMM-UBM, GMM-SVM ve i-vektör sistemleri kullanılarak çocuk konuşmaları için elde edilen yaş grubu tanıma sonuçları sunulmuştur [41]. Çalışmada yaşla ile ilgili önemli bilgileri içeren spektrum bölgelerinin belirlenmesi için çeşitli deneyler yapılmış ve bu deneyler sonucunda çocuk konuşmacıların yaş grubunun (AG1:5-9 yaş, AG2:9-13 yaş, AG3:13-16 yaş) belirlenmesinde 5.5 kHz'nin üzerindeki frekansların en az öneme sahip olduğu görülmüştür. Yaşları 5 ile 16 arasında değişen 1100 konuşmacıdan oluşan OGI çocuk veritabanından rastgele seçilen 334 konuşmacı sistemin eğitimi için kalan 766 konuşmacı ise testi için kullanılmıştır. Yapılan testlerde GMM-UBM ve i-vektör sistemleri ile GMM-SVM sistemine kıyasla daha iyi yaş grubu tanıma performansı sağlanırken en iyi performans ise i-vektör sisteminin 5.5 kHz bant limitli konuşmalara uygulanması sonucunda %85.77 olarak elde edilmiştir.

Oscal ve arkadaşları tarafından yapılan çalışmada uyarım yoğunluğuna göre düşük ve yüksek uyarımlı olarak sınıflandırılan konuşma sinyalleri için farklı özniteliklerin seçilmesine dayalı bir yaş ve cinsiyet tanıma sistemi önerilmiştir [42]. Uyarım seviyesinin belirlenmesinde konuşma çerçevelerinden çıkarılan perde ve enerji öznitelikleri ile oluşturulan Gaussian olasılık yoğunluk fonksiyonları, yaş ve cinsiyet sınıflarının belirlenmesinde ise SVM sınıflandırıcısı kullanılmıştır. Konuşmacıların uyarım seviyesi belirlendikten sonra her uyarım seviyesi ve sınıflandırma görevine (yaş ve cinsiyet) göre farklı bir öznitelik kümesi seçilmiş ve seçilen öznitelik kümesi kullanılarak konuşmacılar yaş ve cinsiyet gruplarına göre sınıflandırılmıştır. Özniteliklerin seçiminde 10 farklı öznitelik (jitter, shimmer, süre, sıfır geçiş oranı (ZCR), ZCR tepesi, temel frekans, spektral yayılma, spektral merkez nokta, formant frekansları MFCC katsayıları) ve bu özniteliklerden türetilen katsayılarla oluşturulan 37 elemanlı bir küme göz önünde bulundurulmuştur. Konuşmacıların cinsiyetlerine göre 2, yaşlarına göre 3 sınıfa ayrıldığı

çalışmada çok dilli Lwazi veritabanı kullanılarak çeşitli testler yapılmıştır. Bu testler sonucunda düşük uyarımlı konuşmacıların cinsiyet ve yaş sınıflandırma başarısı sırasıyla %98.9 ve %61.3 olarak, yüksek uyarımlı konuşmacıların sınıflandırma başarısı ise %98.0 ve %71.6 olarak bulunmuştur.

Barkana ve arkadaşları tarafından yapılan çalışmada perde aralığına dayalı bir öznitelik kümesi tanıtılarak bu öznitelik kümesinin yaş ve cinsiyet sınıflandırma performansı ile MFCC, RASTA_PLP ve temel frekans özniteliklerinin sınıflandırma performansları karşılaştırılmıştır [43]. Çalışmada enerji ve sıfır geçiş oranına dayalı bir ses etkinlik algılaması (VAD) ile ses tespiti yapıldıktan sonra öznitelikler çıkarılmış ve iki farklı sınıflandırıcı kullanılarak (k-enyakın komşuluk ve SVM) öznitelik kümelerinin performans değerlendirmesi yapılmıştır. aGender veritabanı ile yapılan deneylerde her iki sınıflandırıcının da en yüksek yaş + cinsiyet ve yaş sınıflandırma başarısı önerilen perde aralığına dayalı öznitelik kümesinin kullanımıyla elde edilmiştir. Perde frekansının zamanla değişimini temsil eden bu özniteliklerin SVM ile kullanımı sonucunda orta yaşlı kadın konuşmacılar %92.86, yaşlı kadın konuşmacılar %83.61, çocuklar %83.02, orta yaşlı erkek konuşmacılar %73.58, genç kadın konuşmacılar %67.35 ve yaşlı erkek konuşmacılar %34.33 doğrulukla yaş ve cinsiyetlerine göre sınıflandırılmıştır.

Gautam ve arkadaşları tarafından yapılan çalışmada yaşları 4 ile 8 arasında değişen 134 çocuk ve 18 yetişkin konuşmacının sessiz bir odada kaydedilen konuşmaları kullanılarak konuşmacıların yaş tahmini yapılmıştır [44]. Çalışmada yetişkin konuşmacılar için bir yaş grubu (TG6), çocuklar için ise 5 yaş grubu (TG1-TG5) tanımlanmış ve üç farklı sınıflandırma yöntemi (Neuro Fuzzy, K-enyakın komşuluk ve C4.5) kullanılarak konuşmacılar yaş gruplarına göre sınıflandırılmıştır. Çalışmada temel frekans, formant frekansları, LPC katsayıları ve segmental sürelerden oluşan öznitelik vektörleriyle temsil edilen konuşma sinyalleri Neuro fuzzy sınıflandırıcısı ile sınıflandırılmış ve %85 sınıflandırma başarısı sağlanmıştır. Yapılan testlerde yetişkin (TG6) ve 4 yaşındaki konuşmacılar (TG1) %100 başarı ile sınıflandırılırken bu durum genç çocuk ve yetişkin konuşmacıların farklı temporal ve spektral özniteliklere sahip olduğu ve yaşla birlikte bu özniteliklerin yetişkininkine benzemeye başladığı şeklinde yorumlanmıştır.

Feak tarafından yapılan çalışmada 114 Kürt konuşmacının temiz ve gürültülü ortamda kaydedilen konuşmaları kullanılarak otomatik yaş ve cinsiyet tahmini yapılmıştır [45]. Çalışmada ilk dört formant frekansı ve 12 MFCC katsayılarından oluşan bir öznitelik kümesi oluşturulmuş ve bu kümeden seçilen öznitelikler iki farklı sınıflandırıcıya (SVM ve

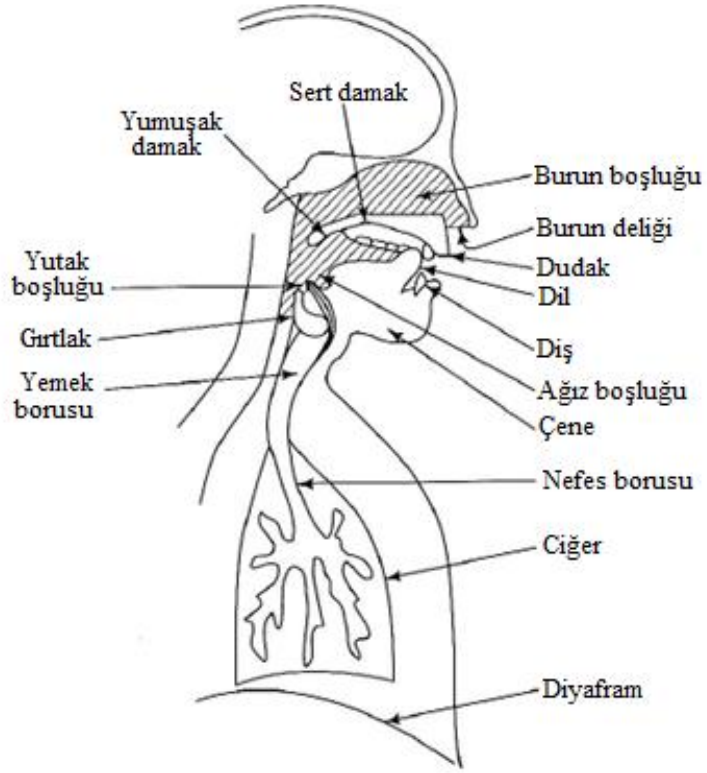
k-NN) uygulanarak konuşmacının yaş ve cinsiyet grubuna karar verilmiştir. Yapılan testlerde iki ve üç sınıflı cinsiyet tanımada en iyi sınıflandırma başarısı ikinci ve üçüncü formant frekanslarının (F2, F3) SVM ile sınıflandırması sonucunda %96 ve %94 olarak elde edilmiştir. Birinci ve dördüncü formant frekansı (F1, F4) ile ilk 12 MFCC katsayılarından oluşan öznitelik vektörlerinin k-NN ile sınıflandırılması sonucunda ise %75.3 yaş tanıma (7 sınıflı) başarı elde edilmiştir. Çalışmada ayrıca bir gürültü bastırma yöntemi ile gürültülü sinyaller işlendikten sonra öznitelikler çıkarılmış ve aynı öznitelik ve sınıflandırıcı kullanılarak yaş tanıma sisteminin başarısı %81.44'e çıkarılmıştır.

1.1.4. Tezin Amacı

Konuşmacıların yaş ve cinsiyet özelliklerine göre sınıflandırılması amaçlanan bu tez çalışmasında farklı modelleme teknikleri ve öznitelik türleri incelenerek bu yöntemlerin yaş ve cinsiyet tanıma performansları araştırılmıştır. Çalışmada ayrıca konuşma süresi ve model büyüklüğünün başarım üzerindeki etkileri incelenerek en uygun konuşma süresi ve model büyüklüğünün belirlenmesine çalışılmıştır. Oturum ve kanal etkilerinin de değerlendirildiği çalışmada farklı modelleme teknikleri ve özniteliklerle geliştirilen sistemler skor seviyesinde birleştirilerek optimum bir sistem önerilmiştir.

1.2. Ses Üretimi

İnsanın ses üretim sistemi fiziksel ve ruhsal yönlerin iç içe geçtiği karmaşık bir süreçtir. Kabaca üç aşamadan oluşan ruhsal süreç kişinin bir mesaj iletme isteğiyle başlar. Daha sonra mesaj dinleyiciler tarafından anlaşılabilen bir şekle (dil) dönüştürülür. Son aşamada ise fiziksel yapıyı kontrol etmek için bir neuromuskular komutlar kümesi çalıştırılır [46]. Fiziksel yapıda kabaca üç bölüme ayrılabilir: ciğerler, ses telleri ve ses yolu. Ciğerler hava akışı ve basınç kaynağı olarak görev yapar. Sesli konuşma gerçekleştirildiği sırada ses telleri periyodik olarak açılıp kapanarak ciğerlerden gelen hava akışı bir darbe dizisine dönüştürülür. Bu darbe dizisi bir akustik uyarım ve sesli konuşma kaynağı olarak işlev görür. Ses yolu ses tellerinin üzerinden başlayıp ağız ve burun deliklerine kadar uzanan boşluklar kümesinden oluşur. Ses yolu bir akustik fitre olarak davranarak ses spektrumunu şekillendirir. Sonunda ses dudak ve burun deliklerinden çevreye yayılır. İnsanın ses üretim mekanizması Şekil 1.3'te gösterilmiştir.

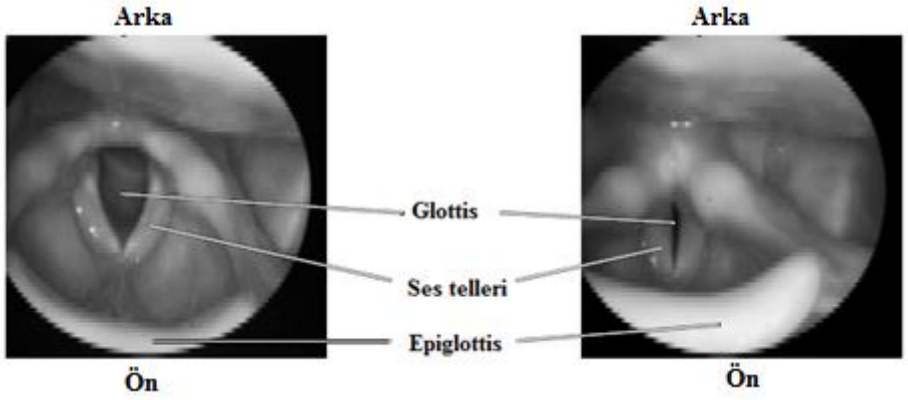


Şekil 1.3. İnsanın ses üretim sistemi [47]

1.2.1. Ses Telleri

Ses telleri gırtlakta yatay olarak yerleşmiş yumuşak ve elastik dokulardır. Ses telleri arasındaki üçgen şeklindeki hava boşluğuna glottis denilir [48, 49]. Diğer bir tanımlamaya göre glottis bu boşluğun etrafını saran yapı olarak da ifade edilir [50].

Bu mekanizma çok sayıda kas ve kıkırdakla desteklenir ve kontrol edilir. Ses telleri ön uçlarında troid kıkırdağına arka uçlarında ise arytenoid kıkırdağına bağlıdır. Arytenoid kıkırdağı gırtlaktaki kaslar tarafından hareket ettirilebilir. Böylece ses telleri arasındaki açıklığın genişliği değişir. Solunum sırasında ses telleri iyice ayrılırken fonlama sırasında ses telleri birbirlerine yakınlaşır. Şekil 1.4'te solunum ve seslendirme sırasında insan gırtlağının üstten resmi gösterilmiştir.



Şekil 1.4. Bir kadın konuşmacının solunum (Sol) ve fonlama (Sağ) sırasında gırtlağının üstten görünüşü

Eğer ciğerlerden gelen hava akışı bu dar açıklıktan geçerse ses telleri titreşime başlar ve bu titreşim hava akışını periyodik bir darbe dizisine dönüştürür. Bu darbeler "glottal flow" veya ses kaynağı, bu süreç ise fonlama olarak adlandırılır.

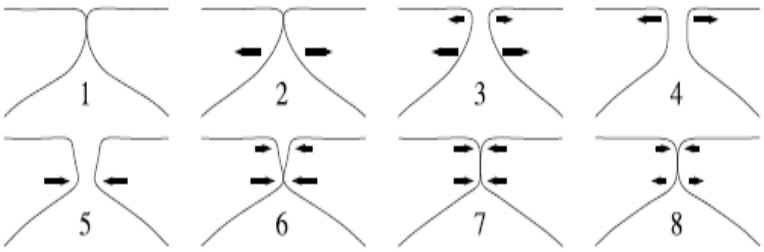
Salınım frekansı ve ses kalitesini düzenlemek için ses tellerinin gerginliği ve uzunluğu kas eylemleri tarafından kontrol edilir. Titreşim yapan ses tellerinin uzunluğu yaklaşık olarak yetişkin erkeklerde 16 milimetre, yetişkin kadınlarda 10 milimetredir. Ancak bu uzunluk gırtlaktaki kasların eylemleri sonucunda birkaç milimetre uzayabilir [51].

Gırtlaksal titreşimin frekansı genellikle f_0 ile gösterilen sesin temel frekansını belirler. Konuşmanın ortalama temel frekansı erkekler için 120 Hz, kadınlar için 200 Hz ve çocuklar için 330 Hz civarındadır [49]. Değişim aralığı geniştir; erkekler için temel frekans 100 Hz'nin altı olağandır fakat tenor şarkıcılar 600 Hz'nin üzerine çıkabilir.

Ses telleri değişik sertlikteki birkaç katmandan oluşur. En üst katman (epithelium) mukozal dokunun (lamina propria) üzerini kaplayan ve üç bölüme ayrılabilen bir katmandır. Mukozal katmanın sertliği derinliğe göre artmaktadır. Ses tellerinin en iç katmanı elastik bir kastır (musculus thyroarytenoideus). Titreşim temel olarak bu dokunun mukozal bölümünde olur. Bu salınımın kendisi kassal bir işlem gerektirmez ve hava basınç değişimleri ve dokuların elastikliği ile sürdürülür. Fakat kaslar ses tellerini bir araya getirmek ve onların titreşim özelliklerini kontrol etmek için kullanılır [52].

Değişik teknikler kullanarak ses tellerinin titreşimi incelendiğinde ses tellerinin üst ve alt bölümlerinin salınım yapmadığı ortaya çıkmıştır. Bir dalga ses tellerinin alt kısmından üst kısmına doğru ilerler. Bu ses telinin dış katmanında yukarı doğru zig zag

çizen dalga benzeri bir hareket yaratır. Bu olay mukozal dalga olarak isimlendirilir ve Şekil 1.5'te izah edilmiştir [53].



Şekil 1.5. İdeal bir ses teli titreşim çevriminin çapraz profili

Ayrıca yatay düzlemde de açılma ve kapanma ses telleri boyunca eşzamanlı olarak oluşmaz. Onun yerine açılma ve kapanma fermuar benzeri bir hareketle bir uçtan diğerine ilerler [54].

1.2.2. Ses Yolu

Ses tellerinin titreşimi spektral olarak zengin akustik bir uyarım sağlar. Bu uyarım gırtlaksal kaynağın üzerindeki kıvrımlar ile şekillendirilir. Gırtlak (Larynx), yutak (pharynx) ve ağızsal kıvrımlarla biçimlenen tüp, ses yolu olarak isimlendirilir. Ses yolunun ortalama uzunluğu yaklaşık olarak erkekler için 17 cm, kadınlar için 15 cm ve çocuklar için 14 cm dir [55]. Değer bir tanımlamaya göre burunsal kıvrımda ses yolu tanımlamasına dahil edilebilir [52].

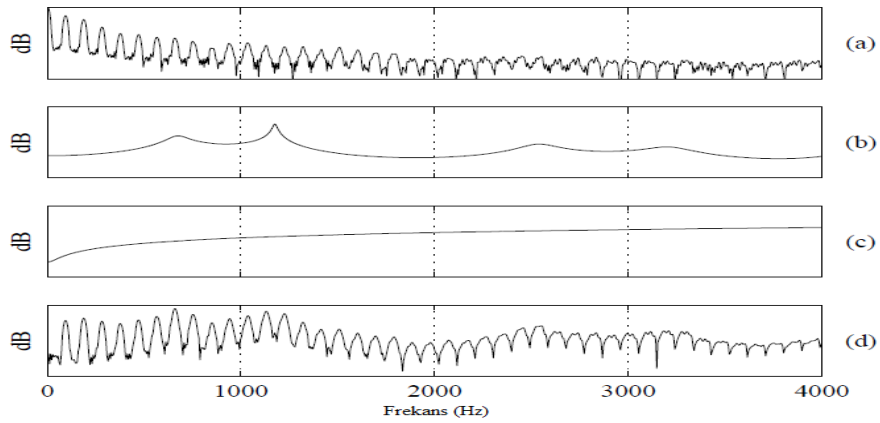
Ses yolu, uyarım sinyalinin spekrumunu değiştirebilen ayarlanabilir bir akustik filtredir. Her sesli ses (vowel sound) ses yolu rezonansları veya formantları tarafından üretilen kendisine ait spektral görünüme sahiptir. Formant frekansları yumuşak düzlem, dil, çene ve dudakların pozisyonuyla belirlenen ses yolunun şekline bağlıdır.

1.2.3. Kaynak-Filtre Teorisi

Fant tarafından önerilen kaynak-filtre teorisine göre insanın ses üretim mekanizması bir uyarım kaynağı ve bir filtre sisteminin seri bir bağlantısı olarak modellenebilir [56]. Kaynak ve filtre birbirlerinden bağımsız olarak değerlendirilir. Sesli konuşma sesinde uyarım, hava akışı tarafından titreşen ses telleri vasıtasıyla sağlanır. Ses yolu fonem bağımlı bir filtre olarak işlev görür.

Ses kaynağının ve ses yolunun bağımsız olduğu varsayımı gerçekten tam olarak geçerli değildir. Çünkü gerçekte gırtlaksal akış ses yolu konfigürasyonundan belli derecede etkilenir. Yine de teorinin geçerliliği ilgilenilen çoğu durum için yeterli olarak değerlendirilebilir. Bu varsayım ses işleme çalışmalarında çok yaygın olmasına rağmen bu teorinin kullanılmadığı durumlar da vardır [57]. Ancak çoğu durumda bağımsızlık varsayımının sağladığı basitlik küçük başarısızlıkları geçersiz kılmaktadır. Şekil 1.6'da ses üretim mekanizması üç ayrık ve bağımsız sürecin seri bir bağlantısı olarak gösterilmiştir: gırtlaksal uyarım, ses yolu ve dudak yayılımı. Şekil 1.7'de ise bu üç sürecin frekans uzayındaki etkisi gösterilmektedir.

Şekil 1.6. Kaynak-filtre modeli



Şekil 1.7. Kaynak-filtre modelinin bileşenlerinin frekans domainindeki etkisi
(a) gırtlaksal uyarım spektrumu, (b) ses yolu fitresinin genlik tepkisi,
(c) dudak yayılımının genlik tepkisi, (d) ses sinyalinin spektrumu

1.3. Biyometri

Suçların çözümünde vücut ölçümlerinin kullanılması fikri Alphonse Bertillon tarafından bir asırdan fazla süre önce tasarlanmış ve daha sonra endüstriyel olarak uygulamıştır [58]. Günümüzde biyometrik tanıma kişilerin tanınması için ayırıcı karakteristiklerinin kullanımı anlamında kullanılmaktadır [59, 60]. Biyometrik tanımlayıcılar veya biyometrikler olarak isimlendirilen bu ayırıcı karakteristikler genellikle fizyolojik ve davranışsal olarak sınıflandırılır.

Parmak izi, yüz, el geometrisi, retina veya iris gibi fizyolojik biyometrikler, farklı zamanlarda ölçülebilen fiziksel karakteristiklerdir. Diğer yandan, imza, ses veya yürüyüş gibi davranışsal biyometrikler zamana uzanan eylemlerden oluşur. Fizyolojik biyometriklerin aksine davranışsal biyometrikler zamanla öğrenilir veya kazanılır ve kasıtlı olarak ve kolaylıkla değiştirilebilir [59].

Biyometrik tanımlayıcıların davranışsal ve fizyolojik olarak sınıflandırılması net gibi görünse de tüm biyometrik belirleyiciler bir şekilde her iki karakteristiğin kombinasyonundan oluşur. Davranışsal biyometrikler hareket veya dinamiklerle ilişkilidir ve kişinin fizyolojik yapısına oldukça bağlıdır. Örneğin ses ve yürüyüş insanın ses mekanizması ve bacaklarına bağlıdır [61].

Bu noktada hangi karakteristiklerin biyometrik tanımlayıcı olarak kullanılabileceği sorusu akla gelir. Maltoni ve arkadaşları [60] bir fizyolojik veya davranışsal karakteristiğin biyometrik tanımlayıcı olarak kullanılabilmesi için aşağıdaki şartları sağlaması gerektiğini belirtmişlerdir.

- Genellik (Universality): Her kişi sahip olmalıdır.
- Ayıt edicilik (distinctiveness): Herhangi iki kişi kendi biyometrik tanımlayıcılarına göre yeterince farklı olmalıdır.
- Süreklilik (permanence): Bir zaman periyodunda yeterince değişmez olmalıdır.
- Toplanabilirlik (collectibility): Nicel olarak ölçülebilir olmalıdır.

Ayrıca uygulamalı biyometrik sistemlerde doğruluk, müdahale durumu (intrusiveness- biyometrik verinin elde edilmesi için kişi ile istenmeyen temas) veya kabul edilebilirlik (acceptability- insanların belirli bir biyometrik tanımlayıcıyı kabul etmeye ne ölçüde gönüllü olduğu) gibi konularda göz önünde bulundurulmalıdır.

Hangi biyometrik belirleyicinin seçileceği uygulamaya bağlıdır. Çünkü her biyometrik belirleyicinin avantaj ve dezavantajları vardır. En yaygın kullanılan

biyometriklerin ana karakteristikleri, güçlü ve zayıf yönleri aşağıda tanıtılarak Tablo 1.1'de özet şeklinde sunulmuştur.

1.3.1. Biyometrik Tanımlayıcılar

Yüz: Yüz en kabul görmüş biyometriklerden birisidir. Çünkü yüz görsel iletişimde kullanılan en yaygın tanıma yöntemlerinden birisidir [60]. Ayrıca parmak izi, retina ve iris gibi daha güvenilir olan biyometrik yöntemlerinin aksine yüz biyometrisi müdahaleci değildir ve katılımcıların işbirliğine gerek duymaz [62].

Video veya hareketsiz resimleri alınan bir sahnedeki bir veya daha fazla kişinin veritabanında kayıtlı yüzleri kullanılarak tanınması istenebilir [62]. Buradaki en büyük zorluk yaşlanma, yüzsel ifade değişimleri, resim çevresindeki hafif değişimler ve kameraya göre yüzün duruğundaki değişiklikler gibi etkileri tolere edebilecek yüz tanıma yöntemlerinin geliştirilmesidir [58, 60].

Parmak izi: Parmak izi, ayırt ediciliği ve performans özelliğinden dolayı en yaygın kullanılan biyometriktir. Her kişinin farklı parmak izine sahip olduğu bilgisi Amerika içişleri bakanlığı tarafından 1893 de kabul edilmiştir. Bu tarihten sonra parmak izi adli uygulamalarda yaygın olarak kullanılmaya başlanmıştır. Yaklaşık elli yıl önce ise ilk makine örüntü tanıma uygulamalarından birisi olarak otomatik parmak izi tanıma sistemleri geliştirilmiştir [60].

Parmak izi orta seviyeli yaygınlığı, kabul edilebilirliği ve toplanabilirliği ile yüksek seviyeli doğruluğuyla karakterize edilir. Buna rağmen parmak izi tanıma problemi halen ilgi çekici ve önemli bir örüntü tanıma problemi olarak değerlendirilir [58, 60].

Ses: Geniş bir konuşmacı vertabanından kişilerin tanınmasında sesin yeterli ayırt ediciliğe sahip olması beklenmez. Ayrıca sesin üç dezavantajı daha vardır; birincisi, bir konuşma sinyalinin kalitesi mikrofon veya iletişi kanalı tarafından bozulabilir; ikincisi, ses bir kişinin sağlık, stres veya ruhsal durumundan etkilenebilir; sonuncusu ise bazı insanlar diğerlerinin seslerini olağandışı şekilde taklit edebilir [59, 60]. Ancak ses yüksek kabul edilebilirliği ile müdahaleci olamayan bir biyomeriktir. Ayrıca günümüzde ses, her yerde kullanılan ve konuşma sinyallerinin oluşturulması ve iletilmesini sağlayan telefon sistemleri üzerinden kişinin tanınmasını gerektiren uygulamalar için uygun olan tek biyometrik tanımlayıcıdır [60, 63].

İris: İris tanıma teknolojisi yüksek doğruluk ve hıza sahiptir. Embriyonik gelişim sırasında kaotik morphogenetik süreçler tarafından belirlenen insan irisinin görsel dokusunun her göz ve kişi için yüksek ayırıcılığa sahip olduğuna inanılır. İris biyometrisinin genelliği ve sürekliliği yüksek, toplanabilirliği orta, kabul edilebilirliği ise düşük düzeydedir. Çünkü bir iris görüntüsünün alınması hem müdahaleci hem de kullanıcının işbirliğini gerektirir [58, 60]. İris tanıma uygulamaları son zamanlarda hastane ve uluslararası sınırlarda erişim kontrolü için yaygın olarak kullanılmaktadır.

El ve parmak geometrisi: Bir insanın eli ile ilişkili olan bu öznitelikler çok ayırıcı olmamasına rağmen el ve parmak geometrisi oldukça yüksek bir süreklilik ile karakterize edilir. Ayrıca kullanıcıların işbirliğini gerektirdiği için orta seviyede kabul edilebilirlik ve oldukça müdahaleci bir yöntem olarak algılanır. Bu biyometriğin gücü toplanabilirliğine dayanmaktadır çünkü elin temsili için gereksinimler oldukça küçüktür. Parmak geometrisi sistemleri daha kompakt boyutundan dolayı zaman zaman tercih edilir [58, 60].

İmza: Her kişinin bir imza karakteristiği vardır ve birçok legal ve ticari işlemde doğrulama yöntemi olarak imza yüksek bir kabul edilebilirlikte kullanılır. Ayrıca imzanın toplanabilirliği de oldukça yüksektir. İmzanın zayıf yanları ise düşük genelliği, ayırıcılığı ve sürekliliğidir. Bu da düşük başarıya yol açar. Aynı kişinin imzası zamanla veya ardışık kullanımlarında değişir. Ayrıca imzaların sahtesi de kolayca üretilebilir [60].

Kulak: Bir kişinin kulağının şeklinin ve kulak kepçesinin kıkırdaklı dokusunun benzersiz olması beklenmez. Ancak bu yapılar biyometrik tanımada kullanılmak için yeterli ayırıcılığa sahiptir. Ayrıca bu yapılar yüksek kabul edilebilirliği ve zaman üzerinde sürekliliği ile karakterize edilir [58].

DNA: DNA her kişi için bir boyutlu benzersiz bir koddur ve kişi tanıma için adli uygulamalarda yaygın olarak kullanılır [58, 60]. DNA'nın yüksek ayırıcılığı ve zaman üzerindeki sürekliliği bu biyometriye dayalı sistemlerin başarısının yüksek olmasını sağlamıştır. Fakat DNA düşük toplanabilirliği ve kabul edilebilirliği ile karakterize edilir. Çünkü DNA'nın edinilmesi yüksek müdahale gerektirmektedir.

Yürüyüş: Bir kişinin yürüyüş şekli pek ayırıcı gibi gözükmez ancak bazı düşük güvenlikli uygulamalar için kişinin doğrulanmasında yeterli özgünlüğe sahip olduğu varsayılır. Yürüyüş özellikle uzun zaman periyodunda sürekli değildir. Ancak yürüyüşün elde edilmesi yüz resminin elde edilmesi ile benzer olduğu için oldukça kabul edilebilir ve müdahaleci olmayan biyometrik olarak değerlendirilir. Yürüyüş aynı zamanda yüksek toplanabilirliği ile karakterize edilir [58, 60].

Retinal tarama: İrisin görsel dokusu gibi retinal damar yapısının her kişi ve göz için bir karakteristiğinin olduğu görülür. Bu yapının kopyasının oluşturulması oldukça zordur. Fakat bu görüntünün elde edilmesi oldukça müdahalecidir ve kullanıcının işbirliği ve çabasını gerektirir. Bu nedenle bu yöntem düşük kabul edilebilirlik ile karakterize edilir [58, 60].

Koku: İnsan vücudu birey için ayırıcı olan kimyasal bir bileşen yayar. Bu ayırıcı özelliğin bir parçası olan koku zaman üzerinde süreklidir ve onun elde edilmesi kullanıcılar tarafında tamamen kabul edilir. Fakat koku verisi kolay toplanamaz ve bu sistemlerin performansları hala sorunludur [58, 60].

Tuş dinamikleri: Kişinin klavye kullanımındaki karakteristiğini temsil eden tuş dinamiklerinin çok sayıda dezavantajı olduğu için muhtemelen en az kullanılan biyometriktir. Tuş dinamikleri ne çok ayırıcıdır nede zaman üzerinde süreklidir. Hatta günlük hayatlarında herkes klavye kullanmadığı için tuş dinamikleri en az genelliğe sahip biyometriktir. Bu karakteristiklerin tamamı düşük performanslı bir sisteme neden olur. Fakat tuş dinamiklerinin elde edilmesi müdahaleci değildir ve bu nedenle kullanıcılar tarafından oldukça kabul edilir bir biyometriktir [58, 60].

Tablo 1.1. Biyometrik belirleyicilerin karşılaştırılması. Y:yüksek, O:orta, D:düşük

Biyometrik tanımlayıcılar	Genellik	Ayıt edicilik	Süreklilik	Toplanabilirlik
Yüz	Y	D	O	Y
Parmak izi	O	Y	Y	O
Ses	O	D	D	O
İris	Y	Y	Y	O
El ve parmak geometrisi	O	O	O	Y
İmza	D	D	D	Y
Kulak	O	O	Y	O
DNA	Y	Y	Y	D
Yürüyüş	O	D	D	Y
Retinal tarama	Y	Y	O	D
Koku	Y	Y	Y	D
Tuş dinamikleri	D	D	D	O

1.4. Konuşma Sinyali İşleme

Bu bölümde konuşma bir sinyal olarak ele alınarak sinyaller ve sistemler, sinyal temsili ve frekans analizi gibi sinyal işleme konusundaki en temel kavramlar gözden geçirilecektir. Daha sonra konuşma sinyallerinin analizi için etkin bir yöntemler kümesi olan kısa süreli analiz tanıtılacaktır. Son olarak ise kepstrum fikri tartışılacaktır.

1.4.1. Sinyaller ve Sistemler

Bir olgunun gözlemlenen ölçümlerine sinyal denilir [64]. Örneğin bir arabanın hızı veya bir hisse senedinin fiyatı farklı uzaylardaki birer sinyal olarak değerlendirilebilir. Normalde bir sinyal çeşitli bağımsız değişkenlerin bir fonksiyonu olarak modellenir. Genellikle bu değişken zamandır ve zamanla değişen sinyal f(t) olarak temsil edebilir. Ancak bir sinyal tek bir değişkenin fonksiyonu olmak zorunda değildir. Örneğin bir resim (x; y) noktasındaki rengi temsil eden bir sinyaldir ve f(x; y) şeklinde gösterilir.

1.4.1.1. Analog ve Sayısal Sinyaller

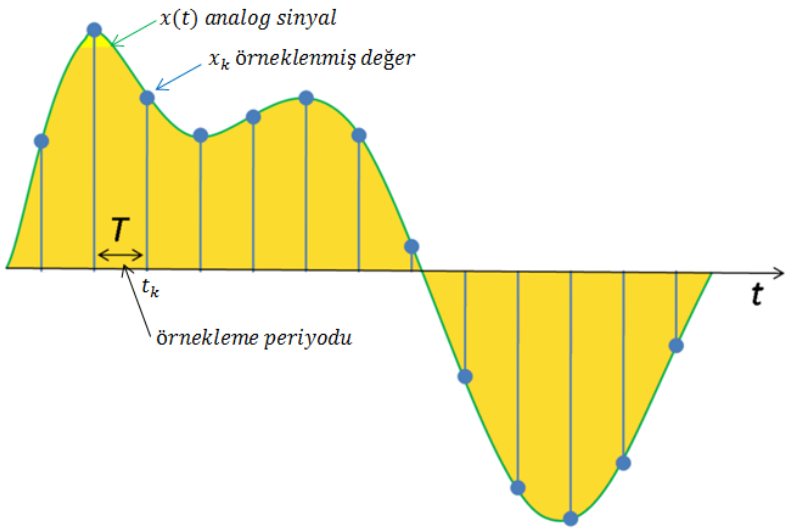
Eğer bir sinyalin aralığı ve uzayı sürekliyse yani bağımsız değişken ve sinyalin değeri keyfi değerler alabiliyor ise o analog bir sinyaldir. Analog sinyaller matematiksel yöntemlerle analiz edilebilir ancak günümüzde çoğu sinyal işleme yöntemlerinin gerçekleştirildiği bilgisayarlar tarafından doğrudan işlenemez. Analog sinyallerin bilgisayarlar tarafından işlenebilmesi için sayısal forma dönüştürülmesi gerekir.

1.4.1.2. Örnekleme ve Nicemleme

Konuşma sinyali analog bir sinyaldir ve bu sinyallerin bilgisayar tarafından işlenebilmesi için sayısal forma dönüştürülmesi gerekir. Sinyal işleme alanında sürekli zaman sinyalleri düzenli zaman aralıklarında örneklenerek ayrık zaman sinyaline dönüştürülür. Örnekleme olarak isimlendirilen bu süreç sonunda $x(t)$ analog sinyalinden sonlu bir x_k serisi elde edilir. Bu serinin her bir örneği Şekil 1.8'de gösterilen t_k anlarına karşılık gelen değerlerdir.

$$t_k = kT \;\; k = 0,1, \ldots \tag{1.1}$$

Buradaki T örnekleme periyodu, örnekleme periyodunun tersi ($f = 1/T$) ise örnekleme frekansı olarak isimlendirilir.



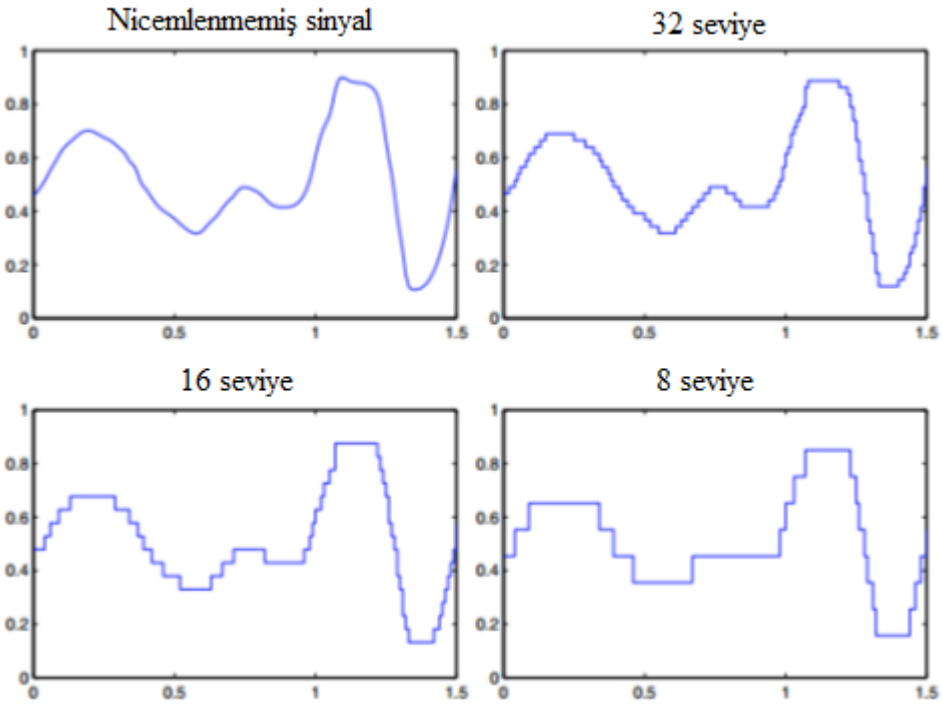
Şekil 1.8. Zaman uzayında örnekleme

Örnekleme periyodunun büyük seçilmesi durumunda elde edilen sayısal sinyal orijinal sinyali istenen kalitede temsil edemeyecektir. Diğer yandan örnekleme periyodunun küçük seçilmesi durumunda ise sinyalden gereksiz örnekler alınacak bu da ekstra bellek ve işlem yüküne neden olacaktır. Bu sorunun çözümü için Nyquist–Shannon örnekleme teoremi önerilmiştir [65]. Bu teoreme göre optimum örnekleme periyodunun belirlenmesi için analog sinyalin bant genişliğinin bilinmesi yeterlidir. Örnekleme teoremine göre eğer $x(t)$ sinyalinin tüm frekansları $[0, W]$ aralığında ise ve örnekleme periyodu $T = 1/2W$ saniye alınırsa orijinal analog sinyal;

$$x(t) = \sum_{k=-\infty}^{\infty} \frac{sin(2\pi W(t-k/2W))}{2\pi W(t-k/2W)} \tag{1.2}$$

ifadesiyle tam olarak temsil edilebilir [66]. Buradaki x_k denklem 1.2'ye göre belirlenen $t_k = k/2W$ noktalarına karşılık gelen sinyal değerlerdir. Oldukça dar bir bant genişliğine sahip (genellikle 100 Hz ile 8 kHz arası) olan konuşma sinyali için 16 kHz'lik bir örnekleme frekansı yeterli bir seviye olarak kabul edilmektedir.

Örnekleme sonucunda gerçek değerlerden oluşan bir $x[n]$ sinyali elde edilir. Ancak bu değerler sürekli bir uzaydadır ve sayısal olarak temsil edilemez. $x[n]$ sinyalinin sayısal olarak temsil edilmesi için her bir örneğinin ayrık bir değere dönüştürülmesi gerekir. Bu amaçla önce sinyalin genlik aralığı alt aralıklara bölünür ve her bir örnek kendisine en yakın seviyeye yuvarlanır. Bu süreç nicemleme olarak isimlendirilir. Daha sonra her bir nicemleme seviyesine farklı bir bit dizisi atanarak analog sinyal sayısala dönüştürülür. Niceleme sonucunda sinyalin gerçek değeri ile atanan ayrık değer arasındaki bir fark oluşur. Bu farkın azaltılması için sinyalin daha küçük alt aralıklara bölünmesi gerekir. Ancak bu durumda da her bir seviyeyi temsil etmek için daha fazla bit kullanılacaktır. Örneğin 8 bit ile temsil edilen sinyal 256 nicemleme seviyesine, 16 bit ile temsil edilen sinyal ise 65537 nicemleme seviyesine bölünecektir. Bit sayısındaki artışla birlikte sinyalin hem kalitesi hem de boyutu artacaktır. Bu nedenle kalite ve boyut arasında bir denge kurulmalıdır. Konuşma tanıma sistemlerinde genellikle 8 veya 16 bitlik nicemleme seviyesi kullanılır. Şekil 1.9'da farklı seviyelerde nicemlenen analog bir sinyal gösterilmiştir.



Şekil 1.9. Farklı seviyelerde nicemlenen analog bir sinyal [64].

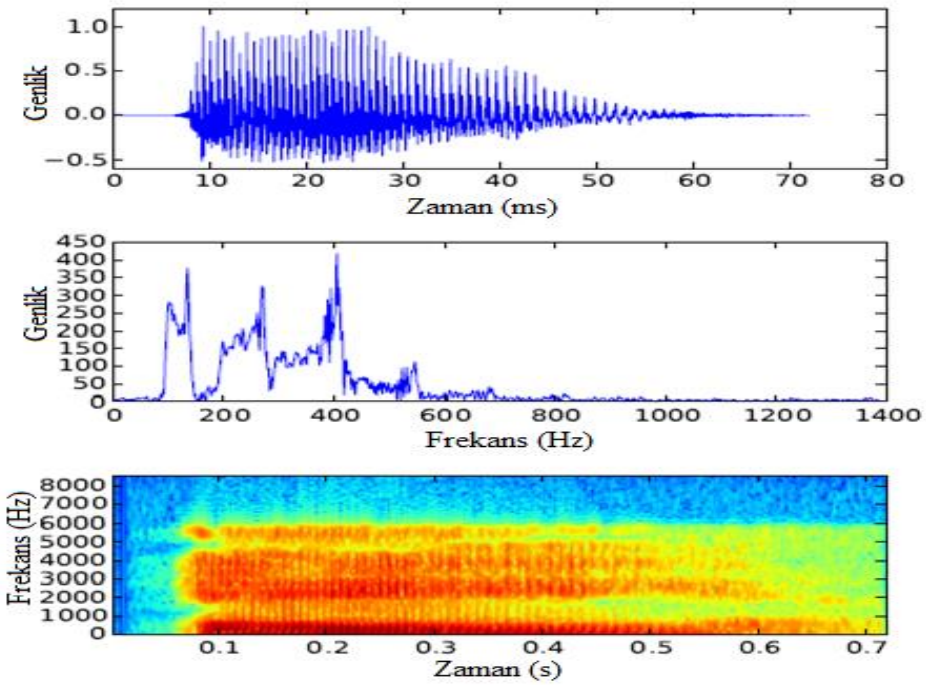
1.4.1.3. Sayısal Sistemler

Giriş olarak aldığı bir sinyale göre belirli görevleri yerine getiren yapılara sistem denilir. Örneğin bir ev termostatı bir anahtar veya kol aracılığı ile alınan girişe göre fırın için elektriksel bir sinyal üreten bir sistem olarak değerlendirilebilir. Sayısal bir sistem bir giriş sinyalinin bir çıkış sinyaline dönüşümü olarak aşağıdaki gibi tanımlanır [64].

$$Y[n] = T\{x[n]\} \tag{1.3}$$

1.4.2. Sinyal Temsili: Zaman Uzayı ve Frekans Uzayı

Konuşma sinyali ses kaynağı ile dinleyici arasındaki ortamda oluşan basınç değişikliklerinin bir dizisidir. Konuşma sinyalini temsil etmek için farklı yöntemler kullanılır. Bu yöntemlerden en yaygın olanı genlik olarak isimlendirilen ses basınç ölçümlerinin zaman ekseninde temsil edildiği osilogramdır. Bu gösterimde yatay eksen zaman olmak üzere sinyaldeki basınç artışı ve azalışı bir eğri ile temsil edilir. Bir konuşma sinyalinin osilogramı Şekil 1.10'da gösterilmiştir.



Şekil 1.10. 44100 Hz'de örneklenen bir yetişkin kadın sesi (a) dalga biçimi, (b) 1400 Hz limitli spektrumu, (c) 0 ile 8000 Hz arası spektogramı

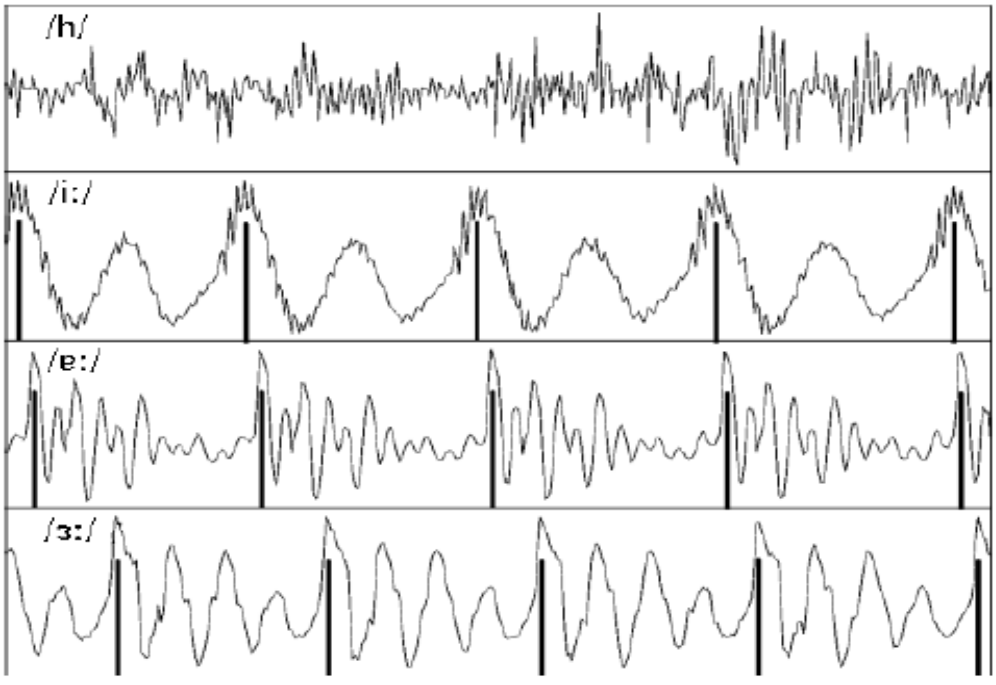
Dalga formunun şekli bize konuşma sinyalin periyodikliği hakkında sezgisel bir yol gösterir. Şekilsel olarak analog bir sinyal

$$x_a(t+T) = x_a(t) \quad \forall t \tag{1.4}$$

şartını sağlaması durumunda periyodiktir. Benzer şekilde sayısal bir sinyalin periyodik olması için

$$x[n+N] = x[n] \quad \forall n \tag{1.5}$$

şartını sağlaması gerekir. Bu şartları sağlamayan bir sinyal periyodik değildir. Periyodik ve periyodik olmayan konuşma sinyallerine ait dalga formları Şekil 1.11'de gösterilmiştir.



Şekil 1.11. Periyodik ve periyodik olmayan konuşma sinyalleri [67]. Üç sesli harfin dalga şekli periyodiktir. /h/ frikatifinin dalga şekli peryodik değildir

Bir sinyali incelemek için diğer bir bakış açısı frekans uzayıdır. Beyaz ışığın bir prizmaya yönlendirilmesi deneyi frekans uzayını açıklamak için kullanılan oldukça meşhur bir örnektir. Newton deneyinde [68] beyaz ışığın bir prizma ile renk bantlarına yani spektrumlarına ayırılabileceği, ikinci bir prizmanın kullanılmasıyla bu renk ışınlarından

beyaz ışığın tekrar oluşturulabileceği gösterilmiştir. Böylece beyaz ışık renk bileşenlerine ayrıştırılarak analiz edilebilir. Her ana rengin bir frekans aralığına karşılık geldiği düşünüldüğünde beyaz ışığın renklere ayrıştırılması frekans analizinin bir şekli olarak değerlendirilir.

Sayısal hesaplama için sinüs dalgası veya sinüzoit oldukça önemli bir sinyaldir ve

$$x_a(t) = Acos(\omega t + \emptyset) \quad -\infty < t < \infty \tag{1.6}$$

ifadesi ile tanımlanır. Burada A sinyalin genliği, w saniyedeki radyan cinsinden açısal frekans, $\emptyset$ ise radyan cinsinden faz açısıdır. Sinyalin hertz cinsinden frekansı f ile açısal frekans arasındaki ilişki

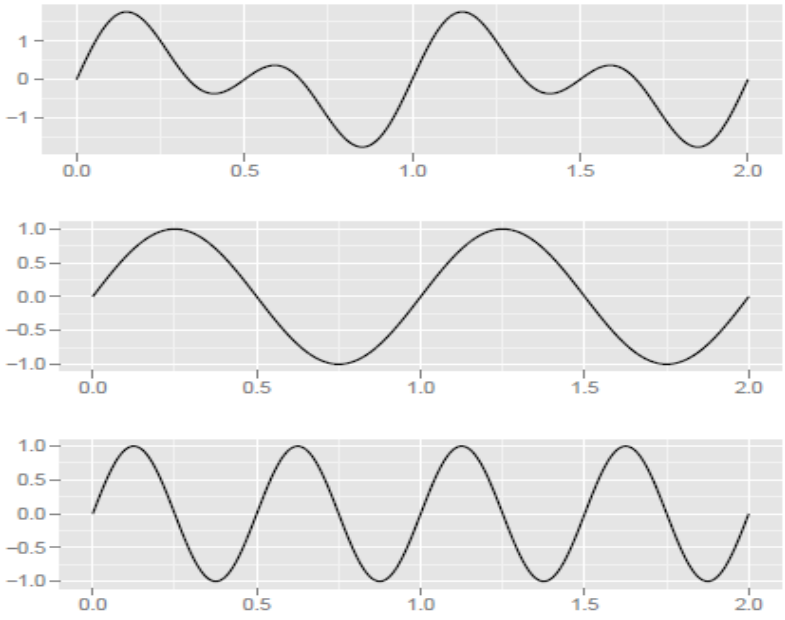
$$\omega = 2\pi f \tag{1.7}$$

ifadesi ile tanımlanır. Denklem 1.4'e göre bir sinüzoid periyodu $T = 1/f$ olan periyodik bir sinyaldir. Sayısal bir sinüzoit ise

$$x[n] = Acos(\omega n + \emptyset) \quad -\infty < n < \infty \tag{1.8}$$

şeklinde tanımlanır. Anacak denklem 1.5'e göre $x[n]$ sinyalinin periyodik olması için $\omega = 2\pi/N$ veya $f = \omega/2\pi$ frekansının rasyonel bir sayı olması gerekir. Bu nedenle denklem 1.8'le tanımlanan sayısal sinyal w'nin her değeri için periyodik değildir.

Bir konuşma sinyalinin frekans analizi onun sinüzoitler toplamı şeklinde ayrıştırılması olarak görülebilir. Şekil 1.12'de sinüzoitlere ayrıştırılan bir konuşma sinyali gösterilmiştir. Bir sinyalin zaman uzayından frekans uzayına değiştirilmesi süreci frekans dönüşümü olarak isimlendirilir.



Şekil 1.12. Bir konuşma sinyalinin sinüzoitlere ayrıştırılması

Spektrum konuşma sinyalinin frekans uzayındaki bir temsil şekli olup her frekansa karşılık gelen genlik iki boyutlu olarak temsil edilmektedir (Şekil 1.10b). Spektrogram ise spektral bilginin üç boyutlu olarak temsil edildiği diğer bir gösterim şeklidir (Şekil 1.10c). Bu gösterimde yatay eksen zamanı, dikey eksen ise frekansı göstermek üzere her bir zaman frekans noktasının genliği gölge tonları ile temsil edilmektedir. Spektrograma yukarıdan bakıldığında her satır sinyalin belirli bir andaki spektrumuna karşılık gelmektedir. Yani spektrogram dikey olarak yerleştirilmiş spektrum dizisi olarak değerlendirilebilir. Sinyalin spektrumundaki yükseklikler spektrogramda koyu noktalarla temsil edilmektedir. Genlik arttıkça noktanın koyuluğu (renkli gösterimde sıcaklığı) artmaktadır. Üç boyutlu temsil özelliği ile spektrogram konuşmanın akustiği hakkında görsel ipuçları sağlayan etkin bir araç olarak kullanılmaktadır.

1.4.3. Frekans Analizi

Fourier analiz yöntemleri bir sinyali frekans uzayına dönüştürmek için kullanılan matematiksel araçlardır. Bu yöntemlerden hangisinin seçileceği sinyalin periyodikliğine ve analog veya sayısal olmasına bağlıdır. Dört farklı Fourier analiz yöntemi Tablo 1.2'de özetlenmiştir.

