BURSA TEKNİK ÜNİVERSİTESİ ❖ FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE SESLER İLE KONUŞMACI KİMLİĞİNİN DOĞRULANMASI/BELİRLENMESİ

YÜKSEK LİSANS TEZİ

Havva ÇELİKTAŞ

Elektrik-Elektronik Mühendisliği Anabilim Dalı

ŞUBAT 2019

BURSA TEKNİK ÜNİVERSİTESİ ❖ FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE SESLER İLE KONUŞMACI KİMLİĞİNİN DOĞRULANMASI/BELİRLENMESİ

YÜKSEK LİSANS TEZİ

Havva ÇELİKTAŞ
(151082307)

Elektrik-Elektronik Mühendisliği Anabilim Dalı

Tez Danışmanı: Doç. Dr. Cemal HANİLÇİ

ŞUBAT 2019

BTÜ, Fen Bilimleri Enstitüsü'nün 151082307 numaralı Yüksek Lisans Öğrencisi Havva ÇELİKTAŞ, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "TÜRKÇE SESLER İLE KONUŞMACI KİMLİĞİNİN DOĞRULANMASI/BELİRLENMESİ" başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı :	**Doç. Dr. Cemal HANİLÇİ** Bursa Teknik Üniversitesi	.....
Jüri Üyeleri :	**Doç. Dr. Hakan GÜRKAN** Bursa Teknik Üniversitesi	.....
	Doç. Dr. Ersen YILMAZ Uludağ Üniversitesi	.....

Savunma Tarihi : 15 Şubat 2019

FBE Müdürü :	**Doç. Dr. Murat ERTAŞ** Bursa Teknik Üniversitesi	..... **...../...../.....**

İNTİHAL BEYANI

Bu tezde görsel, işitsel ve yazılı biçimde sunulan tüm bilgi ve sonuçların akademik ve etik kurallara uyularak tarafımdan elde edildiğini, tez içinde yer alan ancak bu çalışmaya özgü olmayan tüm sonuç ve bilgileri tezde kaynak göstererek belgelediğimi, aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim.

Öğrencinin Adı Soyadı: Havva ÇELİKTAŞ

İmzası :

Aileme,

ÖNSÖZ

Bu tezde Bursa Teknik Üniversitesi yüksek lisans programı boyunca yapmış olduğum, çalışmalarım sonucunda edindiğim bilgileri dikkatlerinize sunmuş bulunmaktayım.

Tez çalışmam boyunca yardımlarını esirgemeyen maddi ve manevi her koşulda bana destek olan, bilgi ve tecrübesiyle yol gösteren çok değerli hocam Sayın Doç. Dr. Cemal HANİLÇİ'ye, bu süreç boyunca destekleriyle hep yanımda olan aileme ve arkadaşlarıma sonsuz teşekkür ederim.

Şubat 2019 Havva Çeliktaş

İÇİNDEKİLER

Sayfa

KISALTMALAR

AKD : Ayrık Fourier Dönüşümü
DGKK : Değiştirilmiş Grup Gecikme Kepstrum Katsayıları
DET : Detection Error Tradeoff
DÖK : Doğrusal Öngörü Katsayıları
DVM : Destek Vektör Makineleri
DZE : Dinamik Zaman Eğirme
EER : Eşit Hata Oranı
EM : Expectation Maximization
GAM : Genel Arkaplan Modeli
GKM : Gauss Karışım Modeli
GMM : Gauss Mixture Model
HFD : Hızlı Fourier Dönüşümü
JFA : Joint Factor Analysis
KB : Konuşmacı Belirleme
KD : Konuşmacı Doğrulama
KSFD : Kısa Süreli Fourier Dönüşümü
LLR : Logarithmic Likelihood Ratio
MAP : Maximum a Posteriori
MFCC : Mel – Frequency Cepstral Coefficients
MFKK : Mel – Frekansı Kepstrum Katsayıları
MGDC : Modified Group Delay Cepstral Coefficients
MLE : Maximum Likelihood Estimation
MLP : Çok Katmanlı Algılayıcı
MODGD : Modified Group Delay
NIST : Natıonal Institute of Standards and Technology
RASTA : Göreceli Spektrum Yöntemi
SMM : Saklı Markov Modeli
SRE : Speaker Recognition Evaluation
SVM : Support Vector Machine
UBM : Universal Background Model
VN : Vektör Nicemleme
YSA : Yapay Sinir Ağları

SEMBOLLER

α_i	: Destek vektör katsayısı
c	: Kanal süpervektörü
$K(:,:)$	: Çekirdek fonksiyonu
M	: Konuşmacıya ve kanala bağlı süpervektörü
m, v, d	: JFA modeli hiperparametreleri
μ	: Gauss karışım ortalaması
μ_i	: i. gauss bileşenine ait ortalama matrisi
$p_i(x)$	: Gauss olasılık yoğunluk fonksiyonu
s	: Konuşmacı süpervektörü
U	: Özkanal matrisi
ω	: Gauss karışım ağırlığı
$\omega(n)$	: Hamming pencere fonksiyonu
ω_i	: i. gauss bileşenine ait ağırlık matrisi
$X_n(\omega)$	: Kısa süreli fourier dönüşümü
$\sum$	: Kovaryans matrisi
$\tau(\omega)$	: Grup gecikme fonsiyonu
σ_i	: i. gauss bileşenine ait kovaryans matrisi
λ_{GAM}	: GAM modeli
λ_{HDF}	: Hedef konuşmacı modeli

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

TÜRKÇE SESLER İLE KONUŞMACI KİMLİĞİNİN DOĞRULANMASI/BELİRLENMESİ

ÖZET

Konuşmacı tanıma sistemleri son yıllarda oldukça popüler hale gelen ancak üzerinde uzun süredir çalışılmasına rağmen hala istenilen performans başarısı elde edilmemiş bir örüntü tanıma problemidir. Konuşmacı tanıma sistemleri, sesli aramadan telefon bankacılığına, çağrı merkezlerinden adli uygulamalara kadar bir çok alanda aktif olarak kullanılmaktadır. Konuşmacı tanıma alanında yapılan çalışmalar, genellikle İngilizce sesler kullanılarak oluşturulan veritabanlarından elde edilen sonuçları göstermektedir. Türkçe sesler kullanılarak oluşturulan veritabanları ile yapılan çalışmalar az sayıda olduğundan dolayı literatürde bilinen ve uygulanan başarılı yöntemlerin Türkçe sesler üzerindeki performansları hala belirsizdir. Bu sebepten dolayı bu tezde konuşmacı tanıma uygulamalarında literatürde çok sık kullanılan sınıflandırma yöntemlerinden, Gauss Karışım Modeli - Genel Arkaplan Modeli (Gaussian Mixture Model - Universial Background Model), Gauss Karışım Modeli - Destek Vektör Makinaları (Gaussian Mixture Model - Support Vector Machine), Birleşik Etmen Analizi (Joint Factor Analysis - JFA), i-vektör yaklaşımı yöntemleri kullanılarak Türkçe metne bağlı konuşmacı doğrulama sistemi üzerindeki başarı performansları incelenmiştir. Kullanılan sınıflandırma yöntemlerinde Mel - Frekansı Kepstrum Katsayıları (Mel - Frequency Cepstral Coefficients) ve Değiştirilmiş Grup Gecikme Kepstrum Katsayıları (Modified Group Delay Cepstral Coefficients) kullanılarak iki farklı öznitelik yönteminin de konuşmacı tanıma sistemi üzerindeki performans etkisi karşılaştırmalı olarak incelenmiştir.

GKM-GAM, GKM-DVM ve JFA sınıflandırıcıları ile yapılan deneysel çalışmalarda 46 konuşmacıdan oluşan Türkçe veritabanı kullanılırken i-vector yaklaşımı kullanılarak yapılan deneysel çalışmalarda ise 59 konuşmacıdan oluşan veritabanı kullanılmıştır. Ayrıca, GKM-GAM sınıflandırıcısıyla yapılan deneylerde, Türkçe sesler ve İngilizce sesler kullanılarak eğitilen arkaplan sesleriyle sistemin dil uyumu arasındaki bağlantının sistem üzerindeki etkisi incelenmiştir.

GKM-GAM, GKM-DVM, JFA, i-vektör sınıflandırıcıları ile yapılan deneysel çalışmalarda MFKK ve DGKK olmak üzere, farklı boyutlardaki öznitelik sayılarının ve farklı sayıdaki gauss bileşenlerinin sistem üzerindeki etkisi de karşılaştırmalı olarak ele alınmıştır.

Deneysel sonuçlara göre sınıflandırıcılar içerisinden en düşük sistem hatasına sahip olan en başarılı sınıflandırıcı % 4,62 EER değeriyle GKM-GAM sınıflandırıcısı olarak bulunmuştur. Aynı zamanda öznitelik yöntemlerinden MFKK özniteliklerinin DGKK özniteliklerine kıyasla sistem üzerinde daha başarılı sonuçlar verdiği gözlenmiştir.

Anahtar kelimeler: Türkçe konuşmacı doğrulama, Gauss karışım modeli, genel arkaplan modeli, birleşik etmen analizi, i-vektör, mel – frekansı kepstrum katsayıları.

VERIFICATION/IDENTIFICATION OF SPEAKER IDENTITY WITH TURKISH VOICES

SUMMARY

Speaker recognition is a pattern recognition problem which has become very popular in recent years but it does not achieve the desired performance although long work on it. Speaker recognition systems are actively used in many areas, from voice calls to telephone banking, from call centers to forensic applications. Studies in the field of speaker recognition generally report the results obtained from databases consisting of English recordings. Because of the less number of studies conducted with the databases created by using Turkish voices, the performances of the aplied and known successful methods on Turkish voices are still uncertain. For this reason, in this thesis, the performance on the Turkish text-based speaker verification system was investigated using Gaussian Mixture Model - Universal Background Model (GMM - UBM), Joint Factor Analysis (JFA) and i-vector approach which are well known methods in speaker recognition systems. In the used classification methods, Mel - Frequency Cepstrum Coefficients and Modified Group Delay Cepstral Coefficients were used as the features and the in performance on the speaker recognition system was analyzed comparatively.

In the experimental studies conducted with GMM-UBM, GMM-SVM and JFA classifiers, the Turkish database consisting of 46 speakers was used, while in the experimental studies using the i-vector approach, the database consisting of 59 speakers was used. In addition, in the experiments conducted with the GMM-UBM classifier, the effects of connection between the background sounds trained by using Turkish and English recordings and system's language compatibility on the system were examined.

In the experimental studies conducted with GMM-UBM, GMM-SVM, JFA, i-vector classifiers, the effect of different number of features and Gaussian components on the system has been discussed comparatively.

According to the experimental results, the most successful classifier having the lowest system error among the classifiers was found as GMM-UBM classifier with the value of 4,62% EER. Besides, it was observed that the MFCC features of the yield better performance on the system than the MODGD features.

Keywords: Turkish speaker verification, Gaussian mixture model, universal background model, joint factor analysis, i-vector, mel – frequency cepstral coefficients.

1. GİRİŞ

Konuşma, insanlar arasındaki en doğal iletişim şekillerinden birisidir. Ses sinyali, parmak izi, retina yapısı ve el geometrisi gibi bireylere özgü biometrik özellikler taşır. Ses sinyali sadece iletilmek istenilen mesaj yada kelime bilgisini değil aynı zamanda konuşan kişinin kimliği hakkında bilgi içerir. Ses sinyali, konuşmacının yaşı, cinsiyeti, duygu durumu gibi fiziksel özellikler hakkında da bilgi içerir [1].

Konuşma tanıma, konuşulan kelimenin anlamı ile ilgilenirken, konuşmacı tanıma ise konuşan kişinin kimliğiyle ilgilenir. Konuşmacı tanıma konuşan kişinin kimliğinin belirlenmesini hedefler [2,3].

Konuşmacı tanıma sistemleri, otomatik konuşmacı tanıma sistemi, yarı otomatik konuşmacı tanıma sistemi ve işitsel konuşmacı tanıma sistemi olmak üzere üç şekilde oluşturulabilir. İşitsel ve yarı otomatik konuşmacı tanıma sistemlerinde kimlik tespiti, uzman bir kişi tarafından yapılırken, otomatik konuşmacı tanıma sistemlerinde kimlik tespiti, uzman bir kişinin müdahalesi olmadan bilgisayarlar aracılığıyla yapılır [4].

Konuşmacı tanıma sisteminde, doğal olarak üretilen ve doğada halihazırda bulunan ses işaretinin varlığı sisteme önemli bir avantaj sağlar. Diğer bir avantajı ise, ses sinyalinin kaydı için basit bir mikrofonun yeterli olmasından kaynaklanır. Diğer bir deyişle, yüksek maliyetli özel bir ekipmana ihtiyaç duyulmadan ses kayıt altına alınabilir [5].

Otomatik konuşmacı tanıma uygulamaları, telefon bankacılığından sesli aramaya, çağrı merkezlerinden bilişim sistemlerine, güvenlik kontrol sistemlerinden adli uygulamalara kadar birçok alanda kullanılmaya başlanılmıştır [2,3]. Bu yüzden son zamanlarda ses araştırmacıları bu konular üzerinde yoğunlaşmıştır.

Bu tezde, Türkçe seslerin kullanıldığı metne bağlı otomatik konuşmacı tanıma sistemi geliştirilmiştir.

Literatürde Türkçe sesler ile oluşturulan veritabanlarıyla yapılan çalışmalar az sayıda olduğundan dolayı bilinen başarılı ve etkili yöntemlerin Türkçe seslerdeki performansları belirsizdir.