Tablo 1.2. Fourier analizi tekniklerinin özeti [69].

Zaman uzayı Özellikleri	Periyodik	Periyodik değil	
Sürekli	Fourier Serisi (FS)	Fourier Dönüşümü(FD)	Periyodik değil
Ayrık	Ayrık Fourier Dönüşümü (AFD)	Ayrık-Zaman Fourier Dönüşümü (AZFD)	Periyodik
	Ayrık	Sürekli	Frekans uzayı Özellikleri

Her yöntem ileri ve ters olmak üzere bir çift dönüşümden oluşmaktadır. Periyodu T olan periyodik sürekli bir $x(t)$ sinyalin Fourier serisi;

$$c_k = \frac{1}{T}\int_T x(t)e^{-j2\pi kt/T}\,dt \qquad ileri\ dönüşüm \tag{1.9}$$

$$x(t) = \sum_{k=-\infty}^{\infty} c_k e^{j2\pi kt/T} \qquad tesr\ dönüşüm \tag{1.10}$$

olarak, periyodik olmayan sürekli bir $x(t)$ sinyalin Fourier serisi;

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}\,dt \qquad \text{ileri dönüşüm} \tag{1.11}$$

$$x(t) = \int_{-\infty}^{\infty} X(\omega)e^{j\omega t}\,d\omega \qquad \text{ters dönüşüm} \tag{1.12}$$

olarak tanımlanır. Periyodu N olan ayrık $x[n]$ sinyalinin ayrık Fourier dönüşümü

$$c_k = \frac{1}{N}\sum_{n=0}^{N-1} x[n]e^{-j2\pi kn/N} \qquad ileri\ dönüşüm \tag{1.13}$$

$$x[n] = \sum_{k=0}^{N-1} c_k\, e^{j2\pi kn/N} \qquad ters\ dönüşüm \tag{1.14}$$

olarak, periyodik olmayan ayrık sinyalin ayrık zamanlı Fourier dönüşümü ise;

$$X(\omega) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n} \qquad ileri\ dönüşüm \tag{1.15}$$

$$x[n] = \frac{1}{2\pi}\int_{2\pi} X(\omega)\, e^{j\omega n} d\omega \qquad ters\ dönüşüm \tag{1.16}$$

olarak tanımlanır.

1.4.4. Kısa Dönem Konuşma İşleme

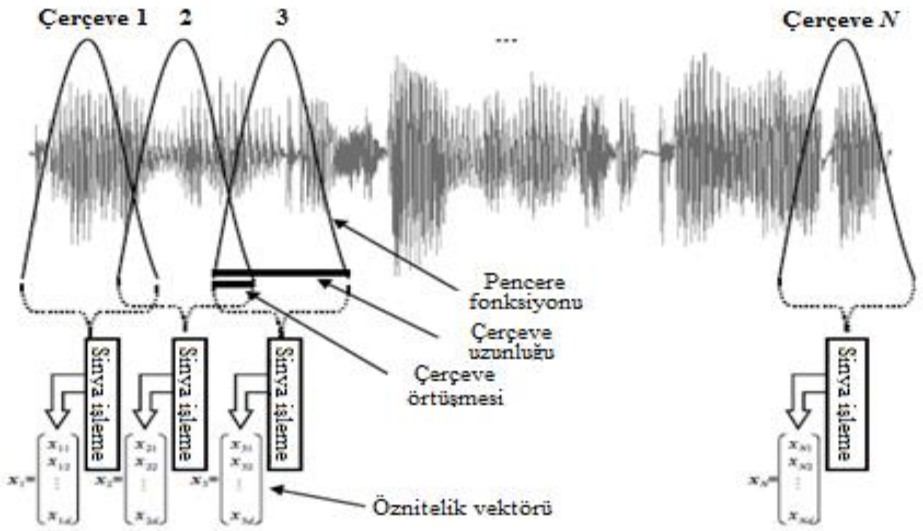
Konuşma sinyali doğası gereği oldukça yavaş değişen bir sinyaldir ve yarı durağan olarak kabul edilir. Ancak yeterince kısa zaman aralıklarında (genellikle 20-30 ms) incelendiğinde konuşma sinyalinin yaklaşık durağan akustik özelliklere sahip olduğu görülmektedir [70]. Bu nedenle konuşma sinyalleri genellikle durağan akustik özelliklere sahip olduğu kısa zaman aralıklarında analiz edilir. Kısa süreli analiz olarak isimlendirilen bu yaklaşımın genel görünümü Şekil 1.13'te verilmiştir. Kısa dönem analizinin ilk aşamasında uzunluğu önceden belirlenmiş bir pencere (genellikle 20-30 ms) komşu çerçeveler arasında bir örtüşme olacak şekilde (genellikle pencere genişliğinin %30-50 si) kaydırılır. Bilgi kaybının önlenmesi için örtüşme zorunludur. Çerçevelerin bitiş noktalarındaki ani değişimleri engellemek için çerçeveler genellikle bir pencere fonksiyonu ile çarpılır. Sinyalin kısa aralıklara bölünmesi işlemi çerçeveleme, her segment ise pencerelenmiş çerçeve (veya sadece çerçeve) olarak isimlendirilir. Konuşma işleme alanında kullanılan çeşitli pencere fonksiyonları vardır [71]. Bunlardan en basit olanı diktörgen çerçeve olup genişliği N olan bir dikdörtgen çerçeve

$$w(n) = \begin{cases} 1, & 0 \le n \le N-1 \\ 0, & diğer\ durumlarda \end{cases} \tag{1.17}$$

ifadesi ile tanımlanır. Zaman uzayındaki bir dikdörtgen çerçevenin kenarlarında ani bir süreksizlik vardır. Bu durum dikdörtgen çerçevenin frekans uzayındaki temsilinde geniş yan loblara ve istenmeyen çınlama etkisine neden olur [72]. Bu geniş salınımdan kurtulmak için zaman uzayında ani süreksizliği olmayan bir pencere fonksiyonu kullanılmalıdır. Bu nedenle konuşma işleme çalışmalarında genellikle denklem 1.18 ile tanımlanan Hamming penceresi kullanılır. Hamming penceresi konik bir pencere olup aslında kabarık bir kosinüs fonksiyonudur.

$$w(n) = \begin{cases} 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right) & 0 \le n \le N-1 \\ 0 & diğer\ durumlarda \end{cases} \tag{1.18}$$

Daha sonra her bir çerçeve üzerinde çeşitli sinyal işleme yöntemleri kullanılarak konuşmanın farklı özelliklerini temsil eden parametrik değerler elde edilir. Özellik vektörü olarak isimlendirilen bu parametrik değerlerin hesaplanması için değişik yöntemler kullanılmaktadır. Konuşma işleme çalışmalarında yaygın olarak kullanılan özniteliklerin bazıları ilerleyen bölümlerde tanıtılacaktır.



Şekil 1.13. Kısa zaman analizi

1.4.4.1. Kısa Dönem Fourier Analizi

Kısa dönem Fourier dönüşümü (KDFD) bir sinyalin frekans ve faz içeriğinin zamanla değişimini belirlemek için kullanılan Fourier tabanlı bir dönüşümdür. Uygulamada KDFD'nin hesaplanma prosedürü şöyledir; uzun süreli bir sinyal önce eşit uzunlukta kısa segmentlere bölünür, daha sonra her bir segment üzerinde Fourier dönüşümü ayrı ayrı hesaplanır. Böylece her segmentin Fourier spektrumu ortaya çıkar. Verilen bir $x[n]$ sinyali için kısa dönem Fourier dönüşümü

$$X(m,w) = X_m(\omega) = \sum_{n=-\infty}^{\infty} x[n]w[m-n]e^{-j\omega n} \tag{1.19}$$

şeklinde her bir zaman segmentinin ayrık Fourier dönüşümlerinin (AFD) bir kümesi olarak veya

$$X(m,\omega) = e^{-j\omega n}\,(x[m] * w[m]e^{j\omega m}) \quad (1.20)$$

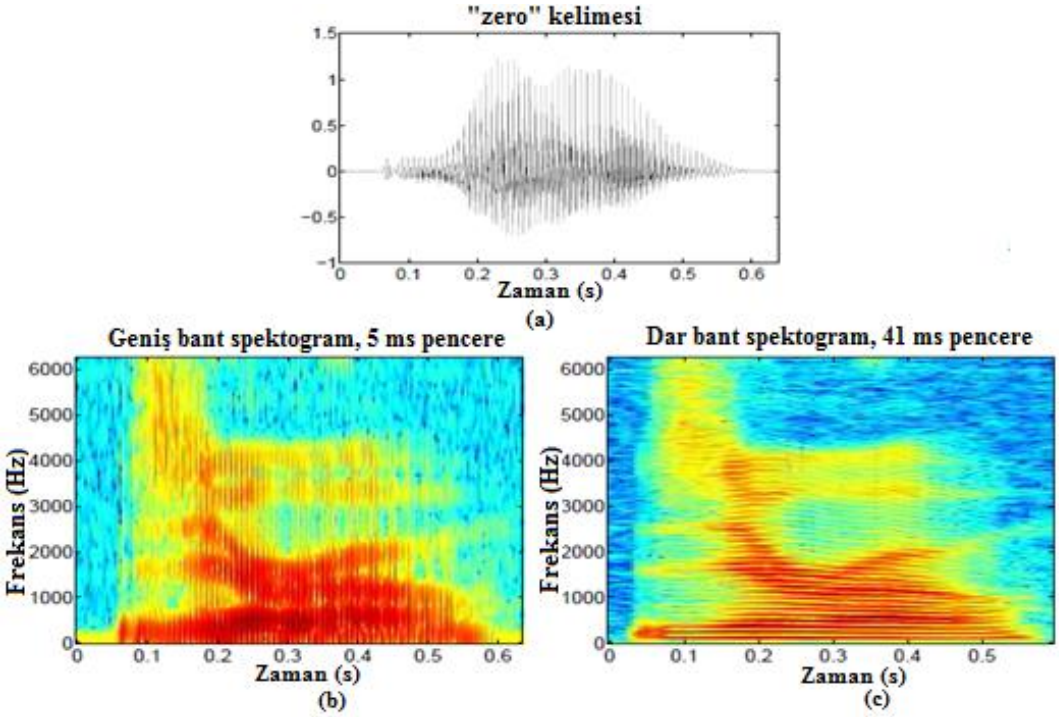
şeklinde $x[m]$'in bir bant geçiren filtre dizisinden (her ω frekansı civarında merkezi olan) geçirilmesi olarak yazılabilir [73].

1.4.4.2. Spektrogramlar

Bir sinyalin spektral karakteristiklerinin zamanla nasıl değiştiğini gösteren üç boyutlu bir temsil şekli olan spektrogram her zaman segmenti farklı kolanda olacak şekilde AFD'lerin bir resme dikey olarak yerleştirilmesiyle oluşturulur [74]. Bu gösterimde genellikle frekans aşağıdan yukarıya doğru, zaman ise soldan sağa doğru artar. Resimdeki her bir nokta belirli bir andaki spektrumun ilgili frekanstaki genliğine karşılık gelir ve parlaklık veya renk olarak temsil edilir. Bir spektrogramın genliği

$$\mathrm{S}(\omega,\mathrm{t}) = |\mathrm{X}(\omega,\mathrm{t})|^2 \quad (1.21)$$

ifadesi ile hesaplanır. Kullanılan pencere uzunluğuna göre spektrogramlar dar ve geniş bant olmak üzere ikiye ayrılır. Bir konuşma sinyalinin dar ve geniş bant spektrogramları Şekil 1.14'te gösterilmiştir. Geniş-bant spektrogramlarda kısa pencere uzunlukları (<10ms) kullanılırken, dar-bant spektrogramlarda ise daha uzun pencereler (>20ms) kullanılır. Pencere süresindeki bu farklılık sinyalin zaman ve frekans çözünürlüğünü etkiler: Geniş-bant spektrogramlar iyi bir zaman çözünürlüğü görünümü sunarken harmonik yapılar için daha az kullanışlıdır. Dar-bant spektrogramlar ise iyi bir frekans çözünürlüğü sunar ancak zaman üzerindeki periyodik değişimleri bulanıklaştırır. Genel olarak geniş-bant spektrogramlar fonetik çalışmalarda daha fazla tercih edilmektedir.



Şekil 1.14. (a) "zero" kelimesinin telaffuzu, (b) geniş-bant spektogram, (c) dar-bant spektogram

1.4.5. Kepstral Analiz

Bölüm 1.2.3'te bahsedilen kaynak filtre modeline göre konuşma sinyali $s(n)$, hızlı değişen bir uyarım kaynağı $u(n)$ ile yavaş değişen ve darbe tepkisi $h(n)$ olan ses yolu filtresinin konvolüsyonu olarak temsil edilir [70].

$$s(n) = u(n) * h(n) \tag{1.22}$$

Birçok konuşma işleme uygulamasında bu bileşenlerin (uyarım ve ses yolu) ayrı olarak işlenmesi gerekmektedir. Ancak yalnızca çıkış sinyali (konuşma) erişilebilir durumdadır ve katlama işleminden dolayı konuşma sinyali kolayca bileşenlerine ayrılamaz. Bir sinyalin toplamsal bileşenlere ayrılması sayısal işaret işleme alanının oldukça zor problemlerinden birisidir. Bu problemin çözüm için Bogert ve arkadaşları tarafından kepstrum yaklaşımı önerilmiştir [75]. Bu yaklaşıma göre sinyal öncelikle zaman uzayından frekans uzayına taşınır. Böylece zaman uzayındaki konvolüsyon işlemi çarpmaya dönüştürülmüş olur.

$$S(\omega) = U(\omega).H(\omega) \tag{1.23}$$

Daha sonra logaritması alınarak çarpıma işlemi toplamaya dönüştürülür.

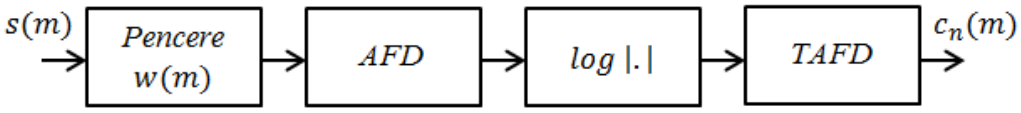
$$Log\ (S(\omega)) = Log(U(\omega).H(\omega)) \tag{1.24}$$

$$Log\ (S(\omega)) = Log(U(\omega)) + Log(H(\omega)) \tag{1.25}$$

Son olarak doğrusal bir işlem olan ters AFD uygulanarak yavaş değişen dürtü tepkisi ile hızlı değişen kaynak sinyalin farklı bölgelerde temsil edildiği yeni bir uzaya geçilir. "Quefrency" olarak isimlendirilen bu uzayda düşük indisli quefrency bileşenleri ses yolu filtresini, büyük indisli quefrency bileşenleri ise uyarı sinyalini temsil etmektedir. Homomorfik bir dönüşüm olan kepstrum (real cepstrum) ayrık zamanda aşağıdaki gibi tanımlanır.

$$c(m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} log|S(\omega)|\, e^{j\omega m} d\omega \tag{1.26}$$

Kepstral analiz yaklaşımı da konuşmanın durağan özelliklere sahip olduğu kısa zaman aralıklarında yapılır. Kısa zamanlı Kepstral analiz olarak isimlendirilen bu sürecin blok diyagramı Şekil 1.15'te gösterilmiştir.



Şekil 1.15. Konuşma sinyalinin kısa zamanlı kepstral analizi

2. YAPILAN ÇALIŞMALAR

2.1. Öznitelik Çıkarma

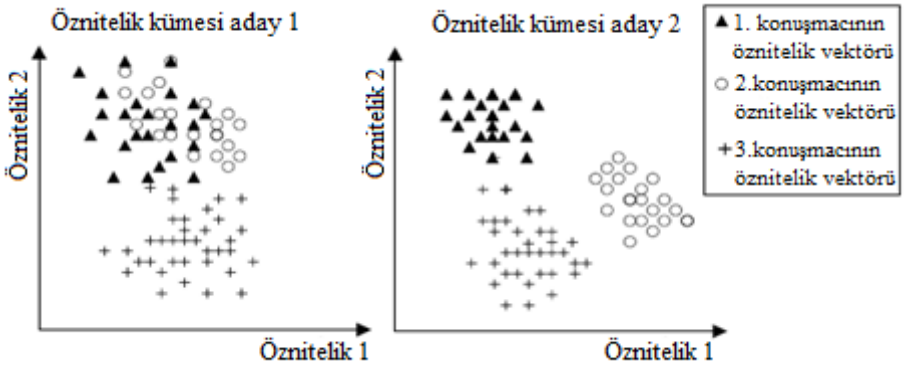
Öznitelik çıkarma biçimsel olarak yüksek boyutlu bir vektörü düşük boyutlu hale dönüştürme süreci olarak bilinir. Bu süreç aslında $f : \mathbb{R}^N \rightarrow \mathbb{R}^d,\ d \ll N$ olan bir haritalamadır. Öznitelik çıkarma iki nedenden dolayı zorunlu bir işlemdir. Birincisi istatistiksel konuşmacı modellerinin güçlü olması için eğitim örneklerinin sayısının yeteri kadar geniş olması gerekir. Boyutla birlikte gerekli olan eğitim vektörünün miktarı üstel olarak artar. Bu olgu boyutluluk sorunu olarak bilir [76-78]. Öznitelik çıkarmanın ikinci nedeni ise azalan hesaplama karmaşıklığıdır.

Konuşmacı tanıma amaçlı kullanılan ideal bir özniteliğin aşağıdaki özelliklere sahip olması gerekir [79].

- Konuşmacılar arasında yüksek değişim,
- Konuşmacı içinde düşük değişim,
- Kolay ölçüm,
- Değiştirmeye ve taklide karşı güçlü,
- Bozulma ve gürültüye karşı güçlü,
- Diğer özniteliklerden maksimum bağımsızlık

İlk iki madde özniteliğin mümkün olduğunca ayırıcı olmasını gerektirir. Şekil 2.1'de iki boyutlu iki farklı öznitelik kümesi gösterilmiştir. Bu iki öznitelik kümesinden ikincisi konuşmacıları daha iyi ayırır. Birincisinde ise ikinci öznitelik, üçüncü konuşmacıyı diğer iki konuşmacıdan ayırır. Ancak bu özniteliklerden hiçbirisi konuşmacıları tek başına tamamen ayıramaz ve bu nedenle her ikisine de ihtiyaç vardır.

Öznitelikler kolayca ölçülebilir olmalıdır. Bu iki faktör içerir. Birincisi öznitelikler konuşma sırasında sıkılıkla ve doğal olarak oluşmalı, böylece kısa konuşma örneklerinden çıkarılabilmelidir. İkincisi öznitelik çıkarmanın kendisi de kolay olmalıdır. Ayrıca otomatik tanımada kullanılan özniteliklerin bir uzman yardımı olmadan ölçülebilir olması gereklidir.



Şekil 2.1. Ayırıcılığı zayıf ve iyi olan iki boyutlu öznitelik örnekleri

İyi bir öznitelik bozulma/gürültü ve ses taklidi gibi çeşitli etkilere karşı güçlü olmalıdır. Son olarak konuşma sinyalinden çıkarılan farklı öznitelikler birbirlerinden maksimum bağımsız olmalıdır. Eğer ilişkili iki öznitelik birleştirilirse bundan hiçbir şey kazanılmaz hatta tanıma oranını düşürebilir.

Uygulamada bu kriterlerin tamamını karşılamak mümkün değildir ve her zaman belirli bir durum için daha önemli olan kriterlere göre öznitelikler arasında bir tercih yapılır. Konuşma ve konuşmacı tanıma sistemlerinde genellikle MFCC gibi spektral öznitelikler kullanılırken konuşmacıları yaş, cinsiyet ve psikolojik durum gibi özelliklerine göre sınıflandıran sistemlerde ise genellikle düşük seviyeli tanımlayıcılarda oluşan (low-level descriptors) geniş bir öznitelik kümesi kullanılır [80]. Aşağıda konuşma ve konuşmacı tanıma sistemlerinde yaygın olarak kullanılan öznitelikler tanıtılmıştır.

2.1.1. Prosodik Öznitelikler

Zamanlama, tonlama ve ritim gibi sesli iletişimde önemli rollere sahip özelliklerle ilişkili olan prosodik öznitelikler özellikle konuşmacı sınıflandırma sistemlerinde yaygın olarak kullanılır. Bu özellikler birden fazla foneme yayıldığı için prosodik öznitelikler suprasegmental (konuşmanın bütünü ile ilgili) olarak kabul edilir. Prosodik özniteliklerin oluşumu ses kaynağına ve ses yolunun şekline bağlıdır [70]. Ses kaynağı etkileri solunum kasları ve ses tellerindeki değişim ile, ses yolu etkileri ise artikulatör hareketleri ile ilişkilidir. Yaygın olarak kullanılan prosodik öznitelikler aşağıda tanıtılmıştır.

2.1.1.1. Perde

Ses tellerinin titreşimi sonucunda oluşan perde sinyali ile ilişkili yaygın olarak kullanılan iki öznitelik vardır; perde frekansı ve gırtlaksal hava akış hızı [81]. Ses tellerinin titreşim oranı temel frekans veya perde frekansı olarak, ses tellerinin titreşimi sırasında gırtlaktan geçen hava akışının hızı ise gırtlaksal hava akışı olarak tanımlanır. Perde sinyalinin tahmini için kullanılan en yaygın yöntem otokorelasyon fonksiyonuna dayalıdır [81]. Öncelikle sinyalin 900 Hz'den yüksek frekansları filtrelenir, daha sonra sinyal $f_s(n;m)$ kısa zaman segmentlerine ayrılır. Son aşamada ise ilk formant frekansının perde ile karışmasını engellemek için lineer olmayan bir kırpma prosedürü uygulanır. Bu işlem,

$$\hat{f}_s(n;m) = \begin{cases} f_s(n;m) - C_{thr} & Eğer\ |f_s(n;m)| > C_{thr} \\ 0 & Eğer\ |f_s(n;m)| < C_{thr} \end{cases} \quad (2.1)$$

ifadesiyle verilir. Burada C_{thr} maksimum $f_s(n;m)$ değerinin yaklaşık %30'u dur. Daha sonra

$$r_s(\eta;m) = \frac{1}{N} \sum_{n=m-N+1}^{m} \hat{f}_s(n;m)\hat{f}_s(n-\eta;m) \quad (2.2)$$

ifadesi ile tanımlanan otokorelasyon fonksiyonu hesaplanır. Burada η gecikmedir. Son olarak ise m'inci çerçevenin perde frekansı

$$\hat{F}_0(m) = \frac{1}{N}\ argmax_{\eta}\ \{|r(\eta;m)|\}_{\eta=N/(F_l/F_s)}^{\eta=N/(F_h/F_s)} \quad (2.3)$$

ifadesiyle hesaplanır. Burada F_s örnekleme frekansına F_l ve F_h ise sırasıyla insanın algıladığı en yüksek ve en düşük perde frekanslarına karşılık gelmektedir. Normalde $F_s = 8000\,Hz$, $F_l = 50Hz$ ve $F_h = 500\,Hz$ olarak kabul edilir [81]. Otokorelasyon fonksiyonun maksimum değeri $max\{|r(\eta;m)|\}_{\eta=N/(F_l/F_s)}^{\eta=N/(F_h/F_s)}$ gırtlaksal hava akış hızını temsil eder.

2.1.1.2. Enerji

Bir konuşma çerçevesinin kısa zaman enerjisi

$$E = log \sum_{n=1}^{N} s_n^2 \tag{2.4}$$

ifadesiyle hesaplanır. Buradaki logaritma fonksiyonu sesin yoğunluğu ile algılanan gürlüğü arasındaki doğrusal olmayan ilişkiyi temsil eder. Ayrıca kaydedilen sinyalin yoğunluğu mikrofon uzaklığı gibi parametrelerden oldukça fazla etkilendiği için genellikle kısa zaman enerji özelliği normalize edilerek kullanılır.

2.1.1.3. Süre

Süre öznitelikleri konuşma oranı, süre ve duraklama gibi konuşma tarzı ile ilişkilidir. En sık kullanılan süreye dayalı öznitelik olan konuşma oranı bir konuşma segmentinin gözlenen süresi ile beklenen sürenin oranına eşittir. Konuşma süresinin duraklama süresine oranı, en uzun duraklama süresi, bir konuşma segmentineki duraklama sayısı ve hatta temel frekansın extramum noktaları süreye dayalı özniteliklerden bazılarıdır. Bu öznitelikler özellikle ruhsal durum tanıma çalışmalarında yaygın olarak kullanılmaktadır [82-85].

2.1.1.4. Sıfır Geçiş Oranı

Sinyaldeki işaret değişimlerinin sayısını temsil eden sıfır geçiş sayısı konuşma analizinde kullanılan faydalı bir özniteliktir. N elemanlı bir çerçeve için kısa dönem sıfır geçiş ölçüsü

$$Z_s(m) = \frac{1}{N0} \sum_{n=m-N+1}^{m} \frac{|sign\{s(n)\} - sign\{s(n-1)\}|}{2} w(m-n) \tag{2.5}$$

ifadesiyle verilir [70]. Buradaki $sign$ fonksiyonu ise denklem 2.6 ile tanımlanır.

$$sign\{s(n)\} = \begin{cases} +1 & Eğer\ s(n) \geq 0 \\ -1 & Eğer\ s(n) < 0 \end{cases} \tag{2.6}$$

2.1.2. Ses Kalitesi Öznitelikleri

Ses kalitesi öznitelikleri kaynak-filtre modelindeki kaynak sinyalin özelliklerini karakterize eder. Bu tezde formant frekansları, jitter, shimmer ve harmonik gürültü oranı gibi doğrudan konuşma sinyalinden hesaplanabilen ses kalitesi öznitelikleri incelenmiştir.

2.1.2.1. Formantlar

Özellikle sesli harflerde belirgin olan ve ses yolunun rezonanslarını temsil eden formant frekansları sinyalin spektrumunda yüksek enerjili tepelerine karşılık gelir. Bu tepe noktalarının konumu ses yolunun şekline ve fiziksel boyutuna bağlıdır. Düşük frekanstan yüksek frekansa doğru formant frekansları genellikle $F1$, $F2$, ve $F3$ olarak isimlendirilir. Bu frekanslardan $F1$ dil gövdesinin yüksekliği ile ters orantılıdır ve dil yüksekliği arttıkça $F1$ frekansı düşer. $F2$ ise tam olarak olmasa da dil gövdesinin ne kadar önde veya arkada olduğuyla ilişkilidir. Dil ne kadar geride ise $F2$ frekansı o kadar düşüktür. Çoğunlukla ilk iki formant frekansı ($F1$ ve $F2$) sesli harfleri ayırmak için yeterlidir. Konuşma spektrumunun önemli özniteliklerinden olan formant frekansı ve onun bant genişliği genellikle doğrusal öngörü analizi ile tahmin edilir [81].

2.1.2.2. Jitter ve Shimmer Öznitelikleri

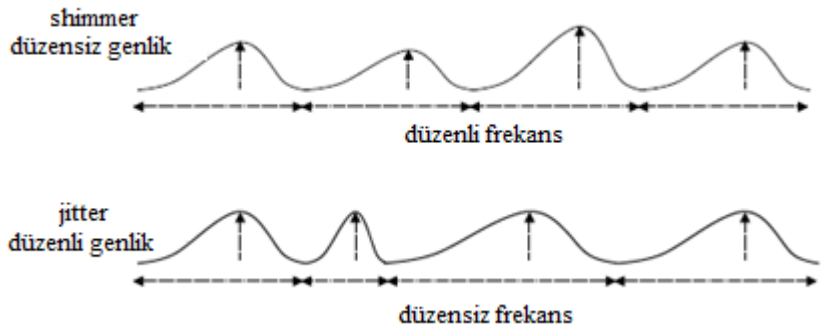
Temel frekans ve genliğin mikro düzeydeki dalgalanmaları temsil eden jitter ve shimmer parametreleri nefeslilik ve sertlik gibi ses kalitesi özellikleriyle ilintilidir [86]. Şekil 2.2'den görüldüğü gibi shimmer genlikteki değişimin, jitter ise frekanstaki değişimin ölçüsüdür. Şekilde verilen iki dalgadan üstteki üçüncü tepede düzensiz genliğe alttaki dalga ise ikinci tepede düzensiz frekansa sahiptir.
Temel frekanstaki değişimleri temsil eden jitter

$$jitter(n) = \frac{|F_0(n+1) - F_0(n)|}{F_0(n)} \tag{2.7}$$

ifadesiyle hesaplanır. Burada $F_0(n)$ n'inci örneğin temel frekansıdır. Shimmer ise enerjideki değişimi temsil eder ve

$$shimmer(n) = \frac{|en(n+1)-en(n)|}{en(n)} \quad (2.8)$$

ifadesiyle hesaplanır. Buradaki $en(n)$ ise n'inci örneğin enerjisidir.



Şekil 2.2. Ses tellerindeki mikro değişimler shimmer ve jitter [86].

2.1.2.3. Harmonik Gürültü Oranı

Ses kalitesini değerlendirmek için kullanılan diğer bir parametre harmonik gürültü oranı (HNR) dir. Temel frekans gibi konuşmanın sesli bölümlerinden hesaplanan HNR sinyalin periyodik bölümünün enerjisi ile çevresel gürültünün enerjisi arasındaki ilişkiyi temsil eder [87]. HNR'nin hesaplanması için öncelikle toplamsal gürültü eklenmiş periyodik sinyal için otokorelasyon fonksiyonunun (ACF) yerel maksimumu belirlenir. ACF'nin T_0 noktasındaki yerel maksimum değeri sinyalin periyodik bölümünün gücüne karşılık gelir. Sinyalin gürültülü parçasının gücü ise sinyalin toplam gücü olan ACF_0^S ile periyodik bölümünün gücü olan $ACF_{T_0}^S$ arasındaki fark olarak kabul edilir. Sonuç olarak HNR dB olarak,

$$HNR = 10\ log\ \frac{ACF_{T_0}^S}{ACF_0^S - ACF_{T_0}^S} \quad (2.9)$$

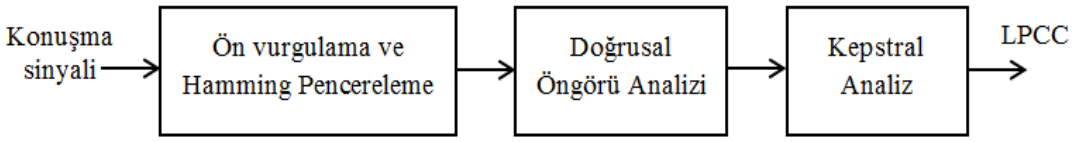
ifadesiyle hesaplanır.

2.1.3. Spektral Öznitelikler

Zamanla değişen bir filtre olarak davranış gösteren ses yolu sistemi rezonans ve anti rezonanslar şeklinde ses yolundaki değişimlerle karakterize edilir. Ses yolunun şekli hem üretilen ses birimi hakkında hem de konuşmacının cinsiyet, yaş ve psikolojik durum gibi kişisel özellikleri hakkında bilgiler taşır. Bu bilgileri temsil etmek için genellikle MFCC, PLP ve LPCC gibi konuşmanın spektrumundan elde edilen öznitelikler kullanılır. Aşağıdaki bölümde bu öznitelikler tanıtılmıştır.

2.1.3.1. Doğrusal Öngörü Kepstral Katsayıları

Doğrusal Öngörü Kepstral Katsayıları (LPCC) konuşmaya dayalı tanıma sistemlerinde yaygın olarak kullanılan bir öznitelik çıkarma yöntemdir. Bu yöntemin arakasındaki temel düşünce herhangi bir konuşma örneğinin kendinden önceki örneklerin doğrusal bir kombinasyonu ile öngörülebileceği fikridir. Blok diyagramı Şekil 2.3'te verilen yöntemin aşamaları aşağıda tanıtılmıştır.



Şekil 2.3 LPCC algoritmasının blok diyagramı

Ön-vurgulama ve Hamming Pencereleme: Algoritmanın ilk aşaması olan ön vurgulamanın amacı konuşma sinyalindeki yüksek frekanslı bileşenleri güçlendirerek spektral enerjiyi eşitlemektir [46]. Bu amaçla genellikle bir dereceli bir FIR filtresi kullanılır. Denklem 2.10 ile verilen ön vurgulama filtresinde a sabittir ve genellikle 0.95 ile 1 arasında seçilir

$$H_p(z) = 1 - az^{-1} \tag{2.10}$$

Ön vurgulamadan sonra konuşma sinyali bir pencere fonksiyonu kullanılarak çerçevelere ayrılır. Bu işlem için genellikle uzunluğu 20-30 ms arasında değişen bir Hamming penceresi kullanılır. Seçilen pencere fonksiyonu sinyal üzerinde belirlenen adım miktarında (genellikle 10 ms) kaydırılarak pencereleme işlemi tamamlanır.

Doğrusal Öngörü Analizi: Doğrusal öngörü yaklaşımı ilk olarak 1971 yılında Atal tarafından ses kodlama amacıyla tanıtılmıştır [88]. Spektral zarfı temsildeki başarısı ve hesaplama kolaylığı yönteme olan ilgiyi arttırmış ve başta ASR olmak üzere birçok ses işleme çalışmasında yaygın olarak kullanılmaya başlanmıştır. Bu yaklaşıma göre herhangi bir andaki konuşma örneği

$$\hat{x}[n] = \sum_{k=1}^{p} a_k x[n-k] \tag{2.11}$$

şeklinde geçmiş p örneğin doğrusal kombinasyonu olarak temsil edilir [46, 47]. Buradaki $\{a_k\}_{k=1}^{p}$ değerleri öngörü katsayıları olup bir konuşma çerçevesi için sabit olduğu kabul edilir. Optimum öngörü katsayıları

$$E_m = \sum_n \left(x[n] - \sum_{k=1}^{p} a_k x[n-k)\right)^2 \tag{2.12}$$

ifadesiyle hesaplanan karesel hata fonksiyonun minimizasyonu ile belirlenir. Bunun için denklem 2.12'nin a_k'ya göre türevi alınır ve sıfıra eşitlenir.

$$\frac{\partial E_m}{\partial a_k} = 0, \quad 1 \leq k \leq p \tag{2.13}$$

Bu işlem sonucunda

$$\emptyset[i,0] = \sum_{k=1}^{p} a_k \emptyset[i,k] \quad 1 \leq i \leq p \tag{2.14}$$

ifadesiyle temsil edilen p bilinmeyenli p adet denklem elde edilir. Burada $\emptyset$ korelasyon fonksiyonu olup

$$\emptyset[i,k] = \sum_n x[n-i]x[n-k] \tag{2.15}$$

ifadesiyle hesaplanır. Denklem 2.12'de seçilen hata aralığına bağlı olarak 2.14 denklemlerinin çözümü için iki farklı yöntem kullanılır; kovaryans ve otokorelasyon [46]. Aslında bu iki yöntem arasında büyük bir fark yoktur. Ancak otokorelasyon yöntemi hem hesaplama yönünden daha verimlidir, hem de her zaman kararlı bir sonuç verir [46, 70]. Bu nedenle öngörü katsayıların belirlenmesinde genellikle otokorelasyon yöntemi tercih edilir.

Otokorelasyon yöntemi;

$$R\ x\ a = r \tag{2.16}$$

ifadesiyle tanımlanan basit bir matris çözümüne karşılık gelir. Burada R Toeplitz olarak isimlendirilen özel bir matris türü (simetrik ve köşegen üzerindeki bütün elemanlar aynı değere sahip), a öngörü katsayılarını içeren vektör, r ise otokorelasyon matrisidir [46, 89]. R matrisi ve r vektörü

$$R[k] = \sum_{n=0}^{N-1-k} s[n]s[n-k] \tag{2.17}$$

ifadesiyle tanımlanan otokorelasyon katsayılarından oluşur ve

$$\begin{bmatrix} \mathrm{R}[0] & \mathrm{R}[1] & \dots & \dots & \mathrm{R}[p-1] \\ \mathrm{R}[1] & \mathrm{R}[0] & \dots & \dots & . \\ \dots & \dots & \dots & \dots & \mathrm{R}[1] \\ \mathrm{R}[p-1] & \dots & \dots & \mathrm{R}[1] & \mathrm{R}[0] \end{bmatrix} \begin{bmatrix} \mathrm{a}_1 \\ \mathrm{a}_2 \\ . \\ \mathrm{a}_\mathrm{p} \end{bmatrix} = \begin{bmatrix} \mathrm{R}[1] \\ \mathrm{R}[2] \\ . \\ \mathrm{R[p]} \end{bmatrix} \tag{2.18}$$

şeklinde tanımlanır. Bu ifadenin sol tarafındaki matrisin toeplitz olması çözümü için özyinelemeli yöntemlerin kullanımına imkan tanır. Lewinson-Durbinin algoritması [46, 89] 2.18 denklemlerinin çözümü için kullanılan özyinelemeli bir yöntemdir ve aşağıdaki adımlardan oluşur.

$$E^0 = R[0]$$

$$k_i = \frac{R[i] - \sum_{j=1}^{i-1} a_j{}^{(i-1)} R[i-j]}{E^{(i-1)}} \qquad 1 \le \mathrm{i} \le \mathrm{p}$$

$$a_i{}^{(i)} = k_i$$

$$a_j{}^{(i)} = a_j{}^{(i-1)} - k a_{i-j}{}^{(i-1)} \qquad 1 \le j \le i-1$$

$$E^{(i)} = \left(1 - {k_i}^2\right)E^{(i-1)} \tag{2.19}$$

Yukarıdaki denklemler $i = 1, 2, \ldots, p$ için özyinelemeli olarak çözülür ve p'inci tekrar sonunda

$$a_j = {a_j}^{(p)} \qquad 1 \leq j \leq p \tag{2.20}$$

ifadesiyle temsil edilen doğrusal öngörü katsayıları hesaplanır.

LPC-Kepstrum: Doğrusal öngörü katsayılarının doğrudan öznitelik olarak kullanımı oldukça nadirdir [70]. Yapılan çalışmalarda ardışık öngörü katsayılarının yüksek ilintiye sahip olduğu ve daha az ilintili bir öznitelik temsilinin daha etkin olacağı görülmüştür [90]. Bu nedenle doğrusal öngörü katsayıları yerine bu katsayılardan türetilen ve daha az ilintili olan doğrusal öngörü kepstrum katsayıları kullanılır. Bu amaçla aşağıdaki özyinelemeli ifadeyle doğrusal öngörü katsayılarından türetilen kepstral katsayılar kullanılır [47].

$$c[n] = \begin{cases} a[n] + \sum_{k=1}^{n-1} \frac{k}{n} c[k]a[n-k], & 1 \leq n \leq p \\ \sum_{k=n-p}^{n-1} \frac{k}{n} c[k]a[n-k], & n > p \end{cases} \tag{2.21}$$

İlk olarak Atal [91] tarafından konuşmacı tanıma için önerilen (2.21) ilişkisi günümüzde hem konuşmacı hem de konuşma tanıma çalışmalarında yaygın olarak kullanılmaktadır.

Uygulamada genellikle p adet LPCC katsayısıyla oluşturulan öznitelik vektörüne çerçevelerin logaritmik enerjisi de eklenir. Böylece logaritmik enerjiyle birlikte öznitelik vektörünün boyu $p + 1$ olur. Uzunluğu N olan bir $s[n]$ konuşma çerçevesinin logaritmik enerjisi

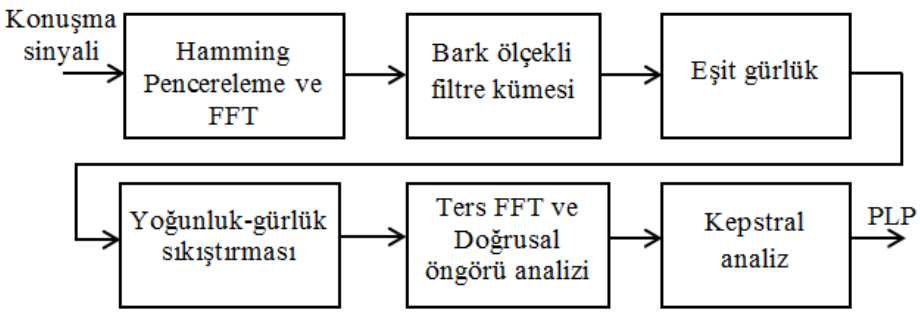
$$Enerji_{log} = ln(\sum_{n=0}^{N-1} s^2[n]) \tag{2.22}$$

ifadesiyle hesaplanır.

2.1.3.2. Algısal Doğrusal Öngörü Katsayıları

Algısal Doğrusal Öngörü Katsayıları (PLP) insanın işitme sistemini taklit etmeye çalışan diğer bir öznitelik çıkarma yöntemidir. PLP'nin arkasında üç temel kavram vardır. Bunlar kritik bant frekans seçiciliği, eşit gürlük eğrisi ve yoğunluk-gürlük güç kanunudur.

Bu yöntemde ilk olarak pencerelenmiş bir çerçevenin FFT spektrumu elde edilir. Daha sonra insan kulağının frekans bant seçiciliğini modelleyen Bark ölçekli bir filtre kümesi uygulanır. Bundan sonra filtre çıkışları insanın duyma hassaslığını tahmin eden eşit-gürlük eğrisine göre ağırlıklandırılır. Ağırlıklandırılmış filtre çıkışları daha sonra sinyal yoğunluğu ve algılanan gürlük arasındaki ilişkiyi tanımlayan yoğunluk-gürlük güç kanununa göre sıkıştırılır. Filtre çıkışlarına ters Fourier dönüşümü uygulanır ve doğrusal öngörü analizi gerçekleştirilir. Son olarak kepstral analiz uygulanır ve PLP katsayıları elde edilir. PLP algoritmasının blok diyagramı Şekil 2.4'te verilmiştir.



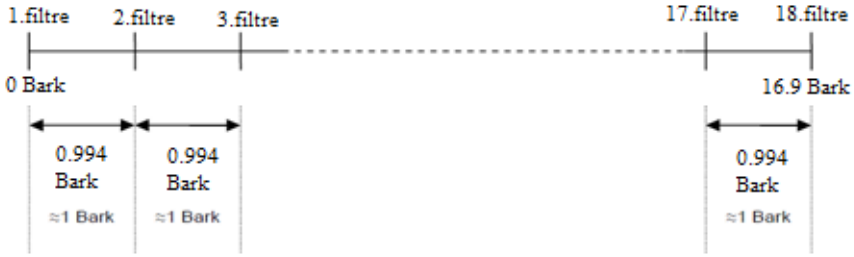
Şekil 2.4. PLP algoritmasının blok diyagramı

Hamming Pencereleme ve FFT: Algoritmanın ilk aşamasında konuşma sinyali Hamming penceresi kullanılarak çerçevelere bölünür. Konuşmanın spektrumunu elde etmek için pencerelenmiş çerçeveler üzerinde hızlı Fourier dönüşümü (FFT) uygulanır. Elde edilen FFT spektrumu $S[k]$, sonraki işlemler için diğer adıma iletilir.

Bark-Ölçekli Filtre Kümesi: PLP analizindeki filtre kümesi Bark ölçeği olarak isimlendirilen doğrusal olmayan bir frekans ölçeğine dayalıdır. Bark ölçeği ile doğrusal frekans arasındaki matematiksel ilişki denklem 2.23 ile gösterilir.

$$f_{Bark} = 6 \ln\left(\frac{f}{600} + \left(\left(\frac{f}{600}\right)^2 + 1\right)^{0.5}\right) \tag{2.23}$$

Burada f Hz cinsinden frekansı, f_{Bark} ise bu frekansa karşılık gelen Bark frekansını temsil eder. Bu filtre kümesindeki filtrelerin merkez frekansları Bark ölçeğinde eşit olarak yayılmıştır. Hermansky [92] bu filtrelerin merkez frekanslarının yaklaşık olarak 1 Bark aralıklı olmasını önermiştir. Hermansky ayrıca ilk ve son filtrenin sırasıyla 0 Bark ve Nyquist frekansına yerleştirilmesini ve bu filtrelerin şekillerinin de en yakın komşu filtreyle aynı olmasını önermiştir. Bu durumda 5 kHz (16.9 Bark) Nyquist frekansı için ilk ve son filtreler sırasıyla 0 ve 16.9 Bark'ta olmak üzere aralarında 16 filtre olacak şekilde yerleştirilecektir. Böylece Şekil 2.5'teki gibi adım genişliği 0.994 Bark olan toplam 18 filtre oluşur.



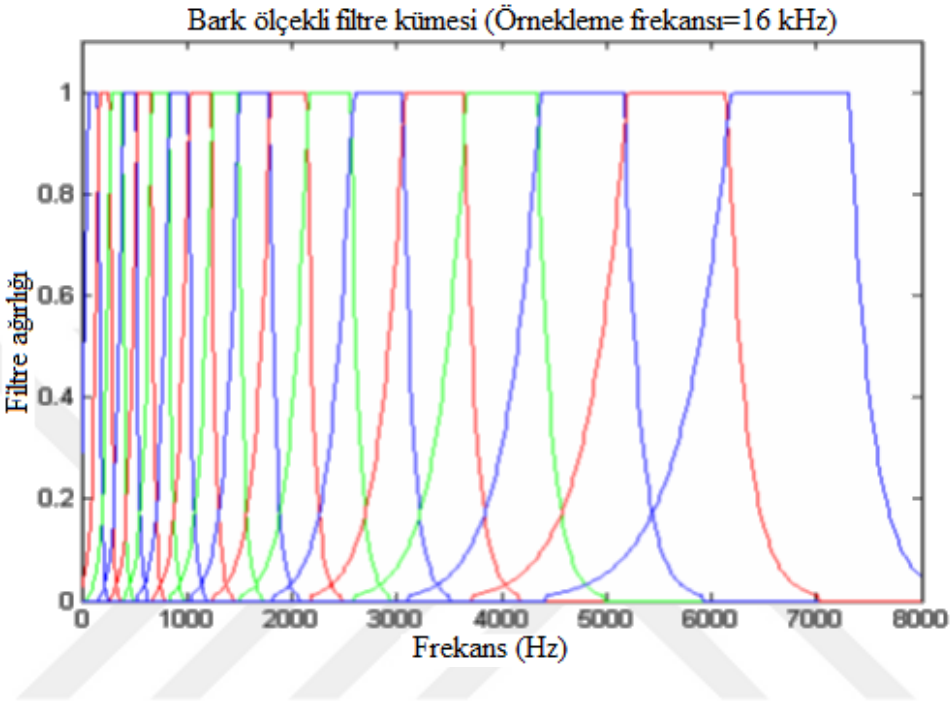
Şekil 2.5. Bark ölçekli bir filtre kümesindeki merkez frekansın dağılımı (Nyquist frekansı = 5 kHz (16.9 Bark))

Filtrelerin şekli Bark ölçeğine göre özdeştir ve aşağıdaki denklemle tanımlanır.

$$\Psi = \begin{cases} 0 & f_{Bark} - f_{c(Bark)} < -2.5 \\ 10^{(f_{Bark} - f_{c(Bark)} + 0.5)} & -2.5 \le f_{Bark} - f_{c(Bark)} \le -0.5 \\ 1 & -0.5 < f_{Bark} - f_{c(Bark)} < 0.5 \\ 10^{(f_{Bark} - f_{c(Bark)} + 0.5)} & 0.5 \le f_{Bark} - f_{c(Bark)} \le 1.3 \\ 0 & f_{Bark} - f_{c(Bark)} > 1.3 \end{cases} \tag{2.24}$$

Burada $f_{c(Bark)}$ bir filtrenin bark ölçeğindeki merkez frekansı, Ψ ise f_{Bark} daki filtre ağırlığıdır. 16 kHz örnekleme frekansı için Bark ölçekli filtre kümesi Şekil 2.6'da gösterilmiştir. Filtre şekilleri Bark ölçeğinde aynı olmasına rağmen doğrusal frekans

ekseninde yüksek frekans bölgesindeki filtrelerin daha geniş olduğu görülmektedir. Bunun nedeni Bark ölçeği ile doğrusal frekans ölçeği arasındaki doğrusal olmayan (2.23) ilişkisidir.



Şekil 2.6. Bark ölçekli filtre kümesi (Örnekleme frekansı = 16kHz)

Her filtrenin çıkışı konuşmanın güç spekturumundaki her bir FFT noktası ile filtre ağırlıklarının çarpımlarının toplamı olup matematiksel olarak;

$$X_m = \sum_{k=0}^{\frac{N}{2}-1} |S[k]|^2 |\Psi_m[k]| \qquad 2 \leq m \leq M-1 \tag{2.25}$$

şeklinde tanımlanır. Burada X_m m'inci filtrenin çıkışı, $|S[k]|^2$ pencerelenmiş konuşma çerçevesinin N noktalı güç spektrumu ve $|\Psi_m[k]|$ da ayrık doğrusal frekans ekseni boyunca m'inci Bark filtresinin filtre katsayılarıdır. İlk ve son filtrelerin çıkışları bu aşamada hesaplanmaz. Çünkü bu iki filtrenin şekilleri kendilerine en yakın komşularının şekliyle aynı olacaktır. Bu filtrelerin çıkışları sonraki aşamada hesaplanacaktır.

Eşit-Gürlük Eğrisi: Eşit gürlük eğrisi insanın farklı frekanslardaki duyma hassaslığını tahmin eder. Nyquist frekansına göre iki farklı tahmin kullanılır [92]. Nyquist frekansı 5 kHz'ye kadar olan uygulamalar için

$$E = \frac{(\omega^2+56.8\times10^6)\omega^4}{(\omega^2+6.3\times10^6)^2(\omega^2+0.38\times10^9)} \tag{2.26}$$

ifadesiyle tanımlanan tahmin kullanılır. Burada $\omega\ radyan^{-1}$ cinsinden açısal frekans ($\omega = 2\pi f$) dır. Nyquist frekansı 5 kHz'nin üzerinde olan uygulamalar için ise

$$E = \frac{(\omega^2+56.8\times10^6)\omega^4}{(\omega^2+6.3\times10^6)^2(\omega^2+0.38\times10^9)(\omega^6+9.58\times10^{26})} \tag{2.27}$$

ifadesiyle tanımlanan tahmin kullanılır. Her bir filtrenin eşit gürlük ağırlığı E onun merkez frekansı (radyan frekansı biçiminde) Nyquist frekansına bağlı olarak yukarıdaki iki eşitlikten birinde yerine koyularak hesaplanabilir.

Filtre kümesindeki her filtrenin çıkışı

$$X_{m(e)} = E_m X_m \quad 2 \le m \le M-1 \tag{2.28}$$

ifadesiyle ağırlıklandırılır. Burada $X_{m(e)}$ eşit-gürlük ağırlığı E_m ile çarpılan m'inci filtre çıkışına karşılık gelir. Alternatif olarak eşit gürlük ağırlıkları filtre ağırlıklarına dolaylı olarak uygulanabilir (denklem 2.29). Böylece filtre çıkışları değiştirilmiş filtre ağırlıklarından doğrudan elde edilebilir (Denklem 2.30).

$$\Psi_{m(e)}[k] = E_m \Psi_m[k] \ \ 2 \le m \le M-1 \tag{2.29}$$

$$X_{m(e)} = \sum_{k=0}^{\frac{N}{2}-1} |S[k]|^2 \left|\Psi_{m(e)}[k]\right| \qquad 2 \le m \le M-1 \tag{2.30}$$

İlk ve son filtrenin şekli kendilerine en yakın komşularıyla aynı olduğu için bu filtrelerin çıkışları

$$X_{1(e)} = X_{2(e)} \tag{2.31}$$

$X_{M-1(e)} = X_{M(e)}$ dir. (2.32)

şeklinde kendilerine en yakın filtre çıkışlarına eşit yapılır.

Yoğunluk-Gürlük Sıkıştırması: Yoğunluk gürlük sıkıştırması belirli bir sinyal yoğunluğu için insanın doğrusal olmayan gürlük algılamasını tahmin eder. PLP'de bu ilişki

$$\emptyset_m = \left(X_{m(e)}\right)^{0.33} \quad 1 \leq m \leq M \tag{2.33}$$

şeklinde küp kök fonksiyonuyla temsil edilir [92]. Burada $\emptyset_m$ yoğunluk-gürlük sıkıştırması sonrası filtre çıkışlarına karşılık gelir.

PLP algoritmasının sonraki aşaması Bölüm 2.1.3.1'de açıklanan Doğrusal Öngörü Analizinin gerçekleştirilmesidir. İlk olarak filtre çıkışlarının otokorelasyon fonksiyonunun hesaplanması gerekmektedir. $\emptyset_m$ filtre çıkışlarına ters fourier dönüşümü uygulanır. $\emptyset_m$ spektrumun yalnızca yarısıdır. Ayrıca ilk ve son $\emptyset_m$ ona en yakın komşusuna eşitlendiği için tüm spektrum ilk ve son $\emptyset_m$ hariç ilk yarımın aynadaki görüntüsü eklenerek elde edilir. Daha basit ifadeyle tüm spektrum aşağıdaki vektörle ifade edilebilir.

$$\Phi = [\emptyset_1\ \emptyset_2 \ldots \ldots\ \emptyset_{M-1}\ \emptyset_M\ \emptyset_{M-1} \ldots\ \emptyset_3\ \emptyset_2] \tag{2.34}$$

Φ üzerinde ters fourier dönüşümü uygulanır ve p doğrusal öngörü derecesi olmak üzere $0 \leq n \leq p$ için ilk $p+1$ katsayı otokorelasyon fonksiyonu $R[n]$ olarak kabul edilir. Otokorelasyon fonksiyonu 2.19 ve 2.20 denklemleri çözülerek doğrusal öngörü katsayılarının hesaplanması için kullanılır.