Bu yüzden bu tezde konuşmacı tanıma sistemlerinde sık kullanılan fakat Türkçe seslerdeki başarı performansları belirsiz yöntemler olan, Gauss Karışım Modeli - Genel Arkaplan Modeli (GKM - GAM), Gauss Karışım Modeli - Destek Vektör Makinaları (GKM - DVM), Birleşik Etmen Analizi (JFA), i-vector yaklaşımı yöntemleri gibi sınıflandırıcılar kullanılarak, Türkçe metne bağlı otomatik konuşmacı tanıma sistemi üzerindeki başarı performansları karşılaştırmalı olarak ele alınmıştır.

Bu tezde, Türkçe seslerle oluşturulan veritabanı kullanılarak kapalı küme, metne bağlı otomatik konuşmacı tanıma sistemi geliştirilmiştir. Sistem, Matlab programı kullanılarak hazırlanmıştır.

Kısaca, bu tezde literatürde bilinen ve kullanılan yaygın sınıflandırıcı algoritmaların Türkçe sesler ile birlikte kullanıldığı durumda performansları analiz edilerek karşılaştırılmıştır. Ayrıca, farklı boyutlarda MFKK ve DGKK öznitelikleri kullanılarak her birinin konuşmacı tanımaya etkisi incelenmiş ve sistem için en iyi parametre değerleri bulunmuştur.

Yapılan deneysel sonuçlara göre, kullanılan sınıflandırıcıların Türkçe sesler üzerindeki konuşmacı tanıma performanslarının oldukça iyi olduğu gözlenmiştir.

2. LİTERATÜR ÖZETİ

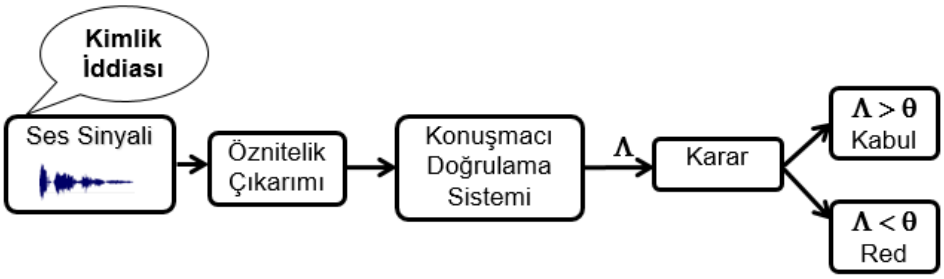
2.1 Konuşmacı Tanımanın Tarihsel Gelişimi

Konuşmacı tanıma problemi ilk kez 1660 yılında bir suçluyu tespit etmek amacıyla ele alınmıştır. Konuşmacı tanımada, ilk bilimsel yaklaşım ise 1930'da Lindbergh'in oğlunun kaçırılmasında kullanılmıştır. 1940'da Potter'in ses spektrograf icadıyla konuşmacı tanıma sistemi için önemli bir adım atılmıştır. 1960'da Kersta ses spektrograflarından konuşmacıların kimlik tespitinin yapılabileceğini ifade ederek Amerikan mahkemelerinde suçlu tespiti için kullanılmıştır [6]. Pruzansky, filtre bankalarını kullanarak ve benzerlik ölçütü için iki dijital spektrogramı ilişkilendirerek araştırma yapmıştır. Pruzansky, Mathews ve Li bu teknik üzerinde çalışmıştır. Doddington ise filtre bankaları yerine formant analiz yöntemini kullanmıştır ancak konuşmacıların ses özelliklerinin değişkenlik göstermesi, konuşmacıların tespit edilmesindeki en ciddi sorunlardan birisi olduğundan Endres ve arkadaşları, Furui, bu problemi ele almışlardır [7]. 1960 ve 1970 yılları arasında araştırmacılar, sesin fonetik özelliklerinden bağımsız olarak konuşmacılara ait öznitelikleri elde etmek amacıyla, ortalama oto-korelasyon, doğrusal öngörü katsayıları gibi birçok istatistiksel metot geliştirdiler [7]. 1980'li yıllarda SMM gibi istatistiksel yaklaşımlar kullanılmaya başlanıldı ve sistem tanıma performasında önemli bir gelişme sağlandı. 1980'lerde yapılan diğer önemli gelişmelerden birisi de normalizasyon işleminin önerilmesiydi. Bu teknikle birlikte sistem, tanıma aşamasında karar vermek için bir eşik değeri belirleyerek, öznitelik parametrelerinin metne bağlı değişkenliklerinden kaynaklanan olabilirlik oranına göre normalize edilmiş ve sistemin tanıma performansı arttırılmıştır [6].

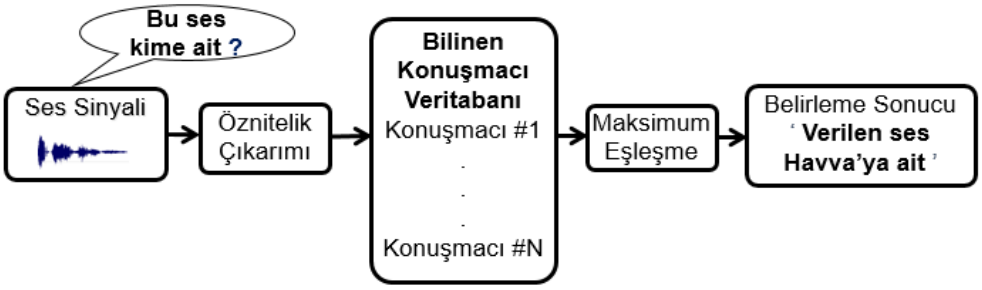
2.2 Konuşmacı Tanıma Sistemi

Konuşmacı tanıma, konuşan kişiye ait bilgileri içeren ses sinyalinden konuşanı otomatik olarak tanımlama sürecidir. Otomatik konuşmacı tanıma sistemi, erişim sistemlerinde kişilerin kimliğinin doğrulanmasına olanak sağlar [3].

Konuşmacı tanıma sistemleri, en genel haliyle konuşmacı doğrulama (KD) ve konuşmacı belirleme (KB) olarak iki sınıfa ayrılır. Kullanıcının kimlik iddiasında bulunduğu, sistemin ise verilen ses örneğine göre bu iddiayı kabul edip etmediği sistem *konuşmacı doğrulama sistemi* olarak bilinir [8]. Konuşmacı belirleme sistemi ise, bilinmeyen bir kişiye ait ses örneğinin *N* adet bilinen konuşmacıdan hangisine ait olduğunu belirleyen sistemdir [9]. Şekil 2.1'de konuşmacı doğrulama sisteminin genel yapısı gösterilmiştir. Şekil 2.2'de konuşmacı belirleme sisteminin genel yapısı gösterilmiştir.



Şekil 2.1 : Konuşmacı doğrulama sistemi



Şekil 2.2 : Konuşmacı belirleme sistemi

Konuşmacı doğrulama ve konuşmacı belirleme arasındaki temel fark karar verme aşamasındaki seçenek sayısıdır. Konuşmacı belirlemede karar verme süreci, sistemdeki konuşmacı sayısıyla ilgilidir. Konuşmacı doğrulamada ise karar verme süreci, sistemdeki kayıtlı konuşmacı sayısına bakmaksızın kabul etme veya reddetme olmak üzere iki şekildedir [2]. Bu yüzden, konuşmacı belirlemede sistemde kayıtlı olan kişi sayısı arttıkça sistem performansı azalırken, konuşmacı doğrulamada sisteme kayıtlı kişi sayısındaki artış performansı olumsuz şekilde etkilemez [2].

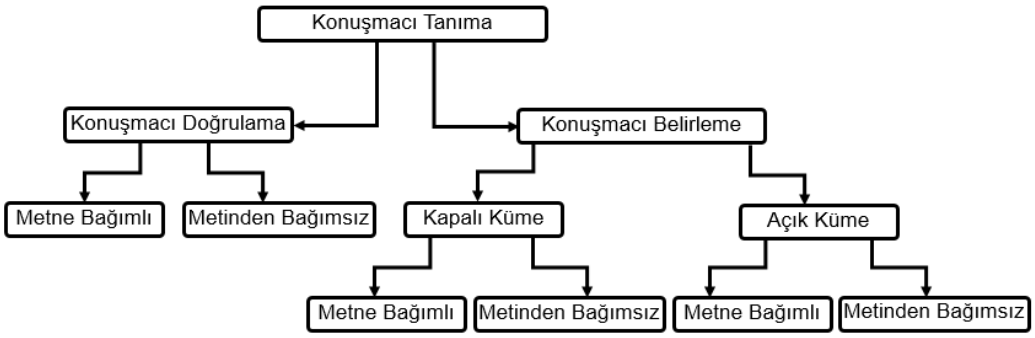
Konuşmacı tanıma sistemleri, sistemin yapısına göre genel olarak üç grup altında toplanmaktadır. Bunlar sırasıyla, metne bağlı sistemler, metinden bağımsız sistemler

ve sistemin kullanıcıdan o anda söylemesini istediği kelime ya da cümleyi bildiren sistemlerdir [9].

Metne bağlı sistemlerde isminden de anlaşıldığı gibi kullanıcı önceden belirlenmiş olan sabit bir metni tekrar eder. Sistem tarafından belirlenmiş olan metin tüm kullanıcılar için aynıdır.

Metinden bağımsız sistemlerde önceden belirlenmiş belli bir metin sınırlaması yoktur. Konuşmacının ses örneğinden kimliğinin belirlenmesi amaçlanır [10]. Metne bağlı sistemler, metin bilgisinin önceden bilinmesinden ve kelime sayısının azlığından dolayı diğerlerine göre performans başarısı yönünden daha üstündür. Bu sebepten dolayı, yüksek başarı performansı istenen konuşmacı tanıma uygulamalarında metne bağlı sistemler tercih edilmektedir [10].

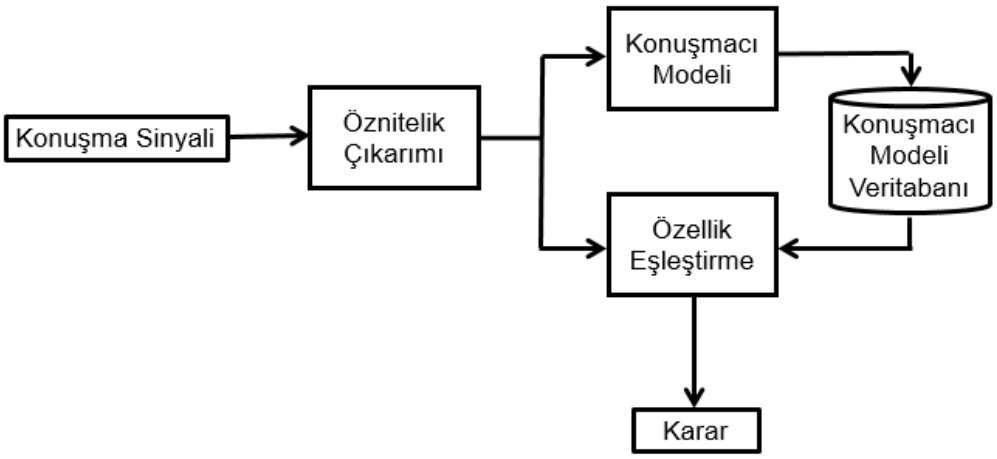
Konuşmacı tanıma sistemleri, açık küme ve kapalı küme olmak üzere iki gruba ayrılır. Kapalı kümede bilinmeyen ses, sistemde kayıtlı olan konuşmacılardan birisine aitken, açık kümede bilinmeyen ses, sistemde kayıtlı olmayabilir. Açık küme ve kapalı küme arasındaki temel fark sisteme erişilip erişilememesiyle ilgilidir. Açık küme konuşmacı tanıma sisteminde, sistemde olmayan hatta yanıltmak isteyen kötü niyetli kişiler sisteme erişim yapabilirken kapalı kümede bu mümkün değildir [2]. Şekil 2.3'te konuşmacı tanıma sisteminin hiyerarşik yapısı gösterilmiştir.



Şekil 2.3 : Konuşmacı tanıma sistemi hiyerarşik yapısı

Konuşmacı tanıma sistemleri, eğitim aşaması ve test aşaması olmak üzere iki farklı aşamadan oluşur. Eğitim aşamasında her konuşmacı için ayrı bir model oluşturulur ve eğitim cümleleriyle konuşmacılar sisteme tanıtılırlar. Bu işlem modelleme olarak bilinmektedir. Test aşamasında ise bilinmeyen ses sinyali referans konuşmacı modelleriyle karşılaştırılarak test sinyalinin hangi konuşmacıya ait olduğu saptanır. Bu işleme sınıflandırma denir [11,12].

Şekil 2.4'de genel bir konuşmacı tanıma sistemi yapısı gösterilmiştir.



Şekil 2.4 : Genel konuşmacı tanıma sistemi

2.3 Öznitelik Çıkarımı

Ses sinyali, konuşmacıyı ve konuşmayı temsil eden büyük oranda bilgi taşır. Bu taşıdığı bilgiden konuşmacıyı temsil edebilecek yeterli bilginin elde edilmesi ve çıkarılması işlemine öznitelik çıkarma işlemi adı verilmektedir.

Öznitelik çıkarma, ses sinyalinden elde edilen fazla miktarda veriyi azaltarak konuşmacıyı ayırt edecek olan ses özelliklerinin korunmasını sağlamaktadır [13]. Öznitelik çıkarma işlemi, tüm konuşmacı tanıma sistemlerinde uygulanan ortak en temel işlemdir [14].

Öznitelik çıkarımı, iki temel nedenden dolayı önemlidir. Bunlardan birincisi, istatistiksel konuşmacı modellerinin gürbüz olması için, ikincisi ise eğitim örneklerinin sayısının ölçülebilir boyutlarda olması içindir. Böylece, işlem fazlalığı da azaltılmış olur [15].