Kepstral Analiz: Kepstral analiz bir önceki aşamada elde edilen LPC katsayılarına uygulanır. PLP deki kepstral analiz LPCC deki ile aynıdır. Denklem 2.21'in çözümü ile LPC katsayıları PLP kepstral katsayılarına dönüşür. PLP katsayılarının derecesi LPC ile aynıdır. Böylece her konuşma çerçevesi için p adet PLP katsayısı üretilmiş olacaktır. Bu katsayılar bir öznitelik vektörü oluşturmak için gruplandırılır.

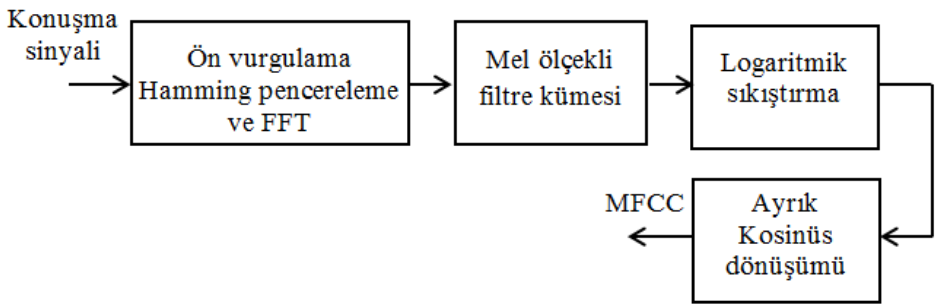
Genellikle her bir öznitelik vektörüne doğrusal öngörü hatasının logaritması eklenir. PLP katsayılarıyla birlikte öznitelik vektörünün boyutu $p+1$ olur. Logaritmik öngörü hatası hepsi kutup filtre kazancının karesinin logaritmasıdır ve

$$PE_{(ln)} = \ln(G^2) = \ln\left(R[0] - \sum_{k=1}^{p} a_k R[k]\right) \quad (2.35)$$

ifadesiyle hesaplanır. Burada $PE_{(ln)}$ logaritmik öngörü hatası, G ise hepsi kutup filtre modelinin filtre kazancıdır.

2.1.3.3. Mel Frekanslı Kepstral Katsayılar

Mel Frekanslı Kepstral Katsayılar (MFCC) konuşma tanıma sistemlerinde yaygın olarak kullanılan bir öznitelik çıkarma yöntemidir [93]. FFT tabanlı olarak bilinen bu yöntemde öznitelik vektörleri pencerelenmiş konuşma çerçevelerinin frekans spektrumlarından çıkarılır. MFCC çıkarma prosedürü Şekil 2.7'de gösterilmiştir.



Şekil 2.7. MFCC algoritmasının blok diyagramı

Ön-Vurgulama, Hamming Pencereleme ve FFT: LPCC deki gibi konuşma sinyali ön vurgulama için bir dereceli bir sayısal filtreden geçirilir. Filtrenin transfer fonksiyonu LPCC'de anlatılan ile aynıdır. Ön vurgulamadan sonra konuşma sinyali Hamming penceresi kullanılarak çerçevelere bölünür. Daha sonra her bir konuşma çerçevesinin spektrumu Fourier dönüşümü (FFT) uygulanarak hesaplanır. N noktalı bir konuşma çerçevesinin FFT spektrumu $S[k]$ $0 \leq k \leq N-1$ ile temsil edilir.

Mel Ölçekli Filtre Kümesi: Mel frekanslı filtre kümesi insanın işitme sistemini taklit eden üçgen şekilli bir bant geçiren filtreler serisidir. Bu filtre kümesi Mel ölçeği olarak isimlendirilen doğrusal olmayan bir frekans ölçeğine dayanmaktadır. Mel frekans ölçeği insan tarafından değerlendirilen perdelerin psikoakustik bir ölçüsüdür. Stevens ve arkadaşlarına göre dinleyici eşiğinin 40 dB üzerindeki 1000Hz lik bir sinyal 1000 mel'lik

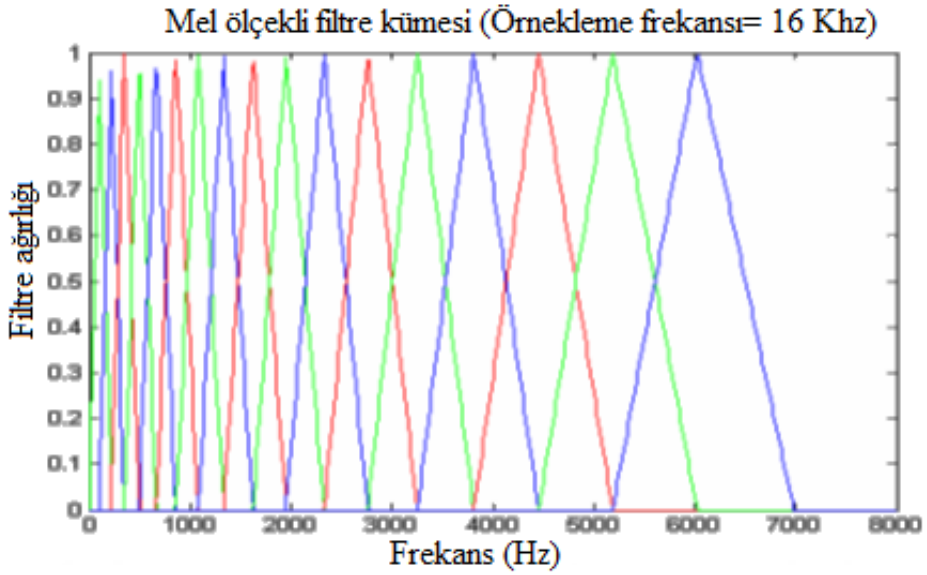
bir perdeye sahip olarak tanımlanır [94]. 1000 Hz'nin altında Mel ölçeği yaklaşık olarak doğrusaldır. 1000 Hz'nin üzerinde ise dinleyiciler aynı perde artışlarını gittikçe artan frekans aralıkları ile algılama eğilimindedir. Bunun için Mel ölçeği ile doğrusal frekans ölçeği arasındaki ilişki doğrusal değildir ve 1000 Hz'nin üzerinde yaklaşık olarak logaritmiktir. Mel ölçeği ile doğrusal frekans ölçeği arasındaki matematiksel ilişki aşağıdaki eşitlikle tanımlanır.

$$f_{Mel} = 1127 \ln\left(\frac{f}{700} + 1\right) \tag{2.36}$$

Burada f_{Mel} Mel cinsinden Mel frekansa f ise Hz cinsinden doğrusal frekansa karşılık gelmektedir. Daha önceden de belirtildiği gibi Mel frekanslı filtre kümesi üçgen şekilli bant geçiren filtreler dizisinden oluşur. Bu filtrelerin alt sınırı önceki filtrenin merkez frekansına, üst sınırı ise sonraki filtrenin merkez frekansına gelecek şekilde üst üste çakışacak şekilde tanımlanır. Üçgen filtrelerin üst köşe noktaları ise filtrelerin merkez frekanslarına karşılık gelir. Mel frekanslı filtre serisinin merkez frekansları Mel ölçeğine göre eşit aralıklıdır. Filtre kümesindeki m'inci filtrenin merkez frekansı aşağıdaki denklemle bulunabilir.

$$f_{cm(Mel)} = f_{L(Mel)} + \frac{m\left(f_{H(Mel)} - f_{L(Mel)}\right)}{M+1} \quad 1 \leq m \leq M \tag{2.37}$$

Burada $f_{cm(Mel)}$ m'inci filtrenin Mel ölçeğindeki merkez frekansı, $f_{L(Mel)}$ ve $f_{H(Mel)}$ ise tüm filtre kümesi tarafından kapsanan frekans aralığının sırasıyla alt ve üst sınırlarıdır. $f_{L(Mel)}$ ile $f_{H(Mel)}$ aralığının içinde M adet üçgensel filtre vardır. Mel ölçekli bir filtre kümesin frekans tepkesi Şekil 2.8'de gösterilmiştir.



Şekil 2.8 Mel ölçekli filtre kümesi (Örnekleme frekansı= 16kHz)

k ayrık frekans indeksi, $|H_m[k]|$ ise m'inci filtrenin frekans yanıtının genliği olmak üzere m'inci filtrenin çıkışı X_m aşağıdaki denklemle ifade edilebilir.

$$X_m = \sum_{k=0}^{\frac{N}{2}-1} |S[k]|^2 |H_m[k]| \quad 1 \leq m \leq M \tag{2.38}$$

Burada $S[k]$ konuşma çerçevesinin N noktalı FFT spektrumudur. FFT spektrumun yarısı diğer yarısının aynadaki görünüşü olduğu için bu toplam ifadesindeki nokta sayısı $\frac{N}{2}$ olacaktır.

Logaritmik Sıkıştırma: Verilen bir sinyal yoğunluğunun algılanan gürlüğünü modellemek için filtre çıkışları logaritmik bir fonksiyon ile sıkıştırılır.

$$X_{m(ln)} = \ln(X_m) \quad 1 \leq m \leq M \tag{2.39}$$

Yukarıdaki eşitlikte $X_{m(ln)}$ m'inci filtrenin logaritmik olarak sıkıştırılmış çıkışıdır.

DCT: Algoritmanın son aşaması filtre çıkışlarının ters katlanmasıdır. Filtre çıkışlarına Ayrık Kosünüs Dönüşümü (DCT) uygulanır ve ilk birkaç katsayı konuşma çerçevesinin öznitelik vektörü olarak gruplandırılır.

Mel ölçekli kepstrumun derecesi p olmak üzere öznitelik vektörü, ilk p DCT katsayısı dikkate alınarak elde edilir. Matematiksel olarak k'ıncı MFCC katsayısı aşağıdaki formülle ifade edilebilir.

$$MFCC_k = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} X_{m(ln)} cos\left(\frac{\pi k(m-0.5)}{M}\right) \quad 1 \leq k \leq p \tag{2.40}$$

Sabit öznitelik vektörüne genellikle logaritmik enerji bileşeni veya sıfır dereceli katsayı yada her ikisi eklenir. Logaritmik enerji bileşeni Bölüm 2.1.3.1'de denklem 2.22 ile tanımlanmıştır. Sıfır dereceli MFCC katsayısı filtre çıkışının sıfırıncı sıradaki DCT katsayısı olup denklem 2.41 ifadesiyle hesaplanır.

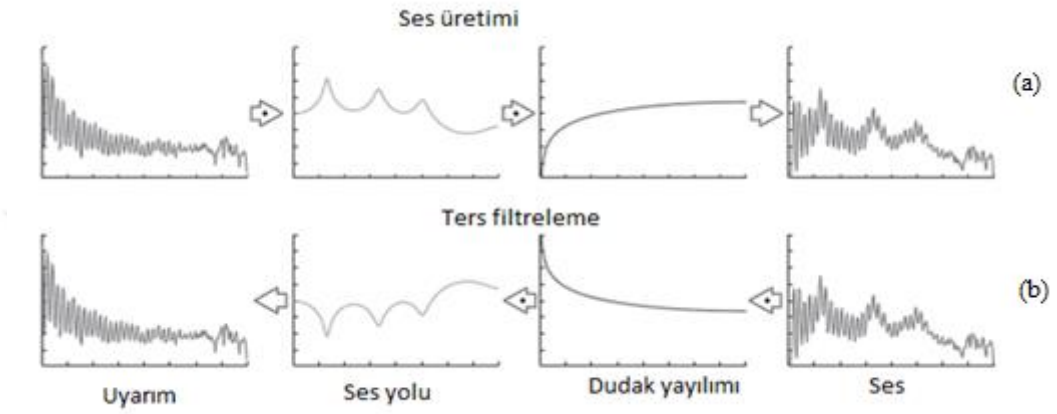
$$MFCC_0 = \sqrt{\frac{1}{M}} \sum_{m=1}^{M} X_{m(ln)} \tag{2.41}$$

2.1.4. Glottal Sinyalden Çıkarılan Öznitelikler

Ses sinyali, özellikle sesli seslerle ilişkili olan ses sinyali (örneğin sesli harfler) akciğerlerdeki daralma ve genişleme ile başlar. Bu daralma ve genişleme akciğerlerdeki hava ile ağıza yakın hava arasında bir basınç farkı oluşturur. Oluşan hava akışı ise ses tellerinin arasından geçer ve ses tellerinin titreşimi sonucunda bir darbe dizisine dönüşür. Hava darbeleri tarafından oluşturulan bu basınç sinyali kısmen periyodiktir ve glottal sinyal olarak isimlendirilir.

Konuşma sırasında ses tellerinin titreşiminin incelenmesi gırtlağın yerinden dolayı zordur. Fakat gırtlaktaki aktiviteleri incelemeyi sağlayan çeşitli teknikler mevcuttur. Bu tekniklerden bazıları cerrahi operasyon veya ek donanımlar gerektirir. Örneğin ağız veya burun boşluğuna yerleştirilen bir kamera sayesinde gırtlaksal davranışlar incelenebilir. Ancak bu yöntemler doğal konuşmayı engellediği ve verinin elde edilmesi zor olduğu için tercih edilmemektedir. EGG'de ses tellerinin titreşimini incelemek için kullanılan diğer bir yöntemdir. Bu yöntemde gırtlağın her iki yanına yerleştirilen elektrotlar arasındaki direnç ölçülerek ses tellerinin aktiviteleri incelenir. Bu yöntemde özel donanımlar gerektirdiğinden ve veri girişinin zorluğundan dolayı pek tercih edilmemektedir.

Gırtlaksal aktiviteyi incelemede kullanılan en yaygın yöntem ters filtrelemedir. Ters filtreleme kaynak-filtre ses üretim modeline dayanmaktadır. Kaynak-filtre teorisine göre ses tellerinin titreşimiyle oluşan darbe dizisi uyarım kaynağı, ses yolu ise fonem bağımlı bir fitre olarak davranır [56]. Gırtlakta oluşan darbe dizisi ses yolunda şekillendikten sonra dudaktan geçerek çevreye ses olarak yayılır (Şekil 2.9-a).

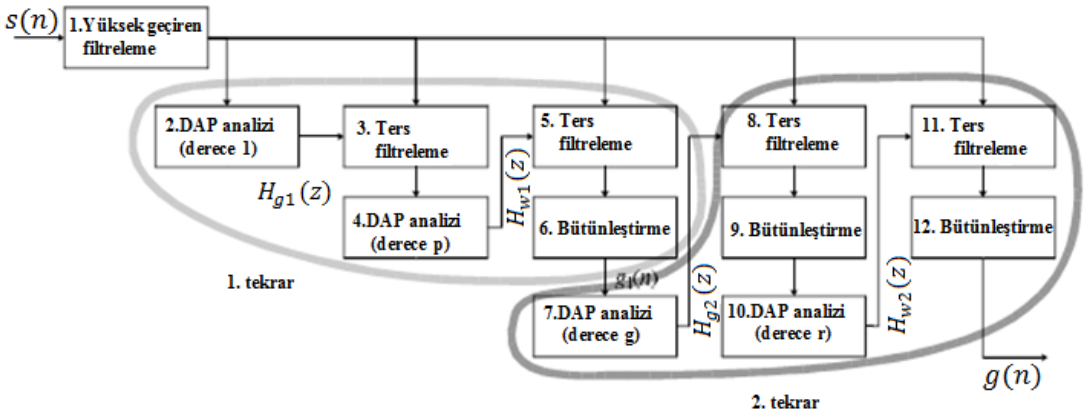


Şekil 2.9. Kaynak-filtre modeline göre ses üretimi ve ters filtreleme süreci

Ters filtreleme, ses sinyalinden ses yolu ve dudağın etkisinin çıkarılması prensibine dayanmaktadır (Şekil 2.9-b). Ters filtrelemede giriş sinyali olarak iki alternatif vardır. Birincisi, bir maske vasıtasıyla elde edilen ağızın dışındaki gerçek hava akışı, diğeri ise belirli bir mesafedeki mikrofonla elde edilen ses basınç sinyali. Maskenin ağız ve burun deliklerine yerleştirilmesi zordur ve doğal konuşmada kısıtlamalara neden olurken günümüzde genlik ve faz tepkileri mükemmel olan mikrofonlar yaygın olarak kullanılmaktadır. Bu nedenle ters filtrelemede giriş olarak genellikle mikrofon kayıtları kullanılmaktadır.

Literatürde ters filtreleme için önerilmiş birçok yöntem vardır. Bu yöntemler otomatik ve manuel olmak üzere ikiye ayrılır. Manuel yöntemlerde ses yolunun formant frekansları elle belirlendiği için hem işlem süresi uzundur hem de sonuçlar kullanıcının kişisel tercihlerine göre değişmektedir. Diğer taraftan otomatik ters filtreleme yöntemlerinde ise tüm işlemler otomatik olarak gerçekleştirilir, böylece hem daha hızlı hem de kullanıcıdan bağımsız sonuçlar elde edilir. Literatürde birçok ters filtreleme yöntemi önerilmiştir. Bu yöntemlerden Paavo Alku tarafından geliştirilen tekrarlamalı uyarlamalı ters filtreleme (IAIF) yöntemi [95], en popüler ters filtreleme yöntemlerinden

birisi olup bu tez çalışmasında da kullanılmıştır. IAIF yöntemi bir ses basınç sinyalini giriş olarak alır ve bu basınç sinyaline karşılık gelen glottal akış sinyalini tahmin eder. Yöntem birbirinin tekrarı olan iki aşamadan oluşur. Şekil 2.10'da gri hatlarla gösterilen bu aşamaların ilkinde gırtlaksal uyarım sinyalinin ilk tahmini üretilir. Daha sonra bu tahmin ikinci aşamaya giriş olarak uygulanır ve bu aşama sonunda daha doğru bir glottal uyarım sinyali elde edilir. IAIF yönteminin işlem basamakları aşağıda tanıtılmıştır.



Şekil 2.10. IAIF yönteminin blok diyagramı

1. Giriş sinyali ilk olarak düşük frekanslı bozucu dalgalanmaları yok etmek için yüksek geçiren filtre ile filtrelenir. Elde edilen sinyal sonraki aşamalarda giriş olarak kullanılır. Bu işlem sırasında konuyla ilgili bilgilerin filtrelenmemesi için kesim frekansının sinyalinin temel frekansından düşük seçilmesi gerekir.
2. Bir dereceli DAP (discrete all-pole) analizi hesaplanır. Bu adım ses spektrumundaki gırtlaksal akış ve dudak yayılımdan kaynaklanan birleşik etkinin başlangıç tahminini verir.
3. Giriş sinyali ikinci aşamada elde edilen filtre ile ters filtrelenir. Bu adım uyarım sinyalinin spektrumu ve dudak yayılım etkisinden kaynaklanan spektral eğimi ortadan kaldırır.
4. Bir önceki aşamanın sonucu DAP ile analiz edilir ve ses yolu transfer fonksiyonunun bir modeli elde edilir. DAP analiz derecesi, "p" frekans bandında modellenecek formant sayısı ile ilişkilidir ve IAIF yönteminin işletmeni tarafından ayarlanabilir. Pratik olarak "p" analiz edilen sinyalin örnekleme

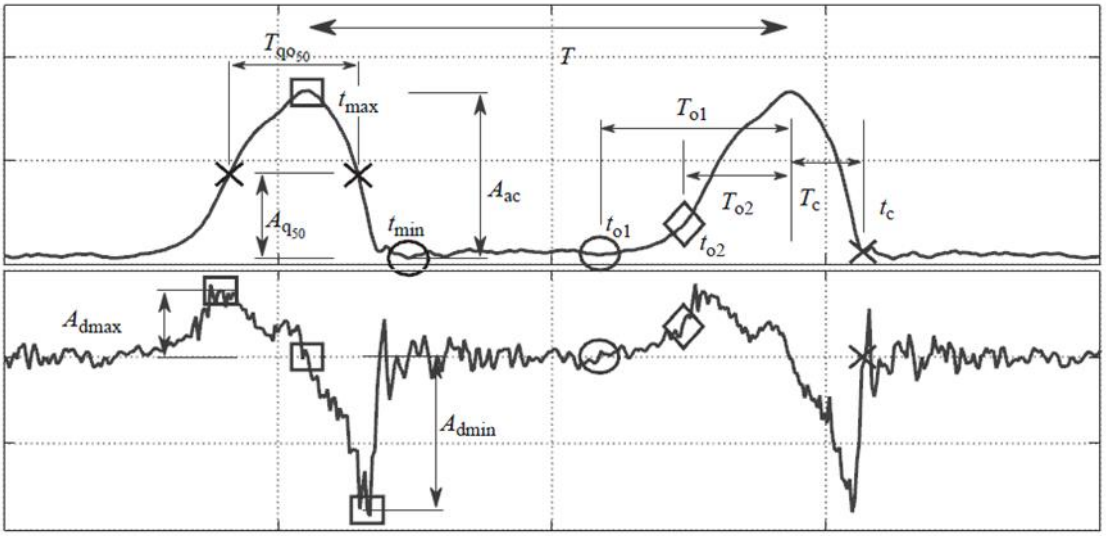
frekansının kHz cinsinden değerine küçük bir sayının eklenmesiyle elde edilen çift bir tamsayı olmalıdır.

5. Giriş sinyali adım 4'te elde edilen "p" dereceli modelin tersi kullanılarak ters filtrelenir.
6. Bir önceki adımın sonucu dudak yayılım etkisini iptal etmek için tümleştirilir. Bu gırtlaksal akışın ilk tahminini üretir ve ilk tekrar tamamlanmış olur.
7. İkinci tekrar elde edilen gırtlaksal akış tahminin "g" dereceli DAP analizi hesaplanarak başlar. Bu işlem gırtlaksal uyarımın ses spektrumundaki etkisinin spektral bir modelini verir. "g" değeri genellikle 2 ile 4 arasında seçilir.
8. Giriş sinyali gırtlaksal katkıyı ortadan kaldırmak için uyarım sinyalinin modeli kullanılarak ters filtrelenir.
9. Önceki aşamanın sonucu tümleştirilerek dudak yayılımı iptal edilir.
10. Ses yolu fitresinin yeni modeli "r" dereceli DAP analiziyle oluşturulur. Kullanıcı tarafından ayarlanabilen "r" değeri genellikle adım 4 deki "p" değerine eşit seçilir.
11. Giriş sinyali bir önceki aşamada elde edilen ses yolu modeli ile ters filtrelenir ve ses yolunun etkisi giriş sinyalinden çıkarılır.
12. Son olarak sinyal tümleştirilerek dudak yayılım etkisi iptal edilir. Bu adım IAIF yönteminin sonucu olan gırtlaksal akışın son tahminini üretir.

Ters filtrelemeyle tahmin edilen gırtlaksal akış sinyali genellikle nicel parametrelere dönüştürülerek analiz edilir. Zaman-uzayı, frekans-uzayı ve model temelli olmak üzere üç sınıfa ayrılan bu parametreler aşağıda tanıtılmıştır.

2.1.4.1 Zaman-Uzayı Parametreleri

Zaman-uzayı parametrelerinin belirlenmesi için öncelikle tahmin edilen gırtlaksal akış çerçevelerinden belirli zaman ve genlik anlarının çıkarılması gerekir. Bu kritik anlar elde edildikten sonra farklı zaman ve genlik temelli parametreler hesaplanır. Gırtlaksal açılma t_0 ve maksimum akış t_{max} gibi farklı kritik anlar Şekil 2.11'de gösterilmiştir.



Şekil 2.11. Gırtlaksal akışın zaman-uzayı parametrelerini hesaplanmasında kullanılan zaman ve genlik anları. Üst bölümde gırtlaksal akış tahmini, alt bölümde ise karşılık gelen türev temsil edilmiştir.

Parametreleştirme işlemi tek bir çerçeve üzerinde gerçekleştirilir. k tane ardışık gırtlaksal akış periyodunu içeren bu çerçevenin uzunluğu genellikle 20 ile 100 ms arasında değişir. İlk olarak sinyal, elde edilen zaman anlarındaki yüksek frekanslı gürültü etkisini ortadan kaldırmak için biraz yumuşatılır. Bu işlem dört vuruşlu doğrusal fazlı düşük geçiren bir FIR filtresi kullanılarak gerçekleştirilir.

Temel periyot uzunluğu T, sinyal çerçevesinin temel frekansı f_0 hesaplanıp ters çevrilerek elde edilir. Daha sonra tüm çerçevenin maksimum örnek değerli zaman anı t^*_{max} elde edilir. $S = \{t_{max,k}\}$ çerçevedeki maksimum tepelerin kümesi olmak üzere $t^*_{max} \in S$ olarak kabul edilir. Bu yüzden diğer tepelerin konumları $t_{max,k}$, t^*_{max} 'dan önce ve sonra T'nin katlarındaki zaman dilimlerinde yerel maksimum aranarak elde edilebilir.

$t_{min,k}$ noktaları $t_{max,k}$ noktalarından sonra aranır. Tepeden tepeye darbe genlikleri $A_{ac,k}$, $A_{max,k} - A_{min,k}$ şeklinde hesaplanır. Çerçevenin türevinin maksimum değeri t_{dmax}, t_{max}'ın solunda, minimum değeri t_{dmin} ise sağında araştırılır. Bulunan genlik değerleri ise A_{dmax} ve A_{dmin} olarak kaydedilir.

Kapanma anı t_c akışın türevinde t_{dmin} 'den sonraki ilk pozitif sıfır geçişi bulunarak tahmin edilir. Ancak gırtlaksal darbenin kademeli açılması nedeniyle açılma anı çelişkilidir ve bu yüzden birincil ve ikincil açılma anları (t_{o1}, t_{o2}) tahmin edilir. Birincil açılma anını tespit etmek için $A_{o,10\%}$ eşik seviyesi $t_{min,k}$ genliğinin %10 fazlası olarak tanımlanır. Buna

karşılık gelen zaman anı elde edilir ve türev pozitif olduğu sürece veya gırtlaksal dönemin önceki %5'i geçerli tarama konumunun altındaki akış aralığının %1'inden daha düşük bir akış değerini içerdiği sürece çerçeve geriye doğru taranır. Sonraki durum algoritmanın yerel bir minimuma takılmadığından emin olmak için kullanılır. İkincil açılma anı $t_{o2,k}$, $t_{o1,k}$'dan sonraki glottal döngü süresinin %5'inden başlayan ve $t_{max,k}$'a kadar uzanan bir zaman penceresinde akışın yumuşatılmış ikinci türevinin en büyük yerel maksimum noktası olarak belirlenir. Tepeden tepeye genlik seviyesinin %50'sini kesen eğri noktaları sözde açılma ve kapanma anları olarak (t_{qo} ve t_{qc}) tanımlanır. Zaman ve genlik anları belirlendikten sonra değişik parametreleri hesaplamak kolaydır. Zaman uzayında hesaplanan bu parametreler aşağıdaki şekilde tanımlanır.

$$OQ_1 = \frac{t_c - t_{o1}}{T} \tag{2.42}$$

$$OQ_2 = \frac{t_c - t_{o2}}{T} \tag{2.43}$$

$$OQ_a = A_{ac}\left(\frac{\pi}{2A_{dmax}} + \frac{1}{A_{dmin}}\right) f_0 \tag{2.44}$$

$$QOQ = \frac{t_{qc} - t_{qo}}{T} \tag{2.45}$$

$$SQ_1 = \frac{t_{max} - t_{o1}}{t_c - t_{max}} \tag{2.46}$$

$$SQ_2 = \frac{t_{max} - t_{o2}}{t_c - t_{max}} \tag{2.47}$$

$$CIQ = \frac{t_c - t_{max}}{T} \tag{2.48}$$

$$AQ = \frac{A_{ac}}{A_{dmin}} \tag{2.49}$$

$$NAQ = \frac{AQ}{T} \tag{2.50}$$

2.1.4.2. Frekans-Uzayı Parametreleri

Zaman uzaylı parametreleştirme yöntemlerinin uygulanması basit olsa da kayıt ekipmanlarının faz tepkisindeki hafif bir non-lineerlik bile gıtlaksal akış tahminin kalitesini olumsuz etkileyebilir. Bu durumlarda tahmin edilen akışın frekans uzayı özelliklerini incelemek faydalı olabilir. Çoklu frekans-uzaylı ses kaynak parametreleri mevcuttur ve bu parametrelerin hepsi esasen spekrumun eğiminin ölçüsüdür.

Harmonik seviye farkı H1-H2, gırtlaksal akışın genlik spektrumunun temel ve ikincil harmoniklerinin dB cinsinden farkı alınarak hesaplanır. Harmonik zenginlik faktörü (HRF) ise,

$$HRF = \frac{\sum_{k\geq 2} H_k}{H_1} \tag{2.51}$$

şeklinde yüksek harmoniklerin genlikleri toplamının birincil harmoniğe oranı olarak tanımlanır. Yüksek harmonikler f_0'ın tam katlarında ($H_k = kf_0$) elde edilirse f_0'ın belirlenmesindeki hafif uyumsuzluklar ve yanlışlıklar süreci tamamen bozabilir. Bu yüzden harmonikler $kf_0 \pm f_0/2$ frekans bölgesindeki yerel maksimum olarak tanımlanır. Parabolik spektral parametre (PSP) ise perde-eşzamanlı olarak hesaplanan logaritmik spektrumun düşük frekanslı bölümünün parabolik bir fonksiyona uydurulması prensibine dayanmaktadır.

2.2. Ses Etkinliği Algılama

Konuşma etkinliği algılama veya konuşma algılama olarak da bilinen ses etkinliği algılama (VAD), insan konuşmasının varlığının veya yokluğunun tespit edildiği bir konuşma işlenme yöntemidir [96]. Kodlama ve tanıma başta olmak üzere birçok konuşma işleme uygulamasında kullanılan VAD, konuşma işlemeyi kolaylaştırdığı gibi konuşma dışı bölümlerde bazı işlemlerin devre dışı bırakılması için de kullanılabilir. VAD ayrıca VoIP (Voice Over Internet Protocol) uygulamalarında sessiz bölümlerin gereksiz kodlanmasını / iletilmesini önleyerek, hesaplama yükünü ve ağ bant genişliğini de azaltır. Diğer taraftan VAD işleminde yapılacak bir hata ses içeren konuşma bölümlerinin atılmasına ve buna bağlı olarak başka hataların oluşmasına neden olacaktır. Tipik bir konuşmada konuşma içeren ses bölümlerinin konuşma dışı ses bölümlerine oranı yaklaşık 40:60 dır [97] . Bu oran göz önünde bulundurulduğunda etkin bir VAD sisteminin kaynak kullanımını önemli ölçüde iyileştireceği açık bir şekilde görülmektedir.

Bir VAD sistemi iki aşamadan oluşur; öznitelik çıkarma ve karar verme aşaması. Öznitelik çıkarma aşamasında konuşma ile gürültü arasındaki ayırıcı karakteristikleri temsil edebilen parametreler çıkarılırken karar aşamasında bu parametreler kullanılarak konuşma/konuşma-dışı ses bölümlerine karar verilir.

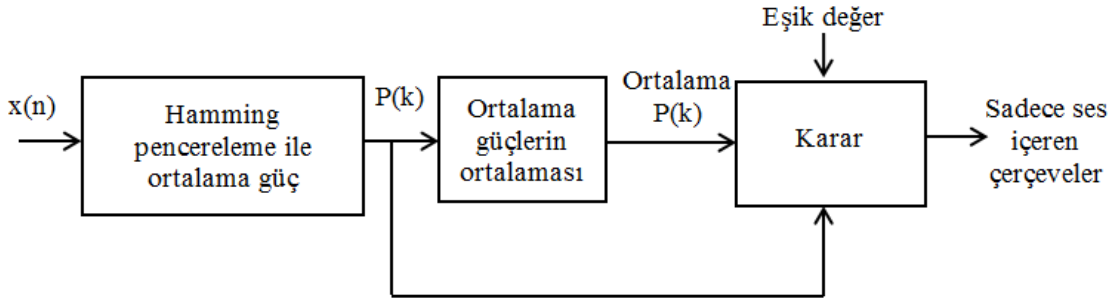
VAD işlemi için kullanılan öznitelikler genellikle 5 gruba ayrılır; enerji-temelli, spektral-uzay temelli, kesptral-uzay temelli, harmonik-temelli ve uzun-süreli öznitelikler. Bu özniteliklerden her biri farklı avantaj ve dezavantajlara sahiptir. Örneğin enerji-temelli öznitelikler basit ve uygulanması kolaydır, ancak gürültüye karşı güçlü değildir. Spektrum ve kepstrum uzayından türetilen öznitelikler gürültü etkilerini azaltmada çeşitli filtreleme tekniklerinden yararlanırken uzun süreli bilgiler ise konuşma ve gürültünün değişkenliğinden yararlanarak özniteliklerin ayrımcı gücünü arttırırlar.

Orijinal konuşma sinyalden bir öznitelik kümesi çıkarıldıktan sonra konuşma/konuşma-dışı ses bölümlerinin belirlendiği karar aşamasına geçilir. Literatürde VAD için önerilmiş birçok yöntem vardır. Bu yöntemler genellikle üç gruba ayrılır; eşik değere dayalı yaklaşımlar, istatistiksel modelleme yaklaşımları ve makine öğrenmesine dayalı yaklaşımlar. Eşik değere dayalı yaklaşımlar basitliği ve düşük karmaşıklığı nedeniyle en yaygın kullanılan teknik olarak öne çıkmaktadır. Bu teknikler genellikle özniteliklerin yeterince ayırt edici ve gürültüye karşı dayanıklı olmasını gerektirmektedir. Ancak SNR oranı düştüğünde öznitelik uzayı doğrusal olarak ayrılamaz ve bu da performansın önemli ölçüde düşmesine neden olur. Bu nedenle daha gelişmiş doğrusal olmayan VAD yaklaşımlara gerek duyulmaktadır. İstatistiksel modelleme yaklaşımlarında, konuşma sinyalinin özelliklerini en iyi yakalayan dağılımın bulunmasına çalışılır. İki taraflı Genelleştirilmiş Gamma Dağılımının, birden fazla konuşma için iyi bir model olduğu gösterilmiştir. Diğer yandan birden fazla dağılımın bir anahtarlama mekanizmasıyla veya karışım şeklinde birlikte kullanıldığı yöntemler de önerilmiştir. Özellikle son yıllarda makine öğrenmesine dayalı VAD yöntemleri kullanılmaya başlanmıştır. Bu yöntemler arasında SVM ve Maksimum Marjin Kümeleme (MMC) en fazla çalışılan ve en iyi performansa sahip yöntemler olarak görülmektedir. Bu yöntemlerin dışında LDA, sinir ağları ve genetik algoritmalar gibi yöntemler de önerilmiş ve standart VAD'lar üzerinde performans artışı sağlanmıştır.

Bu çalışmada enerjiye dayalı bir VAD sistemi kullanılmıştır. Blok diyagramı Şekil 2.12'de gösterilen bu sistemin ilk aşamada 30 ms uzunluğunda bir Hamming penceresi 10 ms aralıklarla konuşma sinyaline uygulanmış ve elde edilen çerçevelerin ortalama gücü, $P(k)$, aşağıdaki ifadeye göre hesaplanmıştır.

$$P_x(k) = \frac{1}{N_{length}} \sum_{n=0}^{N_{length}-1} \left| x(n - k.N_{shift}).v(n) \right|^2 \tag{2.52}$$

Burada $x(n)$ konuşma sinyalini, N_{length}, $v(n)$ Hamming penceresinin örnek uzunluğunu, N_{shift} ise 10 ms kayma parametresine karşılık gelen örnek sayısını temsil etmektedir.



Şekil 2.12. Enerjiye dayalı ses etkinliği algılama sisteminin blok diyagramı

VAD sisteminin ikinci aşamada tüm çerçevelerin ortalama güçlerinin ortalaması, $mean\,P(k)$, üçüncü aşamada ise her bir çerçevenin ortalama gücü, $P(k)$, ile tüm çerçevelerin ortalama gücü arasındaki oran, $P(k)/mean\,P(k)$, hesaplanmıştır. Bu oran önceden belirlenmiş bir eşik seviyeden yüksek ise ilgili çerçeve konuşma içeren bir ses çerçevesi olarak diğer durumda ise konuşma içermeyen bir ses çerçevesi olarak değerlendirilmiştir.

2.3. Konuşmacı Modelleme ve Sınıflandırma Yöntemleri

Genel yapısı Şekil 1.1'de gösterilen konuşmacı sınıflandırma sistemlerinin ilk aşamasında ses sinyalleri çeşitli sinyal işleme yöntemleri kullanılarak öznitelik vektörlerine dönüştürülür. Bu tez çalışmasında kullanılan öznitelik vektörleri ve elde edilme yöntemleri Bölüm 2.1'de anlatılmıştır. Sonraki aşama olan eğitim aşamasında ise konuşmacıların yaş, cinsiyet ve psikolojik durum gibi değişik özelliklerini temsil eden modeller oluşturulur. Son aşamada ise bilinmeyen konuşmacıya ait ses işaretinde elden edilen öznitelik vektörleri konuşmacı modelleri ile karşılaştırılarak konuşmacının ait olduğu sınıfa karar verilir.

Konuşmacı modelleri şablon modeller ve stokastik modeller olarak ikiye ayrılabilir [98]. Bu sınıflandırma parametrik ve parametrik olmayan olarak da bilinir. Şablon modellerde eğitim ve test öznitelik vektörleri birbirlerinin tam olmayan kopyaları olarak

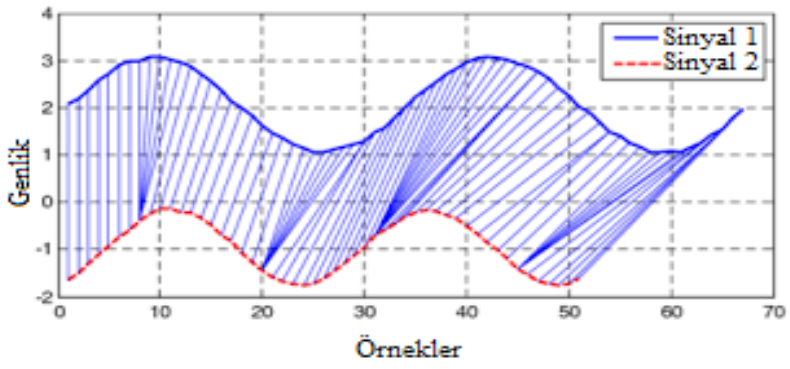
kabul edilir ve doğrudan karşılaştırılır. Bu karşılaştırma sonucunda elde edilen sapma iki örnek aralarındaki benzerliği temsil eder. Vektör nicemleme (VQ) [99] ve Dinamik zaman bükme (DTW) [100] sırasıyla metininden bağımsız ve metne bağımlı şablon modelleme yöntemlerine birer örnektir.

Stokastik modellerde her bir konuşmacı bilinmeyen fakat sabit bir olasılık yoğunluk fonksiyonu ile modellenir. Eğitim aşamasında olasılık yoğunluk fonksiyonunun parametreleri eğitim örneklerinden tahmin edilir. Eşleştirme ise genellikle test cümlesi ile eğitilen modele arasındaki olabilirlik skoruna göre yapılır. Gauss karışım modeli (GMM) [101, 102] ve Hidden Markov modeli (HMM) [103, 104] sırasıyla metin bağımlı ve metin bağımsız tanımada kullanılan oldukça popüler stokastik modelleme yöntemleridir.

Konuşmacı modelleri eğitim bakış açısına göre üretken ve ayırıcı olarak da sınıflandırılabilir. GMM ve VQ gibi üretken modeller özniteliklerin sınıf içi dağılımını tahmin ederken yapay sinir ağları (ANN) [105-107] ve destek vektör makineleri (SVM) [108] gibi ayırıcı modeller ise konuşmacılar arası sınırları modeller. Konuşmacı sınıflandırma sistemlerinde yaygın olarak kullanılan modelleme yöntemleri aşağıdaki bölümde tanıtılmıştır.

2.3.1. Dinamik Zaman Bükme (DTW)

Dinamik zaman bükme (DTW) konuşma tanıma, veri madenciliği ve imza eşleştirme gibi birçok uygulamada yaygın olarak kullanılan bir yöntemdir [109-111]. DTW iki sinyalin öznitelik dizisi arasındaki en iyi zaman hizalamasını aralarındaki uzaklığı en aza indirgeyecek şekilde bulur. İki sinyalin DTW ile hizalanması Şekil 2.13'te gösterilmiştir.



Şekil 2.13. İki sinyalin hizalanma örneği

$$X = x_1, x_2, \dots, x_i, \dots, x_N \tag{2.53}$$

$$Y = y_1, y_2, \dots, y_i, \dots, y_M \tag{2.54}$$

2.53 ve 2.54 ifadesiyle verilen iki zaman serisini hizalamak için DTW yöntemi kullanılabilir. Öncelikle NxM boyutlu bir matris oluşturulur. Bu matrisin (i,j)'inci elemanı $d(x_i, y_j) = \|x_i - y_j\|$ ile hesaplanan (x_i, y_i) noktaları arasındaki uzaklığa eşittir. Daha sonra birikimli maliyet matrisi oluşturulur. Bu matristeki her bir eleman, $C(i,j)$, o noktaya ulaşmanın minimum birikimli maliyetini temsil eder ve 2.55 ifadesiyle hesaplanır.

$$C(i,j) = \begin{cases} \sum_{k=1}^{j} d(x_1, y_k) & , i = 1 \\ \sum_{k=1}^{j} d(x_k, y_1) & , j = 1 \\ d(x_i, y_j) + min\{C(i-1, j-1), C(i-1, j), C(i, j-1)\} & , diğer \end{cases} \tag{2.55}$$

Son aşamada ise en iyi hizalama yolu, W bulunur. İki dizi (X, Y) arasındaki hizalamayı tanımlayan yol $w_k = (i,j)_k$ matris elemanlarının bir kümesidir. En iyi hizalama yolu birikimli maliyeti en az olan noktalar kümesinden oluşur ve aşağıdaki şartları sağlar;

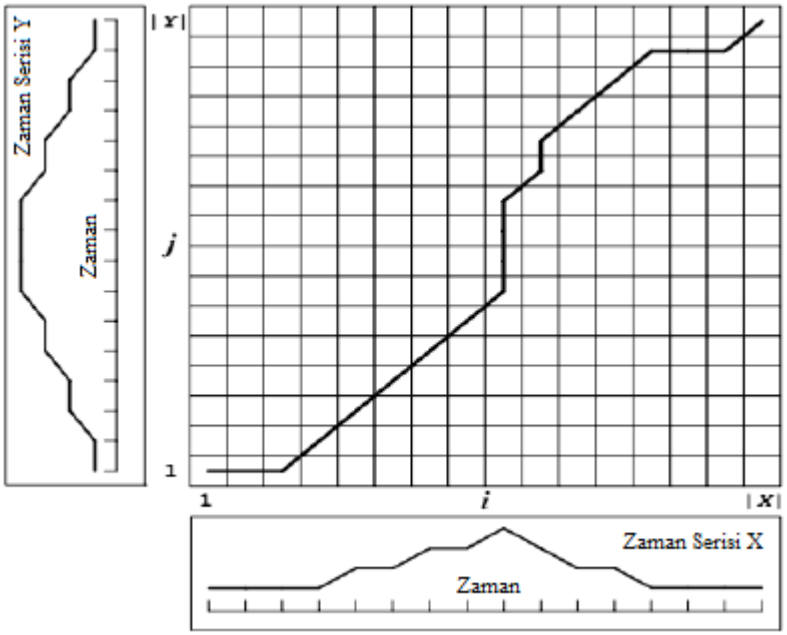
- Sınır şartları: hizalama yolu iki dizinin (X, Y) başlangıç ve bitiş noktalarında başlar ve biter;
  $w_1 = (1,1)\ ve\ w_K = (N, M)$
- Süreklilik: hizalama yolundaki adımlar sadece bitişik hücrelere yöneliktir.
  $w_k - w_{k-1} \in \{(1,1), (1,0), (0,1)\}$
- Monotonluk: Eğer $w_k = (a, b)$ ve $w_{k-1} = (a', b')$ ise
  $a \geq a',\ ve\ b \geq b'$ dir.

En iyi hizalama yolunu bulma algoritması Şekil 2.14'te, bu algoritma sonucunda bulunan bir yol örneği ise Şekil 2.15'te verilmiştir.

Algoritma: En iyi hizalama yolunun bulunması

$yol[] \leftarrow new\ array$
$i = N = rows(C)$
$j = M = columns(C)$
while$(i > 1)\&\ (j > 1)$ ***do***
if $i == 1$ ***then***
$j = j - 1$
else if $j == 1$ ***then***
$i = i - 1$
else
if $C(i-1,j) == min\{(i-1,j); C(i,j-1); C(i-1,j-1)\}$ ***then***
$i = i - 1$
else *if* $C(i,j-1) == min\{(i-1,j); C(i,j-1); C(i-1,j-1)\}$ ***then***
$j = j - 1$
else
$i = i - 1;\ j = j - 1$
end if
$path.add((i,j))$
end if
end while
return $path$

Şekil 2.14. En iyi hizalama yolunu bulan algoritma

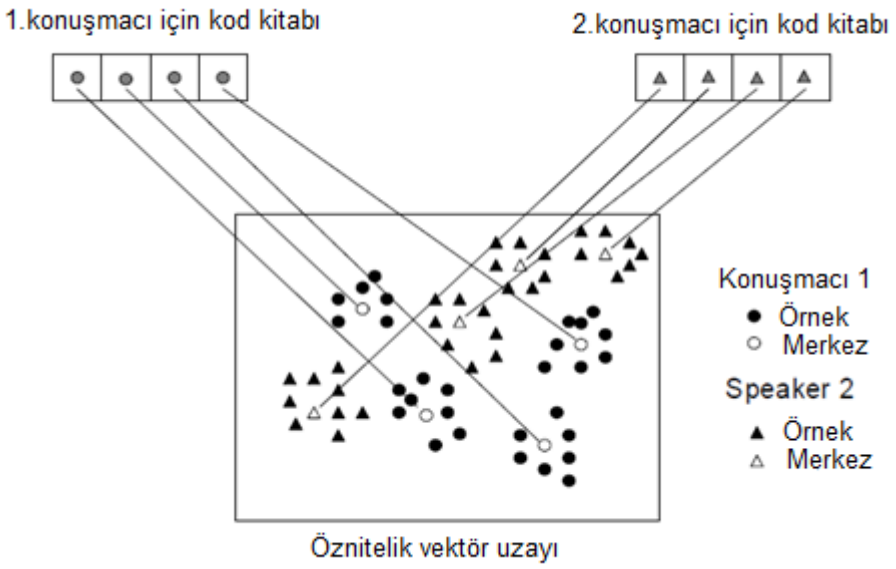


Şekil 2.15. En iyi hizalama yolunun bulunması örneği

2.3.2. Vektör Nicemleme

Vektör niceleme (VQ) geniş bir vektör uzayından sınırlı sayıda bölgeye dönüşüm gerçekleştiren bir tür veri sıkıştırma yöntemidir. Birbirleri ile örtüşmeyen bu bölgeler cluster olarak isimlendirilir ve her cluster ilgili bölgenin merkez noktası ile temsil edilir. Kod kelimesi olarak isimlendirilen bu noktalarından oluşan $C = \{c_1, \dots, c_K\}$ kümesi ise kod kitabı olarak isimlendirilir ve ilgili öznitelik uzayını temsil eder. Kod kelimesi sayısı K, model boyutu olup toplam öznitelik sayısı T'den oldukça küçüktür, $K \ll T$. İki konuşmacı için VQ yönteminin temsili gösterimi Şekil 2.16'da verilmiştir.

VQ yöntemi sonucunda elde edilen kod kelimelerinin dağılımı eğitim vektörlerinin dağılımı ile aynıdır [112]. Yani kod kitabı orijinal verinin sahip olduğu önemli bilgileri koruyarak veri miktarını etkin bir şekilde azaltır. Aslında literatürde VQ yaklaşımı hakkında tam bir anlaşma yoktur. Bazı yazarlar [98] VQ'yu şablon eşleştirme yaklaşımı olarak değerlendirirken bazıları [113, 114] ise stochastic veya olasılıksal bir yaklaşım olarak değerlendirir. VQ'nun şablon eşleştirme olarak değerlendirilmesinin sebebi tüm geçici değişimleri yok sayması ve global ortalamaları kullanmasıdır. VQ'nun stochastic bir yöntem olarak değerlendirilmesinin sebebi ise olasılıksal dağılım modlarını tahmin etmek için kod kelimelerini kullanmasıdır.



Şekil 2.16. İki konuşmacıya ait vektör uzayının VQ yöntemi ile temsili

Verilen bir $X = \{x_1, \dots, x_T\}$ vektör kümesi için VQ yönteminin amacı

$$D = \frac{1}{T}\sum_{t=1}^{T} min_{\ 1\le k\le K}(d(x_t, c_k)) \tag{2.56}$$

ifadesiyle hesaplanan ortalama bozulmayı minimum yapacak $C = \{c_1, \dots, c_K\}$ kod kitabının belirlenmesidir. Buradaki d uzaklık ölçüsü olup en yaygın kullanılan uzaklık ölçüleri Euclidean, city block, weighted Euclidean ve Mahalanobis uzaklıklarıdır [98, 115]. Bu ölçüler;

$$d_c(x, y) = \sum_{i=1}^{N}|x_i, y_i| \quad \text{City block uzaklığı} \tag{2.57}$$

$$d_E(x, y) = \sum_{i=1}^{N}(x_i - y_i)^2 \quad \text{Euclidean uzaklığı} \tag{2.58}$$

$$d_M(x, y) = (x - y)^T . D^{-1} . (x - y) \quad \text{Weighted Euclidean uzaklığı} \tag{2.59}$$

ifadeleri ile temsil edilir. Burada x ve y çok boyutlu öznitelik vektörlerini, D ise ağırlık matrisini temsil eder. D'nin kovaryans matrisi olması durumunda weighted Euclidean uzaklığı Mahalanobis uzaklığı olarak isimlendirilir [98, 115]. Uzaklık ölçüsünün seçimi konusunda Ong ve arkadaşları tarafından yapılan çalışmada kovaryans matrisinin diyagonal elemanlarından oluşan bir D matrisinin kullanıldığı weighted Euclidean uzaklığının konuşmacı belirleme işi için daha uygun olduğuna karar verilmiştir [115]. Böyle bir sonucun nedeni öznitelik vektörlerindeki tüm bileşenlerin eşit öneme sahip olmaması [116] ve bu durumun ağırlıklandırılmış uzaklık ile daha iyi temsil edilebileceği şeklinde açıklanabilir.

Kod kitabının oluşturulması sırasında belirlenmesi gereken iki önemli konu vardır. Bunlardan birincisi kod kitabının boyutu, ikincisi ise kod kitabının oluşturulma yöntemidir. Kod kitabı boyutunun belirlenirken sistemin doğruluğu ile çalışma süresi arasında bir dengenin kuruması gerekir. Çünkü kod kitabının büyük olması tanıma başarısı ile birlikte çalışma süresini de arttıracaktır. Kinnunen ve arkadaşları tarafından yapılan çalışmada konuşmacı tanıma problem için optimum kod kitabı boyutu 64 olarak belirlenmiştir [117].

Kod kitabı oluşturmak için kullanılan çeşitli yöntemler vardır [117]. Bunlardan en popüler olanları K-ortalama (K-means) ve LBG algoritmalarıdır [118]. Bu iki yöntemi birbirinden ayıran temel fark başlangıç aşamasıdır. K-ortalama algoritması rastgele veya belirli bir varsayıma göre seçilen K adet kod kelimesi ile başlar. Daha sonra öznitelik

uzayındaki her bir vektör kendisine en yakın kod kelimesine atanır ve her grubun merkezi yeniden hesaplanarak kod kelimeleri güncellenir. Bu işlem belirli bir sayıya ulaşana kadar veya bozulmadaki değişim belirli bir sınırın altına düşene kadar tekrarlanır. K-ortalama algoritmasının işlem basamakları aşağıda verilmiştir [118].

Adım 1. Başlangıç aşaması: Rastgele veya bazı varsayımlar kullanılarak K adet kod kelimesi oluşturulur.

Adım 2. En yakın komşuluk sınıflandırması: Her bir giriş vektörü x_t

$q_t = argmin_{1\leq k\leq K}\, d(x_t, c_k) \qquad t = 1, \dots, T$

$S_k = \{x_t \,\epsilon\, X | q_t = k\}$

ifadesine göre S_k bölgesine sınıflandırılır.

Adım 3. Kod kitabı güncelleştirmesi: Her bölgenin merkez noktası ilgili bölgedeki vektörler kullanılarak yeniden hesaplanır.