2.3.1 Öznitelik seçimi

Öznitelikler, özellik vektörü yada öznitelik vektörü olarak adlandırılır. Konuşmacıyı temsil eden özellik vektörleri konuşmacı tanıma uygulamalarında oldukça önemli bir yere sahiptir.

Öznitelik seçiminde, nispeten daha düşük boyutlu bir vektör uzayına dönüşüm sağlamak amaçlanır ve önemli bir bilgi kaybı olmadan bu dönüşüm gerçekleştirilir [16]. Öznitelik seçiminde ideal öznitelikler, konuşmacıyı tanımaya yardımcı olabilecek özellikler taşımalıdır.

Bu özellikler şunlardır [1,17]:

- Kolay hesaplanabilir veya ölçülebilir olmalı
- Konuşma anında sıkça oluşmalı ve tabii olarak ortaya çıkmalı
- Taklide karşı dayanıklı olmalı
- Gürültüden ve konuşmacının sağlık koşullarından etkilenmemeli

Konuşma anında sıkça oluşmalı ve doğal olarak ortaya çıkmalıdır. Böylece konuşmacıya ait öznitelikler daha kolay elde edilir. Kolay hesaplanamayan öznitelikler, konuşmacı tanıma sisteminde kullanılmaz yada az kullanılır. Sistem güvenliği için taklide karşı dayanıklı olmalıdır. Konuşmacıdan elde edilen öznitelikler ile konuşmacı tanıma sistemi, gürültüden ve konuşmacının ses değişimlerinden etkilenmeden daima o kişiyi doğru tanıyabilmelidir.

2.3.2 Öznitelik seçim yöntemleri

Öznitelik seçiminde birçok yöntem kullanılmaktadır. Bunlardan bazıları aşağıda belirtilmiştir.

2.3.2.1 Doğrusal öngörü katsayıları

Doğrusal öngörü katsayıları, konuşmacı tanıma sistemlerinde en çok kullanılan öznitelik çıkarma yöntemlerinden birisidir [18]. Doğrusal öngörü katsayıları, o anki ses işaretinin her bir örneğinin geçmiş p adet örneğin ağırlıklandırılmış doğrusal toplamı şeklinde temsil edilir.

DÖK yönteminde amaç, o andaki konuşma ile doğrusal olarak öngörülen konuşma arasındaki karesel hatayı minimum yapan öngörü katsayılarını bulmaktır [19].

Öngörü katsayıları, zamana bağlı olarak değişir. Bir sonraki konuşma örneği, doğrusal öngörü geçmiş örneklerinin ağırlıklandırılmış toplamıdır [19]. Doğrusal öngörü katsayıları, ses yolunu modellemektedir [20].

2.3.2.2 Kepstral katsayılar

Kepstrum, ses sinyalinin uygun bir pencere ile ağırlıklandırılmasıyla elde edilir. Şekil 2.5'te kepstral katsayılarının elde edilişi gösterilmektedir.

Şekil 2.5 : Kepstrum katsayılarının elde edilmesi

Şekil 2.5'te görüldüğü gibi, ses sinyali pencere fonksiyonundan geçirilir. Pencerelenen ses sinyaline ayrık fourier dönüşümü uygulanır. Alınan ses sinyalinin frekans domeninde genliklerinin logaritması alınır. Logaritma sonucu elde edilen değerin ters ayrık fourier dönüşümü alınarak ses sinyalinin kepstral değerleri elde edilir.

2.3.2.3 Formant frekansları

Seslerdeki farklılıklar, ses yolundaki bileşenlerin boyut ve şekil değişikliklerinden kaynaklanmaktadır. Formant frekansları, konuşmacıya ait ses yolu özelliklerini taşır ve ses yolunun rezonans frekansları olarak tanımlanmaktadır [21].

2.3.2.4 Göreceli spektrum yöntemi (RASTA)

Göreceli spektrum yöntemi, öznitelik çıkarımında kullanılan yöntemlerden birisidir. Konuşma içerisindeki gürültü gibi çevresel etkilerin modellenmesine dayanır. Göreceli spektrum yöntemi, insan kulağının bir sesi algılamasının daha önceki seslerden ne derecede etkilendiğidir. İnsan kulağındaki algılama, şu andaki ses ile bir önceki ses arasındaki spektral farka bağımlıdır. Böylece insan kulağı hızlı değişen seslere nazaran, yavaş değişen seslere daha az duyarlıdır [22].

2.3.2.5 Mel frekansı kepstrum katsayıları (MFKK)

MFKK öznitelikleri ilk kez 1980'li yıllarda Davis ve Mermelstein tarafından konuşmacı tanıma için önerilmiştir. MFKK öznitelikleri, yüksek frekans bölgesindeki önemsiz spektral bilgileri baskıladığı için DÖK özniteliklerine göre daha çok avantaj sağlar. MFKK öznitelikleri, insan kulağındaki algısal faktörler dikkate alınarak geliştirilmiştir [14].

İnsan kulağı, 1 kHz' e kadar doğrusal, 1 kHz'den sonra logaritmik olarak değişen algısal bir yapıya sahiptir. Bu sebeple MFKK, insan kulağının algısal bant genişliği ve frekansındaki değişimleri, 1 kHz'e kadarki düşük frekanslarda doğrusal süzgeç kullanarak, 1kHz'den sonrasını logaritmik olarak değişen süzgeçler kullanarak modellemeyi amaçlar. Kullanılan bu ölçek, *mel frekansı ölçeği* olarak adlandırılır [18]. Mel ölçeği şu şekilde ifade edilir ;

$$\text{mel}(f) = 2595 \log(1+\frac{f}{700}) \quad (2.1)$$

Denklem 2.1'de görüldüğü gibi mel ölçeği, 1 kHz'e kadar doğrusal, 1 kHz'den sonra ise logaritmik olarak değişen aralıklarla ifade edilen bir ölçektir.

Son yıllarda MFKK öznitelikleri yüksek başarı performansından dolayı konuşmacı tanıma ve konuşmacı tanıma uygulamalarında en yaygın kullanılan öznitelik çıkarım yöntemlerinden birisi olmuştur. MFKK özniteliklerinin elde edilmesi detaylı olarak Bölüm 3'te anlatılacaktır.

2.3.2.6 Değiştirilmiş grup gecikme kepstrum katsayıları (DGKK)

Ses sinyalinin spektral olarak gösterilmesi Fourier dönüşümü genlik ve faz spektrumları ile temsil edilir. Genel olarak konuşmacı tanıma sistemlerinde öznitelikler, kısa süreli genlik spektrumu kullanılarak elde edilir. Oysa, Fourier dönüşümü faz spektrumu da ses sinyaliyle ilgili önemli bilgilere sahiptir [23]. Bu sebepten dolayı, konuşmacılara ait özniteliklerin elde edilmesinde Fourier dönüşüm fazı da kullanılabilir. Ses sinyalinden doğrudan hesaplanabilen bu yöntem *Grup Gecikme Fonksiyonu* olarak adlandırılır. Grup gecikme fonsiyonu kullanılarak, ses sinyalinden formant bilgisi, perde frekansı gibi bilgiler elde edilebilir. Bu sebeple kepstral öznitelikler, Değiştirilmiş Grup Gecikme Fonksiyonu kullanılarak elde edilebilir. Bu öznitelik çıkarım yöntemi *Değiştirilmiş Grup Gecikme Kepstrum Katsayıları (DGKK)* olarak adlandırılmıştır.

Matematiksel olarak grup gecikmesi kepstral katsayıları, ayrık fourier dönüşümünün faz spektrumunun negatif türevinden elde edilir [23]. Bu yöntem daha sonra Bölüm 3'te ayrıntılı olarak ele alınacaktır.

2.4 Sınıflandırma Yöntemleri

Konuşmacı tanıma uygulamalarında, çeşitli sınıflandırma yöntemleri kullanılmaktadır. Genellikle sınıflandırma yöntemleri, şablon temelli modeller ve istatistiksel temelli modeller olmak üzere iki gruba ayrılır. Metne bağlı sistemler şablon temelli modeller kullanırken, metinden bağımsız sistemler istatistiksel temelli modeller kullanır [24]. İstatistiksel modellerde model eşleşmesi olasılıksal iken şablon temelli modellerde model eşleşmesi deterministiktir [16].

2.4.1 Şablon temelli modeller

Şablon temelli modellerde, bilinmeyen ses sinyali en iyi eşleşmeyi bulmak için önceden kaydedilmiş olan eğitme şablonlarıyla karşılaştırılır. Test cümlesine ait sözcükler, eğitme şablonları ile karşılaştırılarak, en uygun eşleşme şablonu seçilir ve tanıma işlemi gerçekleştirilir. Şablon temelli modeller, zamana bağımlı yada zamandan bağımsız olabilir [25]. Dinamik zaman eğirme ve vektör nicemleme en çok kullanılan yöntemlerdir.

2.4.1.1 Dinamik zaman eğirme (DZE)

Aynı sözcüğü veya cümleyi aynı konuşmacı birkaç kez tekrarladığında dahi seslendiriliş bir öncekiyle aynı hızda ve aynı zamanda olmayabilir, bir önceki seslendirilişe benzemeyebilir. Dinamik zaman eğirme, konuşma hızındaki değişkenlikleri telafi etmek için kullanılan en popüler şablon temelli yöntemlerden birisidir [16].

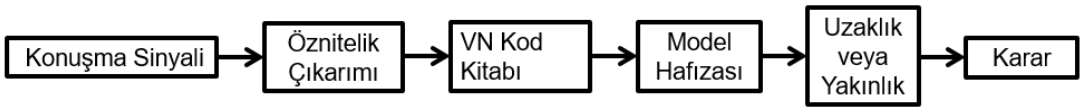
Dinamik zaman eğirme, zaman veya hız bakımından değişebilen iki durum arasındaki benzerliği ölçmek için kullanılan bir algoritmadır [25]. Dinamik zaman eğirme yöntemi, dinamik programlama tekniği kullanılarak gerçekleştirilir. Dinamik programlama için, $M = m_1, m_2, \ldots, m_i$ ve $N = n_1, n_2, \ldots, n_j$ iki farklı zaman dizisinin başlangıçlarını aynı olarak kabul edersek, m_i ve n_j zamanlarının çakıştırılması amaçlanır. Bu sebeple, en iyi eşleşmeyi bulmak için bir fonksiyon tanımlanır. Böylelikle dinamik zaman eğirme, bütün şablonlar arasından optimum eşleşme uzaklıklarını bulur ve en küçük uzaklık veren şablonun hangisi olduğuna karar verilir. Dinamik zaman eğirme, genellikle metne bağlı konuşmacı tanıma uygulamalarında kullanılmaktadır [16].

2.4.1.2 Vektör nicemleme (VN)

Vektör nicemleme, bir eğitim setinden kod sözcükleri olarak bilinen bir dizi vektörü oluşturmak için kullanılan parametrik olmayan veri sıkıştırma veya veri azaltma yöntemidir. Bu sınıflandırma tekniği, güdümsüz (denetimsiz) bir eğitim algoritmasıdır [26]. Vektör nicemleme algoritması, en yakın komşu algoritmasını kullanarak yakın olan vektörleri aynı sınıfa, birbirinden uzak olan vektörleri ise farklı bir sınıfa dahil ederek bir kümeleme işlemi yapar. Eğitim öznitelik vektörlerinden oluşan M adet ayrı küme bir araya gelerek konuşmacı modelleri oluşturulur. Bu kümelere hücre adı

verilir. Bu hücreler, her birinin ortalama vektörleri alınarak kod vektörlerine dönüştürülür. Oluşan kod vektörlerine kod kitabı denir. Tanıma aşamasında ise , giriş ses işareti ile kod kitabındaki referans modeller karşılaştırılarak uzaklık veya yakınlığa göre karar verilir [27].

Vektör nicemleme yöntemi, metne bağlı ve metinden bağımsız konuşmacı tanıma uygulamalarında kullanılabilmektedir [26]. Şekil 2.6'da VN ile konuşmacı tanıma sisteminin temel yapısı gösterilmiştir.



Şekil 2.6 : Vektör nicemleme ile konuşmacı tanıma sistemi temel yapısı

2.4.2 İstatistiksel temelli modeller

İstatistiksel temelli bir sınıflandırıcı modelinde, konuşmacı modelleri olasılığa dayalı olarak oluşturulur. Bir başka deyişle konuşmacı modelleri, olasılık dağılımı kullanılarak modellenir ve sınıflandırmayı olasılığa göre yapar. Metinden bağımsız konuşmacı tanıma uygulamalarında genellikle istatistiksel temelli modeller kullanılmaktadır [28].

Konuşmacılar, ortalama özniteliklerinden ziyade olasılık dağılımları ile modellenir ve sınıflandırma kararlarını ortalama özniteliklere olan mesafeden ziyade olasılıklar üzerine dayandırmayı ifade eder.

Konuşmacıların dağılımlarının bilindiği ve sürekli yoğunluklara sahip oldukları varsayılarak, bir *x* özniteliğinin *i.* konuşmacı tarafından üretilme olasılığı $p_i(x)$' tir. Bayes kuralını kullanarak, konuşmacının *i.* konuşmacı olma olasılığı ;

$$p(x) = \frac{p_i(x)P_i}{p(x)} \tag{2.2}$$

Denklem 2.2'de olduğu gibi ifade edilir. Denklem 2.2'de görüleceği üzere $p_i(x)$, *x* özniteliğinin *i.* konuşmacıdan üretilme olasılığını, P_i, ses sinyalinin *i.* konuşmacıdan olma olasılığını, $p(x)$, herhangi bir konuşmacıdan üretilen özniteliğin olasılık değerini ifade eder.