$\widehat{c_k} = \frac{1}{N_k}\sum_{x_t \epsilon S_k} x_t$

$c_k = \widehat{c_k}$

N_k, S_k bölgesindeki vektör sayısıdır.

Adım 4: Tekrar aşaması: Ortalama bozulmanın yeni ve eski değeri arasındaki fark belirli bir sınırın altına düşene kadar 2 ve 3 adımları tekrarlanır.

K-ortalama algoritması oldukça basit ve hızlı bir yöntem olmasına rağmen başlangıç koşullarına oldukça bağımlıdır. Başlangıç kod kelimelerinin seçimine göre yöntem kısa sürede iyi bir sonuç üretebileceği gibi istenen yakınsamanın sağlanamadığı durumlarda olabilir. Başlangıç kod kelimelerinin seçimi için önerilmiş değişik yöntemler mevcuttur. Random, Forgy, MacQueen ve Kaufman algoritmaları bu yöntemlerden bazıları olup bu yöntemlerin deneysel karşılaştırılması [119] çalışmasında yapılmıştır.

LBG algoritmasında ise giriş olarak tek bir kod kelimesi (tüm vektörlerin ortalaması) kullanılır. Her adımda kod kelimesi ikiye bölünür ve istenen sayıda kod kelimesi elde edilene kadar işlem tekrarlanır. LBG algoritmasının her tekrarında kod kelimesinin sayısı iki katına çıkacağı için bu algoritmayla ancak ikinin katlarında kod kelimesi oluşturulabilir. K-ortalama algoritmasında ise kod kelimesinin sayısı ile ilgili herhangi bir kısıtlama yoktur. LBG algoritmasının işlem basamakları aşağıda verilmiştir [118].

Adım 1: Başlangıç aşaması: $K = 1$ olarak seçilir ve tüm verilerin merkez noktası

$c_k = \frac{1}{T}\sum_{1\le t\le T} x_t$

ifadesi kullanılarak hesaplanır.

Adım 2: Parçalama aşaması: K adet kod kelimesi

$c_k^+ = c_k + \epsilon$

$c_k^- = c_k - \epsilon$

şeklinde $2K$ kod kelimesine parçalanır ve $K = 2K$ yapılır.

Adım 3: Kümeleme aşaması: Bir kümeleme algoritması (örneğin K-ortalama) kullanılarak her kod kelimesi için en iyi merkez nokta belirlenir.

Adım 4: Sonlandırma aşaması: İstenen sayıda kod kelimesi elde edilene kadar 2 ve 3 adımları tekrarlanır.

Bir veri sıkıştırma yöntemi olarak önerilen ve daha sonra çeşitli örüntü tanıma problemlerinde kullanılmaya başlanana VQ yönteminin konuşmacı sınıflandırma problemine uygulanışı şöyledir. Ön işlemler yapılıktan sonra her sınıf için o sınıfa ait eğitim vektörleri kullanılarak M seviyeli bir kod kitabı oluşturulur. Sınıflandırma probleminin eğitim aşamasına karşılık gelen bu işlem sonucunda her konuşmacı sınıfını temsil eden L tane kod kitabı oluşacaktır. Daha sonra test aşamasına geçilir ve bilinmeyen konuşmacılara karşılık gelen test kümeleri ile kod kitapları arasındaki ortalama bozulma hesaplanır. $\{x_1, x_2, \dots, x_N\}$ şeklinde verilen bir test kümesi için kod kitabı l'ye göre ortalama bozulma

$$D^l = \frac{1}{N}\sum_{i=1}^{N} min_{1\le j\le M}(d(x_i, c_j^l)) \tag{2.60}$$

şeklinde hesaplanır. Son aşamada ise L adet ortalama bozulma değerleri karşılaştırılır ve en az bozulmaya sahip kod kitabının sınıfı test kümesinin sınıfı olarak atanır.

$$l^* = argmin_{1\le l\le L} D^l \tag{2.61}$$

2.3.3. Gauss Karışım Modeli

Konuşmacı modellemede kullanılan diğer bir yöntem Gauss karışım modelidir. Gauss karışım modeli (GMM) öznitelik değişimlerinin istatistiksel olarak modellendiği stochastic bir yöntem olup clusterların birbirleri ile örtüştüğü VQ modeli olarak da düşünülebilir [120]. GMM modelinde öznitelikler VQ'daki gibi kendisine en yakın cluster'a atamaz. Onun yerine her bir cluster için öznitelik vektörlerinin ilgili clustera ait olma olasılıklarını gösteren sıfırdan farklı bir değer kullanılır.

Gauss karışım yoğunluğu M adet çok boyutlu Gauss yoğunluk fonksiyonunun ağırlıklandırılmış toplamı olup denklem 2.62 ile ifade edilir [98, 102, 113].

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i b_i(\vec{x}) \tag{2.62}$$

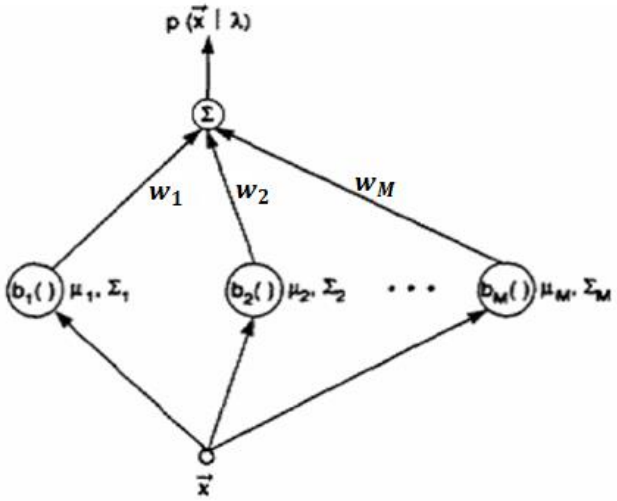
Buradaki M bileşen sayısını, $\vec{x}$ d-boyutlu öznitelik vektörünü, $b_i(\vec{x})$ bileşen yoğunluklarını, w_i ise $\sum_{i=1}^{M} w_i = 1$ şartını sağlayan bileşen ağırlıklarını temsil eder. Her bileşen ise ortalama vektörü $\vec{\mu}_i$ ve kovaryans matrisi Σ_i olan d-değişkenli bir Gauss fonksiyonu olup denklem 2.63 ile ifade edilir.

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}\sqrt{|\sum_i|}}\ exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T {\sum_i}^{-1}(\vec{x}-\vec{\mu}_i)\right\} \tag{2.63}$$

Gauss karışım yoğunluğu tüm bileşenlerin karışım ağırlıkları, ortalama vektörleri ve kovaryans matrisleri ile karakterize edilir ve denklem 2.64 şeklinde gösterilir.

$$\lambda = \{w_i, \vec{\mu}_i, \sum_i\} \quad i = 1, \dots, M \tag{2.64}$$

Konuşmacı sınıflandırma probleminde her sınıf ilgili sınıfa ait konuşmacıların öznitelik vektörleri kullanılarak oluşturulan λ modeli ile temsil edilir. Şekil 2.17'de M bileşenli bir Gauss karışım yoğunluğu grafiksel olarak gösterilmiştir.



Şekil 2.17. M bileşenli Gauss karışım yoğunluğu

Kovaryans matrisinin seçimine bağlı olarak GMM'nin birkaç farklı türü vardır [102]. Bu türlerden "nodal" kovaryanslı GMM modelinde her bileşen için farklı bir kovaryans matrisi kullanılırken (denklem 2.64'teki gibi), "grand" kovaryanslı GMM modelinde her konuşmacı modeli için ayrı bir kovaryans matrisi kullanılır. "Global" kovaryanslı GMM modelinde ise tüm konuşmacı modelleri için tek bir kovaryans matrisi kullanılır. Kovaryans matrisi aynı zamanda tam veya diyagonal olarak da seçilebilir. Tam-kovrayanslı GMM modelinde parametre tahmini için genellikle daha fazla eğitim verisine ihtiyaç duyulur. Ayrıca bu işlemin hesaplama yükü de oldukça fazladır. Diğer yandan tam-kovaryans matrisli bir GMM modeli daha yüksek dereceli diyagonal kovaryans kullanılarak eşdeğer şekilde temsil edilebilir [63]. Bu durum [102, 121] çalışmalarında deneysel olarak da gösterilmiştir. Bahsedilen avantajlarından dolayı bu tez çalışmasında da diyagonal kovaryans matrisli GMM modeli tercih edilmiştir.

2.3.3.1. Parametre Tahmini

GMM parametrelerinin tahmini için kullanılan değişik yöntemler vardır [122]. Bu yöntemlerden en güncel ve iyi kurgulanmış olanı en çok olabilirlik (ML) kestirimidir. ML kestiriminin amacı verilen eğitim verisinden GMM'nin olabilirliğini maksimum yapan model parametrelerini bulmaktır. T adet eğitim vektöründen oluşan bir $X = \{\vec{x}_1, \dots, \vec{x}_T\}$ eğitim kümesi için GMM olabilirliği

$$p(X|\lambda) = \prod_{t=1}^{T} p(\vec{x}_t|\lambda) \tag{2.65}$$

şeklinde yazılır. Bu ifade, λ parametresinin doğrusal olmayan bir fonksiyonudur ve direkt maksimizasyonu mümkün değildir. Bu nedenle ML parametre kestirimi beklenti maksimizasyonu (EM) olarak isimlendirilen tekrarlamalı bir algoritmayla yapılır [123].

EM algoritması verilerin tam olmadığı veya kayıp değerlerin olduğu veri kümelerinden ML parametre kestirimi için kullanılan genel bir yöntemdir. Beklenti ve maksimizasyon olmak üzere iki aşamadan oluşan EM algoritması rastgele veya bazı sezgisel yöntemlerle belirlenen bir λ modelinden $p(X|\bar{\lambda}) \geq p(X|\lambda)$ şartını sağlayan $\bar{\lambda}$ modelini tahmin etmeye çalışır. Belirlenen $\bar{\lambda}$ modeli bir sonraki aşamada başlangıç modeli olarak kullanılarak aynı işlemler tekrarlanır. Süreç belirli bir tekrar sayısına ulaşana kadar yada olabilirlik fonksiyonundaki değişim belirli bir değerin altına düşene kadar tekrarlanarak en uygun parametre tahmini yapılır. EM algoritmasının işlem basamakları şöyledir;

Adım 1. Başlangıç model parametreleri, λ^0, rastgele veya bazı sezgisel yöntemlerle belirlenir ve bu modelin logaritmik olabilirliği denklem 2.66'ya göre hesaplanır.

$$log\ p(X|\lambda) = \frac{1}{T}\sum_{t=1}^{T} log \sum_{i=1}^{M} w_i b_i(\vec{x}) \tag{2.66}$$

Adım 2. E-adımı: O anki λ parametresi kullanılarak her bir $\vec{x}_t$ noktasının tüm karışım bileşenlerine ait olma olasılıkları denklen 2.67'ye göre hesaplanır.

$$p(i|\vec{x}_t, \lambda) = \frac{w_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} w_k b_k(\vec{x}_t)} \qquad 1 \leq i \leq M, \text{ ve } 1 \leq t \leq T \tag{2.67}$$

Böylece her satırının toplamı 1 olan TxM'lık bir matris elde edilir.

Adım 3. M-adımı: Aşağıdaki denklemler kullanılarak yeni model parametreleri hesaplanır.

$$\bar{w}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda) \tag{2.68}$$

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)\vec{x}_t}{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)} \tag{2.69}$$

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)x_t^2}{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)} - \bar{\mu}_i^2 \tag{2.70}$$

Adım 4. *E ve M* adımları belirli bir tekrar sayısına ulaşana kadar veya yeni parametrelerle hesaplanan olabilirlik değerindeki değişim belirli bir sınırın altına düşene kadar tekrarlanır.

2.3.3.2. Konuşmacı Sınıflandırma

Her sınıfın bir konuşmacı olması durumunda konuşmacı belirlemeye karşılık gelen konuşmacı sınıflandırma probleminde $S = \{1,2,\dots,S\}$ sınıflarını temsil etmek için $\lambda_1, \lambda_2, \dots, \lambda_S$ GMM modelleri kullanılır. Konuşmacı sınıflandırmada amaç $X = \{x_1, \dots, x_T\}$ öznitelik vektörü ile verilen bir konuşmacı için maksimum önsel olasılığa sahip sınıfın bulunmasıdır. Bu işlem biçimsel olarak

$$\hat{s} = arg\ max_{1 \le k \le S} Pr(\lambda_k|X) = arg\ max_{1 \le k \le S} \frac{p(X|\lambda_k)Pr(\lambda_k)}{p(X)} \tag{2.71}$$

şeklinde gösterilir. Konuşmacı sınıflarının önsel olasılıklarının eşit olduğu ($Pr(\lambda_k) = 1/S$) varsayıldığında ve tüm konuşmacılar için $p(X)$'in aynı olduğu düşünüldüğünde sınıflandırma kuralı

$$\hat{s} = arg\ max_{1 \le k \le S}\ p(X|\lambda_k) \tag{2.72}$$

şeklinde basitleştirilir. Son olarak gözlemler arasındaki bağımsızlık varsayımı ve logaritma fonksiyonunun kullanımı ile konuşmacı sınıflandırma sistemi

$$\hat{s} = arg\ max_{1 \le k \le S} \sum_{t=1}^{T} log\ \ p(\vec{x}_t|\lambda_k) \tag{2.73}$$

şeklinde temsil edilir. Buradaki $p(\vec{x}_t|\lambda_k)$ denklem 2.62 ile verilen Gauss karışım yoğunluğuna karşılık gelmektedir.

2.3.3.3. Genel Arka Plan Modeli

Gauss karışım modeli güçlü bir yöntem olmasına rağmen bazı dezavantajlara sahiptir [124]. Bunlardan ilki GMM'lerin eğitimi için çok sayıda parametrenin belirlenmesi

gerekir. Bu durum hem hesaplama yükünü hem de eğitim için gereken veri ihtiyacını arttırır. Bu yüzden küçük bir veri kümesi ile eğitilen GMM'lerin performansına güvenilmez. İkinci dezavantaj ise üretken bir model olarak GMM görünmeyen verilerle iyi çalışmaz ve düşük olabilirlik skoru üretir. Bu iki problem konuşmacı adaptasyonu ile ortadan kaldırılabilir. Adaptasyon yaklaşımı tüm konuşmacı verileri ile oluşturulan konuşmacıdan bağımsız bir sistemin konuşmacı verileri ile güncellenerek konuşmacı bağımlı sistemlerin oluşturulması fikrine dayanır [101]. Bu yaklaşımda konuşmacıdan bağımsız sistem tüm konuşmacı verileri ile eğitilen bir GMM modelidir ve genel arka plan modeli (UBM) olarak isimlendirilir.

UBM modelinin eğitimi için farklı yaklaşımlar mevcuttur [125, 126]. Bunlardan en basit olanı tüm verilerin basitçe birleştirilerek tek bir UBM'in EM algoritmasıyla eğitilmesi yaklaşımıdır. Bu yaklaşımda birleştirilen verilerde tüm alt gruplar dengeli olmasına dikkat edilmelidir aksi durumda sonuç model, baskın gruba meyilli olacaktır. Örneğin cinsiyetten bağımsız verilerin kullanıldığı bir durum için erkek ve kadın konuşmaların dengeli olması gerekmektedir. Aksi durumda sonuç model baskın olan cinsiyet grubuna meyilli olacaktır. Özellikle dengesiz veri dağılımın olduğu durumlarda kullanılan diğer bir yaklaşımda ise her alt grup için ayrı bir UBM eğitilip daha sonra bu modeller birleştirilir. Bu yaklaşımda her bileşenin sonuç model üzerindeki etkisi kontrol edilebildiği için sonuç modelinin belirli bir gruba meyilli olması önlenebilmektedir. Bu tez çalışmasında UBM'in eğitimi için birinci yaklaşım tercih edilmiştir.

2.3.3.4. MAP (Maximum A Posteriori) Adaptasyonu

Özellikle yüksek bileşenli GMM'lerin eğitiminde konuşmacıların kısa süreli verileri yeterli olmayacaktır. Bu durumda model parametrelerinin tahmini için doğrudan EM algoritmasının kullanımı yerine UBM'nin uyarlanması yolu tercih edilir. Uyarlama yaklaşımı; konuşmacıdan bağımsız olarak eğitilmiş UBM parametrelerinin yeni verilerle güncellenmesi fikrine dayanmaktadır. UBM ile modeller arasında sıkı bir bağlantı sağlayan uyarlama yaklaşımı ayrık modellere kıyasla hem dahi iyi performans sağlar hem de daha hızlı sonuç verir [101]. Literatürde konuşmacı modellerinin uyarlanması için kullanılan değişik yöntemler vardır [127]. Bu yöntemlerden seyrek veriler için en iyi ve güvenilir sonuç veren MAP uyarlaması EM algoritmasına benzeyen iki aşamalı bir tahmin

algoritmasıdır. MAP algoritmanın ilk aşamasında $X = \{x_1, \dots, x_T\}$ şeklinde verilen konuşmacı öznitelikleri ile UBM bileşenleri arasındaki olasılıksal uyum

$$Pr(i|x_t) = \frac{w_i b_i(x_t)}{\sum_{k=1}^{M} w_k b_k(x_t)} \tag{2.74}$$

ifadesine göre hesaplanır (Şekil 2.18a). Daha sonra $Pr(i|x_t)$ olasılıkları ve x_t vektörleri kullanılarak ağırlık, ortalama ve varyans parametreleri için yeterli istatistikler hesaplanır.

$$n_i = \sum_{t=1}^{T} Pr(i|x_t) \tag{2.75}$$

$$E_i(x) = \frac{1}{n_i} \sum_{\mathrm{t=1}}^{\mathrm{T}} p(i|x_t)\, x_t \tag{2.76}$$

$$E_i(x^2) = \frac{1}{n_i} \sum_{\mathrm{t=1}}^{\mathrm{T}} p(i|x_t)\, x_t^2 \tag{2.77}$$

Son aşamada ise eğitim vektörleri ile elde edilen n_i, $E_i(x)$ ve $E_i(x^2)$ istatistikleri kullanılarak UBM parametreleri güncelleştirilir (Şekil 2.18b).

$$\widehat{w}_i = \left[\frac{\alpha_i^w n_i}{T} + (1 - \alpha_i^w) w_i\right] \gamma \tag{2.78}$$

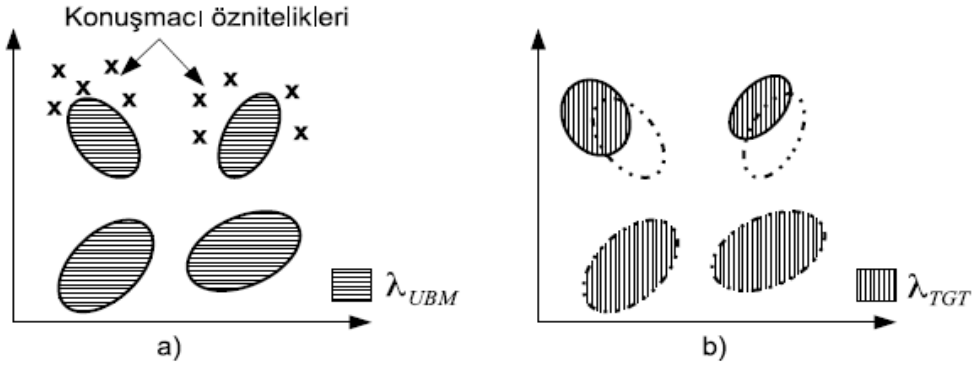
$$\hat{\mu}_i = \alpha_i^m E_i(x) + (1 - \alpha_i^m)\mu_i \tag{2.79}$$

$$\hat{\sigma}_i^2 = \alpha_i^v E_i(x^2) + (1 - \alpha_i^v)({\sigma_i}^2 + {\mu_i}^2) - {\hat{\mu}_i}^2 \tag{2.80}$$

Denklem 2.78-2.80 ile tanımlanan güncelleştirme sonucunda uyarlanmış GMM parametreleri elde edilir. Bu ifadelerdeki $\{\alpha_i^w, \alpha_i^m, \alpha_i^v\}$ katsayıları sırasıyla ağırlık, ortama ve varyans parametrelerinin eski ve yeni tahminleri arasındaki dengeyi belirleyen uyarlama katsayıları olup

$$\alpha_i^\rho = \frac{n_i}{n_i + r^\rho}\ ,\ \ \rho \in \{w, m, v\} \tag{2.81}$$

şeklinde tanımlanır. γ katsayısı ise tüm uyarlanmış karışım ağırlıklarının toplamını bir yapan bir ölçek faktörüdür.



Şekil 2.18. Uyarlama yaklaşımının şekilsel gösterimi. a) Eğitim vektörleri UBM bileşenleriyle olasılıksal olarak eşleştirilir. b) Yeni verilerin istatistikleri ve UBM parametreleri kullanılarak bileşenler uyarlanır.

Deneklem 2.81'deki r^{ρ} değeri ağırlık, ortama ve varyans parametrelerinin uyarlanma oranını belirleyen sabit bir etki faktörüdür. Büyük seçilmesi uyarlama katsayısının küçülmesine o da yeni parametrelerin sonuç model üzerindeki etkisinin azalmasına neden olur. Ağırlık, ortalama ve varyans parametreleri için farklı birer etki faktörü seçilerek her parametrenin farklı oranlarda uyarlanması sağlanabilir. Bu durum uygulamada esneklik sağlasa da elde edilen kazanç oldukça düşüktür ve bu yüzden GMM-UBM sistemlerinde genellikle ortak etki faktörü kullanımı tercih edilir [101, 128]. Ayrıca yapılan deneysel çalışmalarda ağırlık ve varyans parametrelerindeki uyarlanmanın hata oranını arttırdığı görülmüştür. Bu yüzden özellikle konuşmacı verisinin az olduğu durumlarda sadece bileşen ortalamaları uyarlanarak ağırlık ve kovaryans parametreleri UBM'den aynen alındığı kullanım tercih edilir [101, 128]. Bu tez çalışmasında da etki faktörü 14 seçilerek yalnızca bileşen ortalamaları uyarlandığı yaklaşım tercih edilmiştir.

2.3.3.5. MLLR Adaptasyonu

En büyük olabilirlik doğrusal regresyon (MLLR) konuşmacı modellerinin uyarlanması için kullanılan diğer bir yöntemdir. Bu yöntemde UBM ortalamaları doğrusal bir dönüşümle güncellenerek konuşmacı modeli oluşturulur [129, 130]. Bu dönüşümün parametreleri doğrusal regresyonla hesaplanır. A regresyon matrisi ve b toplamsal eğilim vektörü olmak üzere i'inci bileşeninin uyarlanmış ortalama vektörü

$$\hat{\mu}_i = A\mu_i + b \tag{2.82}$$

ifadesiyle hesaplanır. Bu yöntemde konuşmacılar arasındaki temel fark ortalama vektörü ile karakterize edildiği için yüksek dereceli istatistikler uyarlanmaz. MLLR yöntemi özellikle uyarlama verisinin sınırlı olduğu durumlarda iyi performans sağlayan oldukça hızlı ve etkin bir yöntemdir. MAP yöntemi ile tüm model parametreleri değiştirilebildiği için uyarlama verisi arttıkça MAP yönteminin doğruluğu artar [47]. Ayrıca MAP ve MLLR yöntemleri birlikte kullanılarak her iki yöntemin avantajlarından yararlanılabilir [131]. Bu amaçla önce MAP yöntemi ile model parametreleri uyarlanıp, daha sonra ise uyarlanış parametreler MLLR ile düzeltilebilir.

2.3.4. Destek Vektör Makineleri

Önceki bölümlerde anlatılan GMM ve VQ yöntemleri konuşmacı karakteristiklerinin bir parametre kümesi ile temsil edildiği üretken modelleme yöntemleridir. Bu yöntemlerde sınıflandırma işlemi eğitim aşamasında oluşturulan konuşmacı modelleri ile test verilerinin karşılaştırılması sonucunda elde edilen benzerlik skoruna göre yapılır. Benzerlik skoru ise genellikle olasılıksal veya uzaklığa dayalı olarak hesaplanır.

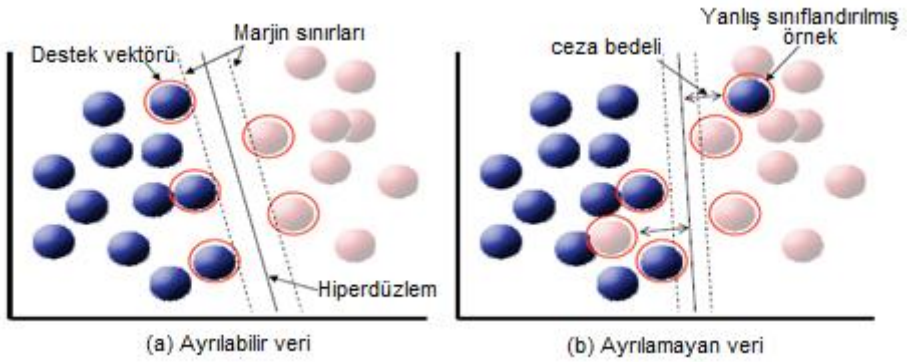
Son zamanlarda makine öğrenme teknikleri örüntü tanıma problemlerine uyarlanmaya başlanmıştır. Bu modelleme teknikleri ile iki sınıfı birbirinden ayıran bir model her iki sınıfın örneklerinden öğrenilerek oluşturulur. Bu tür teknikler ayırımcı modelleme teknikleri olarak isimlendirilir.

Destek vektör makineleri (SVM) örüntü tanıma konusunda oldukça dikkat çeken makine öğrenme tekniklerinden birisidir [132]. Bu tekniğin ayırımcı yapısı el yazısı karakterlerinin tanınması [133, 134], nesne tanıma [135, 136], yüz tanıma [137, 138] ve metin sınıflandırma [139, 140] gibi birçok örüntü tanıma probleminde başarılı bir şekilde uygulanmaktadır. Son yıllarda ise özellikle konuşmacı doğrulama ve belirleme alanında SVM'ye dayalı sınıflandırma yaklaşımı ilgi odağı haline gelmiştir [108, 141-143].

2.3.4.1. Doğrusal Olarak Ayrılabilen Durum

SVM iki sınıfın en yakın örnekleri arasındaki mesafeyi maksimum yapacak ayırıcı hiper düzlemi bulmaya çalışan doğrusal bir sınıflandırıcıdır. SVM sınıflandırıcısını tanımlayabilmek için doğrusal olarak iki sınıfa ayrılabilen l elemanlı bir eğitim vektörü kümesini, $X = \{x_1, x_2, \dots, x_l\}$, göz önünde bulunduralım. Bu gözlemlerin her birisi

$y_i \in \{-1,1\}$ ve $x_i \in R^d$ olmak üzere $\{x_i, y_i\}, i = \{1, \dots, l\}$ şeklinde her gözlemin ait olduğu sınıfı gösteren bir etikete sahip olacaktır. x_i, gözlem vektörlerinin iki boyutlu olduğu varsayılarak X gözlemlerinin doğrusal olarak ayrılabildiği durum Şekil 2.19a'da gösterilmiştir. Bu gösterimde pozitif ve negatif sınıfların ayrımı düz bir çizgi olarak gösterilen ayırıcı bir hiperdüzlem ile gerçekleştirilmiştir. Bu ayırıcı düzlem üzerindeki noktalar $w.x + b = 0$ şeklinde verilir. Burada w hiperdüzlemin normalini, $|b|/\|w\|$ ise hiperdüzlem ile orijin arasındaki uzaklığı temsil eder.



Şekil 2.19. Doğrusal olarak ayrılabilen (a) ve ayrılamayan (b) verileri kümleri ile eğitilen SVM'nin bileşenleri.

Hiperdüzlemin her iki tarafında her sınıfın sınırlarını tanımlayan bir marjin vardır. Şekil 2.19'da kesikli çizgilerle gösterilen bu doğrular,

$$y_i(w.x + b) \geq 1 \tag{2.83}$$

şeklinde tanımlanır. SVM eğitiminde amaç bu marjini en büyük yapacak hiperdüzlemin yerinin belirlenmesidir. $2/\|w\|$ ile verilen marjin genişliğinin en büyüklenmesi 2.83 kısıtlaması altında $\|w\|^2$'nin en küçüklenmesi ile yapılabilir.

Sınıf sınırlarının üzerinde bulunan eğitim örnekleri destek vektörleri olarak isimlendirilir ve genellikle bu noktalar eğitim veri kümesinin yalnızca küçük bir bölümünü oluşturur. Adından da anlaşıldığı gibi destek vektörleri ayırıcı hiperdüzlemi tanımlayan veya destekleyen eğitim örnekleridir. Diğer yandan destek vektörü olarak seçilmeyen eğitim örnekleri SVM'nin eğitiminde herhangi bir role sahip değildir. Yani bu örneklerin eğitim kümesinden çıkarılması ayırıcı hiperdüzlemin yeri değiştirmez.

SVM eğitim algoritması genellikle Lagrange çarpanları, α_i, şeklinde temsil edilir. Böylece ayrılamayan verilerden eğitimin tanımlanması kolaylaştır [132]. Ayırıcı hiperdüzlemin (w normalinin) optimum pozisyonu

$$w = \sum_i \alpha_i y_i x_i \tag{2.84}$$

$$\sum_i \alpha_i y_i = 0 \tag{2.85}$$

kısıtları altında

$$\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i . x_j \tag{2.86}$$

ifadesinin en büyüklenmesi ile belirlenebilir. Bu tanımlamaya göre $\alpha_i > 0$ şartını sağlayan eğitim örnekleri destek vektörleridir ve bu noktalar hiperdüzlemin marjini üzerinde yer alır. Bu nedenle Lagrange çarpanları genellikle destek vektör ağırlıkları veya katsayıları olarak isimlendirilir.

2.3.4.2. Hiperdüzleme Dayalı Sınıflandırma

SVM'nin eğitimi sonucunda sınıf sınırlarını tanımlayan ayırıcı bir hiperdüzlem belirlenir. Bu hiperdüzlem (daha doğrusu hiperdüzlemin normali, w) verilen bir test örneğinin, x, sınıflandırılmasında doğrudan kullanılır. Test örneklerinin hiperdüzlemin hangi tarafında olduğu

$$f(x) = sgn(\sum_i \alpha_i y_i x . x_i + b) \tag{2.87}$$

eşitliğine göre hesaplanabilir. Bu eşitlik 2.84 tanımlamasına dayalı olarak $sgn(w.x + b)$ şeklinde basitleştirilebilir.

SVM'ye dayalı sınıflandırmada amaç test örneklerinin ait olduğu sınıfı bulmaktır. Ancak aşağıda ayrılamayan veri kümeleri için gösterildiği gibi bu sınıflandırma her zaman kolay olmayabilir. Bu durumda genellikle test örneklerinin hiperdüzleme (yani $w.x + b$) olan uzaklığından yararlanılır.

2.3.4.3. Ayrılamayan Veri Kümesi

Ayrılamayan bir veri kümesi söz konusu olduğunda eğitim örneklerinin yanlış sınıflandırılması Lagrange değişkenlerinin son derece büyük olmasına bu da uygulanabilir bir çözümün bulunamamasına neden olur. Bu durumda bazı örneklerin marjini ihlal etmesine izin vermek ve böylece daha iyi genelleştirme performansı sağlamak için bir gevşeklik değişkeni, ε_i, tanımlanır. Gevşeklik parametresinin eklenmesiyle ile 2.83 kısıtlaması

$$y_i(w^T x_i + b) \geq 1 - \varepsilon_i \tag{2.88}$$

şeklinde yumuşatılır ve böylece bazı örneklerin yanlış sınıflandırılmasına izin verilmiş olur. Daha sonra orijinal minimizasyon işlemi gevşeklik parametresi hesaba katılarak

$$min_{w,b} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \varepsilon_i \tag{2.89}$$

şeklinde değiştirilir. 2.88 kısıtlaması altında 2.89 optimizasyonu marjinin yanlış tarafında yer alan noktaların sadece cezalandırılacağı anlamına gelir. Bu durumda C parametresi hatalara karşılık gelen bir ceza olarak düşünülebilir. Eğitim aşamasında belirlenen C parametresinin büyük seçilmesi hatanın maliyetini artacak ve buna bağlı olarak sınıflar arası sınırlar keskinleşecektir. Diğer yandan C parametresinin küçük seçilmesi durumunda ise daha fazla hataya izin verilerek daha yumuşak sınıf sınırları oluşacaktır.

2.3.4.4. SVM Çekirdeği

SVM'ye dayalı sınıflandırmanın özünü oluşturan temel unsur çekirdek fonksiyonudur. SVM çekirdeği açık bir hesaplama olmaksızın giriş öznitelik vektörlerini yüksek boyutlu bir uzaya taşıyan bir fonksiyondur [144]. Bununla birlikte giriş öznitelik uzayında çalışan doğrusal bir çekirdek fonksiyonu tanımlamak da mümkündür.

Çekirdek fonksiyonun kullanılması ile birlikte SVM sınıflandırıcısının karar fonksiyonu

$$f(x) = sgn(\sum_i \alpha_i y_i K(x, x_i) + b) \tag{2.90}$$

şeklinde yazılabilir. Burada x giriş vektörünü, x_i eğitim örneklerini, y_i eğitim örneklerinin sınıfını (± 1), α_i ise ağırlıkları temsil etmektedir. Açıkça belirtilmemesine rağmen 2.87 ifadesinde $K(x, x_i) = x.x_i$ olarak verilen bir çekirdek fonksiyonuna mevcuttur. "lineer çekirdek" olarak isimlendirilen bu çekirdek giriş vektörlerinin doğrusal uzaydaki nokta çarpımlarına karşılık gelmektedir.

Bu zamana kadar anlatılan SVM örneklerinde verilerin doğrusal olarak ayırılması durumu ele alınmıştır. Doğrusal olarak ayırılabilen veri kümeleri için doğrusal SVM çekirdeği uygundur. Ancak SVM yaklaşımı doğrusal olmayan çekirdek fonksiyonlarının kullanımıyla birlikte doğrusal olarak ayrılamayan verileri kümelerine de genişletilebilir.

Doğrusal olmayan SVM çekirdeğinin amacı giriş vektörlerini doğrusal olarak sınıflandırılabileceği yüksek boyutlu bir uzaya taşıyarak veri kümelerinin doğrusal olmayan sınıflandırılmasına imkan sağlamaktır. Bu görev için çekirdek işlevi daha genel olarak şu şekilde ifade edilebilir;

$$K(x_i, x_j) = \Phi(x_i).\Phi(x_j) \tag{2.91}$$

Burada Φ, giriş vektörlerini istenen yüksek boyutlu uzaya taşımak için kullanılan bir haritalama fonksiyonudur. Haritalama fonksiyonları doğrusal sınıflandırmanın gerçekleştirileceği ayrımcı uzayı tanımladığı için uygulamaya özel seçilir.

SVM çekirdeğinin en önemli avantajı açık bir dönüşüm olmadan doğrusal olmayan sınıflandırmaya izin vermesidir. SVM çekirdeği bunu eğitim örnekleri arasındaki ilişkiyi nokta çarpımı şeklinde temsil ederek başarır. Aslında çekirdek fonksiyonu doğrusal sınıflandırıcının hesaplama avantajını koruyarak doğrusal olmayan sınıflandırmanın uygulanması için bir yöntem sağar [132].

SVM çekirdeği gözlemlerin sınıflandırıldığı uzayı ve buna bağlı olarak eğitim algoritması tarafından kullanılan ayırıcı bilgileri tanımlar. Bu yüzden SVM çekirdeklerinin geliştirilmesi genellikle uygulama bağımlıdır ve verilen bir öznitelik kümesi için en ayırımcı uzayı tanımlayan çekirdek fonksiyonun belirlenmesi amaçlanır. SVM çekirdeklerinin uygulama bağımlı olması yeni çekirdeklerin geliştirilmesi için bir motivasyon kaynağı olmuştur. Aşağıdaki bölümde konuşmacı sınıflandırma alanında yaygın olarak kullanılan çekirdek fonksiyonları tanıtılmıştır.

2.3.4.5. Çerçeve Tabanlı Çekirdekler

SVM yaklaşımının konuşmaya dayalı sistemlerde kullanılmaya başlanmasından bu yana problemin uygulamaya özel yönleri hesaba katılarak bir dizi SVM çekirdeği geliştirilmiştir. Konuşmaya dayalı sistemler için önerilen çekirdekler çerçeve-tabanlı ve cümle-tabanlı olarak ikiye ayrılır [145]. Çerçeve tabanlı çekirdekler SVM eğitiminin ve sınıflandırmasının doğrudan çerçeveler veya bir ses segmentinden elde edilen özniteliklerle gerçekleştirilmesine izin verir. Konuşmacı sınıflandırma işinde başarılı bir şekilde uygulanan çerçeve-tabanlı çekirdeklerden bazıları Tablo 2.1'de gösterilmiştir.

Tablo 2.1. Yaygın olarak kullanılan çerçeve tabanlı SVM çekirdeklerinden bazıları

Tanımlama	$K(x_i, x_j)$
Radyal tabanlı fonksiyon (RBF)	$exp\left[-\frac{\left\|x_i - x_j\right\|^2}{\sigma}\right]$
d dereceli polinom	$\left(x_i . x_j + 1\right)^d$

Radyal tabanlı RBF çekirdeği ile sınıfları saracak bir dizi nokta (centroid) belirlenir ve bu noktalar kullanılarak sınıflar arasında ayrım yapılır [146]. Tablo 2.1'deki RBF formülünde σ ile gösterilen parametre çekirdek genişliği olup bu parametrenin değeri sistemin geliştirilmesi sırasında ayarlanır. RBF çekirdeği çekirdek uzayındaki önemli noktaları merkez olarak bulmak için SVM eğitim algoritmasının destek vektör seçim sürecinden yararlanır.

Polinomsal çekirdek performans açısından RBF çekirdeği ile karılaştırılabilir durumdadır [147]. Ancak polinomsal çekirdeğin geniş veri kümeleri ile kullanılması durumunda bazı örnekler için çözüm bulunamayabilir [148]. Bu sorun polinomsal çekirdek Hassian değerleri bir olacak şekilde normalize edilerek çözülür.

Büyük miktarda eğitim verisi kullanılması durumunda oluşan bellek gereksinimi [145] ve konuşmacının farklı uzunluktaki ifadelerinin hesaba katması gereği çerçeve-tabanlı çekirdeklerin geliştirilmesinde ortaya çıkan temel sorunlardır. Bu sorunların çözümü için dizi çekirdek fikri ortaya atılmıştır.

2.3.4.6. Dizi Çekirdekler

Dizi çekirdek olarak da isimlendirilen cümle tabanlı çekirdekler değişken uzunluklu verilerden kaynaklanan etkilerle başa çıkacak şekilde SVM'lerin adaptasyonunu sağlayan bir araştırma alanıdır [145]. Dizi çekirdekler tüm cümleleri tek bir öznitelik vektörü olarak temsil ederek değişken uzunluklu cümlelerin normalize edilmiş bir ölçekte karşılaştırılmasına imkan sağlar. Cümlelerin bu şekilde kompakt temsili ile sınıflandırma sisteminin bellek gereksinimini de azaltır.

Sabit uzunluklu özniteliklerin oluşturulmasında üretici modelleme tekniklerinden faydalanma fikri dizi çekirdeklerin ortak yönünü oluşturmaktadır. Böylece üretici modelleme tekniklerinin genelleme gücü ile SVM'in ayırımcı özellikleri birleştirilmiş olur. Üretici ve ayırımcı modelleme tekniklerini birleştirilmesiyle oluşturulan sistemler genellikle hibrit sistemler olarak adlandırılır. Bu sistemlere örnek olarak GMM ve SVM'nin birlikte kullanılmasıyla oluşturulan sistemler gösterilebilir. Bu sistemlerin otomatik konuşmacı sınıflandırmadaki başarısının altında yatan temel unsur sisteme tanıtılan konuşmacıların ayırt edilmesinde konuşmanın olasılıksal modelinin kullanılmasıdır.

Genelleştirilmiş doğrusal ayırımcı dizi çekirdek fonksiyonu (GLDS) [141] ve GMM süpervektör (GMM-SV) [143] yaklaşımları konuşmacı sınıflandırma sistemlerinde en sık kullanılan iki dizi çekirdek fonksiyonu olup bu yöntemler aşağıda tanıtılmıştır.

2.3.4.7. Genelleştirilmiş Doğrusal Ayırımcı Dizi Çekirdek

Campbell polinom sınıflandırıcılar konusundaki araştırmalarını genişleterek hem hesaplama yükü hem de bellek ihtiyacı açısından verimli bir yöntem olan genelleştirilmiş doğrusal ayırımcı dizi çekirdeğini (GLDS) geliştirdi [141]. GLDS çekirdeği $f(x) = w^t b(x)$ şeklinde tanımlanan genelleştirilmiş ayırımcı bir fonksiyondan türetilir. Burada w konuşmacı modelini, b ise giriş uzayını genişleten skaler fonksiyonlardan oluşan bir vektörü temsil eder ve

$$b(x) = [b_1(x)\ b_2(x)\ \dots\ b_n(x)]^t \tag{2.92}$$

şeklinde gösterilir. Genişleme fonksiyonu olarak genellikle giriş bileşenlerinin belirli bir dereceye kadar monomialleri kullanılır. Bir monomil giriş bileşenlerinin doğrusal kombinasyonundan oluşan bir polinomdur ve $x_{i_1}x_{i_2}\dots x_{i_k}$ şeklinde temsil edilir. m boyutlu bir giriş vektörü için d dereceli monomiallerden oluşan genişleme vektörünün boyutu $\frac{(m+d)!}{m!d!}$ olacaktır. Örneğin $x=[x_1\ x_2]^T$ şeklinde verilen iki boyutlu bir vektörün ikinci dereceden polinomsal açılım vektörü $b(x)=[1\ x_1\ x_2\ x_1^2\ x_1x_2\ x_2^2]^T$ şeklinde tanımlanır. Polinomsal açılım vektörünün amacı $X=\{x_0,x_1,\dots,x_n\}$ şeklinde verilen öznitelik vektörlerini

$$\hat{b}(X)=\frac{1}{n}\sum_i^n b(x_i) \tag{2.93}$$

ifadesine göre $\Re^m$ den $\Re$'ye haritalamaktır. Böylece farklı uzunluktaki konuşmalardan ede edilen değişken boyutlu öznitelik vektörleri eşit uzunlu vektörlere dönüştürülmüş olur.

Haritalama fonksiyonu belirlendikten sonra ilgili sınıfa ait konuşmacı verilerinin genişletilmesiyle edilen diziler birleştirilerek N satırlı bir matris, M_{class}, oluşturulabilir.

$$M_{class}=\begin{bmatrix}\hat{b}(X_0)^t\\ \hat{b}(X_1)^t\\ \vdots\\ \hat{b}(X_N)^t\end{bmatrix} \tag{2.94}$$

Benzer şekilde arka plan kümesindeki diğer (ilgili sınıfa ait olmayan) konuşmacı verileri için de bir matris, M_{non}, tanımlanır ve bu iki matris

$$M=\begin{bmatrix}M_{class}\\ M_{non}\end{bmatrix} \tag{2.95}$$

şeklinde birleştirilerek tüm eğitim verileri temsil edilir. M'in haritalanmış verisi göz önünde bulundurularak GLDS çekirdeği

$$K(X_i,X_j)=\hat{b}(X_i)\,\hat{R}^{-1}\hat{b}(X_j) \tag{2.96}$$

şeklinde tanımlanır. Buradaki $\hat{R} = (1/N_{non})(M^T M)$ olarak tanımlanan korelasyon matrisini, N_{non} ise arkaplan kümesindeki negatif sınıfa ait cümle sayısını temsil eder.

GLDS çekirdeğinin kullanılması durumunda SVM sınıflandırıcısının karar fonksiyonu

$$f(\{x_i\}) = \sum_{i=1}^{N} \alpha_i t_i \bar{b}_i^t \, \bar{R}^{-1} \bar{b}_x + d \tag{2.97}$$

şeklinde yazılabilir. Burada $\bar{b}_i$ destek vektörlerini, $\bar{b}_x$ ise polinomsal açılım vektörünü temsil etmektedir. $d = (d, 0,0, \ldots, 0)^t$ ve genişleme fonksiyonun ilk elemanının $b_1(x) = 1$ olduğu varsayılarak bu ifade

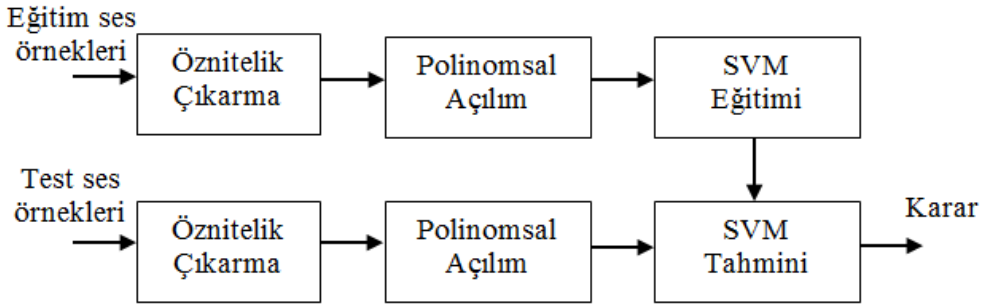
$$f(\{x_i\}) = \left(\sum_{i=1}^{N} \alpha_i t_i \bar{R}^{-1} \bar{b}_i + d\right)^t \bar{b}_x \tag{2.98}$$

şeklinde basitleştirilebilir. Özetle SVM eğitimi sonucunda tüm destek vektörler

$$w = \sum_{i=1}^{N} \alpha_i t_i \bar{R}^{-1} \bar{b}_i + d \tag{2.99}$$

şeklinde tek bir vektöre indirgenebilir ve böylece hedef modellerin saklanması ve puanlanması için kısa bir yol sağlanmış olur. Bu durumda her bir hedef skor basit bir nokta çarpımı ile, $w_{target}^t . \bar{b}_x$, hesaplanır. Burada w_{target} hedef modeli, b_x ise test cümlesinin genişletilmesiyle elde edilen ortalama öznitelik vektörünü temsil etmektedir.

SVM öznitelik uzayında açık bir genişlemeye karşılık gelen GLDS çekirdeğinin kullanımı ile hem hesaplama yükünde hem de modellerin saklanması için gereken bellek miktarında önemli bir azalma sağlanır. Ancak süpervektörlerin boyutlarının kontrolü zordur ve bu durum GLDS yönteminin temel dezavantajı olarak görülür. Bu nedenle uygulamada genellikle iki ya da üç dereceli monomiallerden oluşan polinomsal genişleme vektörü kullanılır. Şekil 2.20'de GLDS-SVM yöntemine dayalı bir sınıflandırma sisteminin işlem basamakları gösterilmiştir.



Şekil 2.20. GLDS-SVM yöntemine dayalı bir sınıflandırma sisteminin işlem basamakları.

2.3.4.8. GMM Ortalama Süpervektör Çekirdeği

Campbell ve arkadaşları tarafından önerilen GMM ortalama süpervektör çekirdeği konuşmaya dayalı tanıma sistemlerinde en fazla ilgi gören SVM çekirdeği olmuştur [143]. Bu çekirdeğin kullanımı ile uyarlanmış GMM ortalama süpervektörlerle temsil edilen akustik gözlemler ayırıcı bir SVM sınıflandırıcı ile birleştirilmiş olur. Ortalama süpervektörler değişken uzunluklu bir konuşma segmentini sabit uzunluklu bir vektöre haritalamak için pratik bir yöntem sunar. Bu haritalama işlemi SVM gibi sınıflandırıcıların değişken uzunluklu konuşmaları kullanabilmesi için zorunludur.

Daha öncede belirtildiği gibi bir GMM modeli,

$$g(x) = \sum_{k=1}^{K} w_k \ \mathcal{N}(x; \mu_k, \Sigma_k) \tag{2.100}$$

şeklinde tanımlanır. Burada w_k Gauss bileşenlerinin karışım ağırlıklarını, μ_k ortalama vektörlerini, Σ_k ise kovaryans matrislerini temsil eder. Ortalama süpervektör ise uyarlanmış bir GMM'den bileşen ortalamaları uç uca eklenerek oluşturulur ve $\mu = [\mu_1^t, \dots, \mu_K^t]^T$ şeklinde gösterilir. K GMM bileşen sayısı ve D ise öznitelik vektörünün boyutu olmak üzere KxD elemanlı bir vektör olan ortalama süpervektör, ayırıcı modelleme süreci ile üretici modelleme süreci arasındaki bilgi bağlantısı olarak görülebilir.

Öznitelik vektörlerinden oluşan bir X kümesi tek bir süpervektörle temsil edilebilir. Bunun için öncelikle bir GMM modeli X öznitelik kümesi kullanılarak eğitilir. Daha sonra eğitilen GMM modelinin bileşen ortalamaları birleştirilerek sonuç süpervektör elde edilir. GMM'lerin eğitimi için genellikle GMM-UBM yapılanması kullanılır. Bu yapılanmaya

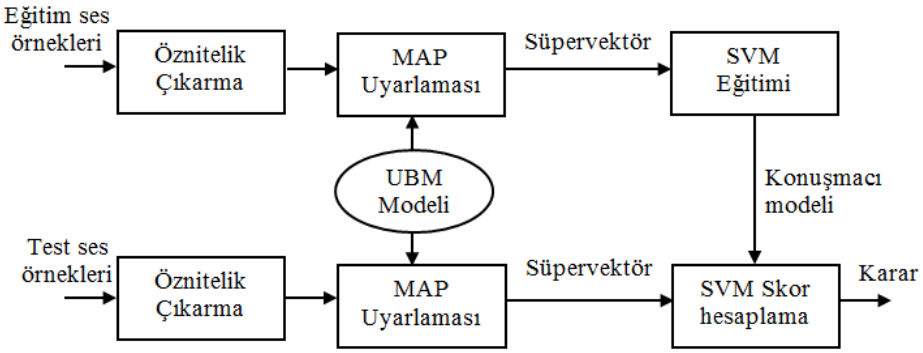
göre bir gözlem kümesini temsil eden GMM modeli, genel arka plan modeli olarak isimlendirilen bir modelden yalnızca bileşen ortalamaları uyarlanarak (MAP yöntemi ile) oluşturulur. Bu uyarlama yaklaşımı verilen bir ortalama süpervektörün tek bir referans noktadan uyarlandığını garanti etmek için gereklidir. Böylece SVM çekirdek uzayında vektörlerin doğrudan karşılaştırılması yapılabilir.

GMM süpervektör çekirdeği GMM'ler arasındaki Kullbach-Leibler (KL) divergence ölçüsü sınırlandırılarak türetilir [143]. $\lambda_{GAM} = \{w_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ şeklinde tanımlanan bir arka plan modeli (UBM) ile $\lambda_a = \{w_k, \mu_k^a, \Sigma_k\}_{k=1}^{K}$ ve $\lambda_b = \{w_k, \mu_k^b, \Sigma_k\}_{k=1}^{K}$ şeklinde MAP uyarlamalı GMM'lerle temsil edilen iki cümleye sahip olduğumuzu varsayarsak. KL divergence çekirdeği;

$$K(\lambda_a, \lambda_b) = \sum_{k=1}^{K} (\sqrt{w_k}\,\Sigma_k^{-1/2}\,\mu_k^a)^T (\sqrt{w_k}\,\Sigma_k^{-1/2}\,\mu_k^b) \tag{2.101}$$

ifadesiyle tanımlanır. Uygulama bakış açısından bu ifade tüm bileşen ortalamalarının SVM eğitiminden önce $\sqrt{w_k}\,\Sigma_k^{-1/2}$ ile normalize edilmesi anlamına gelir. Böylece yalnızca ortalama vektörleri ile oluşturulan süpervektörlere varyans ve ağırlık bilgileri dolaylı olarak dahil edilmiş olur.

GMM ortalama süpervektörlerine dayalı bir sınıflandırma sisteminin blok diyagramı Şekil 2.21'de verilmiştir. Sistemin eğitim aşamasında kayıtlı konuşmacı örneklerinden elde edilen süpervektörler kullanılarak bir SVM eğitilir. Bu işlem sonucunda tüm destek vektörleri, ağırlık ve sapma terimleri denklem 2.84'e göre tek bir vektöre indirgenir ve hedef konuşmacı modeli, w_{target}, olarak saklanır. Sınıflandırma aşamasında ise test konuşmasına karşılık gelen süpervektör GMM-UBM yapılandırmasına göre belirlendikten sonra süpervektör ile hedef model arasındaki eşleştirme skoru basit bir nokta çarpımla, $w_{target}^t . m_{test}$, hesaplanır. Burada w_{target} eğitim aşamasında oluşturulan hedef modeli, m_{test} ise test konuşmasına karşılık gelen ortalama süpervektörü temsil eder. Son aşamada ise elde edilen eşleşme skoruna göre test konuşmasının sınıfına karar verilir.