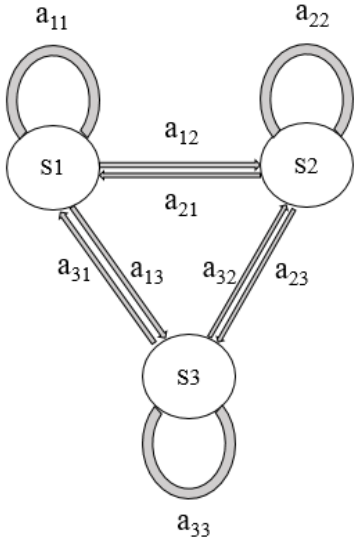
Genellikle her bir konuşmacıya ait önsel olasılıklar eşit olarak kabul edilir. Böylece $p(x)$, konuşmacı olasılık yoğunluklarının ortalaması olarak ifade edilir.

$$p(x) = \sum_{i=1}^{I} p_i(x)P_i \qquad (2.3)$$

Denklem 2.3'de belirtilen I, konuşmacı sayısını ifade eder [29]. I adet kişiyle eğitilen sistemde I adet konuşmacı modeli oluşturulur. Test cümlesi ile sinyalin, I adet kişiden en yüksek olasılığa sahip olan hangi konuşmacıya ait olduğu bulunur.

2.4.2.1 Saklı markov modeli (SMM)

Saklı Markov Modeli hakkındaki temel teori 1960-1970 yılları arasında Baum ve arkadaşları, Baker, Jelinek ve arkadaşları tarafından ortaya atılmıştır. Saklı Markov Modeli son yıllarda oldukça yaygın bir şekilde kullanılmaktadır. Bunun iki temel sebebi vardır. İlk olarak, SMM yöntemi matematiksel yapı bakımından oldukça zengindir ve dolayısıyla geniş bir uygulama alanına, teorik bir temel oluşturur. İkinci olarak ise SMM yöntemi doğru ve uygun bir şekilde uygulandığında başarılı sonuçların elde edilmesine olanak sağlar [30]. SMM, Markov zinciri ve stokastik süreç olmak üzere iki temel unsurdan oluşur. SMM, Markov zincirinin olasılıksal bir fonksiyonudur. Markov süreci, *N* adet sonlu durumdan oluşan $S = \{S_1, S_2, \ldots, S_N\}$ ile ifade edilen ve her durumda ses sinyalinin özellik vektörüne ait olasılık yoğunluk fonksiyonunu kapsayan bir süreçtir. SMM'de sonlu olan durumlar birbirine durum geçiş olasılıkları ile bağlıdır. Her *t* anında *i* durumundan *j* durumuna geçiş rastgele olarak gerçekleşir ve durum geçiş olasılığı a_{ij} ile ifade edilir [31]. Şekil 2.7'de üç durumlu ergodik bir SMM modeli gösterilmiştir.



Şekil 2.7 : Ergodik SMM modeli

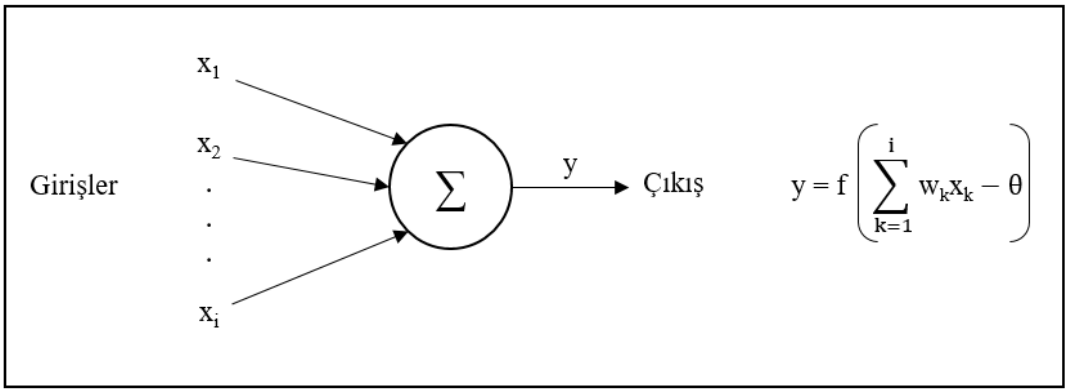
Şekil 2.7'de görüldüğü gibi, S1,S2,S3 üç durumlu bir SMM yapısını ifade eder ve a_{12} 1 durumundan 2 durumuna rastgele geçiş olasılığını ifade eder.

SMM istatistiksel bir yaklaşımdır. SMM, literatürde metne bağlı konuşmacı tanıma uygulamalarında [2] ve metinden bağımsız konuşmacı tanıma uygulamalarında yaygın olarak kullanılmıştır [32].

2.4.2.2 Yapay sinir ağları (YSA)

Yapay Sinir Ağları (YSA), son yıllarda başta konuşmacı tanıma olmak üzere birçok farklı alanda kullanılmaktadır. YSA, yapay sinir hücrelerinin birbiriyle farklı şekilde bağlanmasıyla oluşur. Aslında YSA modeli, insan sinir sistemi yapısı göz önüne alınarak modellenmiştir ancak mevcut en iyi modeller dahi insana ait performansa eşit değildir.

Sinir ağı modelleri, değişken ağırlıklara sahip bağıntılarla birbirine bağlanan pek çok hesaplama parametresinden oluşarak, yüksek bir hesaplama tekniği gerektirir. Yapay sinir ağında kullanılan hesaplama elemanları doğrusal değildir [33]. Yapay bir nöron temel olarak, *i* ağırlıklı girişleri toplayarak bir çıkış üretir. Şekil 2.8' de yapay bir nöron yapısı gösterilmiştir.



Şekil 2.8 : Yapay nöron yapısı

Şekil 2.8'de görüldüğü üzere, y çıkış fonksiyonu i ağırlıklı girişlerin toplamından oluşur ve θ değeri ise eşik değerini temsil etmektedir.

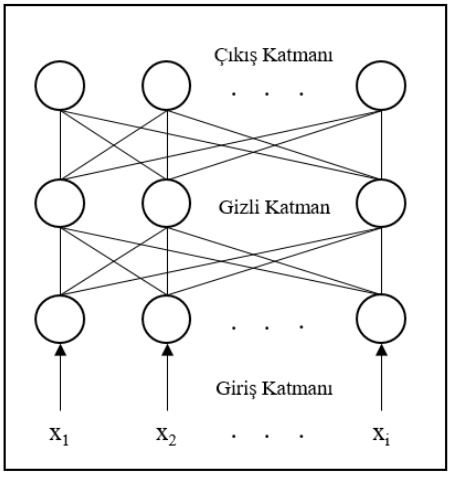
Yapay bir nöron, girişler, ağırlıklar, toplama fonksiyonu, transfer fonksiyonu ve çıkıştan oluşur.

Yapay sinir ağlarının birçok türü vardır. Bunlardan bazıları şunlardır:

- Radyal Temelli Fonksiyon Ağları [34,35]

- Dalgacık Ağları (WN) [36]
- Çok Katmanlı Algılayıcı (MLP) [37]
- Öğretici Vektör Nicemleyici [38]

Çok katmanlı algılayıcı (MLP), yapay sinir ağlarının en sık kullanılan türlerinden biridir. MLP, ileri beslemeli bir sinir ağıdır. Bir MLP, giriş katmanı, belli sayıda gizli katman ve çıkış katmanı olmak üzere üç farklı katmandan oluşur. MLP'ler genellikle yinelemeli bir gradyan algoritmasıyla eğitilir [39]. Şekil 2.9'da MLP'nin genel yapısı gösterilmiştir.



Şekil 2.9 : MLP genel yapısı

Şekil 2.9'da görüldüğü üzere MLP modeli, üç katmandan oluşmaktaktadır ve bir katmandaki bütün nöronlar bir üst katmandaki nöronlara bağlıdır. Bilgi akışı ileri doğrudur. Giriş işareti, giriş nöronları tarafından bir yada birden fazla olan gizli katman nöronlarına, gizli katmandan da çıkış nöronlarına aktarılır. Her bir çıkış nöronu bağlı olduğu sınıf için olasılık bilgisini taşır. Sonuç olarak, giriş işareti en yüksek olasılık bilgisini taşıyan sınıfa atanır.

2.4.2.3 Gauss karışım modeli (GKM)

GKM, konuşmacı tanıma sistemlerinde en sık kullanılan en popüler yöntemlerdendir [40]. GKM yöntemi ilk kez Reynolds tarafından kullanılmıştır [1,41]. GKM yöntemi metinden bağımsız konuşmacı tanıma uygulamalarında oldukça iyi sonuçlar vermektedir. GKM yönteminde gauss bileşenleri konuşmacıya ait sesi temsil etmektedir [40].

Bir GKM, M adet gauss bileşeninden oluşan olasılık dağılımının ağırlıklandırılmış toplamından elde edilmek üzere şu şekilde tanımlanır:

$$p(x|\lambda) = \sum_{i=1}^{M} \omega_i \, p_i(x) \tag{2.4}$$

Denklem 2.4'te görüldüğü üzere $p_i(x)$, ortalaması μ_i ve kovaryans matrisi $\sum_i$ olan i. gauss karışım bileşenini, ω_i, i. gauss bileşeninin karışım ağırlığını, x, konuşmacıya ait öznitelik vektörünü temsil etmektedir.

D-boyutlu bir gauss karışım bileşeni matematiksel olarak şu şekilde ifade edilir;

$$p_i(x) = \frac{1}{2\pi^{D/2}\left|\sum_i\right|^{1/2}} \exp\left\{-\frac{1}{2}\,(x-\mu_i)'(\sum_i)^{-1}(x-\mu_i)\right\} \tag{2.5}$$

$$\sum_{i=1}^{M} \omega_i = 1 \tag{2.6}$$

Denklem 2.6'da ifade edildiği üzere, ağırlık katsayıları toplamı bir olmak zorundadır. Bir konuşmacıya ait GKM modelinin parametreleri, şu şekilde ifade edilir:

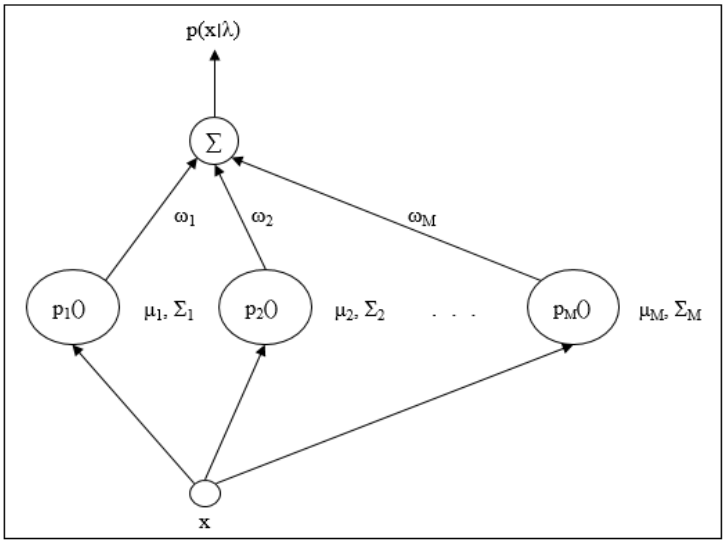
$$\lambda = \{\omega_i, \mu_i, \sum_i\} \;,\; i = 1,\ldots,M \tag{2.7}$$

Denklem 2.7'de görüldüğü üzere, μ_i, gauss olasılık yoğunluk fonksiyonu $p_i(x)$'in ortalamasını ifade etmektedir.

GKM modeli kullanılan bir konuşmacı tanıma sistemindeki her konuşmacı bir GKM modeli ile temsil edilir [40].

Sistemin eğitim aşamasında, GKM model parametreleri elde edilmeye çalışılır. GKM modelinin parametrelerinin elde edilmesi işlemi, yinelemeli bir algoritma yöntemi olan *Beklentinin Maksimumlaştırılması (EM)* algoritması kullanılarak gerçekleştirilir. Konuşmacıya ait parametreler *EM* yöntemi sonucunda elde edilerek, konuşmacı modellerindeki sistem parametreleri eğitilir.

Aslında GKM modeli, tek durumlu bir SMM modeli olarak görülebilir [41]. Şekil 2.10'da M bileşenli Gauss karışım yoğunluğunun yapısı gösterilmiştir.



Şekil 2.10 : M bileşenli Gauss karışım yoğunluğunun gösterimi [40].

GKM, konuşmacı tanıma uygulamalarında oldukça yüksek doğrulukta başarı performansı göstermektedir [1,40]. Bu sebeple, bu tezde kullanılan sınıflandırma yöntemlerinden birisi de GKM sınıflandırıcısıdır ve model Bölüm 3'te ayrıntılı olarak ele alınacaktır.

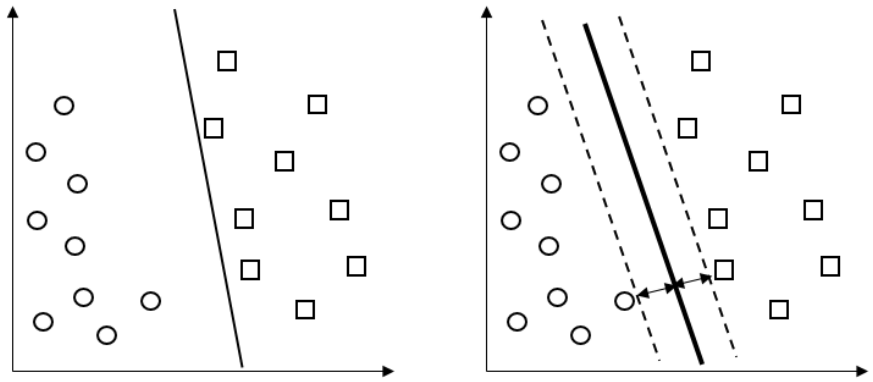
2.4.2.4 Destek vektör makinesi (DVM)

DVM teorisi ilk kez Vapnik tarafından ortaya atılmıştır ve *Yapısal Risk Azaltma* teorisinden geliştirilmiştir [42]. Destek vektör makineleri (DVM), son yıllarda oldukça sık kullanılan güçlü bir sınıflandırıcı türüdür.

DVM, iki sınıfı birbirinden ayırarak örüntü tanıma problemlerinde sık kullanılan yöntemlerden birisi olmuştur. Konuşmacı tanımada DVM, klasik sınıflandırma yöntemlerinin aksine gerçek ve sahte kimlikler arasındaki sınırı belirleyerek modelleme yapar [43].

DVM sınıflandırıcısı, bu iki sınıfa ait öznitelik vektörlerinin birbirinden en iyi şekilde ayrılmasını sağlayan doğruyu bulmayı amaçlar. Bu iki sınıfa ait öznitelik vektörlerine en uzak olan doğruyu bulur [44].

Şekil 2.11'de iki sınıfı birbirinden ayırmak için DVM tarafından oluşturulan doğrular gösterilmiştir. Öznitelik vektörleri Şekil 2.11'de görüldüğü gibi, her zaman doğrusal olarak dağılmayabilir. Bu durumda öznitelikler, çekirdek olarak bilinen matematiksel fonksiyonlar kullanılarak daha yüksek boyutlu olan başka bir uzaya taşınır [44]. Bu fonksiyon Denklem 2.8'de verilmiştir.



(a) (b)

Şekil 2.11 : (a) İki sınıflı veriyi ayıran doğru (b) En iyi doğru.

Çekirdek fonksiyonu şu şekilde ifade edilir [43]:

$$f(x)=\sum_{i=1}^{N} \alpha_i t_i K(x,x_i) + d \tag{2.8}$$

Denklem 2.8'de ifade edildiği üzere, K (:,:) fonksiyonu, çekirdek fonksiyonunu, x_i , destek vektörlerini, t_i , ideal sınıf etiketlerini, N, toplam destek vektör sayısını, α_i , $\sum_{i=1}^{N} \alpha_i t_i = 0$ ve $\alpha_i > 0$ olmak koşuluyla destek vektör katsayılarını ifade etmektedir [43]. Bu sınıflandırıcı ile ilgili daha detaylı bilgi Bölüm 3'te verilecektir.

2.4.2.5 Birleşik etmen analizi (Joint Factor Analysis – JFA)

Birleşik Etmen Analizi (JFA), son yıllarda Amerikan Ulusal Standartlar ve Teknoloji Enstitüsü (NIST) tarafından yapılan Konuşmacı Tanıma Değerlendirme (Speaker Recognition Evaluation - SRE) yarışmalarında, metinden bağımsız konuşmacı tanıma sistemindeki en iyi performansı göstermiştir [45,46,47].

Birleşik etmen analizi, GKM yönteminde olan kanal ve konuşmacının oturum farklılıklarından kaynaklı hataları telafi etmek amacıyla kullanılan bir yöntemdir. Bu modelde her konuşmacı, F, öznitelik vektörlerinin boyutu ve C, genel arkaplan modelinin bileşen sayısı olmak üzere, karışım ağırlığı, kovaryans matrisi ve karışım ortalaması olan çok değişkenli gauss karışım yoğunluğu ile temsil edilir.

Hedef konuşmacı için GKM modeli, Genel Arkaplan Modeli (GAM)'ın ortalama parametrelerinin uyarlanmasıyla elde edilir. Birleşik etmen analizinde, **M** şu şekilde ifade edilir [48]:

$$\mathbf{M} = \mathbf{s} + \mathbf{c} \tag{2.9}$$

Denklem 2.9'da görüldüğü gibi, M, aslında herhangi bir konuşmacıyı temsil eden bir süpervektördür. **M**, konuşmacıya ve kanala bağlı süpervektörlerden oluşur. Denklem 2.9'da s, konuşmacı süpervektörünü, **c**, kanal süpervektörünü temsil etmektedir. Bu vektörler normal dağılımlı vektörlerdir [49].

Konuşmacıya bağlı süpervektörlerin (**s**) dağılımının gizli bir değişken olduğunu varsayarsak konuşmacı süpervektörü şu şekilde ifade edilir:

$$\mathbf{s} = \mathbf{m} + \mathbf{Vy} + \mathbf{Dz} \tag{2.10}$$

Denklem 2.10'da m, GAM'dan elde edilen konuşmacı ve kanaldan bağımsız bir süpervektörü, **V**, düşük ranklı bir dikdörtgen matrisi, **D**, köşegen matrisi, **y** ve **z**, standart gauss dağılımına sahip rastsal vektörleri ifade etmektedir [48]. Bu vektörlerden **y** vektörü konuşmacıya, **z** vektörü ise ortak etmene bağlıdır ve her birinin normal dağılıma $N(0, I)$ sahip rastsal değişken olduğu varsayılır. Konuşmadaki kanal etkisini temsil eden **c** süpervektörü şu şekilde ifade edilir:

$$\mathbf{c} = \mathbf{Ux} \tag{2.11}$$

Denklem 2.11'de görüldüğü gibi, **U**, düşük ranklı bir dikdörtgen matrisi (özkanal matrisi), **x**, standart gauss dağılımına sahip olan bir vektörü temsil etmektedir ve **x**'in bileşenleri JFA'daki kanal etmenleridir. JFA'daki temel amaç büyük bir eğitim seti kullanarak, hiper parametreler olan **V, D, U** parametrelerini eğitmektir. Eğitim aşamasında, Bayes kuramı çerçevesinde birleşik etmenlerin sonsal dağılımları hesaplanabilir. Test aşamasında, test cümlesi X'in olasılığı ise **x**'in önsel dağılımı ile **y** ve **z**'nin sonsal dağılımları birleştilerek elde edilir [48]. Bu model ile ilgili daha detaylı bilgi Bölüm 3'te verilecektir.

2.4.2.6 i - vektör modeli

Birleşik etmen analizi yönteminde kanal etmenlerinin sadece kanal etkilerini modelleyebileceği varsayılmıştı ancak Dehak 2009'da bitirdiği doktora çalışmasında kanal etmenlerinin sadece kanal etkilerini değil aynı zamanda konuşmacılarıda modelleyebildiğini göstermiştir.

Kanala bağlı süpervektörlerin konuşmacılardan elde edilen öznitelikleri de modelleyebildiği görülmüştür. Bunu gerçekleştirmek için yeni bir model ortaya atılmıştır [49].

Bu yeni modelde, daha önce Denklem 2.11'de tanımlanan **M**, konuşmacıya ve kanala bağlı olarak yeniden şu şekilde ifade edilir:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\boldsymbol{\omega} \tag{2.12}$$

Denklem 2.12'de görüldüğü gibi, M, konuşmacıya ve kanala bağlı süpervektörü, **m**, genel arkaplan modelinin ortalama vektörlerinin uç uca eklenmesiyle elde edilen süpervektörü, **T**, konuşmacıları modelleyen düşük ranklı bir matrisi, **ω**, ses sinyalinden elde edilen i-vektör özniteliğini temsil etmektedir.

i - vektör modeli, ses sinyalinden oluşturulan öznitelik vektörlerinin tek bir vektöre dönüştürüldüğü sabit uzunluklu bir vektördür. Eğitim aşaması ve test aşamasında, bütün konuşmacıların ses sinyalleri kullanılarak i - vektör öznitelikleri elde edilir daha sonra bu öznitelik vektörleri kullanılarak karar için skor hesaplaması yapılır ve konuşmacı tespit edilir [50].

3. MATERYAL VE YÖNTEM

Konuşmacı tanıma alanında yapılan çalışmalarda genellikle İngilizce sesler kullanılarak oluşturulan veritabanları ile elde edilen sonuçlar gösterilmektedir. Türkçe sesler kullanılarak oluşturulmuş veritabanları ile yapılan çalışmalar oldukça az sayıda olduğundan dolayı literatürde bilinen, etkili ve başarılı yöntemlerin Türkçe seslerdeki performansları belirsizdir. Bu yüzden, bu tezde Türkçe seslerden oluşturulan Türkçe veritabanı kullanılarak bilinen başarılı yöntemlerin Türkçe seslerdeki performansları karşılaştırmalı olarak ele alınmıştır.

Konuşmacı tanıma sistemi, öznitelik çıkarımı ve konuşmacı modeli oluşturma olmak üzere iki temel kısımdan oluşmaktadır. Bu tezde, MFKK ve DGKK olmak üzere iki farklı öznitelik çıkarım yöntemi kullanılmış ve bu yöntemlerin konuşmacı doğrulama sistemine etkisi analiz edilerek performansları karşılaştırılmıştır.

Türkçe seslerle, GKM - GAM, GKM - DVM, JFA, i-vektör yöntemleri kullanılarak oluşturulan konuşmacı modellerinin konuşmacı doğrulama sistemine etkisi analiz edilerek performansları karşılaştırılmıştır.

GKM-GAM, GKM-DVM ve JFA sınıflandırıcıları ile yapılan deneysel çalışmalarda 46 konuşmacıdan oluşan Türkçe veritabanı kullanılırken, i-vektör yaklaşımı kullanılarak yapılan deneysel çalışmalarda ise 59 konuşmacıdan oluşan Türkçe metne bağlı konuşmacı tanıma veritabanı kullanılmıştır.

Bu bölümde sırasıyla veritabanı özellikleri, öznitelik çıkarımı yöntemleri, konuşmacı doğrulama ve olabilirlik oranı, maksimum olabilirlik ve parametre tahmini ve sınıflandırma yöntemleri ele alınacaktır.

3.1 Veritabanı

GKM-GAM, GKM-DVM ve JFA sınıflandırıcıları ile yapılan deneysel çalışmalarda, 10 bayan ve 36 erkek olmak üzere 46 konuşmacıdan oluşan Türkçe veritabanı kullanılmıştır. Konuşmacılar, sırasıyla "*benim parolam ses kaydımdır*" şeklindeki sabit bir cümleyi tekrarlamıştır. Her bir konuşmacı farklı oturumlarda "*benim parolam*

ses kaydımdır" cümlesini on beş defa tekrarlamıştır. Ses kayıtları arasında bir hafta zaman dilimi vardır ve kayıt işlemi 1 ay boyunca sürmüştür. Konuşmacılardan her birinin eğitilmesi 3 ses kaydı ile gerçekleştirilmiştir. Diğer ses kayıtları, sistemin test aşamasında kullanılmıştır. Çizelge 3.1'de gösterildiği gibi, veritabanında 1690 adet doğru deneme (genuine/target) ve 54080 adet yanlış deneme (impostor/non-target) olacak şekilde toplam 55770 adet deneme verisi ile deneysel çalışmalar yapılmıştır.

Çizelge 3.1 : Konuşmacı doğrulama veritabanı 1.

Konuşmacı Sayısı	Doğru Deneme	Yanlış Deneme	Toplam Deneme
46	1690	54080	55770

i-vektör sınıflandırıcısı ile yapılan deneysel çalışmalarda ise 17 bayan ve 42 erkek olmak üzere toplam 59 kişinden oluşan Türkçe veritabanı kullanılmıştır. Konuşmacıların tümü "*benim parolam ses kaydımdır*" şeklindeki cümleyi tekrarlamıştır. Konuşmacılardan her birinin eğitilmesi 3 ses kaydı ile gerçekleştirilmiştir. Diğer ses kayıtları, sistemin test aşamasında kullanılmıştır. Çizelge 3.2'de görüldüğü gibi veritabanında yer alan konuşmacı sayısı, doğru deneme (genuine/target) ve yanlış deneme (impostor/non-target) sayıları detaylı olarak ifade edilmiştir.

Çizelge 3.2 : Konuşmacı doğrulama veritabanı 2.

	Konuşmacı Sayısı	Doğru Deneme	Yanlış Deneme
Bayan	42	1321	54161
Erkek	17	530	8480
Toplam	59	1851	62641

3.2 Öznitelik Çıkarımı

Ses sinyali, zamanla değişen ve durağan olmayan bir sinyaldir. Bu nedenle, genellikle 20-30 ms uzunluğundaki çerçeve adı verilen kısa parçalara bölünür. Bu aralık içinde sinyalin zamanla değişmediği, sabit kaldığı varsayılır ve her parçadan konuşmacıya ait spektral özellikler elde edilir.

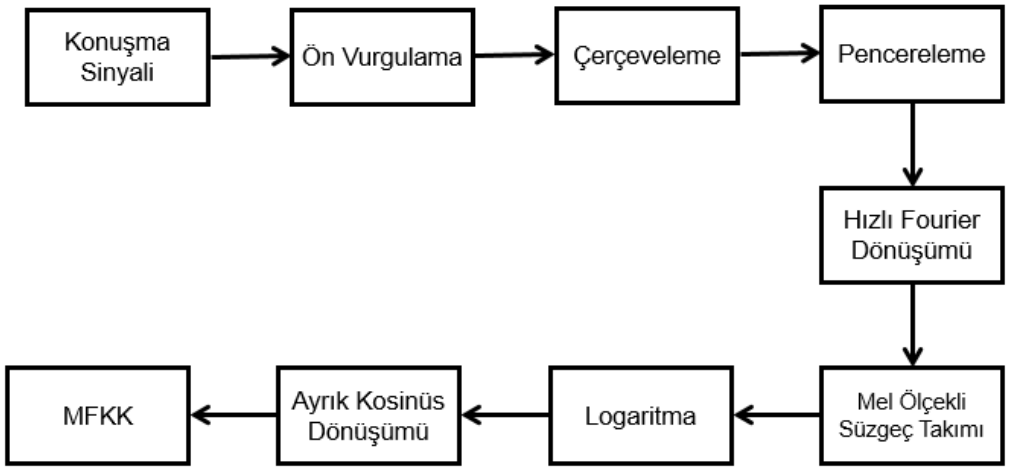
Ses sinyalinin bu şekilde kısa parçalar halinde işlenmesine *kısa dönem analizi* adı verilmektedir. Kısa dönem analizinde ses işaretine, genellikle, ilk olarak ön vurgulama (pre-emphasize) işlemi yapılır.