Şekil 2.21. GMM ortalama süpervektörlerine dayalı bir SVM sisteminin blok diyagramı

2.3.4.9. Çok Sınıflı SVM

Daha önce de belirtildiği gibi SVM ikili bir sınıflandırıcıdır. Ancak bu sınıflandırıcı çok sınıflı sınıflandırma problemine uygulanabilecek şekilde değiştirilebilir. Bu durumda sınıflandırıcının çok sayıda elemandan oluşan sonlu bir L kümesinden belirli bir l etiketini test verilerine ataması gerekir. Diğer bir ifade ile çok sınıflı sınıflandırma problemi aşağıdaki gibi tanımlanabilir;

- x_i eğitim vektörleri, y_i ise öznitelik vektörlerine atanan sınıf etiketleri, $y \in \{1, \dots, n\}$, olmak üzere $T = \{(x_i, y_i)\}_{i=1}^{l}$ şeklinde verilen bir eğitim kümesi için çok sınıflı bir sınıflandırıcı $y = f(x)$ şeklinde x'in sınıf etiketini veren bir karar fonksiyonu, $f(x)$, ile tanımlanır.

Çok sınıflı sınıflandırma probleminin çözümü için farklı çözümler önerilmiştir. Bu önerilerden problemin çok sayıda ikili sınıflandırma problemine ayrıştırılarak elde edilen sonuçlara göre bir karar stratejisinin belirlenmesine dayalı yaklaşımlar özellikle konuşmacı tanıma sistemlerinde yaygın olarak kullanılmaktadır. Bu konuda iki farklı strateji ön plana çıkmıştır; Bire-karşı-bir ve bire-karşı-hepsi. Bu stratejilerden ikincisi daha fazla tercih edilmektedir [108, 149, 150].

- *Bire-Karşı-Bir Stratejisi:*

Çiftli (pairwise) SVM olarak da isimlendirilen bu yaklaşımda sınıfların her olası çifti için bir ikili sınıflandırıcı oluşturulur [151]. Diğer bir ifade ile (c_1, c_2) çifti için bu iki sınıfı birbirinden ayıracak bir SVM bu iki sınıfın verileri ile eğitilir. Bu işlem sırasında eğitim verilerinin

$$y_i = \begin{cases} +1 & c_i = c_1 \; ise \\ -1 & c_i = c_2 \; ise \end{cases} \tag{2.102}$$

şeklinde $+1$ ve -1 etiketleriyle yeniden etiketlendirilmesi gerekir. Bu yaklaşıma göre N sınıflı bir problem için $N * (N-1)/2$ tane ikili SVM sınıflandırıcı eğitilmelidir. Test aşamasında ise farklı stratejiler kullanılır. Bu stratejilerden en basit ve en yaygın olanı en çok kazananlar oylamasıdır (max-wins voting) [152]. Bu stratejiye göre bir test örüntüsü tüm SVM çiftlerine uygulanır ve ikili karşılaştırmalarda daha fazla seçilen sınıfa göre test örüntünün sınıfına karar verilir. Daha açık bir ifadeyle (c_i, c_j) sınıf çifti için çiftli SVM'in karar fonksiyonu $f_{c_i c_j}$

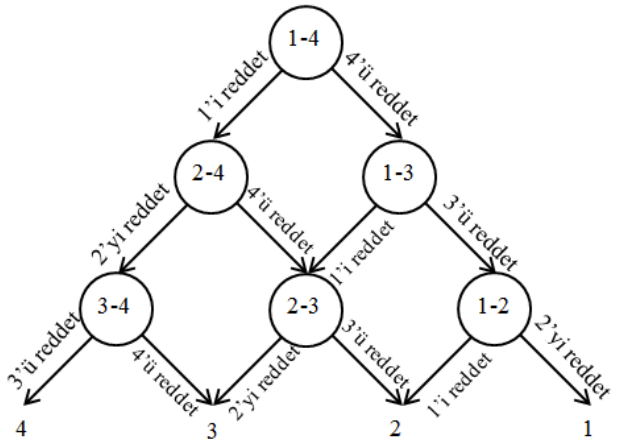
$$f_{c_i c_j}(x) = \begin{cases} 1 & x \; c_i' ye \; ait \; olarak \; sınıflandırılırsa \\ 0 & diğer \; durumda \end{cases} \tag{2.103}$$

olarak tanımlanır. Test örüntüsüne atanan sınıf etiketi ise

$$f(x) = \max_i \sum_{j \neq i} f_{c_i c_j}(x) \tag{2.104}$$

şeklinde ifade edilebilir. İki sınıfın ayın sayıda oy alması durumunda ise karar vermek için ilave bir stratejinin tanımlanması gerekir. Oylama yaklaşımına alternatif olarak Cutzu tarafından önerilen karşı oylama prensibine göre bir girişin bilinmeyen bir sınıfa sınıflandırılmasına izin verilerek yanlış kabul hatalarından kaçınılır [153]. Bu yaklaşımda bir sınıflandırıcı, $f_{c_i c_j}$, bir test örüntüsünü örneğin c_i sınıfa ait olarak puanlarsa, $x \notin c_j$ sonucuna varılır. Böylece oylama sistemi c_j'ye karşı bir oy sağlamış olur.

Bir test örüntüsüne atanacak sınıfı belirlemek için karar graflarının kullanımını içeren daha gelişmiş yöntemler vardır. Örneğin Platt tarafından çok sınıflı sınıflandırma için yönlendirilmiş asiklik grafiği (Directed Acyclic Graphs (DAG)) önerilmiştir (Şekli 2.22)[154].



Şekil 2.22. DAG ile 4 sınıflı bir sınıflandırma problemi

Bu yaklaşımda bir test örüntüsüne atanan sınıf ikili bir DAG kullanılarak belirlenir. Bu grafın her düğümü ikili bir sınıflandırıcıyı temsil eder. Bu durumda test örüntüsünün sınıf etiketinin belirlenmesi için yalnızca $N-1$ karşılaştırma gerekir. Test aşamasına kök düğümden başlanır ve iki sınıf arasında bir karar verilerek birisi reddedilir. Daha sonra seçilen sınıfa göre süreç sağ veya sol düğüme kaydırılır ve yeni bir sınıflandırma bu düğümde gerçekleştirilir. Süreç belirli bir yaprağa ulaşılana kadar devam eder.

- *Bire-Karşı-Hepsi Stratejisi:*

Bire-karşı-kalan olarak da bilinen bu yaklaşımda N tane ikili sınıflandırıcı belirli bir sınıf ile kalan $N-1$ sınıf arasında ayırım yapabilecek şekilde eğitilir [155]. Yani her sınıf için tüm eğitim verileri kullanılarak bir SVM eğitilir. Bu eğitim sırasında n sınıfına ait veriler $+1$, diğer eğitim verileri ise -1 ile etiketlenir. Test aşamasında bir test örüntüsü, x, kazanan-hepsini-alır (winner-takes-all) stratejisine göre sınıflandırılır. Diğer bir ifadeyle bir test örüntüsüne atanan etiket karar fonksiyonu için en yüksek değerin (işaretten bağımsız olarak) elde edildiği sınıfa karşılık gelecektir. n sınıfı için aşağıdaki karar fonksiyonu göz önüne alındığında,

$$f_n(x) = \sum_{i=1}^{m} y_i \alpha_i K(x, x_i) + b \tag{2.105}$$

Test örüntüsü aşağıdaki ifadeye göre sınıflandırılır;

$$f(x) = \text{argmax}_k \, f_k(x) \tag{2.106}$$

Bire-karşı-bir stratejisi ile bire-karşı-hepsi stratejisi arasında önemli bir fark yeni bir sınıfın ne kadar verimli bir şekilde sınıflandırıcıya eklenebileceği konusunda ortaya çıkar. Bu durum örneğin konuşmacı tanıma sistemine yeni bir kullanıcı eklenirken önemlidir. İlk durumda N tane yeni ikili sınıflandırıcının yeniden eğitilmesi gerekirken sonraki durumda ise sistemde mevcut olan tüm ikili sınıflandırıcıların eğitim verisinin tamamı ile yeniden eğitilmesi gerekir.

2.4. Kanal Dengeleme

Konuşmacı tanıma sitemlerinde ortaya çıkan hatalı kararların çoğu, konuşmacıların ses karakteristiklerindeki benzerlikten değil aynı konuşmacının farklı söyleyişlerindeki içsel değişkenlikten kaynaklanır. Bu içsel değişkenlikler ise iletişim kanalı etkileri (sabit hat, hücresel, VoIP, vb), dönüştürücü karakteristikleri, ortam gürültüsü (ev, cadde, ofis, vb) gibi birçok faktöre bağlıdır [156]. Solomonoff ve arkadaşları tarafından önerilen NAP (Nuisance Attribute Projection) yaklaşımı, SVM sınıflandırıcısı ile birlikte kullanılan oldukça popüler bir kanal dengeleme yöntemidir [157]. Bu yöntem belirli bir çekirdeğe özgü değildir ve her tür SVM süpervektörüne uygulanabilir. NAP yöntemindeki temel fikir kanal etkilerini temsil eden bozucu boyutların SVM uzayından çıkarılması ve böylece yalnızca konuşmacı değişkenliğine izin verilmesidir. Bu amaçla uygun bir iz düşüm matrisi kullanılarak orijinal genişletilmiş uzaydan bir alt uzaya iz düşüm yapılır.

NAP izdüşümü çok sayıda farklı konuşmacının ses kayıtlarından oluşan bir arka plan kümesi üzerinden öğrenilerek gerçekleştirilir. Bu amaçla kullanılan en basit yaklaşım her kullanıcı için verilen bir oturum ile oturumların ortalaması arasındaki fark bilgisinin kullanımına dayanmaktadır [149]. Bu fark bilgisi tüm konuşmacılar için bir araya toplanarak bir matris oluşturulur. Daha sonra kovrayans matrisi üzerinde öz değer problemi çözülerek bu küme için değişkenliğin yüksek olduğu boyutlar bulunur.

N_s konuşmacı sayısı olmak üzere i'inci konuşmacının, s_i, h_i farklı oturumdaki kayıtlarından elde edilen süpervektörlerin (veya farklı genişleme biçimlerinin)

$$\left\{\Phi_{(1,s_1)} \ \dots \Phi_{(h_1,s_1)} \ \dots \Phi_{(1,s_{N_s})} \ \dots \Phi_{(h_{N_s},s_{N_s})}\right\} \tag{2.107}$$

şeklinde verildiğini varsayarsak, öncelikle her konuşmacı için ortalama süpervektör hesaplanır ve ilgili konuşmacının tüm örneklerinden çıkarılır;

$$\widetilde{\Phi}_{(l,s_i)} = \Phi_{(l,s_i)} - \bar{\Phi}_{s_i}, \qquad l \in [1\ h_i] \tag{2.108}$$

Böylece $N = h_1 + \cdots + h_{N_s}$ olmak üzere ExN boyutlu aşağıdaki matris elde edilir.

$$M = \left[\widetilde{\Phi}_{(1,s_1)} \ldots \widetilde{\Phi}_{(h_1,s_1)} \ldots \widetilde{\Phi}_{(1,s_{N_s})} \ldots \widetilde{\Phi}_{(h_{N_s},s_{N_s})}\right] \tag{2.109}$$

Bu matris oturumlar arası değişimleri yüksek boyutlu genişletilmiş uzayda temsil eder.

Değişimlerin en yüksek olduğu K boyutlu alt uzayı belirlemek için MM^t kovaryans matrisinin en yüksek öz değerli K tane öz vektörü hesaplanır. Sonuç öz vektörler kanal değişimlerinin en belirgin olduğu alt uzaya bir temel oluşturur ve bu vektörler boyutu ExK olan bir S matrisine indirgenir.

S matrisi X ifadesinin kanal alt uzayına projeksiyonu için kullanılır. I birim matrisi olmak üzere bu projeksiyon işlemi

$$\widehat{\Phi}_X = P(\Phi_X) = (I - SS^t)\Phi_X \tag{2.110}$$

şeklinde tanımlanır. SS^t matrisi ExE boyutundadır ve bu projeksiyon işlemi hesaplama verimliliği için aşağıdaki şekilde hesaplanır.

$$\widehat{\Phi}_X = \Phi_X - S(S^t\Phi_X) \tag{2.111}$$

Bu projeksiyon sonucunda elde edilen vektörler SVM çerçevesine benzer bir şekilde kullanılır.

2.5. Birleştirme Yöntemleri

Herhangi bir tanıma sisteminin genel performansı farklı öznitelikler veya sınıflandırıcılar birleştirilerek iyileştirilebilir. Bunun iki ana nedeni vardır. Bunlardan ilki Bölüm 2.1'de tartışıldığı gibi literatürde önerilen özniteliklerin hiçbirisi ile konuşmacının bütün özellikleri tam olarak temsil edemez. Bu yüzden performans artışı sağlamak için önerilen yöntemlerin çoğunda farklı özniteliklerin birleştirilmesi fikri kullanılmaktadır. Benzer bir durum da sınıflandırıcılar için geçerlidir. Literatürde önerilmiş birçok

sınıflandırma yöntemi vardır. Bu yöntemlerin her birinde farklı teorik alt yapılar kullanılır. Bu yöntemlerin performanslarında ise bir örtüşme söz konusudur. Bu nedenle farklı sınıflandırıcıların birleştirilmesine dayanan yöntemler kullanılarak da performans artışı sağlanabilir.

Birleştirme yöntemlerinin sınıflandırılmasında farklı taksonomiler kullanılmaktadır [158]. Birleştirme yöntemleri tipik olarak dört kategoriye ayrılır; veri seviyeli, öznitelik seviyeli, skor seviyeli ve karar seviyeli. Bu seviyeler, sistem genelinde kullanılan verilerin soyutlama seviyesiyle yakından ilişkilidir. Veri seviyeli birleştirme ses analizinde çok az kullanılmaktadır. Bu yüzden de aşağıdaki bölümde diğer üç kategori incelemiş, veri seviyeli birleştirme konusuna değinilmemiştir.

2.5.1. Öznitelik Seviyeli Birleşim

Öznitelik seviyeli birleşimde farklı akustik öznitelikler, global bir öznitelik vektörü oluşturmak üzere uygun normalleştirme, dönüştürme ve ekleme işlemleri uygulayarak birleştirilir. Daha sonra oluşturulan global öznitelik vektörü öğrenme ve sınıflandırma için bir sınıflandırıcıya uygulanır. Öznitelik seviyeli birleştirme yaklaşımının avantajı hesaplama karmaşıklığını azaltmasıdır. Örneğin F tane sınıflandırıcı kullanmak yerine F tane öznitelik vektörü bir öznitelik kümesine birleştirilir ve tek bir sınıflandırıcı kullanılarak öğrenme gerçekleştirilir. [159] çalışmasında akustik öznitelikleri bir araya getirmek ve genel sınıflandırma performansını arttırmak için basit bir birleştirme işlemi uygulanmıştır. Öznitelik seviyeli birleşimin dezavantajı ise sistemin bireysel özniteliklerin güçlü yönlerinden yararlanamamasıdır.

Öznitelik seviyeli birleşim ile karar seviyeli birleşim [160] çalışmasında karşılaştırılmıştır. Çalışmada akustik ve linguistik öznitelikler çıkarılmış ve bu özniteliklerle oluşturulan sınıflandırıcıların sonuçları karar ağaçları kullanılarak birleştirilmiştir. Daha sonra aynı öznitelik kümeleri birleştirilerek öznitelik seviyeli birleşim gerçekleştirilmiştir. Çalışmada öznitelik düzeyinde birleştirilen parametreleri kullanan sınıflandırıcılar ile karar seviyeli birleşim düzeninde elde edilen sonuçlarından daha üstün sonuçlar elde edildiği gösterilmiştir.

2.5.2. Skor Seviyeli Birleşim

Skor seviyeli birleşimde her bir sınıflandırıcı her bir örnek için bir skor vektörü hesaplar. M uzunluğundaki skor vektörü, girilen örneğin her bir sınıfa ait olma derecesini temsil eder. Farklı sınıflandırıcılar tarafından üretilen skor vektörleri, genel bir skor vektörünü hesaplamak için birleştirilir.

Skor seviyeli birleşim genellikle iki kategoriye ayrılır; kural tabanlı yöntemler ve örüntü sınıflandırma yaklaşımları. Kural tabanlı yöntemlerde maksimum, minimum, toplam [161], ağırlıklı toplam [162], çarpım [163], lojistik regresyon [164] ve çoğunluk oylaması [165] gibi önceden belirlenmiş kurallar kullanılarak normalize edilmiş çoklu sınıflandırma skorları birleştirilir. Örüntü sınıflandırma yaklaşımında ise her bir sınıflandırıcı tarafından üretilen eşleşme skorları öznitelik olarak bir başka sınıflandırıcıya uygulanarak skor seviyeli birleşim gerçekleştirilir. Bu yaklaşımda genellikle destek vektör makineleri, yapay sinir ağları, karar ağaçları, k-en yakın komşuluk sınıflandırıcıları kullanılır. Kural tabanlı yöntemlerin aksine bu sınıflandırıcılar, öznitelik olarak kullanılan vektörlerinin nasıl üretildiğine bakılmaksızın karar sınırlarını öğrenebilirler. Böylece farklı yöntemlerle üretilen, homojen olmayan skorlar normalizasyona gerek duyulmadan bileştirilebilir.

Bu tez çalışmasında farklı sınıflandırıcılar tarafından üretilen bireysel eşleşme skorlarının ağırlıklı toplamlarına dayalı bir bileştirme yaklaşımı kullanılmıştır. Bu yaklaşıma göre her bir sistemin skoru deneysel olarak belirlenen bir ağırlık katsayısı ile çarpılmış ve elde edilen çarpım sonuçları toplanarak tek bir skor vektörü hesaplanmıştır. Bu işlem matematiksel olarak

$$S_f = \alpha_1 s_1 + \alpha_2 s_2 + \cdots + \alpha_N s_N \tag{2.112}$$

şeklinde gösterilir ve bu işlem sonucunda hesaplanan skor vektörüne göre, S_f, giriş konuşmacısının sınıfına karar verilir.

2.5.3. Karar Seviyeli Birleşim

Karar seviyeli birleşim skor seviyeli birleşimin daha yüksek bir soyutlaması olarak düşünülebilir. Skor seviyeli birleşimde her örnek için bir skor vektörü hesaplanırken karar seviyeli birleşimde ise girdi örneğine karşılık gelen skor vektörü bir etiket ile eşleştirilir. Böylece her giriş konuşmasına bir etiket veya karar atanmış olur. Daha sonra bu kararlar birleştirilerek nihai bir karar etiketi belirlenir. Farklı sınıflandırıcı kararlarını birleştirmek için değişik yöntemler önerilmiştir. Bu yöntemlerden en basit ve yaygın kullanılanı çoğunluk sınıfın seçimine dayalı olan yöntemdir [158, 160, 166]. Diğer ilgili yaklaşımlarda borda sayısı [167] veya basit olarak farklı sınıflandırıcı kararları kullanılarak yeni bir öznitelik vektörü oluşturulmuş, bu vektörle eğitilen sınıflandırıcının sonucuna göre de nihai karar verilmiştir. Karar seviyeli birleşim yöntemlerinin ayrıntılı bir incelemesi [168] çalışmasında bulunabilir.

3. BULGULAR VE İRDELEME

3.1. Giriş

Bu bölümde tez çalışmasında incelenen farklı öznitelik çıkarma, konuşmacı modelleme ve sınıflandırma yöntemlerinin konuşmacıların yaş ve/veya cinsiyetlerine göre sınıflandırılmasındaki performansları ayrıntılı olarak incelenmiştir. Öncelikle çalışmalarda kullanılan veri tabanları tanıtılmış daha sonra geliştirilen yaş ve/veya cinsiyet tanıma sistemlerinin uygulama detayları ve elde edilen deneysel sonuçlar verilmiştir. Her bir sistemin avantaj ve dezavantajları vurgulanarak elde sonuçların performans karşılaştırması yapılmıştır. Bu bölümde sunulan sonuçların bir kısmı değişik ulusal/uluslararası dergi ve konferanslarda makale ve bildiri olarak yayınlanarak literatüre dahil edilmiştir.

3.2. Veritabanları

Tez çalışmasında önerilen yöntemlerin değerlendirilmesinde çeşitli veritabanları kullanılmıştır. Bunlardan TIMIT ve ORATOR veritabanlarında yalnızca yetişkin ve yaşlı konuşmacılardan oluşurken, aGender veritabanı ise çocuk, genç, yetişkin ve yaşlı olmak üzere tüm yaş gruplarını kapsayan konuşmacılardan oluşmaktadır. Bu nedenle TIMIT ve ORATOR veritabanları ile yapılan çalışmalarda konuşmacılar yalnızca cinsiyetlerine göre iki sınıfa (erkek ve kadın) ayrılırken, aGender veritabanı ile yapılan çalışmalarda ise konuşmacılar hem yaş hem de cinsiyet gruplarına göre sınıflandırılmıştır. Bu veritabanlarının konuşmacı sayısı, kayıt ortamı, konuşmacıların yaş ve cinsiyet dağılımları gibi detayları aşağıdaki bölümde sunulmuştur.

3.2.1. TIMIT Veritabanı

TIMIT veritabanı, akustik-fonetik bilginin edinilmesi ve otomatik konuşma tanıma sistemlerinin geliştirilmesi ve değerlendirilmesi için konuşma verileri sağlamak üzere tasarlanmıştır. Bu veritabanı Savunma İleri Araştırma Projeleri Ajansı - Bilgi Bilim ve Teknoloji Ofisi (DARPA-ISTO) sponsorluğunda çeşitli birimlerin ortak çabaları sonucunda oluşturulmuştur. Metin tasarımı Massachusetts Institute of Technology (MIT),

Stanford Araştırma Enstitüsü (SRI) ve Texas Instruments (TI) arasındaki ortak çaba sonucunda gerçekleştirilmiştir. Konuşmalar TI'da kaydedilmiş, MIT'de yazılmış ve Ulusal Standartlar ve Teknoloji Enstitüsü (NIST)'de ise CD-ROM üretimi için doğrulanmış ve muhafaza edilmiştir.

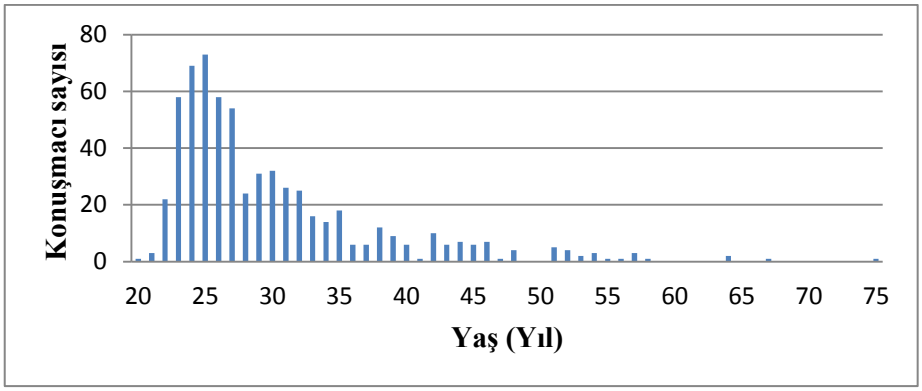
TIMIT veritabanı Amerika Birleşik Devletlerinin 8 ana lehçesinin konuşulduğu bölgelerden seçilen 438 erkek ve 192 kadın toplam 630 konuşmacının ses kayıtlarından oluşmaktadır. Konuşmacıların lehçe bölgeleri çocukluk yıllarında yaşadığı coğrafi bölgeye göre belirlenmekte olup TIMIT veritabanını oluşturan konuşmacıların lehçelerine göre dağılımı cinsiyet ayrımı yapılarak Tablo 3.1'de verilmiştir.

Tablo 3.1. TIMIT veritabanının lehçe dağılımı

Lehçe Bölgesi (DR)	**#Erkek**	**#Kadın**	**Toplam**
1	31 (63%)	18 (27%)	49 (%8)
2	71 (70%)	31 (30%)	102 (16%)
3	79 (67%)	23 (23%)	102 (16%)
4	69 (69%)	31 (31%)	100 (16%)
5	62 (63%)	36 (37%)	98 (16%)
6	30 (65%)	16 (35%)	46 (7%)
7	74 (74%)	26 (26%)	100 (16%)
8	22 (67%)	11 (33%)	33 (5%)
8	438 (70%)	192 (30%)	630 (100%)

Lehçe bölgeleri: dr1: New England, dr2: Northern, dr3: North Midland, dr4:South Midland, dr5:Southern, dr6:New York City, dr7: Western, dr8:Army Brat (moved around)

Konuşmalar, kayıtlar arasında oturum aralığı olmaksızın, 16 kHz'lik örnekleme frekansında ses geçirmeyen bir kabin içinde yüksek kaliteli bir mikrofon kullanılarak kaydedilmiştir. Yaşları 20 ile 75 arasında değişen konuşmacıların yaş dağılımı Şekil 3.1'de gösterilmiştir. Her konuşmacının 10'ar cümle seslendirdiği TIMIT veritabanında toplam kayıt sayısı 6300, konuşmacı bazında ortalama kayıt süresi (10 cümlenin tamamının) ise yaklaşık 30 saniyedir.



Şekil 3.1. TIMIT konuşmacılarının yaş dağılımı

TIMIT veritabanında seslendirilen metinler 2 diyalekt cümle (SA), fonetik olarak yoğun 450 cümle (SX) ve fonetik olarak değişiklik gösteren 1890 cümleden (SI) oluşmaktadır. Bu metinlerden diyalekt cümleler (SA1 ve SA2) tüm konuşmacılar tarafından seslendirilmiş ve farklı lehçelerin ortaya çıkardığı değişkenliği göstermek üzere tasarlanmıştır.

TIMIT veritabanındaki metinler ve konuşmacılar aşağıdaki ölçütler kullanarak eğitim ve test kümelerine bölünmüştür;

- Veri kümesinin yaklaşık % 20-30'u test amaçlı, kalan %70-80'lik bölüm ise eğitim için kullanılmalıdır.
- Hiçbir konuşmacı hem eğitim hem de test bölümlerinde olmamalıdır.
- Bütün lehçe bölgeleri, her lehçeden en az 1 erkek 1 kadın konuşmacı, her iki grupta da temsil edilmelidir.
- İki alt kümedeki metin materyalinin örtüşme miktarı asgariye indirilmelidir; Mümkünse hiçbir metin aynı olmamalıdır.
- Tüm sesbirimleri test materyali kapsamında olmalıdır, tercihen her bir sesbirimi farklı bağlamlarda defalarca ortaya çıkmalıdır.

3.2.2. aGender Veritabanı

aGender, telefon hatları üzerinden okuma ve yarı-spontane konuşma örneklerini içeren bir veritabanıdır. Bu veritabanı ana dili Almanca olan 945 konuşmacının bir ses portalını kendi telefonlarından arayarak okudukları metin ve bazı açık uçlu soruların cevaplarından oluşmaktadır. Konuşmacıların cinsiyet ve/veya yaşını otomatik olarak tespit

eden sistemlerin geliştirilmesine yardımcı olmak amacıyla "InterSpeech 2010 Paralinguistic Challenge" organizasyonu tarafından hazırlanan aGender veritabanı her birinde 18 konuşma olan 6 farklı oturumda kaydedilen toplam 47 saatlik ses kayıtlardan oluşmaktadır. Detayları Tablo 3.2'de verilen sabit ve değişken içerikli ifadelerin seslendirildiği ses kayıtlarının ortalama uzunluğu 2.58 saniye, toplam kayıt sayısı ise 65364 dür. Konuşmalar arasında çeşitliliğin sağlanması için oturumlar arasında bir günlük molalar planlanmış ancak oturumlarının zaman/tarih denetimi yapılmamıştır. Ayrıca konuşmacı başına düşen toplam oturum sayısı da kontrol edilmemiştir. Her kişinin 6 çağrısı iç ve dış mekanlarda bir cep telefonu ile yapılarak farklı kayıt ortamları oluşturulmuştur. Ses sinyalleri standart cep telefonları (GSM standardı) ve sabit hat bağlantıları üzerinden 8000 Hz, 8 bit A-law formatında kaydedilmiş, daha sonra 8000 Hz 16 bit PCM'e genişletilmiştir.

Tablo 3.2. aGeder veritabanında seslendirilen ifadeler

#	Konu	Miktar/açıklama	Örnek/motive edici sorular
1-3	Komut kelimesi	Önceden belirlenmiş kelimeler 18 farklı komut kelimesi	"stop" "hilfe" (yardım)
4-5	Gömülü komut	Önceden belirlenmiş cümleler İçinde 22 farklı komut olan cümleler	"einen Berater bitte" "bir danışman lütfen"
6	Ay	Önceden belirlenmiş kelimeler 12 ay	"Januar" (ocak) "Dezember" (Aralık)
7	Haftanın günleri	Önceden belirlenmiş kelimeler 7 gün	"Montag" (Pazartesi) "Sonntag" (Pazar)
8	Göreceli zaman açıklaması	Önceden belirlenmiş kelimeler 9 farklı kelime/cümle	"heute"(bugün) "nächsteWoche"(Sonraki hafta)
9	Resmi tatiller	Önceden belirlenmiş kelimeler 6 farklı kelime	"Ostern" (paskalya) "Weihnachten" (Noel)
10	Doğum günü	Serbest ifade	"Doğum günün ne zaman?"
11	Zaman	Serbest ifade	"Saat kaç"
12	Tarih	Serbest ifade	"Herhangi bir tarih söylermisiniz"
13	Telefon numarası	Serbest ifade	"Herhangi bir telefon numarasını numara numarası söylermisiniz"
14	Posta kodu	Serbest ifade	"Posta kodunuzu söylermisiniz"
15	Ad	Serbest ifade	"Adınızı söylermisiniz"
16	Soyad	Serbest ifade	"Soyadınızı söylermisiniz"
17-18	Evet/hayır	Serbest ifade	"Şuan yağmur yağıyormu?" "Şuan dışardamısın"

aGender veritabanı eğitim, geliştirme ve test olmak üzere üç bölümden oluşmaktadır. Bu bölümlerden eğitim ve geliştirme yaş/cinsiyet tanıma sistemlerinin geliştirilmesinde, test ise elde edilen sonuçların karşılaştırılmasında kullanılmaktadır. Ayrık konuşmacılardan oluşan bu bölümlerdeki konuşmacı, oturum ve cümle sayıları Tablo 3.3'te verilmiştir.

Tablo 3.3. aGender veritabanındaki konuşmacı, oturum ve cümle sayıları

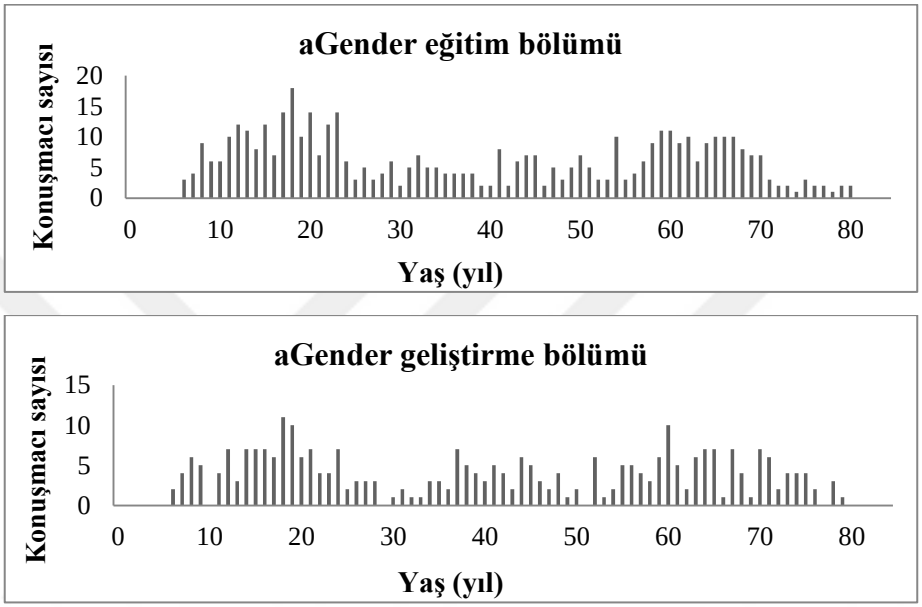
Eğitim/Geliştirme bölümündeki konuşmacı sayısı	770
Test bölümündeki konuşmacı sayısı	175
Eğitim/Geliştirme bölümündeki oturum sayısı	3625
Toplam cümle sayısı	65241
Eğitim/Geliştirme bölümündeki cümle sayısı	53076
Test bölümündeki cümle sayısı	12165

aGender veritabanındaki konuşmacılar çocuk, genç, yetişkin ve yaşlı olmak üzere 4 yaş grubuna ayrılmıştır. Bu yaş gruplarının belirlenmesinde yaşla oluşan fizyolojik değişimlerden ziyade uygulama kaynaklı ihtiyaçlar etkili olmuştur. Yaşlarına göre 4 gruba ayrılan konuşmacılar, çocuklar hariç cinsiyetlerine göre 2 gruba (erkek ve kadın) daha ayrılarak toplamda 7 yaş-cinsiyet grubunda tanımlanmıştır. Eğitim ve geliştirme bölümündeki konuşmacıların yaş ve cinsiyet bilgileri veritabanında paylaşılırken, test bölümündeki konuşmacıların yaş ve cinsiyet bilgileri paylaşılmamıştır. aGender veritabanında tanımlı yaş-cinsiyet grupları ile eğitim ve geliştirme bölümündeki konuşmacı/konuşma sayıları Tablo 3.4'te verilmiştir.

Tablo 3.4. aGender veritabanında tanımlı yaş-cinsiyet sınıfları ile eğitim ve geliştirme bölümündeki konuşmacı/ kayıt sayıları

Sınıf ID	**Yaş Grubu**	**Yaş**	**Cinsiyet**	**#Eğitim**	**#Geliştirme**
1	Çocuk	7-14	X	68/4406	38/2396
2	Genç	15-24	Erkek	63/4638	36/2722
3	Genç	15-24	Kadın	55/4019	33/2170
4	Yetişkin	25-54	Erkek	69/4573	44/3361
5	Yetişkin	25-54	Kadın	66/4417	41/2512
6	Yaşlı	55-80	Erkek	72/4924	51/3561
7	Yaşlı	55-80	Kadın	78/5549	56/3826

Yaşları 7 ile 80 arasında değişen konuşmacıların seçiminde konuşmacıların 7 hedef sınıfa dağılımlarının yaklaşık eşit olmasına dikkat edilmiştir. Ayrıca çocuk sınıfında cinsiyet ayrımı yapılmasa da bu sınıfındaki konuşmacılar da cinsiyetlerine göre dengeli bir şekilde seçilmiştir. aGender veritabanının eğitim ve geliştirme bölümündeki konuşmacıların yaş dağılımı Şekil 3.2'de gösterilmiştir.



Şekil 3.2. aGender veritabanının eğitim ve geliştirme bölümlerindeki konuşmacıların yaş histogramı

3.2.3. Orator Veritabanı

Orator veritabanı 8 Almanca monoloğunun seslendirildiği 145 ses kaydından oluşan bir veri kümesidir. Bu kayıtlardan 117 tanesi profesyonel aktörler (6 kadın, 7 erkek) tarafından seslendirilen ve konuşmacıların farklı sosyal durumlarını (mutlu, mutsuz, kızgın, emin, vb) ifade eden ses kayıtlarından oluşmaktadır. Kalan 28 ses kaydı ise aktör olmayan konuşmacılar tarafından seslendirilen (2 kadın, 12 erkek) doğal konuşmaları içermektedir. Konuşmacılar geniş bir sosyal çevreden olup kayıtlar farklı ortamlarda (işyeri, ev, bir arkadaşın evi gibi) gerçekleştirilmiştir. Ses kayıtları konuşmacının kafasına takılan bir mikrofon vasıtasıyla yapılarak kayıt sırasında ağız ile mikrofon arasındaki mesafesinin sabit kalması sağlanmıştır. Özel bir yükseltici ile güçlendirilen sinyaller 48 kHz, 16 bit mono doğrusal PCM formatında saklanmıştır.

Her konuşmacının farklı sayıda ses kaydının olduğu veritabanında ortalama konuşma süresi yaklaşık 30 saniyedir. Konuşmacılardan üç tanesinin (aktör) hafif aksanı (Bavyera, Avusturya ve İsviçre) vardır. Bazı konuşmacılar oldukça belirsiz yerel lehçeye sahiptir. Ancak konuşmacıların çoğunluğu Almancayı iyi konuşmaktadır. Veritabanındaki aktörlerin yaşı 27 ile 66 arasında olup yaş ortalaması 38'dir. Aktör olmayan konuşmacıların yaş aralığı 21-72, yaş ortalaması ise 41'dir. Toplam 145 ses kaydından oluşan ORATOR veritabanındaki konuşmacıların kayıt sayıları ile birlikte yaş, cinsiyet ve meslek bilgileri Tablo 3.5'te verilmiştir.

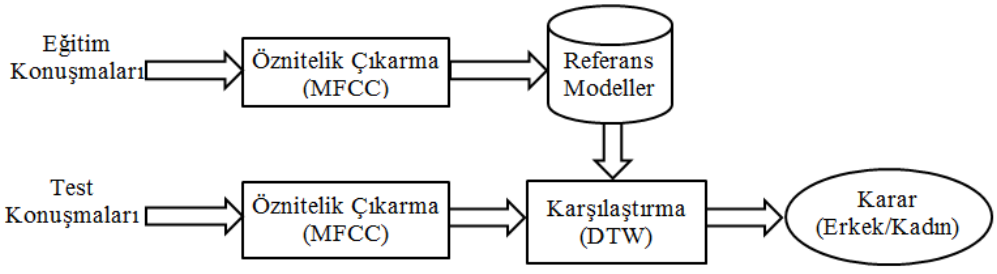
Tablo 3.5. ORATOR veritabanındaki konuşmacıların yaş, cinsiyet ve kayıt sayısı

ID	Cinsiyet	Meslek	Yaş	Kayıt sayısı
AM1	Erkek	Aktör	27	1
AF1	Kadın	Aktör	29	6
AM2	Erkek	Aktör	38	6
AM3	Erkek	Aktör	40	5
AF2	Kadın	Aktör	33	6
AF3	Kadın	Aktör	34	9
AM4	Erkek	Aktör	27	9
AF4	Kadın	Aktör	?	10
AF5	Kadın	Aktör	30+	12
AM5	Erkek	Aktör	36	13
NM1	Erkek	Öğrenci	24	2
AM6	Erkek	Aktör	66	12
NM2	Erkek	İşçi	41	2
NM3	Erkek	İşçi	39	2
NM4	Erkek	Aktör değil	38	2
NM5	Erkek	Grup yöneticisi	31	2
NM6	Erkek	İşçi	30+	2
NM7	Erkek	İşçi	57	2
NM8	Erkek	Satış elemanı	55	2
NM9	Erkek	Öğrenci	25	2
NF1	Kadın	Öğrenci	21	2
NM10	Erkek	Profesör	72	2
AM7	Erkek	Aktör	56	20
AF6	Kadın	Aktör	28	8
NM11	Erkek	Öğretmen	57	2
NF2	Kadın	Kredi uzmanı	55	2
NM12	Erkek	Öğrenci	24	2

3.3. DTW Yöntemi ile Cinsiyet Belirleme Çalışması

Dinamik zaman bükme (DTW) hız ve zamansal olarak değişiklik gösteren iki seri arasındaki benzerliği ölçmede kullanılan bir yöntemdir. Video, ses, grafik gibi doğrusal olarak temsil edilebilen her türlü verinin analizinde kullanılan DTW yönteminin en yaygın kullanıldığı alan ise ses tanıma uygulamalarıdır. Konuşma sinyali konuşmacının aksanı, konuşma stili, konuşma hızı ve konuşmacının ruh hali gibi çeşitli nedenlerden dolayı değişkenlik gösterir. Bu durum aynı konuşmacının değişik zamanlarda seslendirdiği ses kayıtlarda bile olur. Söz konusu değişiklikler doğrusal değildir ve bu nedenle doğrusal bir eşleştirme başarısız sonuç verecektir. DTW yöntemi iki zaman serisi arasındaki optimum eşleşmeyi bulmak için kullanılan bir yöntemdir.

Bu bölümde konuşmacı cinsiyetinin otomatik olarak belirlenmesi amacıyla DTW'ye dayalı olarak geliştirilen cinsiyet tanıma sisteminin uygulama detayları ve elde edilen sonuçlar verilmiştir. Blok diyagramı Şekil 3.3'te verilen DTW'ye dayalı cinsiyet tanıma sistemi eğitim ve test olmak üzere iki aşamadan oluşur. Sistemin eğitim aşamasında cinsiyeti bilinen konuşmacıların ses kayıtlarından çıkarılan öznitelik vektörleri cinsiyet bilgisi ile birlikte saklanarak referans modeller oluşturulur. Test aşamasında ise cinsiyeti tespit edilmek istenen konuşmacının öznitelik vektörü eğitim aşamasındaki gibi çıkarılır. Daha sonra eğitim aşamasında belirlenen referans modelleri ile karşılaştırılarak iki vektör arasındaki uzaklık hesaplanır. Son aşamada ise elde edilen uzaklık değerlerine göre konuşmacının cinsiyetine karar verilir. Farklı veri kümleri ile yapılan testlerde elde edilen sonuçlar aşağıda verilmiştir.



Şekil 3.3. DTW'ye dayalı cinsiyet tanıma sisteminin blok diyagramı

3.3.1. Deneysel Sonuçlar

Blok diyagramı Şekil 3.3'te verilen cinsiyet tanıma sistemi metne bağımlı ve metinden bağımsız olmak üzere iki şekilde test edilmiştir. Metne bağımlı testlerde eğitim ve test için aynı içeriğe sahip ses kayıtları kullanılırken, metinden bağımsız testlerde ise eğitim ve test aşamasında farklı içeriğe sahip ses kayıtları kullanılmıştır. 19 erkek 8 kadın konuşmacının 8 farklı cümleyi farklı sayıda telaffuz ettikleri 145 ses kaydından oluşan ORATOR veritabanı ile yapılan testlerde her konuşmacının birinci cümleyi seslendirdiği ilk örnekler sistemin eğitimi için kullanılmıştır. Eğitim aşamasında toplam 27 ses kaydından oluşan eğitim kümesindeki örneklere karşılık gelen öznitelikler çıkarılmış ve referans model olarak saklanmıştır. Böylece bu aşamanın sonunda 19 erkek ve 8 kadın referans modeli oluşturulmuştur. Çalışmada öznitelik vektörü olarak ise 20 MFCC katsayısından oluşan bir vektör kullanılmıştır. ORATOR veritabanı ile yapılan deneylerde kullanılan eğitim ve test kümelerinin içeriği Tablo 3.6'da verilmiştir.

Tablo 3.6. ORATOR veritabanı ile yapılan deneylerde kullanılan eğitim ve test kümeleri

Eğitim Kümesi		
Konuşmacı sayısı:	19 erkek 8 kadın	
Seçilen örnekler:	***Birinci*** cümlenin bir bölümü	
Örnek sayısı:	27	
Metin içeriği:	"*In der Vergangenheit"*	
	Metne bağımlı test kümesi	**Metinden bağımsız test kümesi**
Konuşmacı sayısı:	19 erkek 8 kadın	19 erkek 8 kadın
Seçilen örnekler:	***Birinci*** cümlenin bir bölümü	***Altıncı*** cümlenin tamamı
Örnek sayısı:	145	145
Metin içeriği:	*"In der Vergangenheit"*	*"Wir erledigen alles Andere"*

Geliştirilen cinsiyet tanıma sistemi ORATOR veritabanındaki 145 ses kaydın tamamı ile test edilmiştir. Öncelikle test konuşmalarına karşılık gelen öznitelik vektörleri çıkarılmış (eğitim aşmasındaki gibi), daha sonra eğitim aşamasında belirlenen referans modellerle karşılaştırılmıştır. Bu işlem sırasında iki vektör (test vektörü ile referans model) arasındaki zamansal kaymalar DTW yöntemi ile giderilmiştir. Zamansal olarak hizalanan vektörlerin arasındaki uzaklığın hesaplanmasında ise iki farklı uzaklık ölçüsü kullanılmıştır;

- Manhattan uzaklığı: $d_M = \sum_{k=1}^{k=N} |x_k - y_k|$
- Öklit uzaklığı: $d_E = \sqrt{\sum_{k=1}^{N} |x_k - y_k|^2}$

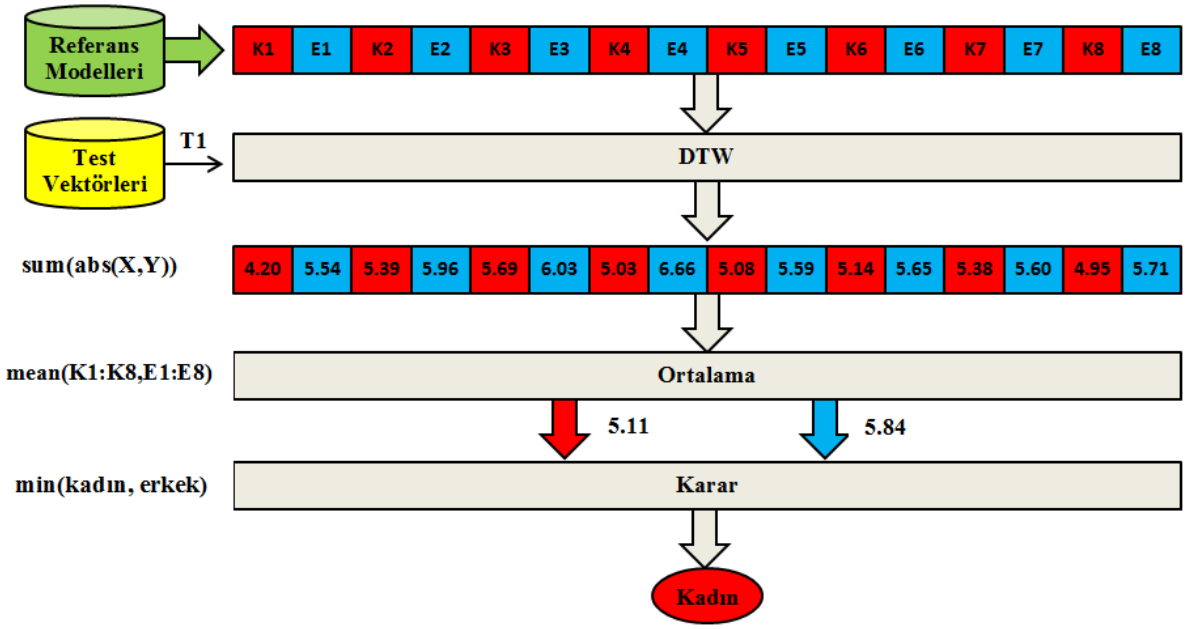
Test vektörü ile referans modeller arasındaki uzaklık hesaplandıktan sonra karar aşamasına geçilir. Karar aşamasında iki farklı yaklaşım kullanılmıştır. Bu yaklaşımların ilkinde test konuşmasının erkek ve kadın referans modellerine olan uzaklıkların ortalaması hesaplanarak minimum ortalama uzaklığa sahip referans modelin cinsiyeti test konuşmacısının cinsiyeti olarak seçilir. İkinci yaklaşımda ise test konuşmacısının cinsiyeti minimum uzaklığa sahip referans modelin cinsiyeti olarak seçilir. Tablo 3.7'de ORATOR veritabanı ile yapılan testlerde elde edilen sonuçlar verilmiştir.

Tablo 3.7. ORATOR veritabanı ile yapılan test sonuçları

Metne bağımlı test sonuçları			
Uzaklık ölçütü/Karar Yöntemi	***Doğru karar sayısı***	***Hatalı karar sayısı***	***Başarı oranı***
Öklit/Ortalama uzaklık	141	4	%97.24
Öklit/Minimum uzaklık	137	8	%94.48
Manhattan/Ortalama uzaklık	142	3	%97.93
Manhattan/Minimum uzaklık	137	8	%94.48
Metinden bağımsız test sonuçları			
Uzaklık ölçütü/Karar Yöntemi	***Doğru karar sayısı***	***Hatalı karar sayısı***	***Başarı oranı***
Öklit/Ortalama uzaklık	133	12	%91.72
Öklit/Minimum uzaklık	106	39	%73.10
Manhattan/Ortalama uzaklık	117	28	%80.68
Manhattan/Minimum uzaklık	102	43	%70.34

ORATOR veritabanı ile yapılan test sonuçlarından önerilen cinsiyet tanıma sisteminin metne bağımlı verilerle daha başarılı olduğu görülmektedir. Ayrıca uzaklık ölçütü olarak öklit uzaklığının, karar yöntemi olarak ise ortalama uzaklığa dayalı yöntemin daha başarılı sonuçlar verdiği elde edilen sonuçlardan görülmektedir.

Önerilen cinsiyet tanıma sistemi TIMIT veritabanından seçilen bir veri kümesi ile de test edilmiştir. 8 farklı diyalekt bölgesinden seçilen 16 konuşmacının (8 erkek, 8 kadın) 10'ar ses kaydından oluşan bu veri kümesi ile yapılan testlerde sistemin eğitimi için tüm konuşmacıların "SA1" isimli ilk kayıtları kullanılırken her konuşmacının kalan 9 konuşması ise test için ayrılmıştır. Çalışmada uzaklık ölçütü olarak ölkit uzaklığı, karar aşamasında ise ortalama uzaklığa dayalı yöntem kullanılmıştır. Bir test konuşmasından çıkarılan öznitelik vektörünün (T1) referans modellerle karşılaştırılarak konuşmacının cinsiyetinin belirlendiği işlem basamakları Şekil 3.4'te gösterilmiştir.



Şekil 3.4. DTW yöntemiyle yapılan bir testin işlem basamakları

Şekil 3.4'ten de görüldüğü gibi "T1" isimli test vektörü 8 erkek (E1-E8) ve 8 kadın (K1-K8) referans modeli ile karşılaştırılarak her bir referans model ile arasındaki uzaklık hesaplanmıştır. Daha sonra bu uzaklıkların cinsiyet gruplarına göre ortalaması alınmış ve minimum ortalama uzaklığın elde edildiği cinsiyet grubu (Kadın) test konuşmacısının cinsiyeti olarak seçilmiştir.

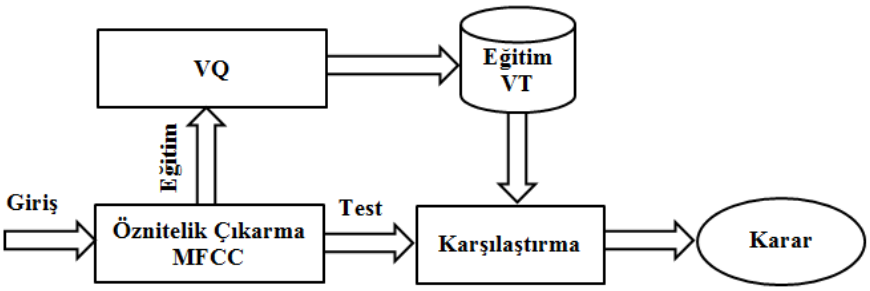
TIMIT veritabanından seçilen veri kümesi ile yapılan test sonuçları Tablo 3.8'de verilmiştir. Elde edilen sonuçlardan önerilen cinsiyet tanıma sisteminin TIMIT veri kümesi ile daha başarılı olduğu görülmektedir. ORATOR veritabındaki ses kayıtlarının farklı ortamlarda yapılmış olması bu durumun nedeni olarak yorumlanabilir.

Tablo 3.8. DTW'ye dayalı cinsiyet tanıma sisteminin TIMIT veri kümesi ile yapılan test sonuçları

	Erkek	**Kadın**	**Toplam**
Konuşmacı sayısı	8	8	16
Örnek Sayısı	80	80	160
Eğitim kümesindeki örnek sayısı	8	8	16
Test kümesindeki örnek sayısı	72	72	144
Doğru karar sayısı	71	70	141
Hatalı karar sayısı	1	2	3
Başarı oranı	%98.61	%97.22	%97.91

3.4. Vekötör Nicemleme (VQ) Yöntemi ile Cinsiyet Belirleme Çalışması

Bu çalışmada metinden bağımsız tanıma sistemlerinde yaygın olarak kullanılan VQ yöntemi cinsiyet tanıma problemine uygulanmıştır. Konuşmacıların erkek ve kadın olarak iki sınıfa ayrılması amaçlanan çalışmada konuşmacı modeli olarak kullanılan VQ kod kitabının boyutu ile başarı arasındaki ilişki incelenmiş ve en uygun kod kitabı boyutu belirlenmiştir. Blok diyagramı Şekil 3.5'te verilen cinsiyet tanıma sistemi üç aşamadan oluşmaktadır; eğitim, karşılaştırma ve karar. Eğitim aşamasında cinsiyeti bilinen konuşmacıların ses örneklerinden çıkarılan öznitelik vektörleri VQ yöntemi ile kod kitaplarına dönüştürülmüş ve cinsiyet bilgisi ile birlikte kaydedilerek eğitim aşaması tamamlanmıştır. Daha sonra cinsiyeti belirlenmek istenen konuşmacının ses kaydından elde edilen öznitelik vektörü veritabanındaki modellerle karşılaştırılarak iki vektör arasındaki öklit uzaklığı hesaplanmıştır. Son aşamada ise elde edilen uzaklıkların cinsiyet gruplarına göre ortalaması hesaplanmış ve ortalama uzaklığı küçük olan grubun cinsiyetine göre test konuşmacısının cinsiyetine karar verilmiştir.



Şekil 3.5. VQ yöntemi ile geliştirilen cinsiyet tanıma sisteminin blok diyagramı

Önerilen sistem iki farklı veri kümesi ile test edilmiştir. Bu testlerin ilkinde TIMIT veritabanından seçilen ses kayıtları kullanılmıştır. Geliştirilen sistem 8 erkek 8 kadın konuşmacının "SA2" isimli ses kayıtlarından oluşan veri kümesi ile eğitilip, 56 kadın, 112 erkek konuşmacının 10'ar cümlesinden oluşan toplam 1680 ses kaydından oluşan veri kümesi ile test edilmiştir. Farklı kod kitabı boyutu (kod vektörü sayısı) kullanılarak yapılan deneylerde elde edilen sonuçlar Tablo 3.9'da verilmiştir.

Tablo 3.9. VQ'ya dayalı cinsiyet tanıma sisteminin TIMIT veri kümesi ile yapılan test sonuçları

Kod kitabı boyutu	**Doğru karar sayısı**	**Yanlış karar sayısı**	**Başarı oranı (%)**
4	1277	403	76,01
8	1554	126	92,50
16	1624	56	96,67
32	1637	43	97,44
64	***1646***	***34***	***97,98***
128	1641	39	97,68

Elde edilen sonuçlardan da görüldüğü gibi konuşmacı modeli olarak kullanılan VQ kod kitabının boyutu arttıkça önerilen cinsiyet tanıma sisteminin başarısı da artmıştır. Ancak 64'ten sonraki boyut artışının başarı üzerinde öneli bir etkisi olmamıştır. Hem başarı oranı hem de işlem yükü ve bellek miktarı düşünüldüğünde önerilen cinsiyet tanıma sistemi için optimum kod kitabının boyutunun 64 olduğu görülmüştür.

VQ'ya dayalı olarak geliştirilen cinsiyet tanıma sistemi Boğaziçi Üniversitesi tarafından hazırlanan Türkçe bir veritabanı [169] ile de test edilmiştir. 11 amatör

tiyatrocunun (7 kadın, 4 erkek) 4 farklı ruhsal durumda (sevinçli, olağan, kızgın ve üzgün) seslendirdiği 11 cümlenin ses kayıtlarından oluşan veritabanı üzerinde iki farklı test yapılmıştır. Bu testlerin ilkinde konuşmacıların yalnızca olağan konuşmaları kullanılmıştır. Sistem her konuşmacının olağan konuşma şekli ile seslendirdiği "beni çok şaşırttın" cümlesi ile eğitilmiştir. Eğitimde kullanılan cümle dışındaki 10 cümle ise sistemin testi için kullanılmıştır. Boğaziçi Üniversitesi duygulu konuşma veritabanının olağan konuşmaları ile yapılan testin uygulama detayları ve elde edilen başarı oranı Tablo 3.10'da verilmiştir.

Tablo 3.10. VQ'ya dayalı cinsiyet tanıma sisteminin Boğaziçi Üniversitesi duygulu konuşma veritabanının olağan konuşmaları ile yapılan test sonuçları

Eğitim kümesindeki konuşmacı sayısı	4 erkek, 7 kadın
Eğitim kümesindeki örnek sayısı	11 (her konuşmacının birer cümlesi)
Eğitim kümesinde seslendirilen içerik	"Beni çok şaşırttın"
Test kümesindeki konuşmacı sayısı	4 erkek 7 kadın
Test kümesindeki örnek sayısı	110 (Her konuşmacının 10'ar cümlesi)
Test kümesinde seslendirilen içerikler	1-Dışarıda kar yağıyor, 2-Sınavdan yetmiş aldım, 3-Hoca bana yetmiş verdi, 4-Galatasaray maçı iki sıfır kazandı, 5-Telefonum çalıyor, 6-Kurs yarın bitiyor, 7-Yarın kar yağacakmış, 8-Dersler iki hafta ertelendi, 9- Kapı açık kalmış, 10- Dersi sadece iki kişi geçemedi
Kod kitabı boyutu	64
Doğru karar sayısı	109
Hatalı karar sayısı	1
Başarı oranı	%99.09

Boğaziçi Üniversitesi duygulu konuşma veritabanı ile yapılan ikinci testte de sistem konuşmacıların normal konuşma şekli ile seslendirdiği "beni çok şaşırttın" içerikli cümleleri ile eğitilmiştir. Test aşamasında ise eğitimde kullanılan cümle dışında kalan 10 cümlenin 4 farklı duygu şekli ile seslendirildiği ses kayıtları kullanılmıştır. Böylece geliştirilen cinsiyet tanıma sisteminin başarısı ile konuşmacının duygusal durumu arasındaki ilişki incelenmiştir. Yapılan test ile ilgili detaylar ve elde edilen başarı oranı Tablo 3.11'de verilmiştir.

Tablo 3.11. VQ'ya dayalı cinsiyet tanıma sisteminin Boğaziçi Üniversitesi duygulu konuşma veritabanı ile yapılan test sonuçları

Eğitim kümesi	Tablo 3.10 ile aynı
Test kümesindeki konuşmacı sayısı	4 erkek 7 kadın
Test kümesindeki örnek sayısı	440 (Her konuşmacının 4 duygu şekli ile seslendirdiği 10'ar cümle)
Test kümesinde seslendirilen içerik	Tablo 3.10 ile aynı
Kod kitabı boyutu	64
Doğru karar sayısı	339
Hatalı karar sayısı	101
Başarı oranı	%70.04

Elde edilen sonuçlardan da görüldüğü gibi önerilen cinsiyet tanıma sisteminin başarısı konuşmacının duygusal durumundan oldukça fazla etkilenmektedir. Bu durum duygusal durumla değişen ses karakteristiklerinin cinsiyet ayrımını zorlaştıran bir etkiye sahip olduğu şeklinde yorumlanabilir. Elde edilen sonuçların konuşmacının duygusal durumuna göre dağılımı Tablo 3.12'de verilmiştir. Bu tablodan da görüldüğü gibi sevinçli konuşmaların ancak %66'sının cinsiyeti doğru olarak tanınmıştır. Kızgın konuşmaların %68'inin, üzüntülü konuşmaların ise %74'ünün cinsiyeti doğru tanınmıştır. Bu sonuçlardan cinsiyet tanıma başarısını en az etkileyen duygunun üzüntü, en fazla etkileyen duygunun ise sevinç olduğu görülmektedir.

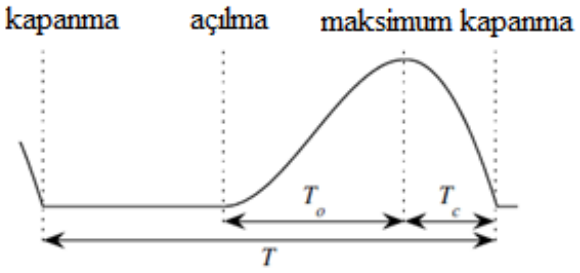
Tablo 3.12. VQ'ya dayalı cinsiyet tanıma sisteminin başarısının konuşmacının duygusal durumuna göre değişimi

	Kızgın	**Olağan**	**Sevinçli**	**Üzüntülü**
Örnek sayısı	110	110	110	110
Doğru karar sayısı	75	109	73	82
Hatalı karar sayısı	35	1	37	28
Başarı oranı	%68.18	%99.09	%66.36	%74.54

3.5. Ses Kaynağına Dayalı Özniteliklerle Cinsiyet Belirleme Çalışması

Günümüzde yaygın olarak kullanılan özniteliklerin (MFCC gibi) çoğu doğrudan ses sinyalinden çıkarılır. Ses sinyalinden elde edilen bu öznitelikler temelde ses yolunun özelliklerini içinde barındırır ve kayıt şartlarından önemli ölçüde etkilenir. Ses yolu ise fonem bağımlı olduğu için ses yolu özelliklerini içinde barındıran özniteliklerde fonem bağımlı olur. Ayrıca bu özniteliklerle güvenilir tanıma sağlanması için tüm fonetik yapıyı kapsayan uzun ses örneklerinden çıkarılması gerekir. Bu nedenle veri miktarına daha az bağımlı olan ve tam olarak metinden bağımsız özniteliklere ihtiyaç vardır. Metinden bağımsız öznitelik olarak ses kaynağı öznitelikleri iyi bir alternatiftir. Ses kaynağı öznitelikleri "glottal volume velocity waveform" veya kısaca "glottal flow" olarak isimlendirilen sesli seslerin (voiced sound) kaynağını karakterize eden parametrelerdir. En yaygın kullanılan ses kaynağı özniteliği ses tellerinin titreşim hızı olan temel frekans parametresidir. Gırtaksal akışın şeklini (kapanma süresi gibi) ve frekans uzayı etkilerini temsil eden parametrelerde öznitelik olarak kullanılmaktadır. Gırtlaksal akış sinyalinden çıkarılan parametreler üç ana gruba ayrılır; zaman-uzayı parametreleri, frekans uzayı parametreleri ve modele dayalı parametreler. Bu çalışmada gırtlaksal akış sinyalinden zaman uzayında çıkarılan açılma (OQ), kapanma (CIQ) ve hız (SQ) oranı ile frekans uzayında çıkarılan harmonik seviye farkı (H1-H2) parametrelerinin konuşmacının cinsiyeti ile ilişki incelenmiştir. Gırtlaksal akış sinyalini tahmin etmek için Paavo Alku tarafından geliştirilen tekrarlamalı uyarlamalı ters filtreleme (IAIF) yöntemi kullanılmıştır. Bir konuşma sinyalini giriş olarak alan ve bu sinyale karşılık gelen gırtlaksal akış sinyalini tahmin eden IAIF yönteminin detayları Bölüm 2.14'te verilmiştir.

Gırtlaksal akış sinyalinden zaman uzayı parametrelerinin çıkarılması için gırtlaksal açılma, kapanma ve maksimum akış gibi kritik zaman noktalarının belirlenmesi gerekir. Bir gırtlaksal akış sinyalinin temsili şekli ve kritik zaman noktaları Şekil 3.6'da gösterilmiştir.



Şekil 3.6. Gırtlaksal akış sinyalinin temsili şekli ve kritik zaman noktaları

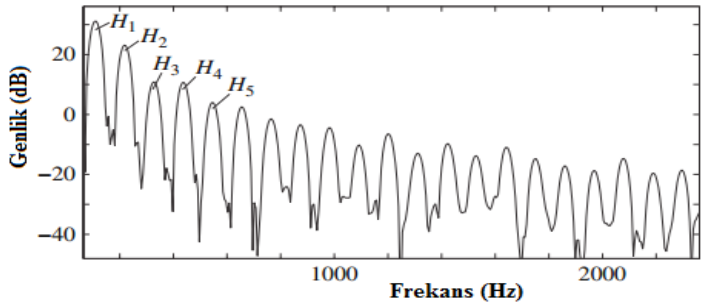
Ters filtreleme (IAIF) yöntemi ile tahmin edilen gırtlaksal akış sinyalinin kritik zaman noktaları belirlendikten sonra gırtlaksal açılma oranı (OQ), kapanma oranı (CIQ) ve hız oranı (SQ) parametreleri aşağıdaki şekilde hesaplanır.

$$Açılma\ oranı(OQ) = \frac{T_o+T_c}{T} \quad (3.1)$$

$$Kapanma\ oranı(CIQ) = \frac{T_c}{T} \quad (3.2)$$

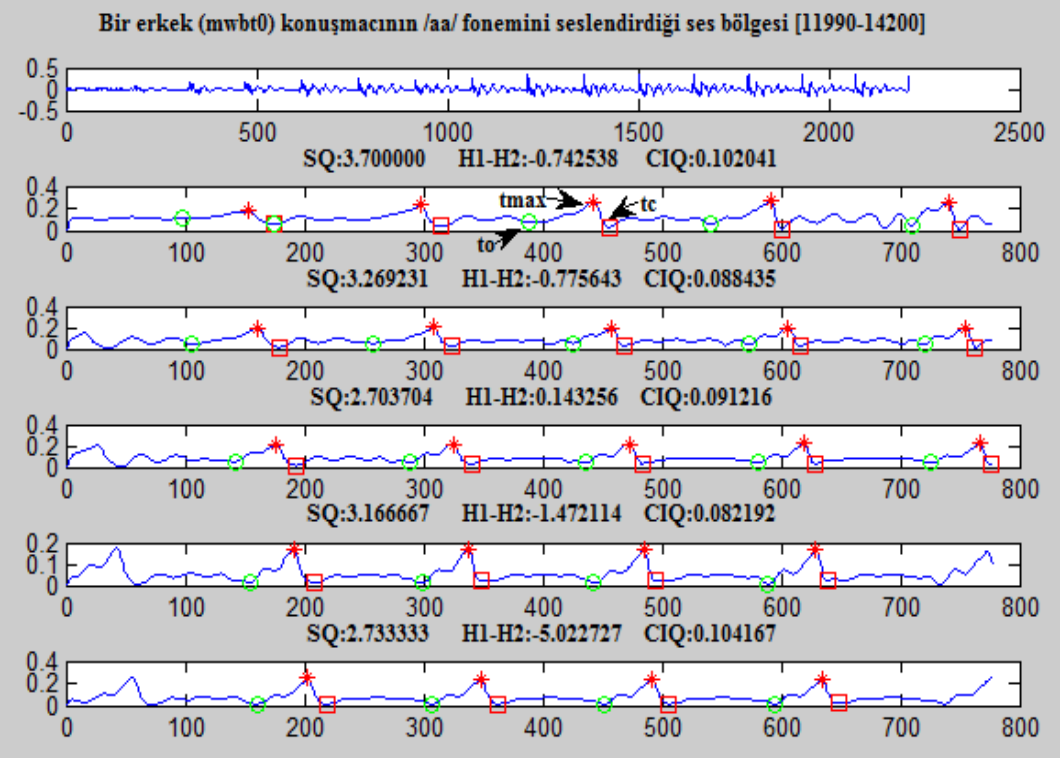
$$Hız\ oranı\ (SQ) = \frac{T_o}{T_c} \quad (3.3)$$

Çalışmada incelenen harmonik seviye farkı parametresi (H1-H2) ise gırtlaksal akış sinyalinin spektrumundan elde edilen ilk iki harmoniğin desibel cinsinden farklı alınarak hesaplanır. Şekil 3.7'de bir akış sinyalinin temsili gösterimi verilmiştir. Bu gösterimde ilk beş harmonik seviye H1'den H5'e kadar sıralı numaralar ile temsil edilmiştir.

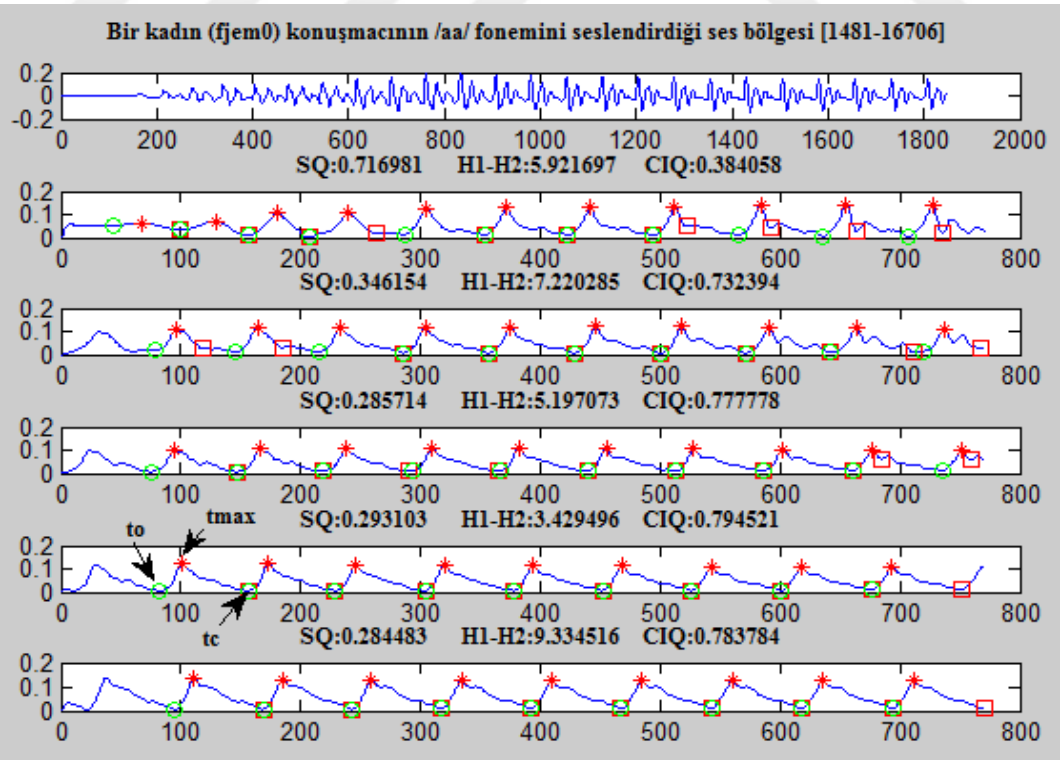


Şekil 3.7. Bir akış spektrumunun temsili şekli

Gırtlaksal akış sinyalinden elde edilen OQ, CIQ, SQ ve H1-H2 parametreleri ile konuşmacının cinsiyeti arasındaki ilişkinin araştırıldığı bu çalışmada TIMIT veritabanından seçilen ses kayıtları kullanılmıştır. 438 erkek, 192 kadın konuşmacının 10'ar cümlesinden oluşan TIMIT veritabanından her konuşmacının /a/ fonemini seslendirdiği konuşma bölgeleri seçilerek 50 ms uzunluğunda 5 analiz bölgesi oluşturulmuştur. Daha sonra bu bölgeler üzerinde IAIF yöntemi kullanılarak her bölgeye karşılık gelen gırtlaksal akış sinyali elde edilmiştir. Son aşamada ise gırtlaksal akış sinyalinin kritik zaman noktaları belirlenerek bu noktalara göre zaman uzayı parametreleri (OQ, CIQ ve SQ) hesaplanmıştır. Çalışmada gırtlaksal akış sinyalinin tepe noktaları t_{max} olarak, türevinin minimum olduğu noktanın sağındaki ilk sıfır geçiş noktası t_c ve t_{max}'ın solundaki son sıfır geçiş noktası da t_o olarak belirlenmiştir. Ancak gırtlaksal akış sinyalindeki küçük bir çukur bile bu noktaların yerini değiştireceğinden gırtlaksal sinyaldeki çukurların derinliğine ve genişliğine bakılarak kritik zaman noktalarına karar verilmiştir. Çalışmada incelenen bir erkek ve bir kadın konuşmacının /a/ fonemini seslendirdiği ses bölümünden elde edilen gırtlaksal akış sinyali ve belirlenen kritik zaman noktaları Şekil 3.8 ve 3.9'da gösterilmiştir.



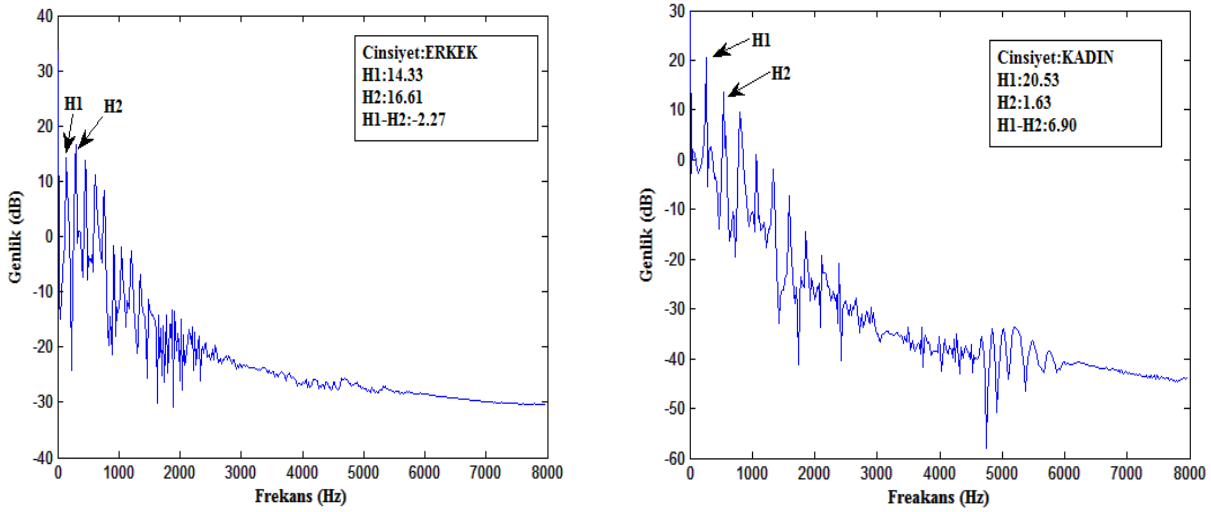
Şekil 3.8. Bir erkek konuşmacının konuşmasından çıkarılan gırtlaksal akış sinyali ve belirlenen kritik zaman noktaları



Şekil 3.9. Bir kadın konuşmacının konuşmasından çıkarılan gırtlaksal akış sinyali ve belirlenen kritik zaman noktaları

Şekil 3.8 ve 3.9'dan da görüldüğü gibi her analiz bölgesinde konuşmacının temel frekansına göre farklı sayıda gırtlaksal darbe oluşmaktadır. Çalışmada her analiz bölgesindeki gırtlaksal darbelerin kritik zaman noktaları (t_o, t_c ve t_{max}) ayrı ayrı tespit edilerek bu noktaların ortancası seçilmiş ve ilgili bölgenin zaman uzayı parametreleri (OQ, CIQ ve SQ) hesaplanmıştır. Daha sonra aynı işlem beş analiz bölgesi için uygulanarak ilgili konuşma için nihai zaman parametreleri elde edilmiştir. Böylece kritik zaman noktalarının yanlış tespit edilmesinden kaynaklanabilecek hatalı sonuçlar en aza indirgenmiştir. Yapılan çalışmada erkek ve kadın konuşmacıların konuşmalarından elde edilen gırtlaksal akış sinyalinin şeklinde ve bu sinyalden elde edilen zaman uzayı parametreleri arasında belirgin bir fark olduğu görülmüştür. Bu farklılık gırtlaksal akış sinyalinin ve bu sinyalden çıkarılan parametrelerin konuşmacının cinsiyetinin tanınmasında iyi bir öznitelik olarak kullanılabileceği anlamına gelmektedir.

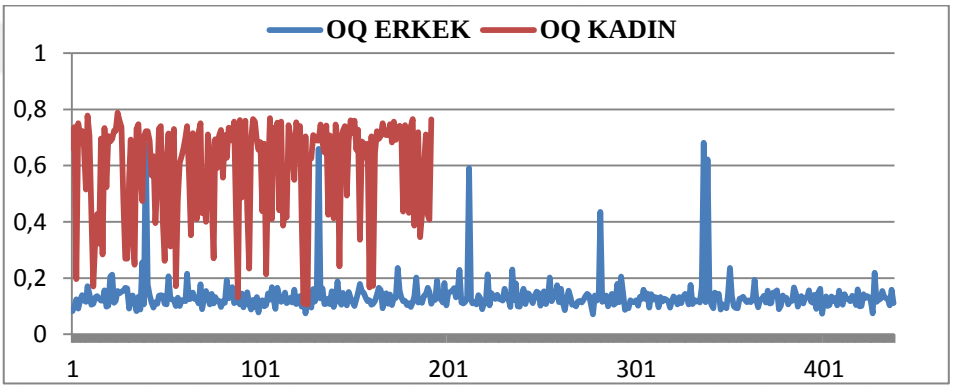
Çalışmada kullanılan harmonik seviye farkı parametresi (H1-H2) ise glottal akış sinyalinin spektrumundan ilk iki harmonik seviye arasındaki fark desibel cinsinden hesaplanarak elde edilmiştir. Bir erkek ve bir kadın konuşmacının seslendirdiği /a/ fonemine ait gıtlaksal sinyalin harmonik yapısı Şekil 3.10'da gösterilmiştir.



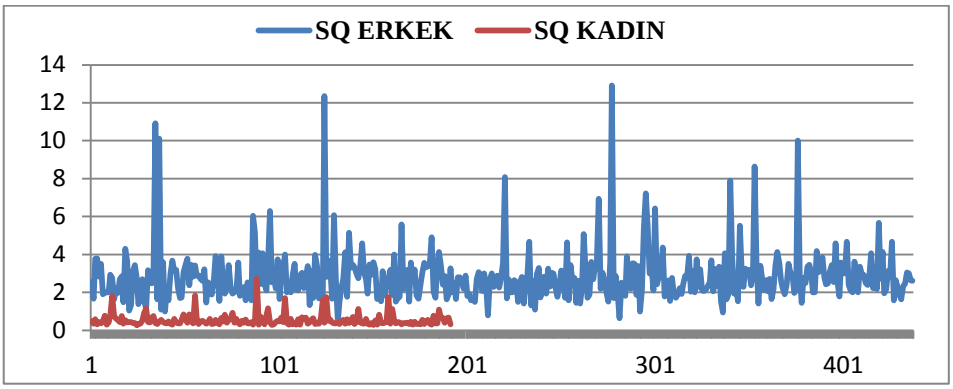
(a) (b)

Şekil 3.10. Bir erkek (a), bir kadın (b) konuşmasından elde edilen glottal sinyalin harmonik yapısı

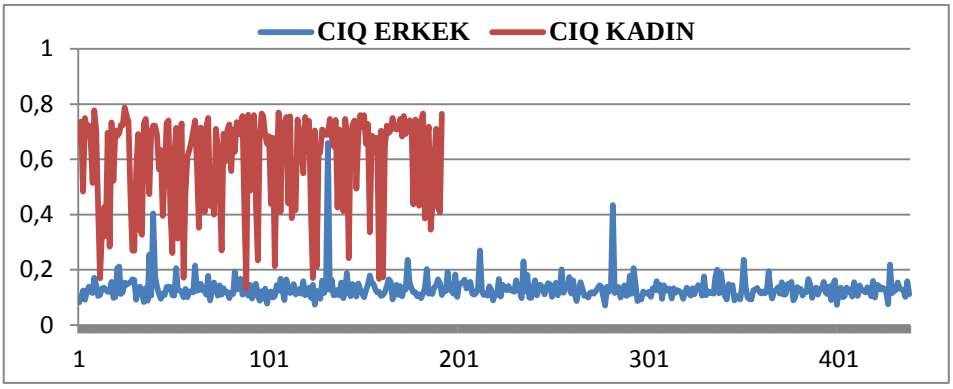
Şekil 10'dan da görüldüğü gibi erkek konuşmacıya ait spektrumda ilk harmonik seviye ikinci harmonik seviyeden küçükken kadın konuşmacının spektrumda ise ilk harmonik seviye ikinci harmonik seviyeden büyüktür. Bu durum harmonik seviye farkı parametresinin de konuşmacının cinsiyetine belirlemede kullanılabileceği anlamına gelmektedir. Çalışmada TIMIT veritabanındaki tüm konuşmacıların /a/ fonemini seslendirdiği ses bölümleri kullanılarak gırtlaksal akış sinyali çıkarılmıştır. Daha sonra bu sinyallerden OQ, SQ, CIQ zaman uzayı parametreleri ve H1-H2 frekans uzayı parametreleri çıkarılmıştır. Bu parametrelerin konuşmacının cinsiyetine göre dağılımları Şekil 3.11-3.14'te gösterilmiştir.



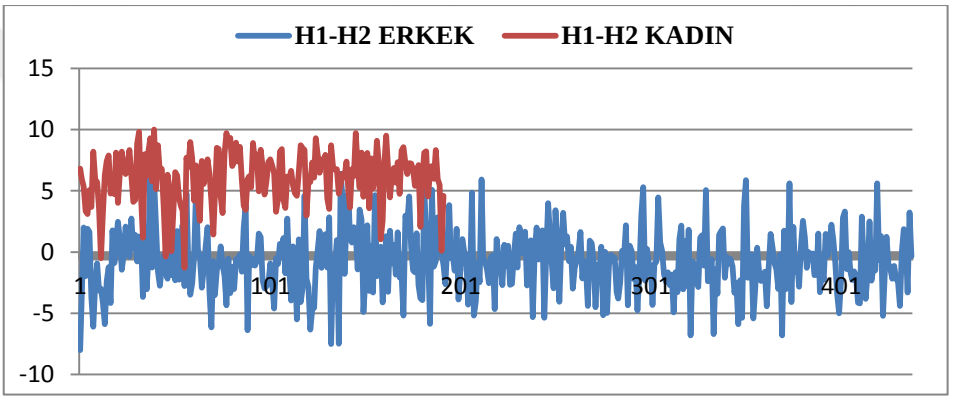
Şekil 3.11. Gırtlaksal açılma oranı parametresinin (OQ) konuşmacının cinsiyetine göre dağılımı



Şekil 3.12. Gırtlaksal hız oranı parametresinin (SQ) konuşmacının cinsiyetine göre dağılımı



Şekil 3.13. Gırtlaksal kapanma oranı parametresinin (CIQ) konuşmacının cinsiyetine göre dağılımı



Şekil 3.14. Harmonik seviye farkı parametresinin (H1-H2) konuşmacının cinsiyetine göre dağılımı

Gırtlaksal akış sinyalinden elde edilen OQ, SQ, CIQ ve H1-H2 parametreleri ile konuşmacının cinsiyeti arasında bir ilişkinin olduğu bu parametrelerin dağılımlarından kolaylıkla görülebilir. Çalışmada OQ, SQ, CIQ ve H1-H2 parametrelerinin dağılımları incelenerek her parametre için birer eşik seviye belirlenmiş ve bu eşik seviyeye göre konuşmacının cinsiyetine karar verilmiştir. TIMIT veritabanındaki 630 konuşmacının /a/ fonemini seslendirdiği konuşmaları ile yapılan testlerde belirlenen eşik seviyeler ve bu seviyeye göre elde edilen cinsiyet tanıma oranları Tablo 3.13'te verilmiştir.

Ses kaynağından (gırtlaksal akış sinyali) türetilen zaman ve frekans uzayı parametrelerinin konuşmacının cinsiyetini belirlemede oldukça başarılı olduğu elde edilen sonuçlardan görülmektedir. Yapılan testlerde en yüksek cinsiyet tanıma başarısı %97.77 ile gırtlaksal kapanma oranı (CIQ) parametresi ile elde edilirken, en düşük başarı ise harmonik seviye farkı parametresi (H1-H2) ile %93.33 olarak elde edilmiştir.

Tablo 3.13. CIQ, SQ, OQ ve H1-H2 parametrelerine göre cinsiyet tanıma oranları

Parametre	**CIQ**	**SQ**	**OQ**	**H1-H2**
Eşik değer	0.25	1	0.25	3 dB
Doğru karar sayısı	616	614	613	588
Hatalı karar sayısı	14	16	17	42
Başarı oranı (%)	97.77	97.46	97.30	93.33

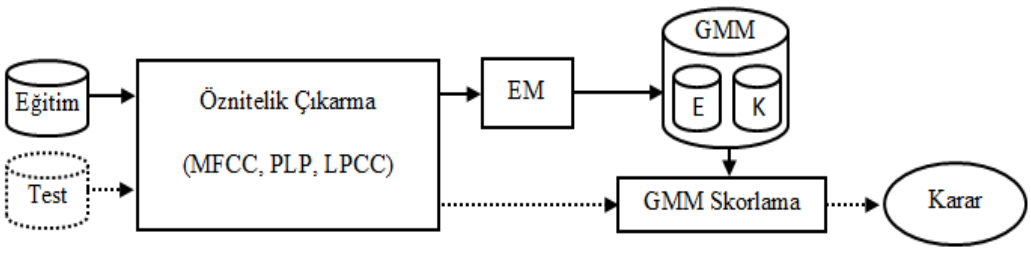
3.6. GMM Yöntemi ile Cinsiyet Belirleme Çalışması

Gauss karışım modeli özellikle karmaşık verileri modellemede kullanılan güçlü ve esnek bir araçtır. Bu çalışmada ses kayıtlarından çıkarılan öznitelik vektörleri GMM ile modellenmiş ve konuşmacının cinsiyetlerini otomatik olarak tanıyan bir sistemin geliştirilmesinde kullanılmıştır. Blok diyagramı Şekil 3.15'te verilen cinsiyet tanıma sitemi eğitim, test ve karar olmak üzere üç aşamadan oluşmaktadır. Eğitim aşamasında cinsiyeti bilinen konuşmacıların seslendirdiği ses kayıtlarından çıkarılan öznitelikler kullanılarak erkek ve kadın konuşmacılar için iki farklı GMM modeli, P_{erkek} ve $P_{kadın}$, oluşturulur. Test aşamasında ise cinsiyeti belirlenmek istenen konuşmacının ses kaydından çıkarılan öznitelikler eğitim aşamasında oluşturulan cinsiyet modelleri ile karşılaştırılır ve

$$LL_{erkek}(x) = \frac{1}{N}\sum_{i=1}^{N} log(P_{erkek}(x_i|w,\mu,\sum) \quad (3.4)$$

$$LL_{kadın}(x) = \frac{1}{N}\sum_{i=1}^{N} log(P_{kadın}(x_i|w,\mu,\sum) \quad (3.5)$$

ifadelerine göre iki farklı olabilirlik skoru hesaplanır. Karar aşamasında ise elde edilen logaritmik olabilirlik skorları karşılaştırılarak test konuşmacısının cinsiyetine karar verilir.

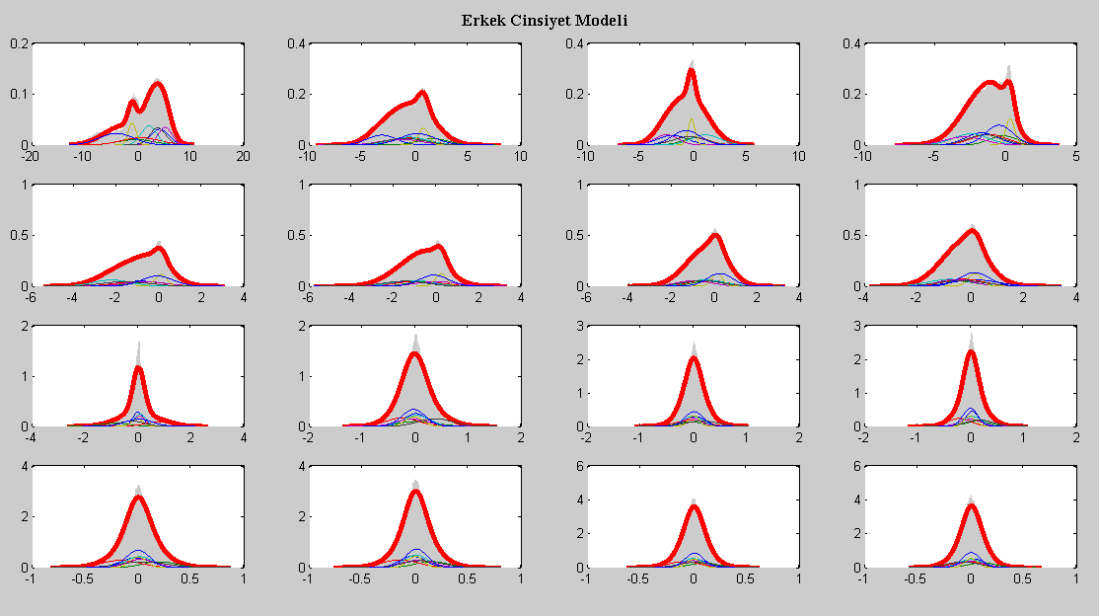


Şekil 3.15. GMM'ye dayalı cinsiyet tanıma sisteminin blok diyagramı

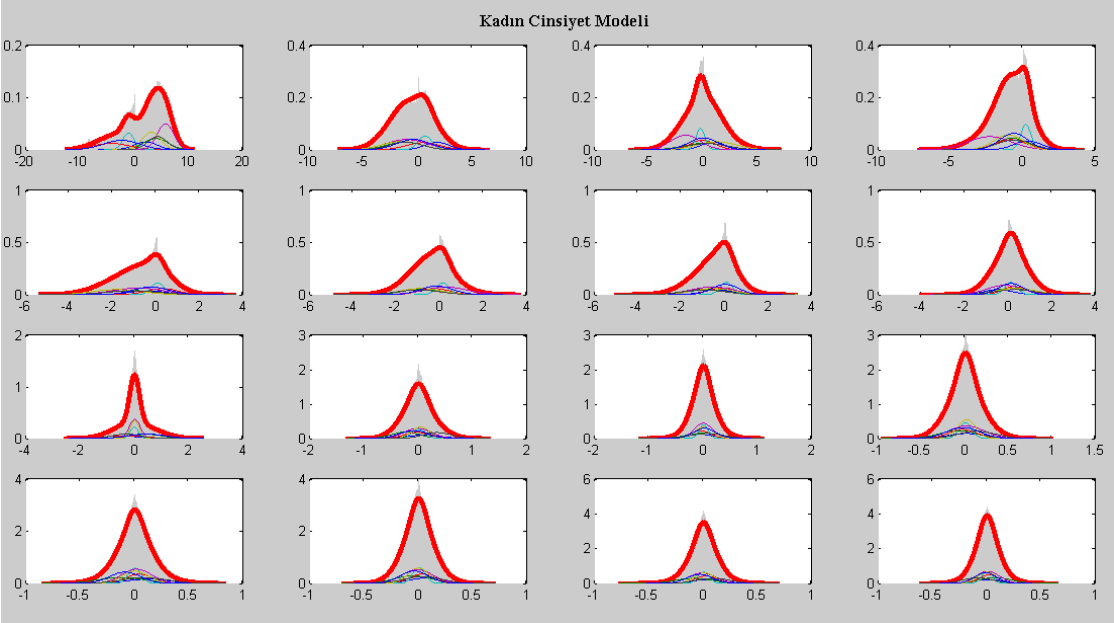
Çalışmada önerilen cinsiyet tanıma sistemi TIMIT veritabanından seçilen ses kayıtları ile test edilmiştir. Sistem veritabanından rastgele seçilen 72 konuşmacının 720 konuşmasıyla oluşturulan veri kümesiyle eğitilmiştir. Bu verilerin seçiminde konuşmacıların yaş ve cinsiyet dağılımlarının yaklaşık eşit olmasına dikkat edilmiştir. Böylece bu verilerle eğitilen cinsiyet modellerinde belirli bir yaş grubunun baskın olması önlenmiştir. Test aşamasında ise eğitim aşamasında kullanılan konuşmacılar dışındaki diğer 558 konuşmacının 5580 konuşması kullanılmıştır.

Çalışmada MFCC, LPCC ve PLP olmak üzere üç farklı öznitelik çıkarma yöntemi kullanılarak bu yöntemlerin konuşmacının cinsiyetini tanımadaki başarıları araştırılmıştır. Her bir yöntemle elde edilen kepstral katsayılar türevlerinden elde edilen delta katsayılarıyla (Δ, ΔΔ) birleştirilerek öznitelik vektörü olarak kullanılmıştır. Çalışmada ayrıca kullanılan özniteliklerin boyutunun ve bu öznitelikleri modellemede kullanılan GMM modelinin bileşen sayısı ile başarı arasındaki ilişki de araştırılmıştır. Bu amaçla çeşitli testler yapılmış ve bu testler sonucunda önerilen cinsiyet tanıma sistemi için en uygun öznitelik türü, öznitelik boyutu ve GMM bileşen sayısı belirlenmiştir.

Çalışmada cinsiyet modeli olarak kullanılan iki Gauss karışım modeli Şekil 3.16 ve 3.17'de gösterilmiştir. Bu modellerde konuşma sinyallerinden çıkarılan 16 boyutlu öznitelik vektörleri 8 Gauss bileşeninin toplamı şeklinde temsil edilmektedir.



Şekil 3.16. 8-bileşenli GMM erkek modeli

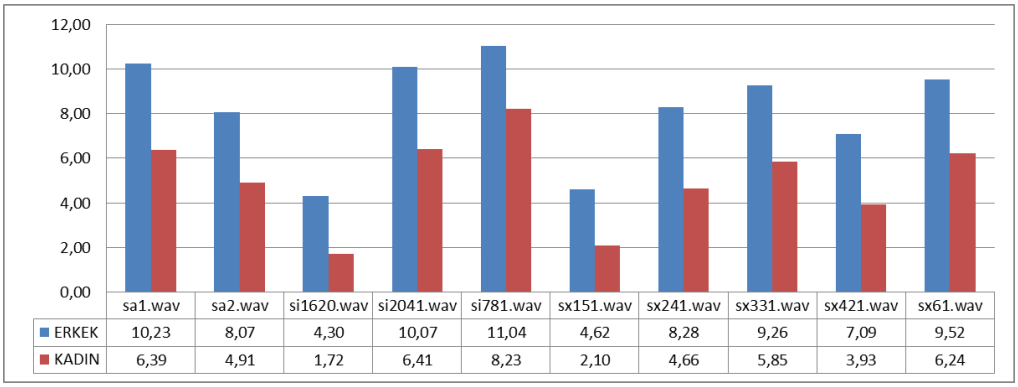


Şekil 3.17. 8-bileşenli GMM kadın modeli

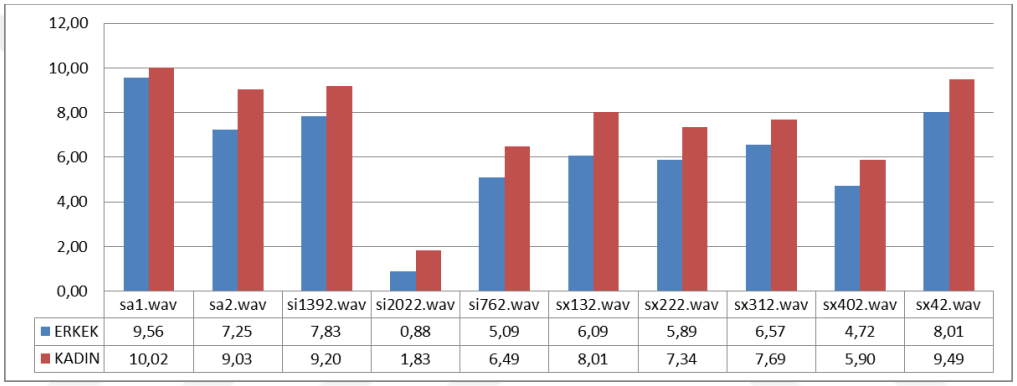
Şekil 3.16 ve 3.17'den de anlaşıldığı gibi GMM'yi oluşturan bileşen sayısı arttıkça konuşma sinyallerinden çıkarılan öznitelikler daha iyi temsil edilecektir. Ancak bileşen sayısıyla birlikte işlem yükü ve bellek miktarı da artacaktır. Bu nedenle bileşen sayısı seçilirken işlem yükü ve başarı arasında bir denge kurulmalıdır.

Erkek ve kadın cinsiyet modellerinin eğitimi tamamlandıktan sonra test aşamasında geçilir. Test aşamasında cinsiyeti bilinmeyen konuşmacının ses kayından çıkarılan öznitelik vektörü ile eğitim aşamasında oluşturulan cinsiyet modelleri karşılaştırılır. Bu karşılaştırma sonucunda test konuşması ile ilgili cinsiyet modeli arasındaki benzerliği ifade eden olabilirilik skoru hesaplanır. Son aşamada ise elde edilen olabilirilik skoruna göre test konuşmacısının cinsiyetine karar verilir. Bir erkek ve bir kadın konuşmacının onar 10'ar konuşması ile yapılan testlerde elde edilen logaritmik olabilirlik skorları Şekil 3.18 ve 3.19'da gösterilmiştir.

Elde edilen sonuçlar incelendiğinde test edilen konuşmacının cinsiyetine uygun olarak ilgili cinsiyet modelinin olabilirlik skorunun daha büyük olduğu görülmektedir. Bu skorlar göz önünde bulundurularak yapılacak bir tanıma işleminde test edilen 20 konuşmanın tamamı doğru cinsiyet grubuna atanacaktır.



Şekil 3.18. Bir erkek konuşmacının (mabc0) 10 konuşmasına ait logaritmik olabilirlik skorları



Şekil 3.19. Bir kadın konuşmacının (faem0) 10 konuşmasına ait logaritmik olabilirlik skorları

GMM'ye dayalı olarak geliştirilen cinsiyet tanıma sistemi TIMIT veritabanından seçilen 558 konuşmacının 5580 konuşması ile test edilmiştir. Üç farklı öznitelik vektörü kullanılarak yapılan bu testlerde elde edilen sonuçlar Tablo 3.14'te verilmiştir.

Elde edilen sonuçlardan GMM bileşen sayısındaki ve öznitelik boyutundaki artışın cinsiyet tanıma oranını arttırdığı, ancak belirli bir değerden sonraki artışın başarı üzerinde olumlu bir etkisinin olmadığı görülmektedir. Çalışmada kullanılan üç öznitelik türü ile de %95'in üzerinde cinsiyet tanıma başarısı sağlanmıştır. Düşük boyutlu öznitelik vektörlerin yüksek bileşenli GMM'lerle modellenmesi durumunda başarı oranı %80'lere kadar düşerken, yüksek boyutlu özniteliklerin düşük bileşenli GMM'lerle modellenmesi durumunda başarı oranı %95'ler seviyesinde kalmıştır. Bu durum önerilen cinsiyet tanıma sisteminin başarısında kullanılan öznitelik vektörünün boyutunun daha önemli olduğu şeklinde yorumlanabilir.

Tablo 3.14. GMM'ye dayalı cinsiyet sınıflandırma sisteminin bileşen sayısı, öznitelik türü ve boyutuna göre sınıflandırma başarıları

Öznitelik Türü	***Boyutu***	***GMM 1***	***GMM 2***	***GMM 4***	***GMM 8***	***GMM 16***
MFCC	4	87.83	89.92	90.89	93.6	93.78
LPCC		70.78	74.15	77.52	75.59	64.06
PLP		88.47	89.03	88.45	90.89	92.40
MFCC	8	88.56	93.08	91.75	94.98	95.94
LPCC		81.27	84.06	90.08	88.06	88.85
PLP		94.42	98.81	95.71	94.78	93.88
MFCC	16	98.2	98.92	98.76	***99.37***	99.22
LPCC		92.84	93.87	96.54	96.66	97.06
PLP		95.98	95.51	96.50	95.19	96.46

Çalışmada en yüksek tanıma oranı 16 elemanlı MFCC özniteliklerinin 8 bileşenli GMM ile modellenmesi sonucunda %99.37 olarak elde edilmiştir. Elde edilen sonuçlara göre incelenen öznitelik çıkarma yöntemlerinden MFCC en başarılı yöntem olurken öznitelik boyutu ve GMM bileşen sayısının düşük olması durumunda PLP yönteminin başarısı MFCC'den daha yüksek olmuştur. Diğer yandan özellikle düşük boyutlu LPCC öznitelik vektörleri ile yapılan testlerde önerilen cinsiyet tanıma sisteminin başarısı %70 seviyelerine kadar düşmektedir. Çalışmada en yüksek cinsiyet tanıma başarısının (%99.37) sağlandığı durumun ayrıntıları Tablo 3.15'te verilmiştir. Elde edilen sonuçlardan da görüldüğü gibi geliştirilen cinsiyet tanıma sistemi ile erkek ve kadın konuşmacıların cinsiyetleri yaklaşık eşit doğrulukta tanınmıştır. Çalışmada ayrıca kullanılan öznitelik türüne göre sisteminin işlem süresi de incelenmiştir. En yüksek başarı oranının elde edildiği MFCC öznitelikleri işlem süresine göre de en başarılı yöntem olurken MFCC ile geliştirilen cinsiyet tanıma sisteminin işlem süresi PLP'ye göre 3.2, LPCC ye göre ise 3.5 kat daha kısa sürmüştür.

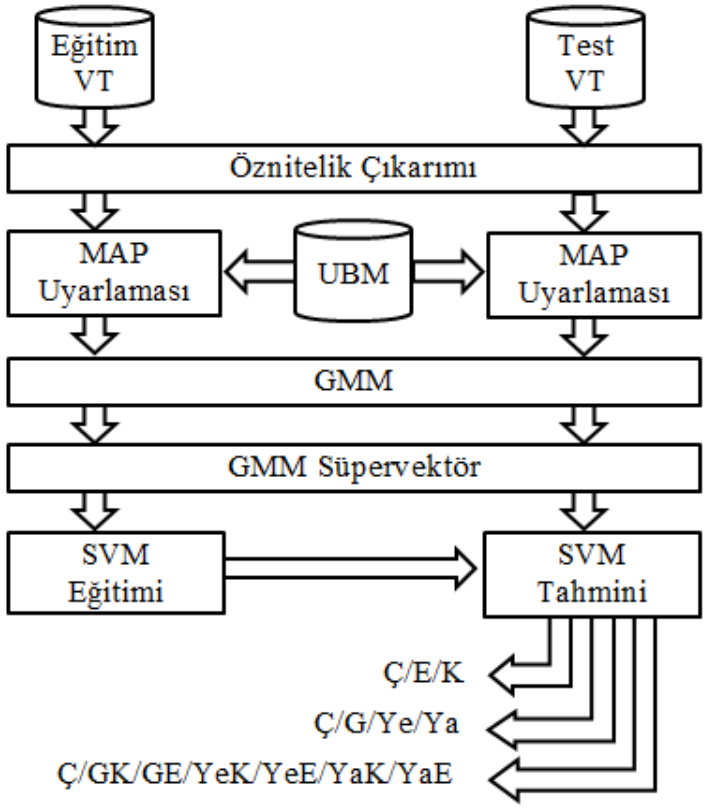
Tablo 3.15. 16 MFCC katsayısından oluşan özniteliklerin 8 bileşenli GMM ile modellenmesi sonucunda elde edilen cinsiyet tanıma başarısı

	Erkek	Kadın
Test edilen örnek sayısı	4010	1570
Doğru karar sayısı	3981	1564
Hatalı karar sayısı	29	6
Başarı oranı	%99.27	%99.61

3.7. GMM-SV SVM ile Yaş ve Cinsiyet Tanıma Çalışması

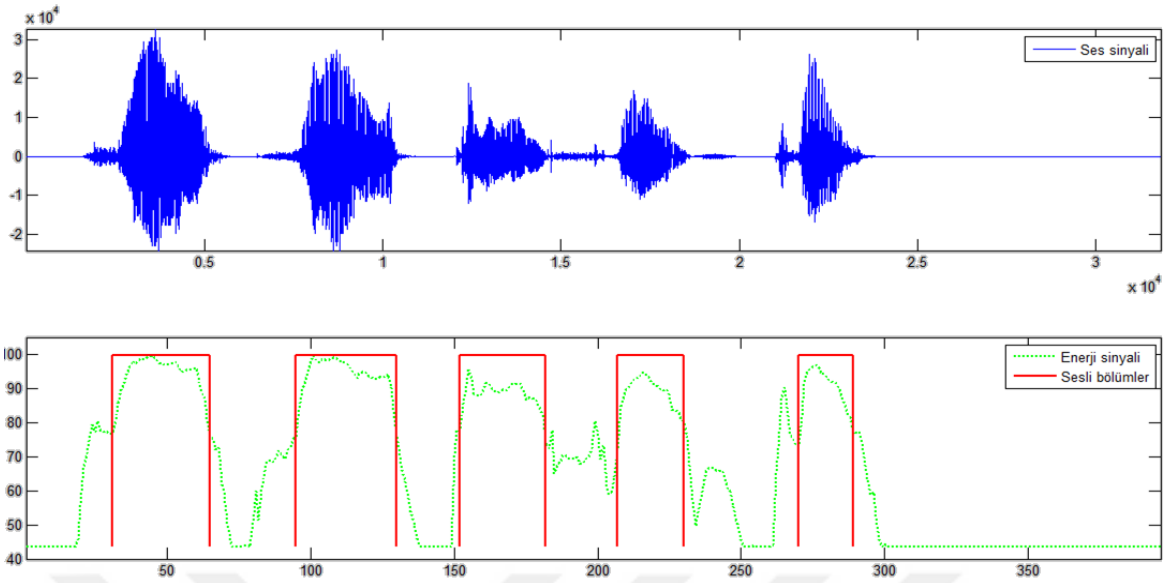
GMM'nin genelleyici gücü ile SVM'nin ayırıcı özelliklerini birleştiren GMM süpervektörlerine dayalı SVM yaklaşımı (GMM-SV SVM) ilk olarak Campbell tarafından konuşmacı doğrulama amacıyla kullanılmıştır. Konuşmacı doğrulamada sağladığı yüksek başarı yönteme olan ilgiyi arttırmış ve birçok alanda kullanılmaya başlamıştır. GMM-SV SVM yaklaşımı bu çalışmada konuşmacıların yaş ve cinsiyet gruplarına göre sınıflandırılması amacıyla kullanılmıştır. Bu amaçla geliştirilen sisteminin blok diyagramı Şekil 3.20'de verilmiştir. Eğitim ve test olmak üzere iki aşamadan oluşan sınıflandırma sisteminin eğitim aşamasında konuşmacıların ses kayıtlarından çıkarılan öznitelik vektörleri kullanılarak her konuşma için bir GMM modeli oluşturulur. Daha sonra bu GMM'ler süpervektörlere dönüştürülür ve bir SVM sınıflandırıcısı bu süpervektörlerle eğitilir. Test aşamasında ise eğitim aşamasındaki gibi GMM ortalama süper vektörlerine dönüştürülen konuşma sinyalleri eğitilen SVM'ye uygulanarak test konuşmacısının yaş ve/veya cinsiyet sınıfına karar verilir.