Kısa dönem analizinde ön vurgulama, çerçeveleme ve pencereleme olarak yapılan ön işlemler tüm öznitelik çıkarma yöntemleri için ortaktır [8].

Bu tezdeki deneysel çalışmalar sırasında konuşmacı tanıma sistemleri için kullanılan en popüler öznitelik yöntemlerinden birisi olan MFKK öznitelikleri ve DGKK öznitelikleri kullanılmıştır.

3.2.1 Mel frekansı kepstrum katsayıları

MFKK, konuşmacı tanımada en çok bilinen ve kullanılan yöntemdir [3]. MFKK özniteliklerinin elde edilmesi Şekil 3.1'de gösterilmektedir.



Şekil 3.1 : MFKK özniteliklerinin elde edilmesinin blok şeması

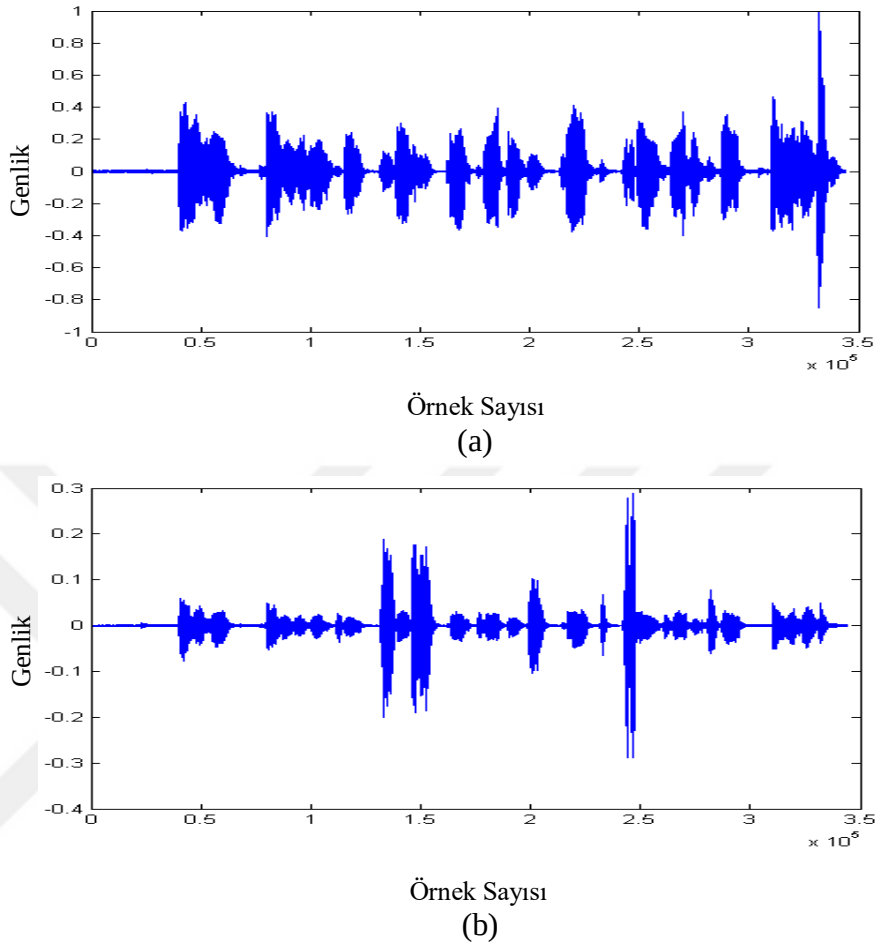
Şekil 3.1'de gösterildiği gibi, ses sinyaline öncelikle ön-vurgulama (pre-emphasis) filtresi uygulanır ve sonra ses sinyali 15 ms'lik kısımları örtüşen 30 ms uzunluğunda çerçeveler halinde bölünür. Daha sonra çerçevelenmiş ses sinyali Hamming pencere fonksiyonu ile çarpılır. Bu çerçevelerin hızlı fourier dönüşümü alınır ve genlik spektrumu hesaplanır. Hesaplanan genlik spektrumu, mel ölçekli üçgen süzgeçlerden geçirilir. Süzgeç çıkışlarının logaritması alınır ve ayrık kosinüs dönüşümü uygulanarak böylelikle spektral uzaydan kepstral uzaya geçilir. Böylece MFKK öznitelik vektörleri oluşturulur.

3.2.1.1 Ön vurgulama

Ön vurgulamada amaç, bir sinyalin yüksek frekans bileşenlerini daha baskın hale getirmektir ve genellikle konuşmacı tanıma uygulamalarında bir ön işlem olarak yapılır [51]. Ön vurgulama süzgecinin transfer fonksiyonu şu şekildedir;

$$H(z) = 1 - 0.97\, z^{-1} \tag{3.1}$$

Şekil 3.2 (a)'da orjinal ses sinyalini (b)'de ön vurgulama işlemi yapıldıktan sonra süzgeç çıkışından elde edilen ses sinyalini göstermektedir.



Şekil 3.2 : (a) Orijinal ses sinyali, b) Ön vurgulama filtresi uygulanan ses sinyali.

3.2.1.2 Çerçeveleme

Çerçeveleme işlemi sırasıyla şu şekilde gerçekleşir;

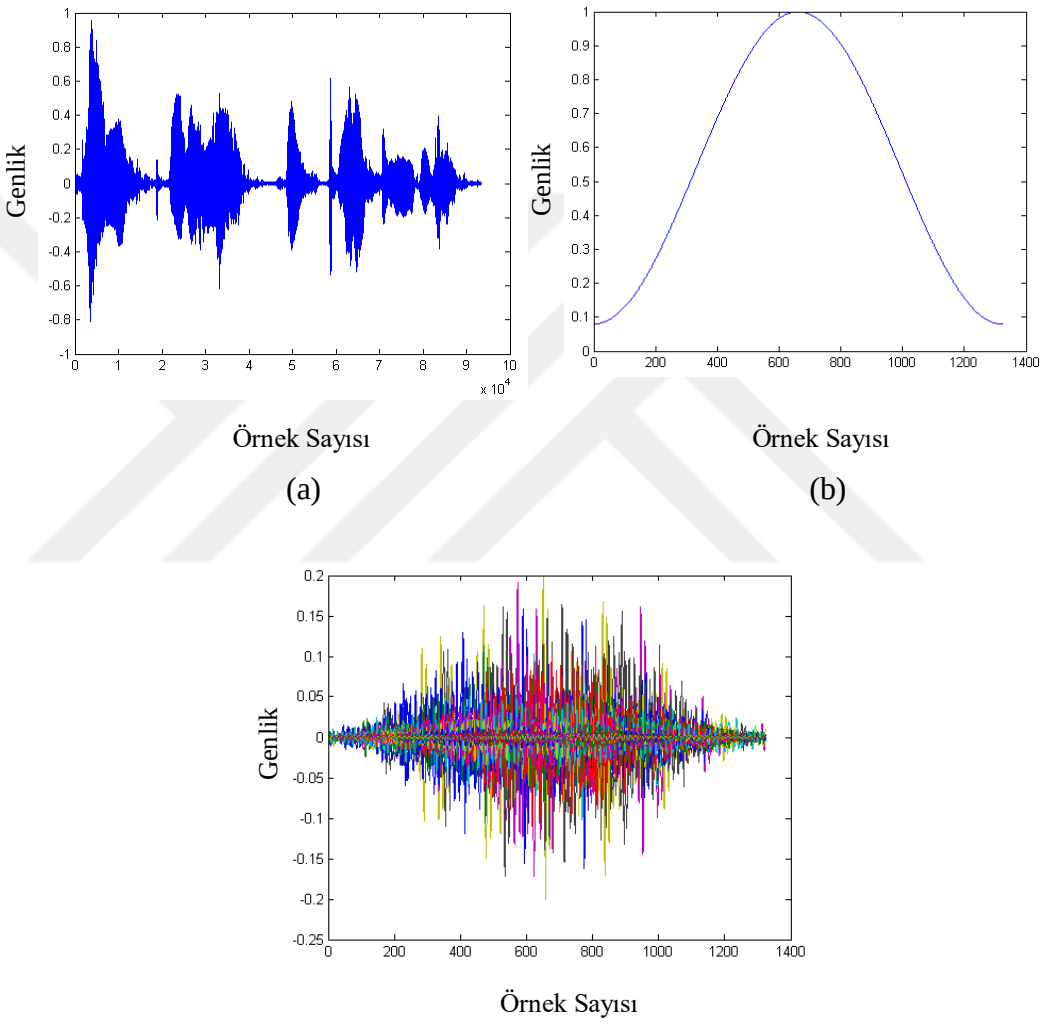
- Ses sinyali, N adet örnekten oluşan çerçevelere ve M adet örtüşen komşu örneklerden oluşan çerçevelere bölünür (M<N).
- İlk çerçeve N adet örnekten oluşan çerçeve ile başlar.
- İkinci çerçeve ilk çerçeveden M adet örnek sonra başlar ve ilk çerçevenin üzerine N-M kadar örtüşür [18].

Bu tezde, 15 ms'lik kısımları örtüşen 30 ms'lik çerçeveler kullanılmıştır. Böylece, örtüşme ile çerçeve sonundaki bilgi kaybı önlenmiş olur.

3.2.1.3 Pencereleme

Ses sinyaline uygulanan çerçeveleme işleminden sonra pencereleme işlemi yapılır. Pencereleme işlemindeki amaç, çerçeveleme işleminden kaynaklı spektral bozulmaları önlemek ve sinyalde oluşacak süreksizlikleri gidermektir [18].

Şekil 3.3 (a)'da orjinal ses sinyali, (b)'de hamming pencere fonksiyonu (c)'de ise hamming pencere fonksiyonu ile pencerelenmiş sinyal gösterilmiştir.



Şekil 3.3 : (a) Orijinal ses sinyali, b) Hamming pencere fonksiyonu, (c) Hamming pencere fonksiyonu ile pencerelenmiş ses sinyali

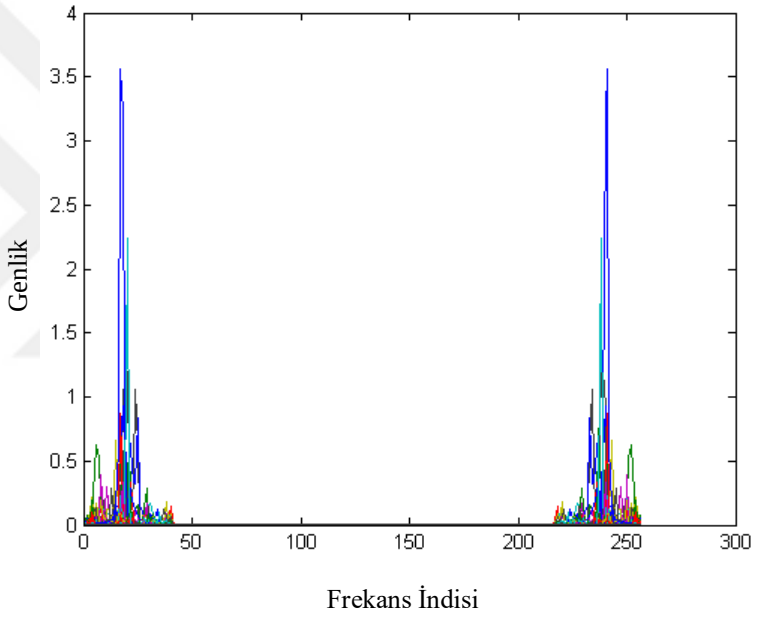
Yaygın olarak kullanılan birçok farklı pencereleme fonksiyonu vardır. Bu tezde Hamming pencere fonksiyonu kullanılmıştır. $1 \leq n \leq N$ olmak üzere, Hamming pencere fonksiyonu şu şekildedir;

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad (3.2)$$

Denklem 3.2'de ifade edilen *N*, çerçeve uzunluğunu ifade etmektedir.

3.2.1.4 Hızlı fourier dönüşümü (HFD)

N adet örnekten oluşan ses sinyalinin zaman domeninden frekans domenine dönüştürmek amacıyla HFD yapılır. Pencerelenen sinyalin genlik spektrumu HFD ile hesaplanır. HFD, ayrık fourier dönüşümünden türetilmiştir [52]. Şekil 3.4'te ön vurgulama yapılarak hamming pencere fonksiyonu ile çarpılan ses sinyalinin HFD'si gösterilmiştir.