Önerilen sınıflandırma sistemi giriş olarak verilen konuşmaları konuşmacının yaş ve/veya cinsiyet özelliklerine göre üç kategoride sınıflandırır. Bu kategorilerin ilki olan cinsiyet kategorisinde konuşmacılar erkek, kadın ve çocuk olarak üç sınıfa ayrılırken, ikinci kategori olan yaş kategorisinde ise konuşmacılar çocuk, genç, yetişkin ve yaşlı olarak dört sınıfa ayrılır. Yaş ve cinsiyet kategorilerinin birleşimi olan son kategoride ise konuşmacılar çocuk, genç erkek, genç kadın, yetişkin erkek, yetişkin kadın, yaşlı erkek ve yaşlı kadın olmak üzere yedi sınıfa ayrılmaktadır.



Şekil 3.20. GMM-SV SVM ile yaş ve cinsiyet sınıflandırma sisteminin blok diyagramı

Çalışmada konuşmacıların ses karakteristiklerini temsil etmek için MFCC katsayılarından oluşan bir vektör kullanılmıştır. 13 MFCC katsayısı ile bu katsayıların birinci ve ikinci dereceden türevlerinden (Δ ve ΔΔ) oluşan 39 elemanlı bu vektörler konuşmaların yalnızca ses içeren bölümlerinden çıkarılmıştır. Çalışmada konuşmanın ses içeren bölümlerinin belirlenmesi için sinyalin enerjisine dayalı olarak geliştirilen bir yöntem kullanılmıştır. Bu amaçla 20 ms genişliğindeki bir Hamming penceresi 10 ms kaydırılarak konuşma sinyalleri çerçevelere bölünmüş ve her bir çerçevenin enerjisi hesaplanmıştır. Daha sonra ise belirlenen bir eşik seviyeye göre ilgili çerçevenin konuşma içerip içermediğine karar verilmiştir. Bir çocuğun seslendirdiği konuşma sinyali ve bu sinyal için enerjiye dayalı yöntem ile tespit edilen konuşma içerikli bölümler Şekli 3.21'de gösterilmiştir.



Şekil 3.21. Enerjiye dayalı ses tespiti

Ses içeren bölümleri belirlendikten sonra bu bölümlerden çıkarılan öznitelikler kullanılarak her konuşma için bir GMM modeli eğitilir. Ancak özellikle yüksek bileşenli GMM'lerin eğitiminde konuşmacıların kısa süreli konuşmaları yeterli olmayacaktır. Bu nedenle GMM'lerin doğrudan öznitelik vektörleri ile eğitimi yerine konuşmacıdan bağımsız olarak eğitilen bir arka plan modelinin (UBM) konuşmacı verileri ile uyarlanması yaklaşımı (GMM-UBM) ile GMM'ler eğitilmiştir. Çalışmada yalnızca bileşen ortalamaları uyarlanırken, kovaryans ve ağırlık parametreleri UBM'den doğrudan alınmıştır. Uyarlama yöntemi olarak MAP (Maximum A Posteriori Adaptation), etki faktörü olarak ise 14 değeri seçilmiştir.

Blok diyagramı Şekil 3.20'de verilen yaş ve cinsiyet sınıflandırma sistemi aGender veritabanından seçilen konuşmalarla test edilmiştir. aGender veritbanının eğitim bölümünden rastgele seçilen 35 konuşmacının 2250 konuşması genel arka plan modelinin eğitiminde kullanılırken kalan 436 konuşmacının 30276 konuşması ise SVM sınıflandırıcısının eğitiminde kullanılmıştır. UBM'nin eğitiminde kullanılan verilerin seçiminde her yaş ve cinsiyet grubundan eşit sayıda konuşmacının seçilmesine dikkat edilmiştir. Böylece bu verilerle eğitilen UBM'nin herhangi bir cinsiyet veya yaş grubuna meyilli olması engellenmiştir. UBM'nin eğitiminde kullanılan verilerin sınıf içi dağılımları Tablo 3.16'da verilmiştir.

Tablo 3.16. UBM'nin eğitiminde kullanılan verilerin sınıf içi dağılımı

Yaş-cinsiyet Sınıfı	Konuşmacı Sayısı	Konuşma Sayısı	Toplam Süre (Saniye)
Çocuk	5	231	662
Genç Erkek	5	330	835
Genç Kadın	5	338	820
Yetişkin Erkek	5	314	806
Yetişkin Kadın	5	346	874
Yaşlı Erkek	5	312	838
Yaşlı Kadın	5	379	1038

Genel arka plan modelinin eğitimi tamamlandıktan sora SVM sınıflandırıcısının eğitimine geçilir. Bu amaçla UBM'nin eğitiminde kullanılmayan 436 konuşmacının 30276 konuşmasına karşılık gelen GMM'ler UBM'den uyarlanarak oluşturulur. Daha sonra bu modellerin ortalama bileşenleri uç uca eklenerek sabit uzunluklu süpervektörlere dönüştürülür. Son aşamada ise bu vektörler ve ait oldukları konuşmacının sınıf etiketi kullanılarak SVM'in eğitimi tamamlanır. Çalışmada geliştirilen SVM sınıflandırıcısının eğitiminde çekirdek fonksiyonu olarak RBF fonksiyonu, en uygun RBF parametrelerinin belirlenmesi için ise ızgara tarama yöntemi kullanılmıştır. Ceza parametresi C -5 ile 15 arasında, gamma parametresi ise -15 ile 3 arasında taranmış ve elde edilen sonuçlara göre optimum RBF parametreleri belirlenmiştir.

Eğitilen SVM sınıflandırıcısı aGender veritabanının geliştirme bölümündeki 299 konuşmacının 20548 konuşması ile test edilerek konuşmacıların yaş ve/veya cinsiyet sınıfları tahmin edilmiştir. Çalışmada ayrıca eğitim ve test aşamasında kullanılan konuşmaların süresi ile konuşmaları modellemede kullanılan GMM'lerin bileşen sayısının başarı üzerindeki etkisi de incelenmiştir. Bu amaçla her konuşmacının seslendirdiği konuşmalar bir bütün olarak değerlendirilmiş ve sabit uzunluklu parçalara bölünmüştür. Daha sonra bu parçalardan çıkarılan öznitelikler farklı bileşen sayılı GMM'lerle modellenerek konuşmacının yaş ve cinsiyet sınıfı tahmin edilmiştir. Çalışmada 2, 4, 8, 16 ve 32 saniye olmak üzere 5 farklı konuşma süresi ve 16, 32, 64, 128 ve 256 olmak üzere 5 farklı bileşen sayısı kullanılarak testler yapılmış ve elde edilen sonuçlara göre en uygun konuşma süresi ve GMM bileşen sayısı belirlenmiştir. Eğitim ve test aşamasında kullanılan örnek sayıları seçilen konuşma süresine göre değişmektedir. Çalışmada belirlenen 5

konuşma süresi için eğitim ve test aşmasında kullanılan örnek sayıları Tablo 3.17'de verilmiştir.

Tablo 3.17. Konuşma süresine göre eğitim ve test aşamasında kullanılan örnek sayısı

Süresi	**Eğitimde kullanılan örnek sayısı**	**Testte kullanılan örnek sayısı**
2 s	12504	8392
4 s	6357	4270
8 s	3285	2205
16 s	1742	1173
32 s	982	663

Geliştirilen sistem yaş, cinsiyet ve yaş&cinsiyet olmak üzere üç kategoride test edilmiştir. Bu testlerde sadece SVM'ye uygulanan eğitim ve test vektörlerinin sınıf etiketleri değiştirilmiş diğer işlemlerde herhangi bir değişiklik yapılmıştır. Yani cinsiyet kategorisinde 3 sınıf etiketi ile temsil edilen vektörler yaş kategorisinde 4, yaş&cinsiyet kategorisinde ise 7 sınıf etiketi ile tanımlanmış ve bu vektörler kullanılarak konuşmacıların ilgili kategorideki sınıflandırılması yapılmıştır.

Konuşmacıların erkek, kadın ve çocuk olarak üç sınıfa ayrıldığı cinsiyet kategorisinde yapılan testlerde elde edilen sonuçlar Tablo 1.18'de verilmiştir.

Tablo 3.18. GMM-SV SVM ile cinsiyet sınıflandırma başarısı

Süre	**GMM16**	**GMM32**	**GMM64**	**GMM128**	**GMM256**
2 s	86,96	87,24	87,06	87,89	88,34
4 s	88,83	89,18	90,05	89,91	90,23
8 s	89,30	90,88	91,61	91,29	90,11
16 s	90,62	91,47	***92,42***	91,39	90,28
32 s	90,95	92,06	92,15	90,08	84,16

Bu sonuçlardan konuşma süresi ve GMM bileşen sayısındaki artışın cinsiyet sınıflandırma başarısını arttırdığını ancak belirli bir değerden sonraki süre ve bileşen artışının sonuç üzerinde önemli bir etkisinin olmadığı görülmektedir. Yapılan testlerde en yüksek cinsiyet sınıflandırma başarısı 16 saniyelik konuşmaların 64 bileşenli GMM'lerle modellenmesi sonucunda %92.42 olarak elde edilmiştir. Bu sonucun elde edildiği durum için karışıklık matrisi ise Tablo 3.19'da verilmiştir.

Tablo 3.19. GMM-SV SVM cinsiyet sınıflandırıcısı için karışıklık matrisi (Ç:Çocuk, K:Kadın, E:Erkek)

	Ç	**K**	**E**
Ç	**64,44**	28,85	6,71
K	4,18	**95,64**	0,18
E	0,63	1,89	**97,48**

Karışıklık matrisinden hatalı kararların çoğunun çocuk konuşmacıların kadın olarak sınıflandırılmasından kaynaklandığı, yetişkin konuşmacılar arasındaki hatalı kararların ise oldukça düşük olduğu görülmektedir. Çocuk ile kadın sınıfları arasında oluşan yüksek karışıklık temel frekans özelliklerindeki yakınlıkla açıklanabilir. Normal bir konuşmada çocuk ile kadın konuşmacıların ortalama temel frekans özelliğinde bir çakışma bölgesi oluşurken çocuk ile erkek konuşmacılar aralıklarında böyle bir çakışma bölgesi oluşmaz [170]. Bu durum çalışmada elde edilen sonuçlarla benzerlik göstermektedir.

Konuşmacıların çocuk, genç, yetişkin ve yaşlı olarak 4 grupta sınıflandırıldığı yaş kategorisinde yapılan testlerde elde edilen sonuçlar Tablo 3.20'de verilmiştir.

Tablo 3.20. GMM-SV SVM ile yaş sınıflandırma başarısı

Süre	**GMM16**	**GMM32**	**GMM64**	**GMM128**	**GMM256**
2 s	49,34	50,15	50,21	50,68	50,66
4 s	51,99	53,86	54,38	54,05	53,41
8 s	54,74	57,05	57,10	56,69	49,52
16 s	57,97	58,82	**60,10**	56,78	50,89
32 s	54,30	58,22	58,52	53,24	42,08

Bu sonuçlardan konuşma süresi ve GMM bileşen sayısının cinsiyet modeli üzerindeki etkisinin yaş modelinde de geçerli olduğu görülmektedir. Yapılan testlerde en yüksek sınıflandırma başarısı cinsiyet kategorisinde olduğu gibi 16 saniyelik konuşmaların 64 bileşenli GMM'lerle modellenmesi sonucunda elde edilmiştir. %60.10 olan bu oran için oluşan karışıklık matrisi Tablo 3.21'de verilmiştir.

Tablo 3.21. GMM-SV SVM yaş sınıflandırıcısı için karışıklık matrisi (Ç:Çocuk, G:Genç, Ye:Yetişkin, Ya:Yaşlı)

	Ç	**G**	**Ye**	**Ya**
Ç	**67,79**	16,77	10,07	5,37
G	9,40	**60,90**	18,80	10,90
Ye	1,58	27,44	**45,11**	25,87
Ya	1,36	7,26	23,58	**67,80**

Karışıklık matrisinden çocuk ve yaşlı konuşmacıların genç ve yetişkin konuşmacılara kıyasla daha iyi sınıflandırıldığı, en düşük sınıflandırma başarısının ise yetişkin yaş grubunda olduğu görülmektedir. Diğer iki yaş grubuna kıyasla çocuk ve yaşlı konuşmacı gurubunda sağlanan yüksek başarın insan sesindeki değişimin ergenlik ve yaşlılık döneminde daha fazla olmasıyla ilişkili olabileceği değerlendirilmiştir.

Konuşmacıların yaş ve cinsiyet özelliklerine 7 sınıfa ayrıldığı yaş&cinsiyet kategorisinde elde edilen sonuçlar Tablo 3.22'de verilmiştir.

Tablo 3.22. GMM-SV SVM ile yaş&cinsiyet sınıflandırma başarısı

Süre	**GMM16**	**GMM32**	**GMM64**	**GMM128**	**GMM256**
2 s	47,35	48,18	48,98	50,42	49,91
4 s	51,22	52,69	53,54	53,70	53,13
8 s	52,70	54,47	56,96	55,92	49,47
16 s	55,84	57,97	***60,02***	56,44	49,10
32 s	55,80	57,62	59,88	53,99	39,81

Yaş&cinsiyet kategorisinde de en yüksek başarı 16 saniyelik konuşmaların 64 bileşenli GMM'lerle modellenmesi sonucunda elde edilmiştir. Bu durum için oluşan karışıklık matrisi Tablo 3.23'te verilmiştir.

Tablo 3.23. GMM-SV SVM yaş&cinsiyet sınıflandırıcısı için karışıklık matrisi (Ç:Çocuk, GK:Genç kadın, GE:Genç erkek, YeK:Yetişkin kadın, YeE:Yetişkin erkek, YaK:Yaşlı kadın, YaE:Yaşlı erkek)

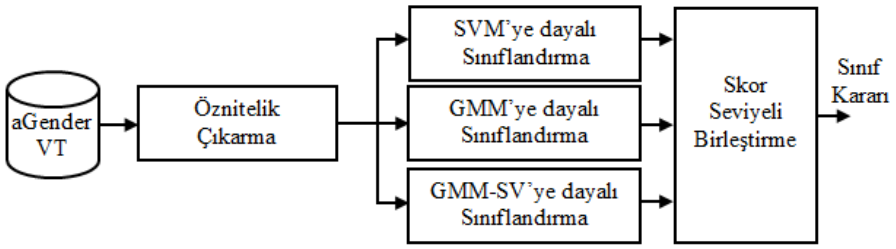
	Ç	**GK**	**GE**	**YeK**	**YeE**	**YaK**	**YaE**
Ç	**69,2**	11,4	4,7	6	2	6,7	0
GK	13,5	**65,5**	0	12,5	0	7,9	0,6
GE	1,7	0	**51,5**	1,7	28,7	4,3	12,1
YeK	4,4	24,1	0	**44,6**	0	26,9	0
YeE	0	0	25,2	0	**51,1**	0,7	23
YaK	2,2	6,5	0,4	25	0,5	**65,4**	0
YaE	1,8	0	7,1	0	21,4	1,8	**67,9**

Elde edilen sonuçlar cinsiyet ve yaş kategorisinde elde edilen sonuçlarla benzerlik göstermektedir. Cinsiyet kategorisinde çocuk ve kadın sınıfları arasında oluşan yüksek karışıklık yaş&cinsiyet kategorisinde kadın yaş gruplarına dağılmıştır. Benzer şekilde yetişkin yaş grubunda görülen düşük başarının yaş&cinsiyet kategorisine etkisi yetişkin erkek ve yetişkin kadın sınıflarında görülmüştür. Ayrıca erkek ve kadın sınıfları arasındaki en yüksek karışıklığın genç erkek ile yaşlı kadın sınıfı arasında olması insan algılamasıyla benzerlik göstermektedir.

3.8. Skor Seviyeli Birleştirme Yaklaşımı ile Yaş ve Cinsiyet Tanıma Çalışması

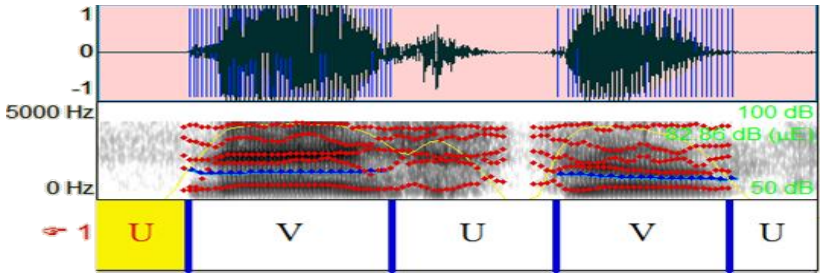
Bu çalışmada kısa süreli telefon konuşmalarını konuşmacının yaş ve/veya cinsiyetine göre sınıflandıran bir sistem geliştirilmiştir. Blok diyagramı Şekil 3.22'de verilen sistem üç alt sistemin skor seviyesinde birleşiminden oluşmaktadır. Geliştirilen sistemin ilk aşamasında konuşmaların ses içeren bölümleri perde frekansına dayalı olarak belirlenir ve yalnızca bu bölümler kullanılarak öznitelik vektörleri çıkarılır. Çalışmada MFCC, PLP ve prosodik özniteliklerden oluşan üç fraklı öznitelik vektörü kullanılmıştır. Daha sonra bu

öznitelik vektörleri kullanılarak üç farklı sınıflandırma sistemi oluşturulmuştur. Bu alt sistemlerin her birisinde farklı sınıflandırma yaklaşımları ve özniteliklerin faklı kombinasyonları kullanılmıştır. Son aşamada ise üç alt sistemin sonuçları skor seviyesinde birleştirilerek giriş konuşmaları konuşmacının yaş ve cinsiyet grubuna göre sınıflandırılmıştır.



Şekil 3.22. Skor seviyeli birleştirme yaklaşımına dayalı sistemin blok diyagramı

Çalışmada konuşmaların sesli/sessiz bölümlerinin belirlenmesi için Praat [171] programında yazılan bir script kullanılmıştır. Bu script ile konuşmanın perde frekansları otokorelasyona dayalı olarak tahmin edilmiş ve belirlenen bir eşik seviyeye göre konuşmanın sesli ve sesiz bölümlerine karar verilmiştir. Bu programla yapılan bir analizin sonucu Şekil 3.23'te gösterilmiştir.



Şekil 3.23. Konuşmanın sesli/sessiz bölümlerinin tespiti

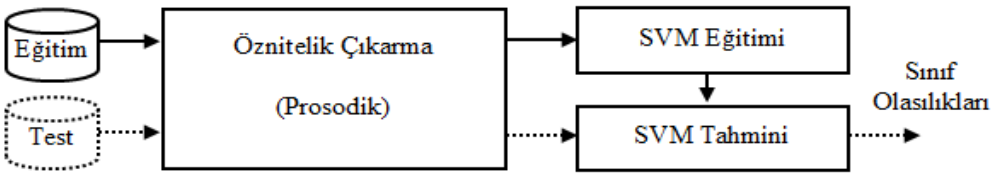
Üç farklı alt sistemin birleşiminden oluşan bu sistemin geliştirilmesinde aGender veritabanındaki ses kayıtları kullanılmıştır. Sistemin eğitiminde veritabanının eğitim bölümdeki 471 konuşmacının 32526 konuşmasının tamamı kullanılırken test aşamasında ise veritabanının test bölümündeki 299 konuşmacının 20548 konuşmasından uzunluğu 0.5 saniyeden büyük olan 10715 tanesi kullanılmıştır. Geliştirilen sistemin eğitim ve test aşamasında kullanılan veri kümesi ile ilgili detaylar Tablo 3.24'te verilmiştir.

Tablo 3.24. Eğitim ve test aşamasında kullanılan veri kümesi

Eğitim aşamasında kullanılan konuşmacı sayısı	471
Eğitim aşamasında kullanılan örnek sayısı	32526
Eğitim aşamasında kullanılan en kısa konuşmanın uzunluğu	0.1 s
Eğitim aşamasında kullanılan en uzun konuşmanın uzunluğu	5 s
Eğitim aşamasında kullanılan konuşmaların ortalama uzunluğu	0.75 s
Test aşamasında kullanılan konuşmacı sayısı	299
Test aşamasında kullanılan örnek sayısı	10715
Test aşamasında kullanılan en kısa konuşmanın uzunluğu	0.5 s
Test aşamasında kullanılan en uzun konuşmanın uzunluğu	5 s
Test aşamasında kullanılan konuşmaların ortalama uzunluğu	1 s

3.8.1. SVM'ye Dayalı Yaş ve Cinsiyet Sınıflandırma Sistemi

Çalışmada önerilen birleşik sistemi oluşturan üç alt sistemden birincisi olan SVM'ye dayalı sınıflandırma sistemi beş farklı akustik parametreden türetilen 42 prosodik öznitelikle temsil edilen konuşmaların bir SVM sınıflandırıcısı ile sınıflandırılması fikrine dayanır. Blok diyagramı Şekil 3.24'te verilen sistemin eğitim aşamasında eğitim için belirlenen konuşmalardan çıkarılan prosodik öznitelik vektörleri kullanılarak bir SVM sınıflandırıcısı eğitilir. Test aşamasında ise test konuşmalarından çıkarılan prosodik öznitelikler eğitilen SVM sınıflandırıcısına uygulanır ve öznitelik vektörlerinin SVM uzayındaki konumuna göre ilgili konuşmacının yaş ve/veya cinsiyet grubuna karar verilir. Çalışmada oluşturulan SVM sınıflandırıcısının eğitiminde RBF çekirdek fonksiyonu kullanılırken çoklu sınıflandırma için ise bire karşı bir yaklaşımı kullanılmıştır.



Şekil 3.24. SVM'ye dayalı sınıflandırma sisteminin blok diyagramı

Çalışmada kullanılan prosodik özniteliklerin listesi Tablo 3.25'te verilmiştir. Bu özniteliklerin konuşma sinyallerinden çıkarılmasında Praat ses analiz programında yazılan bir script kullanılmıştır. Bu script ile tüm konuşma sinyalleri analiz edilmiş ve her konuşmaya karşılık gelen 42 prosodik öznitelik 5 akustik parametreden türetilmiştir. Bu parametrelerden perde frekansının tahmininde orokorelasyon yöntemi, formant frekansının tahmininde ise Burg algoritması kullanılmıştır.

Tablo 3.25. Çalışmada kullanılan prosodik öznitelikler

Öznitelik grubu	**İstatistikler**	**Öznitelik sayısı**
Perde frekansı	Ortalama, ortanca, standart sapma, maksimum, minimum, Jitter (local), Jitter (local, absolute), Jitter(rap), Jitter(ddp) Shimmer(local), Shimmer (local, dB)	11
Formant frekansı	Birinci, ikinci ve üçüncü formantın ortalaması (F1, F2, F3) F1, F2 ve F3'ün standart sapması ve bant genişliği	9
Yoğunluk	Ortalama, maksimum, minimum ve standart sapma	4
Harmoniklik	Ortalama otokorelasyon, Ortalama harmonik gürültü oranı ve Ortalama gürültü harmonik oranı	3
Uzun süreli ortalama spektrum	Ortalama, maksimum, minimum, eğim, basıklık (kurtosis) ve çarpıklık (skewness), birinci ve ikinci harmoniklerin genliğin genliği (H1, H2), birinci, ikinci ve üçüncü formantın genliği (A1, A2, A3), H1-H2, H1-A1, H1-A2, H1-A3	15

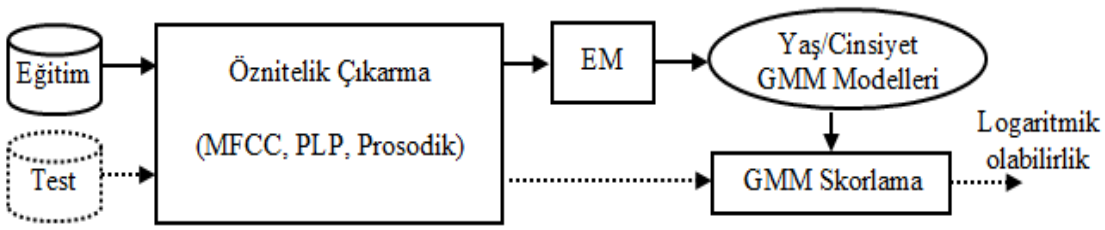
Çalışmada geliştirilen SVM'ye dayalı sınıflandırma sistemi aGender veritabanın eğitim bölümündeki 471 konuşmacının 32526 konuşması ile eğitilmiş ve test bölümündeki 10715 konuşma ile test edilmiştir. Yaş, cinsiyet ve yaş&cinsiyet olmak üzere üç kategorilerinde yapılan testlerde elde edilen sonuçlar Tablo 3.26'da verilmiştir. Elde edilen sonuçlardan prosodik özniteliklerle geliştirilen bu sistemin konuşmacının yaş ve cinsiyetini belirlemede oldukça başarılı olduğu görülmektedir.

Tablo 3.26. Prosodik özniteliklerle geliştirilen SVM'ye dayalı sistemin sınıflandırma başarısı

Öznitelik vektörünün türü	Prosodik
Öznitelik vektörünün boyutu	42
Üç sınıflı cinsiyet sınıflandırma başarısı (%)	87.31
Dört sınıflı yaş sınıflandırma başarısı (%)	47.79
Yedi sınıflı yaş&cinsiyet sınıflandırma başarısı (%)	47.24

3.8.2. GMM'ye Dayalı Yaş ve Cinsiyet Sınıflandırma Sistemi

Çalışmada önerilen birleşik sistemi oluşturan üç alt sistemden ikincisi olan GMM'ye dayalı sınıflandırma sisteminde üç farklı öznitelik vektörü (MFCC, PLP ve prosodik) ile oluşturulan Gauss karışım modelleri kullanılarak konuşmacının yaş ve/veya cinsiyet sınıfı tahmin edilmiştir. Blok diyagramı Şekil 3.25'te verilen sistemin eğitim aşamasında her yaş ve/veya cinsiyet sınıf için bir GMM modeli eğitilir. 256 bileşenden oluşan bu modellerin eğitiminde aGender veritabanının eğitim bölümünden seçilen 20 konuşmacının yaklaşık birer dakikalık konuşmaları kullanılmıştır. Her sınıfa ait 20 konuşmacı eğitim veritabanından rastgele seçilmiş ve bu konuşmacıların ses kayıtlarından çıkarılan öznitelik vektörleri ile ilgili sınıfı temsil eden birer GMM modeli eğitilmiştir. Böylece eğitim aşaması sonunda erkek, kadın ve çocuk olmak üzere 3 cinsiyet modeli, çocuk, genç, yetişkin ve yaşlı olmak üzere 4 yaş modeli ve bu modellerin birleşiminden oluşan 7 yaş&cinsiyet modeli oluşturulmuştur. Çalışmada yaş ve cinsiyet modeli olarak kullanılan GMM'lerin eğitiminde EM algoritması kullanılırken tekrar sayısı olarak ise 20 seçilmiştir.



Şekil 3.25. GMM'ye dayalı sınıflandırma sisteminin blok diyagramı

Yaş ve cinsiyet modellerin eğitimi tamamlandıktan sonra sistem aGender veritabanının test bölümündeki 299 konuşmacının 10715 konuşması ile test edilmiştir. Ortalama uzunluğu 1 saniye olan konuşmalar ile yapılan bu testlerde konuşmadan çıkarılan öznitelik vektörleri eğitim aşamasında oluşturulan GMM modelleri ile karşılaştırılmış ve

$$LL_{model}(x) = \frac{1}{N}\sum_{i=1}^{N} log(P_{model}(x_i|w, \mu, \sum) \quad (3.6)$$

ifadesine göre hesaplanan logaritmik olabilirlik skoruna göre konuşmacının yaş ve/veya cinsiyet sınıfına karar verilmiştir. Çalışmada geliştirilen GMM'ye dayalı sistem yaş, cinsiyet ve yaş&cinsiyet olmak üzere üç kategoride test edilmiş ve elde edilen sonuçlar Tablo 3.27'de verilmiştir.

Tablo 3.27. GMM'ye dayalı sistemin sınıflandırma başarısı

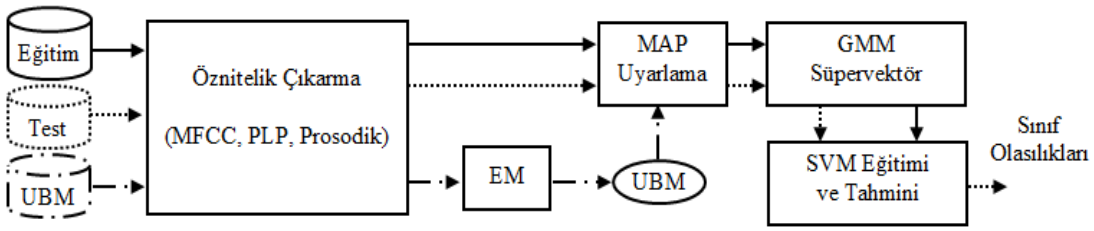
Kullanılan öznitelik türü ve boyutu	Cinsiyet (3 sınıflı)	Yaş (4 sınıflı)	Yaş&cinsiyet (7 sınıflı)
13 MFCC +Δ+ Δ Δ	84.76	44.43	43.27
15 PLP +Δ+ Δ Δ	84.58	45.52	44.41
42 prosodik öznitelik	80.42	40.67	40.20

Elde edilen sonuçlardan MFCC ve PLP öznitelikleri ile geliştirilen GMM sisteminin her üç kategoride de prosodik öznitelikleriyle geliştirilen sistemden daha başarılı olduğu görülmektedir. Ayrıca daha önceki çalışmada elde edilen sonuçlar göz önünde bulundurulduğunda GMM'ye dayalı yaş ve cinsiyet tanıma sisteminin SVM'ye dayalı sisteme kıyasla daha başarısız olduğu görülmektedir.

3.8.3. GMM-SV SVM ile Yaş ve Cinsiyet Sınıflandırma Sistemi

Çalışmada önerilen birleşik sistemi oluşturan üç alt sistemden sonuncusu olan GMM-SV SVM yaklaşımında üç farklı öznitelik vektörü (MFCC, PLP ve prosodik) ile oluşturulan GMM süpervektörleri birer SVM sınıflandırıcısına uygulanmış ve elde edilen olasılıksal sonuçlara göre konuşmacının yaş ve/veya cinsiyet grubuna karar verilmiştir. Blok diyagramı Şekil 3.26'da verilen sistemin ilk aşamasında genel arka plan modeli

(UBM) olarak isimlendirilen bir GMM modeli konuşma sinyallerinden çıkarılan öznitelik vektörleri ile oluşturulmuştur. Bu modelin eğitiminde her yaş ve cinsiyet grubundan yaklaşık eşit miktarda konuşma verisi kullanılmıştır. Böylece bu modelin herhangi bir yaş veya cinsiyet grubuna meyilli olması önlenmiş olur. 32 bileşenli bir GMM modeli olan UBM'nin eğitiminde aGender veritabanın eğitim bölümünden rastgele seçilen 35 konuşmacının (her sınıftan beşer konuşmacı) yaklaşık 30 dakikalık konuşmaları kullanılmıştır. UBM parametrelerinin tahmininde EM algoritması kullanılırken tekrar sayısı olarak ise 20 seçilmiştir.



Şekil 3.26. GMM–SV SVM sisteminin blok diyagramı

UBM'nin eğitimi tamamlandıktan sonra eğitim veri kümesindeki her konuşmaya karşılık gelen GMM modeli UBM'den uyarlanarak oluşturulur. Çalışmada uyarlama yaklaşımı olarak MAP yöntemi, uyarlama oranını belirleyen etki faktörü olarak ise 14 değeri seçilmiştir. Daha sonra GMM'lerin ortalama bileşenleri uç uca eklenir ve oluşturulan GMM süpervektörleri kullanılarak bir SVM sınıflandırıcısı eğitilir. SVM sınıflandırıcısının eğitiminde çekirdek fonksiyonu olarak RBF fonksiyonu, en iyi SVM parametrelerinin (ceza ve gamma parametreleri) belirlenmesi için ise ızgara tarama yöntemi kullanılmıştır. SVM'nin eğitiminden sonra test aşmasına geçilir. Önce test konuşmasına karşılık gelen GMM süpervektörleri UBM'den uyarlanarak oluşturulur. Daha sonra bu vektörler eğitilen SVM sınıflandırıcısına uygulanır ve test vektörlerinin SVM uzayındaki konumuna göre ilgili konuşmacının yaş ve/veya cinsiyet sınıfına karar verilir.

Çalışmada önerilen GMM-SV SVM sistemi aGender veritabanın test bölümündeki 299 konuşmacının 10715 konuşma ile test edilmiştir. Yaş, cinsiyet ve yaş&cinsiyet olmak üzere üç kategorilerinde yapılan testlerde elde edilen sonuçlar Tablo 3.28'de verilmiştir.

Tablo 3.28. GMM-SV SVM sisteminin sınıflandırma başarısı

Kullanılan öznitelik türü ve boyutu	Cinsiyet (3 sınıflı)	Yaş (4 sınıflı)	Yaş&cinsiyet (7 sınıflı)
13 MFCC +Δ+ΔΔ	87.05	49.72	48.31
15 PLP+Δ+ ΔΔ	87.88	49.72	48.37
42 prosodik öznitelik	87.36	48.30	47.75

Elde edilen sonuçlardan üç farklı öznitelik vektörü ile geliştirilen GMM-SV SVM sistemlerinin yaş ve/veya cinsiyet sınıflandırma başarısının yaklaşık aynı olduğu görülmektedir.

3.8.4. Skor Seviyeli Birleştirilme Yaklaşımı

Çalışmada önerilen sistemin son aşamasında üç alt sistemin sonuçları skor seviyesinde birleştirilir. Bu sistemlerin ikisinde sınıflandırıcı olarak SVM, birinde ise GMM yöntemi kullanılmıştır. GMM'nin sonucu logaritmik olabilirlik skoru, SVM'nin ise tahmin edilen sınıfın etiketidir. Bu durumda bu sonuçlar doğrudan birleştirilemez. Bu sonuçların birleştirilebilmesi için ortak bir ölçeğe çevrilmesi gerekir. Çalışmada her bir sistemin sonuçları önce olasılıksal skorlara dönüştürülmüş daha sonra birleştirilmiştir.

GMM'ye dayalı sistemin sonuçlarının olasılıksal skorlara dönüştürülmesi için puanlamaya dayalı bir yöntem kullanılmıştır. Bu yöntemde ilgili yaş-cinsiyet modelleri ile yapılan karşılaştırma sonucunda hesaplanan logaritmik olabilirilik skorları büyükten küçüğe doğru sıralanmış ve bu sıralamaya göre her sınıfa bir puan verilmiştir. Çalışmada en büyük olabilirlik skoruna sahip sınıfa 10, ikinciye 8 ve diğerlerine sırasıyla 5, 4, 3, 2, 1 puan verilmiştir. Puanlama işlemi birleştirilecek alt sistemlerin hepsinde yapılır. Daha sonra her sınıfın toplam puanı hesaplanır ve tüm puanların toplamına bölünerek olasılıksal skorlar elde edilir. Son aşamada ise hesaplanan olasılıksal skorlara göre test konuşmasının sınıfına karar verilir. Üç farklı sistem tarafından verilen puanların olasılıksal skorlara dönüştürüldüğü örnek bir hesaplama Tablo 3.29'da verilmiştir.

Tablo 3.29. Puanlamaya dayalı yöntemle olasılık skorlarının hesaplanması

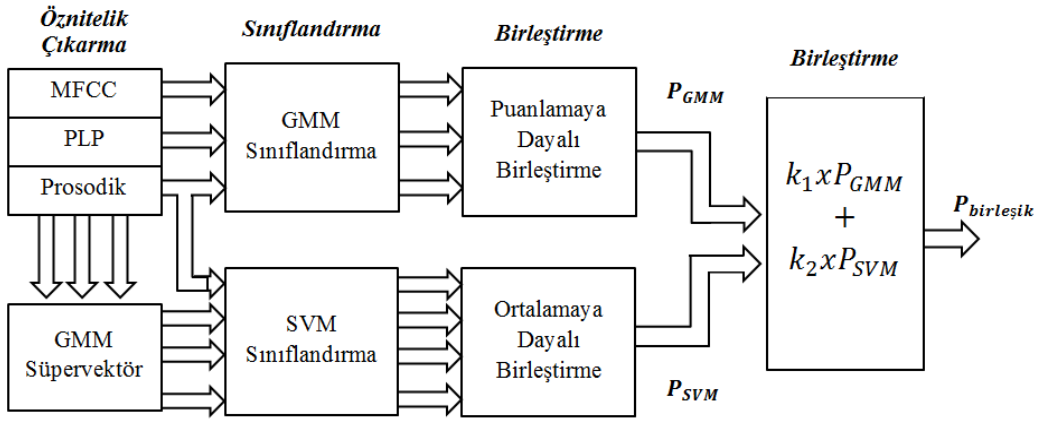
	S1	S2	S3	S4	S5	S6	S7	Toplam
Sistem1'e göre puanlama	8	10	5	1	2	4	3	33
Sistem2'ye göre puanlama	5	2	10	10	1	3	4	33
Sistem3'e göre puanlama	1	3	10	5	2	4	8	33
Toplam puan	14	15	25	16	5	11	15	99
Olasılıksal skor	0.14	0.15	***0.25***	0.16	0.05	0.11	0.15	1

SVM sonuçlarının olasılıksal skorlara dönüştürülmesi için farklı yöntemler geliştirilmiştir. Bu yöntemlerden Platt [172] tarafından geliştirilen ölçekleme yaklaşımı en yaygın kullanılan yöntem olup bu çalışmada da kullanılmıştır. Yaygın olarak kullanılan açık kaynak makine öğrenme kütüphaneleri tarafından da desteklenen Platt'ın ölçekleme yaklaşımı libSVM kütüphanesinde bir parametrenin (-b) setlenmesiyle uygulanmaktadır. Bu çalışmada da libSVM kütüphanesi kullanılmış ve ilgili parametre setlenerek SVM'lerin olasılıksal skorlar üretmesi sağlanmıştır. Son aşamada ise farklı SVM'ler tarafından üretilen olasılıksal skorlarının ortalaması alınmış ve en yüksek olasılığa göre test konuşmasının sınıfına karar verilmiştir.

Çalışmada dördü SVM'ye üçü ise GMM'ye dayalı olmak üzere toplam yedi farklı sınıflandırma sistemi geliştirilmiştir. Bu sistemler önce kendi içinde birleştirilerek P_{SVM} ve P_{GMM} ile gösterilen iki farklı olasılıksal skor elde edilir. Daha sonra bu iki olasılıksal skor

$$P_{birleşik} = k_1 x P_{GMM} + k_2 x P_{SVM} \tag{3.7}$$

ifadesine göre birleştirilerek tek bir skora dönüştürülür. Burada k_1 ve k_2 katsayıları sırasıyla GMM ve SVM sistemlerinin sonuç üzerindeki ağırlıklarını belirleyen parametreler olup değeri 0 ile 1 arasında değişen gerçel sayılardır. Toplamı 1 olan bu sayıların belirlenmesi için sistem 0.1 adım arakları ile test edilmiş ve elde edilen başarı oranına göre en uygun ağırlık katsayıları $k_1 = 0.3$, $k_2 = 0.7$ olarak belirlenmiştir. Çalışmada kullanılan skor seviyeli birleştirme yaklaşımının blok diyagramı Şekil 3.27'de verilmiştir.



Şekil 3.27. Önerilen skor seviyeli birleştirme yaklaşımının blok diyagramı

Çalışmada önerilen birleşik sistem aGender veritabanının geliştirme bölümündeki 299 konuşmacının 20549 konuşmasıyla test edilmiştir. Yaş, cinsiyet ve yaş&cinsiyet olmak üzere üç kategoride yapılan testlerde elde edilen sonuçlar Tablo 3.30'da verilmiştir.

Tablo 3.30. Önerilen birleşik sistemin üç kategorideki sınıflandırma başarıları

ID	Sınıflandırma Yaklaşımı	Öznitelik Türü	Yaş&Cinsiyet	Yaş	Cinsiyet
1		MFCC	48.31	49.72	87.05
2	GMM-SV SVM	PLP	48.37	49.72	87.88
3		Prosodik	47.75	48.30	87.36
4	***Skor seviyeli birleştirme 1,2,3***		***52.77***	***53.64***	***90.31***
5		MFCC	43.27	44.43	84.76
6	GMM	PLP	44.41	45.52	84.58
7		Prosodik	40.20	40.67	80.42
8	***Skor seviyeli birleştirme 5,6,7***		***44.56***	***45.25***	***85.21***
9	SVM	Prosodik	47.24	47.79	87.31
10	***Skor seviyeli birleştirme 4,9***		***53.33***	***54.02***	***90.32***
11	***Skor seviyeli birleştirme 8,10***		***53.50***	***54.10***	***90.39***

Elde edilen sonuçlardan GMM-SV SVM yaklaşımının her üç kategoride de en başarılı yöntem olduğu görülmektedir. Diğer iki yöntemden prosodik özniteliklerle oluşturulan SVM yaklaşımı ile bu yönteme yakın başarı elde edilirken başarısı en düşük yöntem ise GMM'ye dayalı sınıflandırma yaklaşımı olmuştur. SVM yaklaşımı ile geliştirilen dört sistem kendi arasında birleştirilmiş ve cinsiyet kategorisinde %2.44, yaş kategorisinde %4.3 ve yaş&cinsiyet kategorisinde ise %4.96 başarı artışı sağlanmıştır. GMM'ye dayalı sistemlerin birleştirilmesi durumunda ise çok az bir başarı artışı sağlanmıştır. Çalışmada en yüksek başarı oranı tüm alt sistemlerin birleştirilmesi sonucunda elde edilmiştir. Tablo 3.30'da 11 numarasıyla gösterilen bu sistem bireysel başarısı en yüksek olan 2 numaralı sistemle göre cinsiyet kategorisinde %2.51, yaş kategorisinde %4.38, yaş&cinsiyet kategorisinde ise %5.13 başarı artışı sağlamıştır.

3.8. Kanal Dengeleme

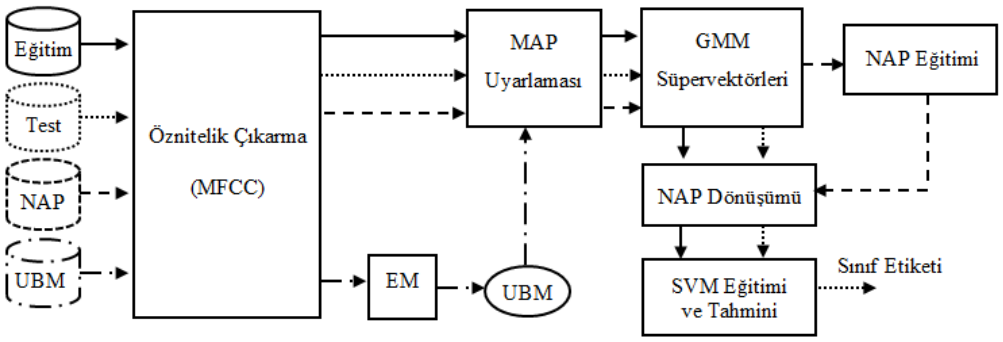
Solomonoff ve arkadaşları [157] tarafından geliştirilen NAP (Nuisance Attribute Projection) yöntemi kanal etkilerinin giderilmesi için SVM sınıflandırıcısı ile birlikte kullanılan oldukça popüler bir yöntemdir. Bu yöntem belirli bir çekirdeğe özgü değildir ve her tür SVM süpervektörüne uygulanabilir. NAP yöntemiyle SVM eğitimi öncesinde oturum değişkenliğinden kaynaklanan istenmeyen yönler süpervektörlerden çıkarılır. Bir süpervektör için NAP dönüşümü

$$s' = s - U(U^T s) \tag{3.8}$$

şeklinde verilir. Burada U öz kanal matrisi olup bu matris her biri çeşitli eğitim konuşmalarına (oturumlar) sahip çok sayıda konuşmacıdan oluşan bir geliştirme veri kümesi kullanılarak eğitilir. Her konuşmacının süpervektörlerinden o konuşmacının tüm süpervektörlerinin ortalaması çıkarılır ve elde edilen vektörler birleştirilerek öz kanal matrisi oluşturulur. Sadece kanal değişimlerini temsil eder bu matris konuşmacı bilgisi içermez ve bu matristen elde edile öz vektörlerden değişimin en yüksek olduğu K tanesi ile kanal değişim alt uzayı temsil edilir. Son olarak bu alt uzay konuşmacı verilerinden çıkarılarak konuşmacı süpervektörlerindeki kanal etkileri azaltılmış olur.

Bu çalışmada GMM-SV SVM yaklaşımı ile geliştirilen yaş ve/veya cinsiyet tanıma sistemine NAP yöntemi uygulanarak bu yöntemin yaş ve/veya cinsiyet tanıma üzerindeki

etkisi incelenmiştir. Blok diyagramı Şekil 3.28'de verilen sistemin geliştirilmesinde aGender veritabanından seçilen telefon konuşmaları kullanılmıştır.

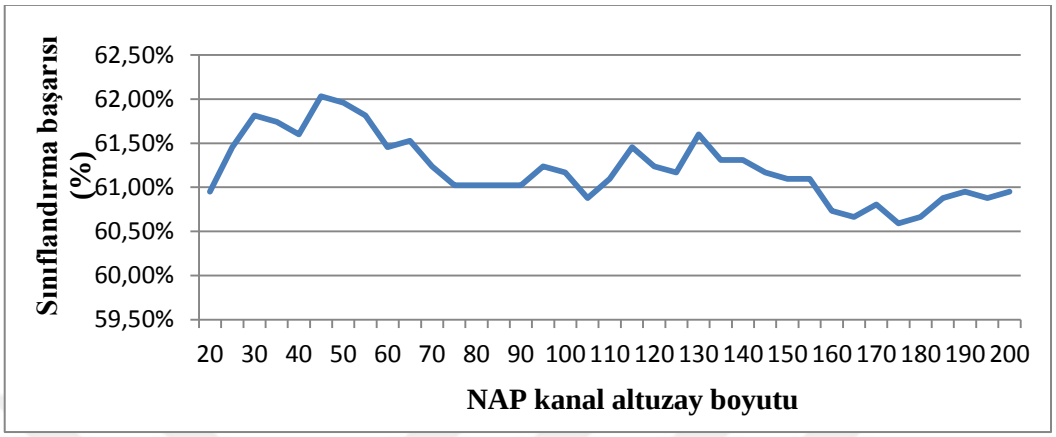


Şekil 3.28. Kanal dengelemeye dayalı sınıflandırma sisteminin blok diyagramı

aGender veritabanının eğitim bölümünden rastgele seçilen 70 konuşmacı genel arka plan modelinin eğitiminde, diğer 70 konuşmacı ise NAP'ın öz kanal matrisinin eğitiminde kullanılmıştır. Bu konuşmacıların seçiminde her yaş ve cinsiyet sınıfından eşit sayıda (5'er tane) konuşmacının seçilmesine dikkat edilmiştir. UBM ve NAP'ın eğitimi için seçilen 140 konuşmacının dışında kalan 331 konuşmacı ise SVM'nin eğitiminde kullanılmıştır. Çalışmada öznitelik olarak 13 MFCC katsayılarının birinci ve ikinci türevlerinden oluşan 39 elemanlı bir öznitelik vektörü kullanılmıştır. GMM bileşen sayısı olarak 128, UBM'nin eğitimi için ise EM algoritması kullanılmıştır. Çalışmada konuşmacıların her oturumda seslendirdiği konuşmalar bir bütün olarak kabul edilmiş ve bu konuşmalara karşılık gelen GMM'ler UBM'den uyarlanarak eğitilmiştir. Daha sonra bu modeller süpervektörlere dönüştürülmüş ve UBM parametreleri ile $\sqrt{w_k}\,\Sigma_k^{-1/2}$ ifadesine göre normalize edilmiştir. Son aşamada ise NAP yöntemi ile belirlenen kanal alt uzayı bu vektörlerden çıkarılmış ve elde edilen vektörler doğrusal çekirdekli bir SVM sınıflandırıcısına uygulanarak konuşmacıların yaş ve cinsiyet grubu tahmin edilmiştir.

Geliştirilen sistemin test aşamasında aGender veritabanının test bölümündeki 299 konuşmacı tarafından 1388 oturumda seslendirilen 20548 konuşma verisi kullanılmıştır. Test aşamasında da her oturumdaki konuşmalar bir bütün olarak kabul edilmiş ve oluşturulan 1388 cümle ile sistem test edilmiştir. Ortalama 10 saniye konuşma içeren cümlelerle yapılan bu testlerde değişik NAP kanal alt uzay boyut için sistemin başarısı ölçülmüş ve elde edilen başarıya göre geliştirilen sistem için en uygun kanal alt uzay

boyutu belirlenmiştir. 20 ile 200 arasındaki kanal alt uzay boyutu için elde edilen 7 sınıflı sınıflandırma başarımının değişimi Şekil 3.29'da gösterilmiştir.



Şekil 3.29. NAP kanal alt uzay boyutunun yaş ve cinsiyet sınıflandırma başarısına etkisi

Şekil 3.29'dan da görüldüğü gibi kanal alt uzay boyutunun 45 seçilmesi durumunda en yüksek yaş ve cinsiyet sınıflandırma başarısı elde edilmiştir. Çalışmada uygulanan kanal dengeleme yaklaşımın başarı üzerindeki etkisinin belirlemesi için sistem aynı şartlarda NAP'lı ve NAP'sız olarak test edilmiştir. Kanal dengeleme yapılmadan gerçekleştirilen test sonucunda 1388 örneğin 840 tanesinde konuşmacının yaş ve cinsiyet grubu doğru tahmin edilerek %60.51 sınıflandırma başarısı elde edilirken NAP yönteminin kullanılmasıyla doğru sınıflandırılan örnek sayısı 861'e başarı oranı ise %62.03'e çıkmıştır. Bu sonuçlardan NAP yöntemi ile gerçekleştirilen kanal dengeleme yaklaşımının GMM süpervektörlerine dayalı olarak geliştirilen yaş ve cinsiyet sınıflandırma sisteminin başarısını arttırdığı görülmektedir.

4. SONUÇLAR

Bu tez çalışmasında konuşmacıları yaş ve cinsiyet grubuna göre otomatik olarak sınıflandıran bir sistemin geliştirilesi amaçlanmıştır. Konuşmacının yaş ve cinsiyet sınıfının otomatik olarak belirlenmesi başta ticari, medikal ve adli olmak üzere geniş bir uygulama alanına sahiptir. Otomatik olarak tespit edilen yaş ve cinsiyet bilgisi doğrudan bir hizmetin (reklam, müzik, ürün gibi) seçiminde kullanılabileceği gibi konuşmacı ve konuşma tanıma sistemleri başta olmak üzere konuşmaya dayalı birçok sistemde ön bilgi olarak da kullanılır. Özellikle cinsiyet bilgisi konuşmacı tanıma sistemlerinde araştırma uzayını aynı cinsiyetli konuşmacılarla sınırlandırarak; konuşma tanıma sistemlerinde ise cinsiyet bağımlı modellerin tanımlanmasına imkan sunarak performans artışı sağlamaktadır.

Özellikle sağladığı kolay kullanım, düşük maliyet, yüksek güvenirlik ve uzaktan kullanılabilme imkanı konuşmaya dayalı sistemlere olan ilgiyi artmıştır. Ancak bu tür sistemlerin geliştirilmesinde karşılaşılan bazı zorluklar da vardır. Bu zorlukların başında konuşma sinyalinin oldukça değişken bir içeriğe sahip olması gelir. Bu değişkenliğin kaynağı kullanılan mikrofon, iletişim kanalı, ortam gürültüsü gibi dış faktörler olabileceği gibi seslendirilen metnin içeriği, kişinin hasta veya yorgun olması, ruhsal durumundaki değişiklikler gibi içsel faktörler de olabilir. Ayrıca iç ve dış faktörlerde hiçbir değişim olmasa dahi kişinin farklı zamanlarda seslendirdiği iki konuşmasına karşılık gelen ses sinyali tamamen aynı olmaz. Ses biyometrisindeki bir diğer zorluk da sesin taklit edilebilmesidir. Bazı insanlar bir diğer kişinin sesini olağan dışı şekilde taklit edebilmektedir. Ses taklidi özellikle konuşmacı doğrulama sitemlerinde ele alınması gereken bir konu olup bu tez çalışmasında değerlendirilmemiştir.

Konuşmaya dayalı yaş ve cinsiyet sınıflandırma sistemleri genellikle üç aşamadan oluşur; öznitelik çıkarma, modelleme ve karar aşaması. Sistemin genel performansını doğrudan etkileyen bu aşamaların her birisi için önerilmiş değişik yöntemler mevcuttur. Ancak bu yöntemlerin hiçbiri tüm şartlar için en iyi sonucu üretmez. Bu nedenle her aşama için en uygun yöntemin belirlenmesi ve bu yöntemler kullanılarak ilgili sınıflandırma sisteminin geliştirilesi gerekir.

Bu tez çalışmasında öznitelik çıkarma, modelleme ve karar aşamalarında kullanılan çeşitli yöntemler incelenerek bu yöntemlerle geliştirilen yaş ve cinsiyet sınıflandırma

sistemlerinin performans değerlendirmeleri yapılmıştır. Her bir sistemin avantaj ve dezavantajları ortaya koyulmuş ve bu sistemler için en uygun model büyüklüğü, konuşma süresi, öznitelik boyutu gibi parametreler belirlenmiştir. Çalışmada ayrıca üç farklı sistemin skor seviyeli birleştirilmesine dayalı yeni bir sistem önerilmiştir. Önerilen bu sistemde farklı öznitelik ve sınıflandırma yöntemleri ile geliştirilen alt sistemlerin sonuçları önce olasılıksal skorlara dönüştürülmüş daha sonra ise deneysel olarak belirlenen ağırlık katsayıları ile çarpılarak birleştirilmiştir. Yapılan deneyler sonucunda çalışmada önerilen skor seviyeli birleştirme yaklaşımının konuşmacıların yaş ve cinsiyet gruplarına göre sınıflandırılmasında %5 civarında bir başarı artışı sağladığı görülmüştür.

Bu çalışmada geliştirilen yaş ve cinsiyet sınıflandırma sistemlerine ilişkin elde edilen sonuçlar aşağıda özetlenmiştir.

1. Konuşmaya dayalı tanıma sistemleri genellikle metne bağımlı ve metinden bağımsız olarak iki gruba ayrılır. Metne bağımlı sistemlerde kullanılan metnin içeriği sabit olup sistemin hem eğitim aşamasında hem de test aşamasında aynı içerik kullanılırken metinden bağımsız sistemlerin eğitim ve test aşamalarında farklı içeriğe sahip konuşmalar kullanılır. Metinden bağımsız sistemler daha esnektir ve geliştirilen sistemlerin genellikle metinden bağımsız olması istenir. Ancak bu sistemlerin geliştirilmesi daha zordur ve bu sistemlerde konuşma içeriğinden kaynaklanan değişimlerden daha az etkilenen yöntemlerinin kullanılması gerekir. Metne bağımlı ve metinden bağımsız sistemlerde kullanılan çeşitli yöntemler vardır. Bu yöntemlerden dinamik zaman bükme (DTW) özellikle metne bağımlı sistemlerde kullanılan bir yöntem olup bu çalışmada konuşmacının cinsiyet grubunun (erkek ve kadın) belirlenmesi için kullanılmıştır. Metne bağımlı ve metinden bağımsız olarak geliştirilen DTW sistemi iki farklı veritabanı ile test edilmiştir. Veritabanlarından ilki gürültüsüz bir ortamda tek bir oturumda kaydedilen TIMIT veritabanı, diğeri ise farklı ortamlarda farklı iletişim hatları üzerinden kaydedilen telefon konuşmalarından oluşan ORATOR veritabanıdır. Yapılan testlerde geliştirilen cinsiyet tanıma sisteminin eğitim ve test aşamasında kullanılan konuşmaların içeriğinden ziyade kayıt ortamı ve iletişim kanalı gibi dış faktörlerden daha fazla etkilendiği görülmüştür. Gürültüsüz bir ortamdaki metinden bağımsız olarak geliştirilen sistemin başarısı (%97.91) gürültülü ortamdaki metne bağımlı sistemin başarısından (%97.24) daha yüksek olmuştur.

2. Çalışmada incelenen modelleme yöntemlerden bir diğeri de Vektör Nicemle (VQ) yaklaşımıdır. Özellikle metinden bağımsız tanıma sistemlerinde yaygın olarak kullanılan VQ yöntemi ile metinden bağımsız bir cinsiyet tanıma sistemi geliştirilerek kod kitabı boyutunun tanıma başarısına etkisi araştırılmıştır. TIMIT veritabanı ile yapılan testlerde kod kitabı boyunun cinsiyet tanıma başarısını arttırdığı ve en yüksek tanıma oranın (%97.98) ise 64 elemanlı kod kitabı ile elde edildiği görülmüştür. VQ yöntemiyle geliştirilen aynı sistem konuşmacıların farklı ruhsal durumlarda seslendirdiği konuşmalarından oluşan Boğaziçi Üniversitesi duygulu konuşma veritabanı ile de test edilmiştir. İlgili veritabandaki konuşmacıların olağan konuşmalarıyla %100'e yakın başarı elde edilirken her konuşmacının 4 farklı duygu halinde (sevinçli, olağan, kızgın ve üzgün) seslendirdiği konuşmalarla yapılan testlerde cinsiyet tanıma başarısında %30'a yakın azalma olmuştur. Elde edilen bu sonuçlara göre seslendirilen metnin içeriği, kayıt ve iletişim ortamı gibi etkenlerin dışında konuşmacının ruhsal durumu da cinsiyet tanıma başarısını etkilemektedir. Her bir duygu durumu için elde edilen başarı oranları ayrı ayrı incelendiğinde en yüksek cinsiyet tanıma başarısı üzgün ruh halinde (%74.54), en düşük başarı ise sevinç ruh halinde (%66.36) elde edilmiştir.
3. Konuşmaya dayalı tanıma sistemlerinde yaygın olarak kullanılan özniteliklerin çoğu (MFCC gibi) ses yolunun özelliklerini içinde barındırır ve kayıt şartlarından önemli derecede etkilenir. Ses yolu ise fonem bağımlıdır ve bu durumda ses yolu özelliklerini içinde barındıran özniteliklerde fonem bağımlı olur. Ayrıca bu özniteliklerle güvenilir tanıma sağlanması için tüm fonetik yapıyı kapsayan uzun konuşmaların kullanılması gerekir. Bu nedenle veri miktarına daha az bağımlı olan ve tam olarak metinden bağımsız özniteliklere ihtiyaç vardır. Ses tellerinin titreşimi sonucunda gırtlakta oluşan ve ses kaynağı olarak değerlendirilen hava akışı sinyali bu ihtiyacı karşılamada iyi bir alternatiftir. Bu çalışmada gırtlaksal akış sinyalinden zaman uzayında çıkarılan açılma (OQ), kapanma (CIQ) ve hız (SQ) oranı ile frekans uzayında çıkarılan harmonik seviye farkı (H1-H2) parametrelerinin konuşmacının cinsiyeti ile ilişkisi araştırılmıştır. TIMIT veritabanındaki 438 erkek ve 192 kadın konuşmacının /a/ fonemine karşılık gelen gırtlaksal akış sinyali tekrarlamalı uyarlamalı ters filtreleme (IAIF) yöntemi ile belirlenmiş ve her bir konuşmacı için ilgili ses kaynağı parametreleri

hesaplanmıştır. Elde edilen sonuçlardan hesaplanan dört ses kaynağı parametresinin konuşmacının cinsiyeti ile ilişkili olduğu görülmüştür. Her parametre için bir eşik seviye belirlenerek bu seviyeye göre sınıflandırması yapıldığında, zaman uzayında çıkarılan üç parametre ile %97 seviyesinde, frekans uzayından çıkarılan parametreyle ise %93 seviyesinde cinsiyet (erkek ve kadın) sınıflandırma başarısı elde edilmiştir. Bu sonuçlara göre çalışmada incelenen zaman uzayı parametreleri konuşmacının cinsiyetini tanınmada frekans uzayı parametresine göre daha başarılı olmuştur. Elde edilen sınıflandırma başarıları göz önünde bulundurulduğunda incelenen ses kaynağı parametrelerinin (OQ, CIQ, SQ ve H1-H2) konuşmacının cinsiyetinin tanınmasında öznitelik olarak kullanılabileceği değerlendirilmiştir.

4. Gauss karışım modeli (GMM) özellikle karmaşık verileri modellemede kullanılan güçlü ve esnek bir araçtır. Bu çalışmada üç farklı öznitelik vektörü (MFCC, PLP ve LPCC) ile geliştirilen GMM'ye dayalı sınıflandırma sistemlerinin konuşmacının cinsiyet (erkek ve kadın) grubunu belirlemedeki başarıları araştırılmıştır. Geliştirilen sistemler farklı bileşen sayısı ve farklı öznitelik boyutları ile test edilerek en uygun bileşen sayısı, öznitelik türü ve boyutu belirlenmiştir. TIMIT veritabanı ile yapılan testlerde hem öznitelik boyutunun hem de GMM bileşen sayısının cinsiyet tanıma başarısını arttırdığı ancak belirli bir değerden sonraki artışın başarı üzerinde olumlu bir etkisinin olmadığı görülmüştür. Kullanılan özniteliklerden MFCC en başarılı yöntem olurken öznitelik boyutu ve GMM bileşen sayısının düşük olması durumunda PLP yönteminin başarısı MFCC'den daha yüksek olmuştur. En yüksek cinsiyet tanıma oranı 16 MFCC katsayısından oluşan özniteliklerin 8 bileşenli GMM ile modellenmesi sonucunda %99.37 olarak elde edilmiştir. Düşük boyutlu öznitelik vektörlerin yüksek bileşenli GMM'lerle modellenmesi durumunda başarı oranı %70'lere kadar düşerken, yüksek boyutlu özniteliklerin düşük bileşenli GMM'lerle modellenmesi durumunda başarı oranı %95'ler seviyesinde kalmıştır. Bu durum GMM'ye dayalı cinsiyet sınıflandırma sisteminin başarısında kullanılan öznitelik vektörünün boyutunun GMM bileşen sayısından daha önemli olduğu şeklinde değerlendirilmiştir. Çalışmada en düşük başarı 4 LPCC katsayısından oluşan özniteliklerin 1 bileşenli GMM ile modellenmesi sonucunda %70.74 olarak bulunmuştur. Ancak bu oran öznitelik boyunun ve bileşen

sayısının artması ile birlikte %97'ye kadar çıkmıştır. Bir bileşenli GMM'lerle yapılan testlerde bile cinsiyet tanıma başarısı %98 başarı seviyesi yakalandığı düşünüldüğünde konuşmacının yalnızca cinsiyetinin tanınmasında yüksek bileşenli GMM'lere gerek olmadığı değerlendirilmiştir.

5. GMM'nin genelleyici gücü ile SVM'nin ayırıcı özelliklerini birleştiren GMM süpervektörlerine dayalı SVM (GMM-SV SVM) yaklaşımı başta konuşmacı doğrulama sistemleri olmak üzere birçok ses işleme uygulamasında yaygın olarak kullanılmaktadır. Bu çalışmasında GMM-SV SVM yaklaşımı konuşmacının yaş ve cinsiyet sınıfın belirlemesi amacıyla kullanılmıştır. Geliştirilen sistem farklı konuşma süresi ve GMM bileşen sayısı ile test edilmiş ve elde edilen sonuçlara göre en uygun konuşma süresi ve GMM bileşen sayısı belirlenmiştir. aGender veritabanı ile yapılan testlerde konuşmacılar cinsiyetlerine 3 (erkek, kadın ve çocuk), yaşlarına göre 4 (çocuk, genç, yetişkin ve yaşlı) yaş ve cinsiyetlerine göre ise 7 sınıfa ayrılmıştır. Her üç kategoride de konuşma süresi ve bileşen sayısındaki artış sınıflandırma başarısını arttırmış ancak belirli bir değerden sonraki süre ve bileşen artışının sonuç üzerinde önemli bir etkisinin olmamıştır. Yapılan testler sonucunda en yüksek sınıflandırma başarısı 16 saniyelik konuşmaların 64 bileşenli GMM'lerle modellenmesi sonucunda elde edilmiştir. Yetişkin konuşmacıların cinsiyet grubunun belirlenmesinde %95'in üzerinde başarı sağlanırken çocuk konuşmacıların için bu oran %64 seviyesine kadar düşmüştür. Hatalı kararların büyük bir bölümü çocuk ve kadın cinsiyet grubu arasında olmuştur. Bu durumun çocuk ve kadın konuşmacıların temel frekans özelliğindeki yakınlıkla ilişkili olabileceği değerlendirilmiştir. Yaş gruplarına göre sınıflandırmada ise çocuk ve yaşlı konuşmacılar genç ve yetişkin konuşmacılara kıyasla daha başarılı sınıflandırılmıştır. Bu durumun sesteki değişimin ergenlik ve yaşlılık döneminde daha fazla olmasıyla ilişkili olabileceği değerlendirilmiştir. Konuşmacının yaş ve cinsiyeti birlikte değerlendirildiğinde ise çocuk ve kadın sınıfları arasında görülen yüksek hata oranı başta genç kadın olmak üzere kadın yaş grupları arasında dağılmıştır. Erkek ve kadın sınıfları arasındaki en yüksek hata oranı ise genç erkek ile yaşlı kadın sınıfı arasında olmuştur. Bu durum insan algılamasıyla benzerlik göstermektedir.
6. Bir sınıflandırma sisteminin performansı farklı öznitelikler veya sınıflandırıcılar kullanılarak geliştirilen alt sistemlerin sonuçları birleştirilerek arttırılabilir. Bu

çalışmada da üç farklı öznitelik türü (MFCC, PLP ve prosodik) ve üç farklı sınıflandırma yaklaşımı (SVM, GMM ve GMM-SV SVM) ile geliştirilen toplam 7 alt sistemin sonuçları birleştirilerek konuşmacının yaş ve cinsiyet sınıfının tanınmasında performans artış sağlanmıştır. Aynı sınıflandırma yaklaşımının kullanıldığı alt sistemler önce kendi aralarında birleştirilmiş daha sonra ise elde edilen sonuçlar birleştirilerek nihai sonuç elde edilmiştir. Çalışmada GMM sınıflandırıcısı ile geliştirilen alt sistemlerin kendi aralarındaki birleştirme sonucunda çok az bir performans artışı sağlanırken (%0.15), SVM'ye dayalı sistemlerin birleştirilmesi sonucunda %5 civarında bir performans artışı sağlanmıştır. En yüksek sınıflandırma başarısı ise tüm alt sistemlerin hiyerarşik birleştirilmesi sonucunda elde edilmiştir. Bu sistem ile konuşmacıların cinsiyet grubunun belirlenmesinde %2.5, yaş grubunun belirlenmesinde %4.4, yaş ve cinsiyet grubunun belirlenmesinde ise %5.1 oranında performans artışı sağlanmıştır.

7. Çalışmada son olarak konuşmacının yaş ve cinsiyet grubunun belirlenmesinde iletişim kanalı ve oturum değişkenliğinden kaynaklanan etkilerin performans üzerindeki etkisi araştırılmıştır. NAP yöntemi ile gerçekleştirilen kanal dengeleme sonucunda GMM-SV SVM sisteminin yaş ve cinsiyet sınıflandırma başarısında %1.5 oranında artış sağlanmıştır.
8. Bu tezde elde edilen sonuçlardan konuşmacının yaş ve cinsiyet grubunun tanınmasında, kullanılan verinin uzunluğu, iletişim ortamı, öznitelik çıkarma yöntemi, sınıflandırma algoritması ve karar aşamasında kullanılan yöntem gibi birçok parametrenin etkili olduğu görülmüştür.

5. ÖNERİLER

1. Konuşmacıları yaş ve cinsiyetlerine göre otomatik olarak sınıflandıran sistemlerinin geliştirilmesinde her yaş ve cinsiyet grubundan yaklaşık eşit miktarda konuşma verisine ihtiyaç duyulur. Ancak mevcut veritabanlarının çoğu yalnızca yetişkin konuşmacıların ses kayıtlarından oluşur ve bu yüzden yaş ve cinsiyet tanıma sistemleri için uygun değildir. Her yaş ve cinsiyet grubundan yaklaşık eşit sayıda konuşmacının farklı iletişim ve kayıt ortamlarında seslendirdiği konuşmalardan oluşan veritabanlarının geliştirilmesi bu eksikliğin giderilmesi anlamında faydalı olacaktır.
2. Yaşlanmanın ses üzerindeki etkileri kişiden kişiye değişiklik gösteren oldukça karmaşık bir süreçtir. Kişinin cinsiyeti, ırkı, sigara içip içmediği, gibi birçok faktör bu değişimi etkileyebilir. Bu etkilerin araştırılarak ortaya koyulması gelecek çalışmalara yön vermek adına faydalı olacaktır. Örneğin kişinin önce cinsiyet grubunun daha sonra yaş grubunun belirlendiği iki seviyeli bir sınıflandırma yaklaşımı kullanılarak bu yaklaşımın yaş ve cinsiyet tanıma başarısına etkisi araştırılabilir.
3. Yetişkin konuşmacıların cinsiyetlerine göre sınıflandırılmasında oldukça yüksek başarı sağlanmaktadır. Ancak sisteme çocuk konuşmacılar dahil edildiğinde başarı oranı %10 civarında düşmektedir. Bu durum özellikle çocuk konuşmacıların kadın olarak sınıflandırılmasından kaynaklanır. Kadın ve çocuk konuşmacıların ses özelliklerinin daha ayrıntılı incelenerek bu iki sınıfı birbirinden ayıracak yeni özniteliklerin araştırılması faydalı olacaktır. Bu konuda konuşmanın ritim, hız ve vurgu ses özellikleri incelenebilir.
4. Çalışmada ses kaynağından çıkarılan parametrelerden yalnızca 4 tanesinin konuşmacının cinsiyeti ile ilişkisi araştırılmıştır. Ses kaynağından çıkarılan parametrelerin tamamının konuşmacının yaş ve cinsiyet grubu ile ilişkisi araştırılabilir.
5. Çalışmada kullanılan aGender veritabanında tanımlanan yaş gruplarının sınırlarının belirlenmesinde kişinin yaşla oluşan fizyolojik değişimlerden ziyade çağrı merkezi gibi ticari uygulamaların ihtiyaçları göz önünde bulundurulmuştur.

Yaşla oluşan fizyolojik değişimlere göre belirlenmiş yaş aralıklarının başarı üzerindeki etkisi araştırılabilir.

6. Çalışmada önerilen skor seviyeli birleştirme yaklaşımı ile yaş ve cinsiyet sınıflandırma performansında artış sağlanmıştır. Benzer bir yaklaşım öznitelik seviyesinde uygulanarak öznitelik ve skor seviyeli birleştirme yaklaşımlarının karşılaştırılması yapılabilir.
7. Ayrıca öznitelik ve skor seviyeli birleştirme yaklaşımlarının birlikte kullanıldığı hibrit bir sistem geliştirilerek performans değerlendirmesi yapılabilir.
8. LDA ve PCA gibi yöntemlerle düşük boyutlu uzayda temsil edilen özniteliklerin başarı üzerindeki etkisi araştırılabilir.

6. KAYNAKLAR

1. Schultz, T., Speaker Characteristics, Speaker Classification I, Springer, Berlin Heidelberg, 47-74, 2007.

2. Dobry, G., Hecht, R.M., Avigal, M. ve Zigel, Y., Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal, IEEE Transactions on Audio, Speech, and Language Processing, 19,7 (2011) 1975-1985.

3. Tanner, D.C. ve Tanner, M.E., Forensic aspects of speech patterns: voice prints, speaker profiling, lie and intoxication detection, Lawyers & Judges Publishing Company, 2004.

4. Li, M., Han, K.J. ve Narayanan, S., Automatic speaker age and gender recognition using acoustic and prosodic level information fusion, Computer Speech & Language, 27,1 (2013) 151-167.

5. Schötz, S., Perception, analysis and synthesis of speaker age, 47, Lund University, 2006.

6. Kelly, F., Drygajlo, A. ve Harte, N., Speaker verification in score-ageing-quality classification space, Computer Speech & Language, 27,5 (2013) 1068-1084.

7. Bahari, M.H. ve Van Hamme, H., Speaker age estimation and gender detection based on supervised non-negative matrix factorization, IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS), Eylül 2011, Milan, Italy, 1-6.

8. Mysak, E.D., Pitch and duration characteristics of older males, Journal of Speech & Hearing Research, 2 (1959) 46-54.

9. Linville, S.E., Vocal aging, Singular Thomson Learning, 2001.

10. Minematsu, N., Sekiguchi, M. ve Hirose, K., Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Mayıs 2002, Orlando, Florida, USA, I-137-I-140.

11. Bahari, M.H., Speaker age estimation using Hidden Markov Model weight supervectors, 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), Temmuz 2012, Montreal, QC, Canada, 517-521.

12. Muller, C., Wittig, F. ve Baus, J., Exploiting speech for recognizing elderly users to respond to their special needs, Eighth European Conference on Speech Communication and Technology, Eylül 2003, Geneva, Switzerland, 1305–1308.

13. Shafran, I., Riley, M. ve Mohri, M., Voice signatures, IEEE Workshop on Automatic Speech Recognition and Understanding, 2003. ASRU'03, Kasım 2003, Virgin Islands, 31-36.

14. Mahmoodi, D., Marvi, H., Taghizadeh, M., Soleimani, A., Razzazi, F. ve Mahmoodi, M., Age estimation based on speech features and support vector machine, 3rd Computer Science and Electronic Engineering Conference (CEEC), Temmuz 2011, Colchester, United Kingdom, 60-64.

15. Chen, C.C., Lu, P.T., Hsia, M.L., Ke, J.Y. ve Chen, O.T.C., Gender-to-Age hierarchical recognition for speech, IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS), Ağustos 2011, Seoul, Korea, 1-4.

16. van Heerden, C., Barnard, E., Davel, M., van der Walt, C., van Dyk, E., Feld, M. ve Müller, C., Combining regression and classification methods for improving automatic speaker age recognition, 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Mart 2010, Dallas, Texas, USA, 5174-5177.

17. Dobry, G., Hecht, R.M., Avigal, M. ve Zigel, Y., Dimension reduction approaches for SVM based speaker age estimation, Tenth Annual Conference of the International Speech Communication Association, Eylül 2009, Brighton, United Kingdom, 2031–2034.

18. Acero, A. ve Huang, X., Speaker and gender normalization for continuous-density hidden Markov models, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996 (ICASSP-96), Mayıs 1996, Atlanta, Georgia, USA, 342-345.

19. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C.A. ve Narayanan, S.S., The INTERSPEECH 2010 paralinguistic challenge, InterSpeech, Eylül 2010, Makuhari, Japan, 2795-2798.

20. Vergin, R., Farhat, A. ve O'Shaughnessy, D., Robust gender-dependent acoustic-phonetic modelling in continuous speech recognition based on a new automatic male/female classification, Proc. Fourth International Conference on Spoken Language (ICSLP 96), Ekim 1996, Philadelphia, PA, USA, 1081-1084.

21. Parris, E.S. ve Carey, M.J., Language independent gender identification, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96), Mayıs 1996, Atlanta, GA, USA, 685-688.

22. Slomka, S. ve Sridharan, S., Automatic gender identification optimised for language independence, Proc. of IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications (TENCON'97), Aralık 1997, Brisbane, Queensland, Australia, 145-148.

23. Zeng, Y.-M., Wu, Z.-Y., Falk, T. ve Chan, W.-Y., Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech, International

Conference on Machine Learning and Cybernetics, Ağustos 2006, Dalian, China, 3376-3379.

24. Hu, Y., Wu, D. ve Nucci, A., Pitch-based gender identification with two-stage classification, Security and Communication Networks, 5,2 (2012) 211-225.

25. Harb, H. ve Chen, L., Voice-based gender identification in multimedia applications, Journal of intelligent information systems, 24,2 (2005) 179-198.

26. Shue, Y.-L. ve Iseli, M., The role of voice source measures on automatic gender classification, International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), Nisan 2008, Las Vegas, NV, USA, 4493-4496.

27. Ting, H., Yingchun, Y. ve Zhaohui, W., Combining MFCC and pitch to enhance the performance of the gender recognition, 8th International Conference on Signal Processing, Kasım 2006, Beijing, China, 16-20.

28. Djemili, R., Bourouba, H. ve Korba, M.C.A., A speech signal based gender identification system using four classifiers, International Conference on Multimedia Computing and Systems (ICMCS), Mayıs 2012, Tangier, Morocco, 184-187.

29. Gaikwad, S., Gawali, B. ve Mehrotra, S., Gender identification using SVM with Combination of MFCC, Advances in Computational Research, 4,1 (2012).

30. Phoophuangpairoj, R. ve Phongsuphap, S., Two-Stage Gender Identification Using Pitch Frequencies, MFCCs and HMMs, IEEE International Conference on Systems, Man, and Cybernetics, Ekim 2015, Kowloon, China, 2879-2884.

31. Bakir, C., Automatic Speaker Gender Identification for the German Language, Balkan Journal of Electrical & Computer Engineering 4,2 (2016) 79-83.

32. Müller, C., Automatic recognition of speakers' age and gender on the basis of empirical studies, Ninth International Conference on Spoken Language Processing, Eylül 2006, Pittsburgh, Pennsylvania, USA, 2118-2121.

33. Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Muller, C., Huber, R., Andrassy, B. ve Bauer, J.G., Comparison of four approaches to age and gender recognition for telephone applications, International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), Nisan 2007, Honolulu, HI, USA, IV-1089-IV-1092.

34. Bocklet, T., Maier, A. ve Nöth, E., Age Determination of Children in Preschool and Primary School Age with GMM-Based Supervectors and Support Vector Machines/Regression, Proc. of the 11th international conference on Text, Speech and Dialogue, Eylül 2008, Brno, Czech Republic, 253-260.

35. Bocklet, T., Stemmer, G., Zeissler, V. ve Nöth, E., Age and Gender Recognition Based on Multiple Systems - Early vs. Late Fusion, INTERSPEECH 2010, Eylül 2010, Makuhari, Chiba, Japan, 2830-2833.

36. Porat, R., Lange, D. ve Zigel, Y., Age Recognition Based on Speech Signals Using Weights Supervector, INTERSPEECH 2010, Eylül 2010, Makuhari, Chiba, Japan, 2814-2817.

37. Meinedo, H. ve Trancoso, I., Age and gender classification using fusion of acoustic and prosodic features, INTERSPEECH 2010, Eylül 2010, Makuhari, Chiba, Japan, 2818-2821.

38. Nguyen, P., Le, T., Tran, D., Huang, X. ve Sharma, D., Fuzzy Support Vector Machines for Age and Gender Classification, INTERSPEECH 2010, Eylül 2010, Makuhari, Chiba, Japan, 2806-2809.

39. Kockmann, M., Burget, L. ve Cernocký, J., Brno University of Technology system for Interspeech 2010 Paralinguistic Challenge, NTERSPEECH 2010, Eylül 2010, Makuhari, Chiba, Japan, 2822-2825.

40. Kenny, P., Ouellet, P., Dehak, N., Gupta, V. ve Dumouchel, P., A study of interspeaker variability in speaker verification, IEEE Transactions on Audio, Speech, and Language Processing, 16,5 (2008) 980-988.

41. Safavi, S., Russell, M.J. ve Jancovic, P., Identification of Age-Group from Children's Speech by Computers and Humans, INTERSPEECH 2014, Eylül 2014, Singapore, 243-247.

42. Chen, O.T.-C. ve Gu, J.J., Improved gender/age recognition system using arousal-selection and feature-selection schemes, International Conference on Digital Signal Processing (DSP), Temmuz 2015, Singapore, 148-152.

43. Barkana, B.D. ve Zhou, J., A new pitch-range based feature set for a speaker's age and gender classification, Applied Acoustics, 98 (2015) 52-61.

44. Gautam, S. ve Singh, L., Developmental pattern analysis and age prediction by extracting speech features and applying various classification techniques, International Conference on Computing, Communication & Automation (ICCCA), Mayıs 2015, Noida, India, 83-87.

45. Faek, F.K., Objective Gender and Age Recognition from Speech Sentences, ARO-The Scientific Journal of Koya University, 3,2 (2015) 24-29.

46. Rabiner, L. ve Juang, B.-H., Fundamentals of speech recognition, Prentice Hall, New Jersey, 1993.

47. Huang, X., Acero, A. ve Hon, H.-W., Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall, New Jersey, 2001.

48. Lieberman, P. ve Blumstein, S.E., Speech physiology, speech perception, and acoustic phonetics, Cambridge University Press, Cambridge, 1988.

49. Laver, J., Principles of phonetics, Cambridge University Press, Cambridge, 1994.

50. Merriam-Webster, Merriam-Webster's collegiate dictionary, Merriam-Webster, Massachusetts, 2004.

51. http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/changes.html, National Center for Voice and Speech, Voice Changes Throughout Life. 19 Ekim 2016.

52. Laukkanen, A.-M. ve Leino, T., Ihmeellinen ihmisääni [The wonderful human voice. The basics, assessment, measurement and development of voice usage and speech technique], Helsinki: Gaudeamus, 1999.

53. Story, B.H., An overview of the physiology, physics and modeling of the sound source for vowels, Acoustical Science and Technology, 23,4 (2002) 195-206.

54. Baken, R.J., Electroglottography, Journal of Voice, 6,2 (1992) 98-110.

55. Claes, T., Dologlou, I., ten Bosch, L. ve Van Compernolle, D., A novel feature transformation for vocal tract length normalization in automatic speech recognition, IEEE transactions on speech and audio processing, 6,6 (1998) 549-557.

56. Fant, G., Acoustic Theory of Speech Production, The Hague, Mouton, 1960.

57. Rabiner, L.R. ve Gold, B., Theory and application of digital signal processing, Prentice-Hall, New Jersey, 1975.

58. Jain, A.K., Ross, A. ve Prabhakar, S., An introduction to biometric recognition, IEEE Transactions on circuits and systems for video technology, 14,1 (2004) 4-20.

59. Bolle, R.M., Connell, J., Pankanti, S., Ratha, N.K. ve Senior, A.W., Guide to biometrics, Springer, New York, 2004.

60. Maltoni, D., Maio, D., Jain, A. ve Prabhakar, S., Handbook of fingerprint recognition, Springer, New York, 2009.

61. Brand, J., Mason, J.S. ve Colomb, S., Visual speech: A physiological or behavioural biometric, International Conference on Audio-and Video-Based Biometric Person Authentication, Haziran 2001, Halmstad, Sweden, 157-168.

62. Zhao, W., Chellappa, R., Phillips, P.J. ve Rosenfeld, A., Face recognition: A literature survey, ACM Comput. Surv., 35,4 (2003) 399-458.

63. Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D. ve Reynolds, D.A., A tutorial on text-independent speaker verification, EURASIP journal on applied signal processing, 2004 (2004) 430-451.

64. Beigi, H., Fundamentals of Speaker Recognition, Springer Publishing Company, Incorporated, 2011.

65. Shannon, C.E. ve Weaver, W., The mathematical theory of communication, University of Illinois press, 2015.

66. https://en.wikipedia.org/wiki/Nyquist-Shannon_sampling_theorem, Nyquist–Shannon sampling theorem. 20 Ekim 2016.

67. http://clas.mq.edu.au/speech/acoustics/waveforms/speech_waveforms.html, Speech Waveforms. 20 Ekim 2016.

68. http://www.webexhibits.org/colorart/bh.html, WebExhibits. Newton and the color spectrum. 20 Ekim 2016.

69. http://www.princeton.edu/~cuff/ele201/kulkarni_text/frequency.pdf, ELE 201: Information Signals - Course Notes, Chapter 4 Frequency Domain and Fourier Transforms. 20 Ekim 2016.

70. Deller, J.R., Hansen, J.H. ve Proakis, J.G., Discrete-time processing of speech signals, IEEE Press, New Jersey, 2000.

71. Furui, S., Digital Speech Processing, Synthesis, and Recognition Marcel Dekker, New York, 2001.

72. Prokis, J.G. ve Manolakis, D.G., Digital signal processing: principles, algorithms and applications, 11, Prentice Hall, New Jersey, 1996.

73. Quatieri, T.F., Discrete-time speech signal processing: principles and practice, Prentice Hall Press, 2001.

74. https://engineering.purdue.edu/VISE/ee438L/lab9/pdf/lab9a.pdf, Purdue University: ECE438 - Digital Signal Processing with Applications, Laboratory 9: Speech Processing (Week 1). 6 Kasım 2010.

75. Bogert, B.P., Healy, M.J. ve Tukey, J.W., The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking, Proceedings of the symposium on time series analysis, 1963, Wiley, NY, 209-243.

76. Bishop, C.M., Neural networks for pattern recognition, Oxford university press, New York, NY, USA, 1995.

77. Jain, A.K., Duin, R.P.W. ve Mao, J., Statistical pattern recognition: A review, IEEE Transactions on pattern analysis and machine intelligence, 22,1 (2000) 4-37.

78. Jain, A. ve Zongker, D., Feature selection: Evaluation, application, and small sample performance, IEEE transactions on pattern analysis and machine intelligence, 19,2 (1997) 153-158.

79. Rose, P., Forensic Speaker Identification, Taylor & Francis, London, 2002.

80. Schuller, B., Wöllmer, M., Eyben, F. ve Rigoll, G., Prosodic, spectral or voice quality? Feature type relevance for the discrimination of emotion pairs, Linguistic Insights, Studies in Language and Communication, 97 (2009) 285-307.

81. Ververidis, D. ve Kotropoulos, C., Emotional speech recognition: Resources, features, and methods, Speech communication, 48,9 (2006) 1162-1181.

82. Steidl, S., Batliner, A., Nöth, E. ve Hornegger, J., Quantification of segmentation and F0 errors and their effect on emotion recognition, International Conference on Text, Speech and Dialogue, Eylül 2008, Brno, Czech Republic, 525-534.

83. Krajewski, J. ve Kröger, B.J., Using prosodic and spectral characteristics for sleepiness detection, INTERSPEECH 2007, Ağustos 2007, Antwerp, Belgium, 1841-1844.

84. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N. ve Kessous, L., Combining efforts for improving automatic classification of emotional user states, Fifth Slovenian and First International Language Technologies Conference, Kasım 2006, Ljubljana, Slovenia, 240-245.

85. Graciarena, M., Shriberg, E., Stolcke, A., Enos, F., Hirschberg, J. ve Kajarekar, S., Combining prosodic lexical and cepstral systems for deceptive speech detection, 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Mayıs 2006, Toulouse, France, I-I.

86. Schötz, S., Acoustic analysis of adult speaker age, Speaker Classification I, Springer-Verlag, Berlin, Heidelberg, 88-107, 2007.

87. Steidl, S., Automatic classification of emotion related user states in spontaneous children's speech, PhD thesis, University of Erlangen-Nuremberg Erlangen, Germany, 2009.

88. Atal, B.S. ve Hanauer, S.L., Speech analysis and synthesis by linear prediction of the speech wave, The Journal of the Acoustical Society of America, 50,2B (1971) 637-655.

89. Makhoul, J., Linear prediction: A tutorial review, Proceedings of the IEEE, 63,4 (1975) 561-580.

90. Rosenberg, A. ve Sambur, M., New techniques for automatic speaker verification, IEEE Transactions on Acoustics, Speech, and Signal Processing, 23,2 (1975) 169-176.

91. Atal, B.S., Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, the Journal of the Acoustical Society of America, 55,6 (1974) 1304-1312.

92. Hermansky, H., Perceptual linear predictive (PLP) analysis of speech, the Journal of the Acoustical Society of America, 87,4 (1990) 1738-1752.

93. Davis, S.B. ve Mermelstein, P., Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech and Signal Processing, 28,4 (1980) 357-366.

94. Stevens, S.S., Volkmann, J. ve Newman, E.B., A scale for the measurement of the psychological magnitude pitch, The Journal of the Acoustical Society of America, 8,3 (1937) 185-190.

95. Alku, P., Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering, Speech communication, 11,2-3 (1992) 109-118.

96. Ramirez, J., Górriz, J.M. ve Segura, J.C., Voice activity detection. fundamentals and speech recognition system robustness, Robust Speech Recognition and Understanding, InTech, 2007.

97. Beritelli, F., Casale, S. ve Ruggeri, G., Performance evaluation and comparison of ITU-T/ETSI voice activity detectors, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01), Mayıs 2001, Salt Lake City, UT, USA, 1425-1428.

98. Campbell, J.P., Speaker recognition: a tutorial, Proceedings of the IEEE, 85,9 (1997) 1437-1462.

99. Soong, F.K., Rosenberg, A.E., Juang, B.-H. ve Rabiner, L.R., Report: A vector quantization approach to speaker recognition, AT&T technical journal, 66,2 (1987) 14-26.

100. Furui, S., Cepstral analysis technique for automatic speaker verification, IEEE Transactions on Acoustics, Speech, and Signal Processing, 29,2 (1981) 254-272.

101. Reynolds, D.A., Quatieri, T.F. ve Dunn, R.B., Speaker verification using adapted Gaussian mixture models, Digital signal processing, 10,1 (2000) 19-41.

102. Reynolds, D.A. ve Rose, R.C., Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE transactions on speech and audio processing, 3,1 (1995) 72-83.

103. BenZeghiba, M.F. ve Bourlard, H., User-customized password speaker verification using multiple reference and background models, Speech Communication, 48,9 (2006) 1200-1213.

104. Naik, J.M., Netsch, L.P. ve Doddington, G.R., Speaker verification over long distance telephone lines, International Conference on Acoustics, Speech, and Signal Processing (ICASSP-89), Mayıs 1989, Glasgow, Scotland, 524-527.

105. Farrell, K.R., Mammone, R.J. ve Assaleh, K.T., Speaker recognition using neural networks and conventional classifiers, IEEE Transactions on speech and audio processing, 2,1 (1994) 194-205.

106. Heck, L.P., Konig, Y., Sönmez, M.K. ve Weintraub, M., Robustness to telephone handset distortion in speaker recognition by discriminative feature design, Speech Communication, 31,2 (2000) 181-192.

107. Yegnanarayana, B. ve Kishore, S.P., AANN: an alternative to GMM for pattern recognition, Neural Networks, 15,3 (2002) 459-469.

108. Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer, E. ve Torres-Carrasquillo, P.A., Support vector machines for speaker and language recognition, Computer Speech & Language, 20,2 (2006) 210-229.

109. Senin, P., Dynamic time warping algorithm review, Information and Computer Science Department University of Hawaii, 855 (2008) 1-23.

110. Keogh, E., Exact indexing of dynamic time warping, Proc. of the 28th international conference on Very Large Data Bases, Ağustos 2002, Hong Kong, China 406-417.

111. Muda, L., Begam, M. ve Elamvazuthi, I., Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques, Journal of Computing, 2,3 (2010) 138-145.

112. Gersho, A. ve Gray, R.M., Vector quantization and signal compression, 159, Springer Science & Business Media, 2012.

113. Gish, H. ve Schmidt, M., Text-independent speaker identification, IEEE signal processing magazine, 11,4 (1994) 18-32.

114. Naik, J.M., Speaker verification: A tutorial, IEEE Communications Magazine, 28,1 (1990) 42-48.

115. Ong, S., Sridharan, S., Yang, C.-H. ve Moody, M., Comparison of Four Distance Measures for Long Time Text-Independent Speaker Identification, Fourth International Symposium on Signal Processing and Its Applications (ISSPA 96), Ağustos 1996, Gold Coast, Australia, 369-372.

116. Zhen, B., Wu, X., Liu, Z. ve Chi, H., On the Importance of Components of the MFCC in Speech and Speaker Recognition, Acta Scientiarum Naturalium-Universitatis Pekinensis, 37,3 (2001) 371-378.

117. Kinnunen, T., Kilpeläinen, T. ve FrÄnti, P., Comparison of clustering algorithms in speaker identification, Proc. IASTED Int. Conf. Signal Processing and Communications (SPC), Eylül 2000, Marbella, Spain, 222-227.

118. Xuedong, H., Acero, A. ve Hon, H.-W., Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall PTR, 2001.

119. Pena, J.M., Lozano, J.A. ve Larranaga, P., An empirical comparison of four initialization methods for the k-means algorithm, Pattern recognition letters, 20,10 (1999) 1027-1040.

120. Dhonde, S.B. ve Jagade, S.M., Pattern-Matching for Speaker Verification: A Review, Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), Cilt 1, Springer International Publishing, Cham, 19-25, 2015.

121. Reynolds, D.A., Speaker identification and verification using Gaussian mixture speaker models, Speech communication, 17,1 (1995) 91-108.

122. McLachlan, G. ve Peel, D., Finite mixture models, John Wiley & Sons, 2004.

123. Dempster, A.P., Laird, N.M. ve Rubin, D.B., Maximum likelihood from incomplete data via the EM algorithm, Journal of the royal statistical society. Series B (methodological), 39,1 (1977) 1-38.

124. Togneri, R. ve Pullella, D., An overview of speaker identification: Accuracy and robustness issues, IEEE Circuits and Systems Magazine, 11,2 (2011) 23-61.

125. Isobe, T. ve Takahashi, J.-i., Text-independent speaker verification using virtual speaker based cohort normalization, Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99), Eylül 1999, Budapest, Hungary, 987–990.

126. Rosenberg, A.E. ve Parthasarathy, S., Speaker background models for connected digit password speaker verification, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96), Mayıs 1996, Atlanta, GA, USA, 81-84.

127. Ferras, M., Leung, C.-C., Barras, C. ve Gauvain, J.-L., Comparison of speaker adaptation methods as feature extraction for SVM-based speaker recognition, IEEE Transactions on Audio, Speech, and Language Processing, 18,6 (2010) 1366-1378.

128. Van Vuuren, S., Speaker verification in a time-feature space, Citeseer, 1999.

129. Leggetter, C.J. ve Woodland, P.C., Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Computer Speech & Language, 9,2 (1995) 171-185.

130. Jurafsky, D. ve Martin, J.H., Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice-Hall International, Upper Saddle River, N.J. London, 2000.

131. Wang, H., Zhang, X., Xiao, X., Zhang, J. ve Yan, Y., Combining MAP and MLLR approaches for SVM based speaker recognition with a multi-class MLLR technique, Second International Symposium on Information Science and Engineering, Aralık 2009, Shanghai, China, China, 447-450.

132. Burges, C.J., A tutorial on support vector machines for pattern recognition, Data mining and knowledge discovery, 2,2 (1998) 121-167.

133. Choisy, C. ve Belaid, A., Handwriting recognition using local methods for normalization and global methods for recognition, Sixth International Conference on Document Analysis and Recognition, Eylül 2001, Seattle, WA, USA, 23-27.

134. Teow, L.-N. ve Loe, K.-F., Robust vision-based features and classification schemes for off-line handwritten digit recognition, Pattern Recognition, 35,11 (2002) 2355-2364.

135. Gao, D., Zhou, J. ve Xin, L., SVM-based detection of moving vehicles for automatic traffic monitoring, IEEE Intelligent Transportation Systems, Ağustos 2001, Oakland, CA, USA, 745-749.

136. Li, Z., Weida, Z. ve Licheng, J., Radar target recognition based on support vector machine, 5th International Conference on Signal Processing Proceedings (ICSP 2000), Ağustos 2000, Beijing, China, China, 1453-1456.

137. Guo, G., Li, S.Z. ve Chan, K., Face recognition by support vector machines, Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Mart 2000, Grenoble, France, France, 196-201.

138. Shih, P. ve Liu, C., Face detection using discriminating feature analysis and support vector machine, Pattern Recognition, 39,2 (2006) 260-276.

139. Ye, Q., Huang, Q., Gao, W. ve Zhao, D., Fast and robust text detection in images and video frames, Image and Vision Computing, 23,6 (2005) 565-576.

140. Lan, M., Tan, C.-L., Low, H.-B. ve Sung, S.-Y., A comprehensive comparative study on term weighting schemes for text categorization with support vector machines, Special interest tracks and posters of the 14th international conference on World Wide Web, Mayıs 2005, Chiba, Japan, 1032-1033.

141. Campbell, W.M., Generalized linear discriminant sequence kernels for speaker recognition, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Mayıs 2002, Orlando, FL, USA, I-161-I-164.

142. Campbell, W.M., Campbell, J.P., Reynolds, D.A., Jones, D.A. ve Leek, T.R., Phonetic speaker recognition with support vector machines, Advances in neural information processing systems, Aralık 2003, Whistler, British Columbia, Canada, 1377-1384.

143. Campbell, W.M., Sturim, D.E. ve Reynolds, D.A., Support vector machines using GMM supervectors for speaker verification, IEEE signal processing letters, 13,5 (2006) 308-311.

144. Bennett, K.P. ve Campbell, C., Support vector machines: hype or hallelujah?, ACM SIGKDD Explorations Newsletter, 2,2 (2000) 1-13.

145. Wan, V. ve Renals, S., Evaluation of kernel methods for speaker verification and identification, IEEE International Conference on Acoustics, Speech, and Signal Processing, Mayıs 2002, Orlando, FL, USA I-669-I-672.

146. Scholkopf, B., Sung, K.-K., Burges, C.J., Girosi, F., Niyogi, P., Poggio, T. ve Vapnik, V., Comparing support vector machines with Gaussian kernels to radial basis function classifiers, IEEE transactions on Signal Processing, 45,11 (1997) 2758-2765.

147. Campbell, W.M. ve Assaleh, K.T., Polynomial classifier techniques for speaker verification, IEEE International Conference on Acoustics, Speech, and Signal Processing, Mart 1999, Phoenix, AZ, USA, USA, 321-324.

148. Wan, V. ve Campbell, W.M., Support vector machines for speaker verification and identification, Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop, Aralık 2000, Sydney, NSW, Australia, 775-784.

149. Fauve, B.G., Matrouf, D., Scheffer, N., Bonastre, J.-F. ve Mason, J.S., State-of-the-art performance in text-independent speaker verification through open-source software, IEEE Transactions on Audio, Speech, and Language Processing, 15,7 (2007) 1960-1968.

150. Karam, Z.N. ve Campbell, W.M., A new kernel for SVM MLLR based speaker recognition, 8th Annual Conference of the International Speech Communication Association (Interspeech 2007), Ağustos 2007, Antwerp, Belgium, 290-293.

151. Kreßel, U.H.-G., Pairwise classification and support vector machines, Advances in kernel methods, MIT Press, 255-268, 1999.

152. Friedman, J., Another approach to polychotomous classification, Technical report, Department of Statistics, Stanford University, 1996.

153. Cutzu, F., Polychotomous classification with pairwise classifiers: A new voting principle, Proc. of the 4th international conference on Multiple classifier, Haziran 2003, Guildford, UK, 115-124.

154. Platt, J.C., Cristianini, N. ve Shawe-Taylor, J., Large Margin DAGs for Multiclass Classification, nips, 547-553.

155. Vapnik, V.N. ve Vapnik, V., Statistical learning theory, 1, Wiley New York, 1998.

156. Domınguez, J.G., Session variability compensation in speaker and language recognition, Ph.D. thesis, UniversidadAut´onoma de Madrid, Madrid, Spain, 2011.

157. Solomonoff, A., Quillen, C. ve Campbell, W.M., Channel compensation for SVM speaker recognition, in Proc. Odyssey04, Haziran 2004, Toledo, Spain, 219-226.

158. Xu, L., Krzyzak, A. ve Suen, C.Y., Methods of combining multiple classifiers and their applications to handwriting recognition, IEEE Transactions on systems, man, and cybernetics, 22,3 (1992) 418-435.

159. Balti, H. ve Elmaghraby, A.S., Speech emotion detection using time dependent self organizing maps, IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Aralık 2013, Athens, Greece, 470-478.

160. Planet, S. ve Iriondo, I., Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition, 7th Iberian Conference on Information Systems and Technologies (CISTI), Haziran 2012, Madrid, Spain, 1-6.

161. Indovina, M., Uludag, U., Snelick, R., Mink, A. ve Jain, A., Multimodal biometric authentication methods: a COTS approach, Workshop on Multimodal User Authentication (MMUA 2003), Aralık 2003, Santa Barbara, CA, 99-106.

162. Ong, M.G.K., Connie, T., Jin, A.T.B. ve Ling, D.N.C., A single-sensor hand geometry and palmprint verification system, Proceedings of the 2003 ACM SIGMM workshop on Biometrics methods and applications, Kasım 2003, Berkley, California, 100-106.

163. Alkoot, F.M. ve Kittler, J., Experimental evaluation of expert fusion strategies, Pattern recognition letters, 20,11 (1999) 1361-1369.

164. Marcialis, G.L. ve Roli, F., Fingerprint verification by fusion of optical and capacitive sensors, Pattern Recognition Letters, 25,11 (2004) 1315-1322.

165. Kittler, J., Matas, J., Jonsson, K. ve Sánchez, M.R., Combining evidence in personal identity verification systems, Pattern Recognition Letters, 18,9 (1997) 845-852.

166. Sayedelahl, A., Araujo, R. ve Kamel, M.S., Audio-visual feature-decision level fusion for spontaneous emotion estimation in speech conversations, IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Temmuz 2013, San Jose, CA, USA, 1-6.

167. Van Erp, M. ve Schomaker, L., Variants of the borda count method for combining ranked classifier hypotheses, 7th International Workshop on Frontiers in Handwriting Recognition, Eylül 2000, Amsterdam, 443-452.

168. Mangai, U.G., Samanta, S., Das, S. ve Chowdhury, P.R., A survey of decision fusion and feature fusion strategies for pattern classification, IETE Technical review, 27,4 (2010) 293-307.

169. Meral, H., Ekenel, H. ve Ozsoy, A., Analysis of emotion in Turkish, XVII National Conference on Turkish Linguistics, Mayıs 2003, Eskisehir, Türkiye.

170. Williamson, G., Human communication: A linguistic introduction (2nd edition), Billingham, UK, 2006.

171. Boersma, P., Praat, a system for doing phonetics by computer, Glot international, 5,9/10 (2002) 341-345.

172. Platt, J., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Advances in large margin classifiers, 10,3 (1999) 61-74.

ÖZGEÇMİŞ

04 Mart 1978 yılında Kayserinin Pınarbaşı ilçesinde doğdum. İlköğrenimimi 1989 yılında Samsun 50. Yıl İlkokulu'nda, ortaöğrenimimi 1992 yılında Samsun Gülsüm Sami Kefeli İlköğretim Okulu'nda tamamladım. 1995 yılında ise Ordu Atatürk Lisesi'ni bitirdim. 1999 yılında Karadeniz Teknik Üniversitesi Bilgisayar Mühendisliği Bölümünden mezun oldum. Aynı yıl Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda başladığım yüksek lisans eğitimime devam ederken 2000-2003 yılları arasında araştırma görevlisi olarak görev yaptım. 2003 yılında Ordu Üniversitesi Meslek Yüksekokulu Bilgisayar Teknolojileri Bölümüne öğretim görevlisi olarak atandım. 2004 yılında yüksek lisans eğitimimi tamamladıktan sonra askerlik görevimi yaptım. 2006 yılında Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda doktora eğitimime başladım. Halen Ordu Üniversitesi Meslek Yüksekokulu Bilgisayar Programcılığı Bölümünde öğretim görevlisi olarak devam etmekteyim. Evli ve 2 çocuk babasıyım. Yayınlarım aşağıda verilmiştir.

SCI-Expanded Dergi:

Yücesoy, E. and Nabiyev, V.V., A new approach with score-level fusion for the classification of a speaker age and gender, Computers & Electrical Engineering, 53 (2016) 29-39.

Yücesoy, E. and Nabiyev, V.V., Konuşmacı Yaş ve Cinsiyetinin GKM Süpervektörlerine Dayalı Bir DVM Sınıflandırıcısı ile Belirlenmesi, Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, 31,3 (2016).

Ulusal Hakemli Dergi:

Nabiyev, V.V. and **Yücesoy, E**., Gender Identification of the Speaker Using VQ Method, Turkish Journal of Computer and Mathematics Education (TURCOMAT), 1,1 (2009).

Uluslararası Sempozyum:

Nabiyev, V.V. and **Yücesoy, E.**, VQ Yöntemiyle Konuşmacı Cinsiyetinin Belirlenmesi, Turkish Journal of Computer and Mathematics Education Vol, 1,1 (2009) 35-47.

Yücesoy, E. and Nabiyev, V.V."Gender identification of a speaker using MFCC and GMM", in Electrical and Electronics Engineering (ELECO), 2013 8th International Conference on. 2013. IEEE.

Ulusal Sempozyum:

Nabiyev, V.V. and **Yucesoy, E.**"Konuşmacı Cinsiyetinin Temel Frekansa Göre Belirlenmesi", Çankaya Üniversitesi Birinci Mühendislik ve Teknoloji Sempozyumu. 2008. Ankara.

Yucesoy, E. and Nabiyev, V.V."Gender identification of the speaker using DTW method", in Signal Processing and Communications Applications Conference, 2009. SIU 2009. IEEE 17th. 2009. IEEE.

Yücesoy, E. and Nabiyev, V.V."Gender identification of a speaker from voice source", in Signal Processing and Communications Applications Conference (SIU), 2013 21st. 2013. IEEE.

Yucesoy, E. and Nabiyev, V.V."Comparison of MFCC, LPCC and PLP features for the determination of a speaker's gender", in Signal Processing and Communications Applications Conference (SIU), 2014 22nd. 2014. IEEE.

Yucesoy, E. and Nabiyev, V.V."Age and gender recognition of a speaker from short-duration phone conversations", in Signal Processing and Communications Applications Conference (SIU), 2015 23th. 2015. IEEE.