Şekil 3.4 : Pencerelenmiş sinyalin HFD alınmış hali

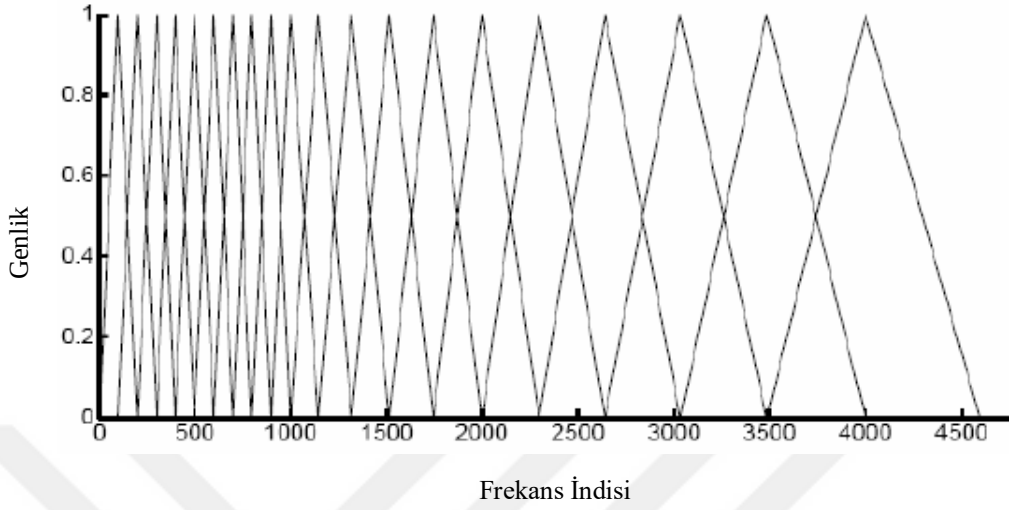
Şekil 3.4'de görüldüğü gibi, pencerelenmiş işaretin N/2'ye göre simetrik olduğu görülmektedir. Bu çalışmada N, 256 olarak alınmıştır.

3.2.1.5 Mel ölçekli süzgeç takımı

HFD ile genlik spektrumu hesaplanan ses sinyali, mel ölçekte yerleştirilmiş üçgen süzgeç takımından geçirilir. Mel ölçekli süzgeçler, 1 kHz'e kadar doğrusal, 1 kHz'den daha yüksek frekanslarda logaritmik aralıklarla değişmektedir.

Bu çalışmada, 27 adet mel ölçekte yerleştirilmiş üçgen süzgeç dizisi kullanılmıştır. Şekil 3.5'te mel ölçek yapısı gösterilmektedir.

Şekil 3.5'te görüldüğü gibi ilk 10 süzgeç, merkez frekansları bakımından doğrusal olarak, sonraki 10 süzgeç ise logaritmik aralıklarla yerleştirilmiştir. Her süzgecin bitiş noktası, kendinden bir sonraki süzgecin merkez frekansına bağlıdır [14].



Şekil 3.5 : Mel ölçekte dizilmiş üçgen süzgeç yapısı [14].

3.2.1.6 Logaritma alma

Ses sinyalinin süzgeç çıkışında elde edilen işaretin logaritması alınarak özniteliklerin dinamik değişimlere karşı daha az hassas olması sağlanır [53].

3.2.1.7 Ayrık kosinüs dönüşümü (AKD)

Logaritmik süzgeç çıkışlarına yapılan AKD ile frekans domeninden tekrar zaman domenine geçilerek MFKK katsayıları elde edilir.

3.2.2 Değiştirilmiş grup gecikme kepstrum katsayıları (DGGK)

Kısa Süreli Fourier Analizi, ses sinyalini işlemek için kullanılabilir. Kısa Süreli Fourier Dönüşümü (KSFD) şu şekilde ifade edilir;

$$X_n(\omega) = |X_n(\omega)|\, e^{-j\,\theta_n(\omega)} \tag{3.3}$$

Denklem 3.3'te görüldüğü gibi $|X_n(\omega)|$ kısa süreli genlik spektrumunu, $\theta_n(\omega)$, kısa süreli faz spektrumunu temsil etmektedir. Genlik spektrumunun karesi olan $|X_n(\omega)|^2$ ise kısa süreli güç spektrumu olarak adlandırılır.

Grup gecikmesi, Fourier dönüşümünün fazının negatif türevi olarak tanımlanır. Grup gecikme fonksiyonu matematiksel olarak şu şekilde ifade edilir;

$$\tau(\omega) = -\frac{d\left(\theta(\omega)\right)}{d\omega} \tag{3.4}$$

Denklem 3.4'te görüldüğü gibi, bir sinyalin faz spektrumu olan $\left(\theta(\omega)\right)$, ω'nın sürekli bir fonksiyonu olarak tanımlanır [23].

Sabit bir değerden sapan grup gecikme fonksiyonunun değerleri, fazın doğrusal olmama derecesini gösterir. Grup gecikme fonksiyonu Denklem 3.4 kullanılarak ses sinyalinden hesaplanabilir.

$$\tau_x(\omega) = -\mathrm{Im}\frac{d\left(\log\left(x(\omega)\right)\right)}{d\omega} \tag{3.5}$$

$$\tau_x(\omega) = \frac{X_R(\omega)Y_R(\omega)+Y_I(\omega)X_I(\omega)}{|X(\omega)|^2} \tag{3.6}$$

Denklem 3.6'da ifade edilen R ve I alt indisleri sırasıyla Fourier dönüşümünün reel ve imajiner kısımlarını ifade etmektedir. X(ω) ve Y(ω) sırasıyla, x(n) ve nx(n)'in Fourier dönüşümleridir [23].

Grup gecikme fonksiyonu, pencereleme ve gürültüden kaynaklı sıfırlar nedeniyle [54] ses sinyalinin fazının minimum veya transfer fonksiyonun kutuplarının birim çember içinde olmasını gerektirir [55]. Bu sebeple grup gecikme fonksiyonunda değişiklik yapılarak *Değiştirilmiş Grup Gecikme Fonksiyonu* elde edilir ve şu şekilde ifade edilir;

$$\tau_m(\omega) = \left(\frac{\tau(\omega)}{|\tau(\omega)|}\right)|\tau(\omega)|^{\alpha} \tag{3.7}$$

Denklem 3.7'de belirtilen τ(ω) şu şekilde ifade edilir;

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega)+Y_I(\omega)X_I(\omega)}{|S(\omega)|^{2\gamma}} \tag{3.8}$$

Denklem 3.8'de görüldüğü gibi, S(ω), |X(ω)|'nin yumuşatılmış (smoothed version) şeklidir. α ve γ parametreleri, doğal ses sinyalindeki ani artışlardaki genliğini azaltmak ve konuşma spektrumunun dinamik aralığını iyileştirmek için kullanılır.

α ve γ parametreleri, $0 < \alpha \leq 1.0$ ve $0 < \gamma \leq 1.0$ aralığında değişir [23]. Bu tezde farklı α, γ değerleri sırasıyla 0.3 ve 0.1 olarak alınmıştır.

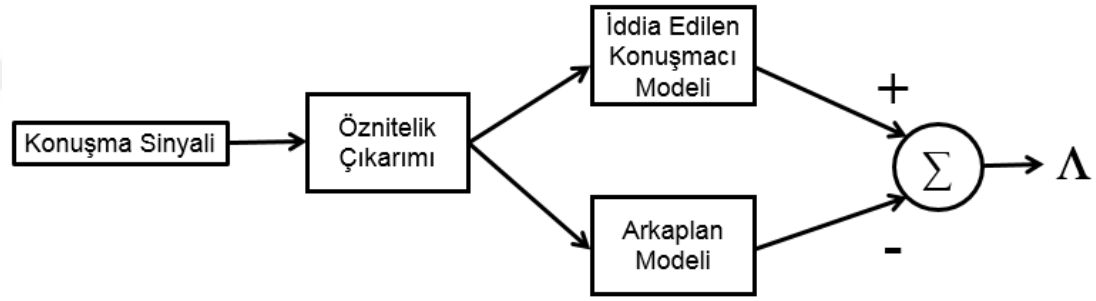
3.3 Konuşmacı Doğrulama ve Olabilirlik Oranı

Konuşmacı doğrulama daha önceki bölümlerde de bahsedildiği gibi, verilen bir ses sinyalinin iddia edilen kişiye ait olup olmadığına karar verme işlemidir. Verilen bir ses sinyali Y, ve iddia edilen konuşmacı S olmak üzere, konuşmacı doğrulama H_0 ve H_1'den oluşan temel bir hipotez testidir. Bu hipotez testi şu şekilde tanımlanır;

H_0 : Y ses sinyali iddia edilen S konuşmacısına aittir.

H_1 : Y ses sinyali iddia edilen S konuşmacısına ait değildir.

Şekil 3.6'da olabilirlik oran testi ile konuşmacı doğrulama sistemi gösterilmiştir.



Şekil 3.6 : Olabilirlik oran testi ile konuşmacı doğrulama

Bu iki hipotez arasından bir karar vermek için kullanılabilecek en uygun yöntem olabilirlik oran testidir. Olabilirlik oran testi şu şekilde gösterilir;

$$\frac{p(\mathrm{Y} \mid \mathrm{H}_0)}{p(\mathrm{Y} \mid \mathrm{H}_1)} \begin{cases} \geq \theta & \text{kabul } \mathrm{H}_0 \\ < \theta & \text{ret } \mathrm{H}_0 \end{cases} \tag{3.9}$$

Denklem 3.9'da görüldüğü gibi, $p(Y \mid H_i)$, i=0,1 olmak üzere, H_i hipotezi için ses sinyalinin hesaplanan olasılık yoğunluk fonksiyonunu ifade etmektedir.

H_i hipotezi, ses sinyalinin olabilirlik oranı olarakta ifade edilmektedir. θ ifadesi, H_0 hipotezi için karar vermek amacıyla kullanılan karar eşiğini temsil etmektedir. Bir konuşmacı tanıma sisteminde temel amaç, ses sinyalinin iddia edilen konuşmacıya ait kabul ve ret olabilirlik oranlarının belirlenmesidir.

Şekil 3.6'da olabilirlik oran testine göre konuşmacı doğrulama sistemindeki temel bileşenleri göstermektedir. İlk aşamada elde edilen konuşmacıya ait öznitelikler çıkartıldıktan sonra, konuşmacıyı ve ses sinyalini temsil eden $X=\{x_1,\ldots,x_T\}$ öznitelik vektör kümesi elde edilir. Bu öznitelik vektörleri, H_0 ve H_1 hipotezlerinin olabilirlik oranlarını hesaplamak için kullanılır.

Matematiksel olarak H_0 hipotezi, x öznitelik uzayında S konuşmacısını karakterize eden $\boldsymbol{\lambda_{hyp}}$ olarak adlandırılan bir modelle temsil edilir. Gauss karışım modeli, H_0 hipotezi için öznitelik vektörlerinin dağılımını temsil edebilen en uygun yöntemdir. Bu sebeple, $\boldsymbol{\lambda_{hyp}}$, ortalama vektörleri, karışım ağırlıkları ve kovaryans matrislerinden oluşan bir GKM'yi temsil eder.

Matematiksel olarak H_1 hipotezi ise x öznitelik uzayında S konuşmacısına ait olmamayı karakterize eden $\boldsymbol{\lambda_{\overline{hyp}}}$ olarak adlandırılan bir modelle temsil edilir. Olabilirlik oranı ise şu şekilde hesaplanır;

$$\frac{p(X\,|\lambda_{hyp})}{p(X\,|\lambda_{\overline{hyp}})} \tag{3.10}$$

Genellikle, logaritmik olabilirlik oranı kullanılmakta olup şu şekilde hesaplanır;

$$\Lambda(X) = \log p(X|\lambda_{hyp}) - \log p(X|\lambda_{\overline{hyp}}) \tag{3.11}$$

H_0 hipotezine ait λ_{hyp} modeli iyi tanımlanmış olup S konuşmacısına ait eğitim ses sinyalleri kullanılarak model parametreleri tahmin edilebilir olmasına rağmen H_1 hipotezine ait model olan $\lambda_{\overline{hyp}}$ modelini oluşturmak için S konuşmacısı dışında olası bütün ihtimallerin temsil edilmesi gerekir. Bunu gerçekleştirmek için iki farklı yaklaşım mevcuttur.

İlk yaklaşımda amaç fazla sayıda başka konuşmacı modelleri kullanarak $\lambda_{\overline{hyp}}$ modelini hesaplamaktır. İkinci yaklaşımda ise amaç, en çok kullanılan yaklaşım olmakla birlikte S konuşmacısından faklı olmak koşuluyla çok sayıda farklı konuşmacıların ses işaretlerinin kullanılarak $\lambda_{\overline{hyp}}$ modelini eğitmektir. Bu yöntem literatürde, *dünya modeli, genel model, genel arkaplan modeli (GAM – universial background model – UBM)* gibi farklı isimlerle bilinmektedir. Tanıma sırasında, konuşmacılardan elde edilen konuşma örneklerinin toplamıyla, $\lambda_{\overline{hyp}}$ modelini temsil etmek için tek bir λ_{GAM} modeli eğitilir.

Özetle GAM yönteminde, farklı konuşmacılardan alınan ses örnekleriyle elde edilen öznitelikler kullanılarak H_1 hipotezini temsil eden λ_{GAM} modeli oluşturulur. Daha sonra, S konuşmacısına ait eğitim ses örneklerinden elde edilen özniteliklerle H_0 hipotezine ait λ_{hyp} modelini ifade eden hedef konuşmacı modeli λ_{HDF}, GAM'dan en büyük sonsal olasılık (MAP) adaptasyonu ile oluşturulur [28].

3.4 Maksimum Olabilirlik ve Parametre Tahmini

GKM modeli, Denklem 2.6'da daha önce bahsedildiği gibi, $\lambda = \{\omega_i, \mu_i, \sum_i\}$, i = 1,…,M olmak üzere, ortalama vektörleri, karışım ağırlıkları ve kovaryans matrislerinden oluşan üç farklı parametreden oluşur.

GKM yönteminde konuşmacı modeli oluşturulurken, ses eğitim örneklerinden elde edilen öznitelik vektörlerinin dağılımına göre en uygun GKM parametrelerinin yani λ'nın tahmin edilmesi amaçlanır.

GKM parametrelerini tahmin etmek için uygulanabilir birçok farklı yöntem vardır. Bu yöntemlerden en sık kullanılan ve bilinen yöntem *En Büyük Olabilirlik Kestirimi (Maximum Likelihood Estimation - MLE)* yöntemidir. MLE'de, GKM'nin ses eğitim örneklerinden elde edilen konuşmacı öznitelik vektörlerinin dağılımının olabilirliğini en büyük yapan parametreleri bulmak amaçlanır. T adet eğitim vektöründen oluşan $X = \{\vec{x_1},\ldots,\vec{x_T}\}$ dizisi için GKM olabilirliği şu şekilde yazılabilir;

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\vec{x_t} \mid \lambda) \tag{3.12}$$

Denklem 3.12'de görülen ifade, λ parametrelerinin doğrusal olmayan bir fonksiyonudur ve doğrudan MLE yöntemi uygulanarak yapılamaz. Bundan dolayı, MLE yöntemi, yinelemeli bir algoritma olan *Beklentinin Maksimumlaştırılması (EM)* algoritmasını kullanarak λ parametrelerini tahmin eder.

EM algoritması, λ başlangıç modelini, $\bar{\lambda}$ yeni modeli veya bir sonraki modeli temsil etmek üzere $p(X \mid \bar{\lambda}) \geq p(X \mid \lambda)$ oluncaya kadar yinelemeli olarak devam eder. Yeni model bir sonraki iterasyon için başlangıç modeli olarak kabul edilir ve bu işlem yakınsama değerine ulaşıncaya kadar devam eder. Bu temel yaklaşım SMM parametrelerini tahmin etmek için kullanılan bir yöntem olan Baum - Welch algoritmasıyla aynıdır [40].

EM algoritması, gözlenen öznitelik vektörlerini *k* ve *k+1*'inci iterasyonlar için $p(X \mid \lambda^{k+1}) \geq p(X \mid \lambda^{k})$ yinelemesi için tahmin edilen GKM parametrelerini yinelemeli olarak düzenler. Genellikle 5 iterasyon parametrelerin yakınsaması için yeterlidir [28]. Bu tezde iterasyon sayısı olarak 10 seçilmiştir.

3.5 Sınıflandırma Yöntemleri

Öznitelik vektörlerinin elde edilmesinden sonraki aşamada, konuşmacıyı temsil eden özniteliklerle konuşmacı modeli oluşturulur. Bu bölümde eğitim ve test aşamasında konuşmacı modellerinde kullanılan sınıflandırma yöntemlerinden detaylı bir şekilde bahsedilmiştir.

3.5.1 GKM – GAM modeli

GKM – GAM modeli, $p(X \mid \lambda_{\overline{hyp}})$'ı ifade etmek amacıyla konuşmacıdan bağımsız olarak tek bir arkaplan modeli kullanılması temeline dayanır. GAM, konuşmacıdan bağımsız öznitelik vektörlerinin dağılımını ifade etmek amacıyla eğitilen büyük bir GKM modelidir. Konuşmacı modelinin uyarlandığı GKM – GAM modelinde, konuşmacıya ait eğitim ses özniteliklerini ve MAP adaptasyonunu kullanarak GAM parametrelerinin adaptasyonu yapılır ve sonuçta yeni bir hipotez konuşmacı modeli oluşturulur. Konuşmacı modelinin uyarlanmasındaki temel mantık, adaptasyon aracılığla GAM'daki iyi eğitilmiş parametreleri güncelleyerek konuşmacı modelini oluşturmaktır. Bu şekilde konuşmacı modeli ile GAM arasında sıkı bir bağ kurulmuş olur. Adaptasyon, EM algoritması gibi iki aşamadan oluşan bir tahmin sürecidir. İlk aşamada, GAM'daki her bir karışım için hesaplanan konuşmacının eğitim verisi için yeterli istatistiklerin (λ) hesaplandığı EM algoritmasının beklenti (expectation) aşaması ile aynıdır. İkinci aşamada ise EM algoritmasının aksine, adaptasyon için elde edilen yeni yeterli istatistik tahminleri, daha sonra verilere bağlı bir karışım katsayısı kullanılarak GAM karışım parametrelerinden gelen eski istatistiklerle birleştirilir. Veriye bağlı karışım katsayısı konuşmacıdan gelen fazla miktarda bilgi içeren karışımların, son parametre tahmini için yeni yeterli istatistiklere dayandığı ve konuşmacıdan gelen az veriye sahip karışımların son parametre için eski yeterli istatistiklere dayandığı şeklinde tasarlanmıştır. Verilen bir GAM modeli, λ_{GAM}, ve hipotez konuşmacıya ait eğitim öznitelik vektörleri $X = \{x_1,\ldots,x_T\}$ için GKM – GAM yöntemi ile konuşmacı modeli parametreleri elde edilirken ilk olarak, öznitelik vektörlerinin GAM modelindeki her bir gauss bileşenine olasılıksal olarak karşılığı, GAM'daki *i.* karışım için şu şekilde hesaplanır;

$$\Pr(i \mid x_t) = \frac{\omega_i \, p_i(x_t)}{\sum_{j=1}^{M} \omega_j \, p_j(x_t)} \tag{3.13}$$

Denklem 3.13'te görülen $p_i(x_t)$, GAM modelindeki Denklem 2.5'te belirtilen D-boyutlu *i.* gauss bileşenini ifade etmektedir. Daha sonra $Pr(i \mid x_t)$ ve x_t kullanılarak yeterli istatistikler hesaplanır. Bunlar, her bir gauss bileşeni için atanan vektör sayısı, birinci moment değeri, ikinci moment değerleridir ve sırasıyla şu şekilde hesaplanırlar;

$$n_i = \sum_{t=1}^{T} Pr(i \mid x_t) \tag{3.14}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i \mid x_t)\, x_t \tag{3.15}$$

$$E_i(x^2) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i \mid x_t)\, x_t^2 \tag{3.16}$$

Son olarak bu yeni yeterli istatistikler kullanılarak, hedef konuşmacı modelindeki λ_{HDF} için i. gauss bileşeninin uyarlanmış güncel parametreleri hesaplanır (Şekil 3.7).

$$\hat{\omega}_i = [\alpha_i^{\omega} n_i / T + (1 - \alpha_i^{\omega})\, \omega_i]\, \gamma \tag{3.17}$$

$$\hat{\mu}_i = \alpha_i^{m} E_i(x) + (1 - \alpha_i^{m})\, \mu_i \tag{3.18}$$

$$\hat{\sigma}_i^2 = \alpha_i^{v} E_i(x^2) + (1 - \alpha_i^{v})(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2 \tag{3.19}$$

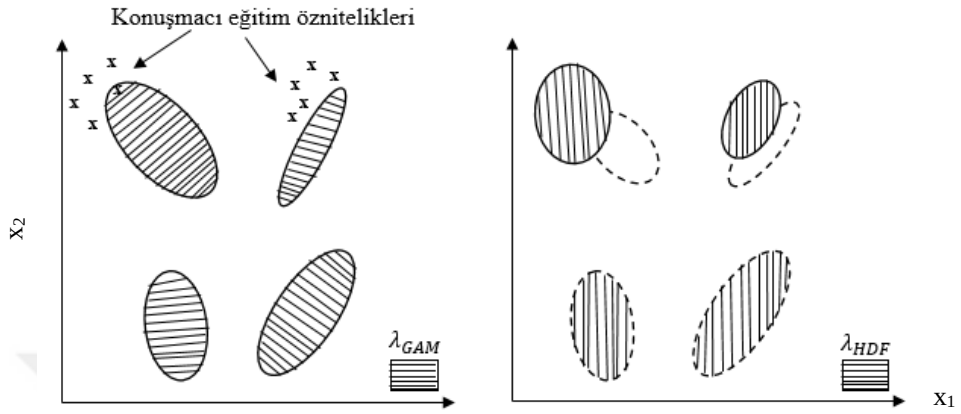
Denklem 3.17, 3.18 ve 3.19'da görüldüğü gibi sırasıyla, $\omega_i, \mu_i, \sigma_i^2$, λ_{GAM} modelindeki *i.* gauss bileşenine ait ağırlık, ortalama ve kovaryans matrisinini ifade etmektedir. $\hat{\omega}_i, \hat{\mu}_i, \hat{\sigma}_i^2$ ise MAP adaptasyonu ile hesaplanmış konuşmacı modeli λ_{HDF} için *i.* gauss bileşenine ait ağırlık, ortalama ve kovaryans matrisini ifade etmektedir. Buradaki $\{\alpha_i^{\omega}, \alpha_i^{m}, \alpha_i^{v}\}$ sırasıyla, ağırlık, ortalama ve kovaryans matrisine ait eski ve yeni tahminler arasındaki dengeyi belirleyen adaptasyon parametreleridir. γ ise ölçek faktörünü (scale factor) temsil etmektedir.

Her bir karışım bileşeni ve parametresi için adaptasyon katsayısı α_i^{ρ}, $\rho \in \{\omega, m, v\}$ olmak üzere şu şekilde tanımlanmıştır;

$$\alpha_i^{\rho} = \frac{n_i}{n_i + r^{\rho}} \tag{3.20}$$

Denklem 3.20'de görüldüğü gibi, *r* sabit bir ilgililik faktörünü (relevance factor) temsil eder [28].

Bu tezde, ilgililik faktörü 16 olarak alınmıştır. Şekil 3.7'de örnek bir konuşmacı modelinin iki boyutlu uzayda (x_1,x_2) uyarlanması gösterilmektedir.



Şekil 3.7 : GKM – GAM yöntemi ile konuşmacı modelinin uyarlanması.

Şekil 3.7'de görüldüğü gibi konuşmacı eğitim öznitelikleri GAM modelindeki sadece iki gauss bileşenine atandığı için (Şekil 3.7a) uyarlanmış konuşmacı modelinde sadece iki gauss bileşeni değişmektedir. Geriye kalan bileşenler ise GAM modelinden aynen aktarılmaktadır (Şekil 3.7b).

GKM – GAM yöntemi ile konuşmacı doğrulama sisteminin test aşamasında logaritmik olabilirlik oranı daha önce Denklem 3.11'de anlatıldığı üzere şu şekildedir;

$$\Lambda(Y) = \log p(Y|\lambda_{HDF}) - \log p(Y|\lambda_{GAM}) \tag{3.21}$$

Burada $Y=\{y_1,\ldots,y_T\}$ bilinmeyen konuşmacıya ait öznitelik vektörleri olmak üzere, λ_{HDF}, iddia edilen hedef konuşma modelini, λ_{GAM}, arkaplan modelini temsil eder.

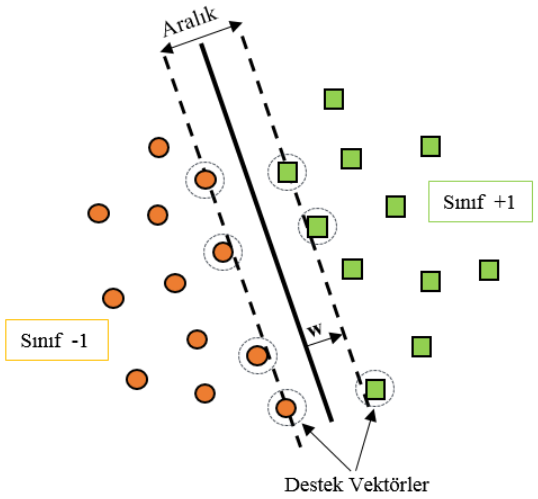
GKM – GAM yöntemini kısaca özetlemek gerekirse ilk olarak, çok fazla sayıda konuşmacıdan alınan ses örnekleri kullanılarak EM algoritması ile M adet gauss bileşeninden oluşan genel arkaplan modeli λ_{GAM} elde edilir. Daha sonra eğitim aşamasında iddia edilen hedef konuşmacı modeli λ_{HDF}, konuşmacının eğitim sesinden elde edilen öznitelik vektörleri kullanılarak λ_{GAM} modelinden türetilerek elde edilir.

Test aşamasında ise Denklem 3.21'de ifade edildiği gibi λ_{HDF} ve λ_{GAM} modelleri kullanılarak logaritmik olabilirlik oranı hesaplanarak karar verme işlemi gerçekleştirilir.

3.5.2 GKM – DVM modeli

DVM, örüntü sınıflandırma problemleri için kullanılan güçlü bir sınıflandırıcı tekniğidir. Konuşmacı tanıma uygulamalarında, gerçek konuşmacıyı (target speaker) ve sahte konuşmacıyı (impostor speaker) birbirinden ayıran, ayırt edici bir yöntemdir. Bir DVM daha önce Denklem 2.8'de anlatıldığı gibi çekirdek fonksiyonunun toplamından oluşan iki sınıflı bir sınıflandırıcıdır. Bu iki sınıf birbirinden hiperdüzlem (hyperplane) adı verilen ayırıcı bir doğru ile ayrılır.

DVM sınıflandırıcılarında temel amaç, aralığı (margin) maximum yapan bir optimum ayırıcı hiperdüzlemi oluşturmaya çalışmaktır. Şekil 3.8'de optimum ayırıcı hiperdüzlem ile DVM yapısı gösterilmiştir. Burada bahsedilen aralık kavramı ayırıcı hiperdüzlemden en yakın veriye olan minimum uzaklıktır [43]. i = 1,…,N olmak koşuluyla, N adet örnekten oluşan eğitim veri kümesinde, $x_i \in R^d$, eğitim öznitelik vektörlerini, $y_i \in \{-1,1\}$ sınıf etiketlerini temsil etmek üzere, her bir örneğin hangi sınıfa ait olduğu $\{x_i, y_i\}$ şeklinde belirtilir [44]. İki sınıfın birbirinden doğrusal olarak ayrılabildiği durum Şekil 3.8'de gösterilmiştir. DVM sınıflandırıcılarında, ayırıcı hiperdüzlem üzerinde bulunan noktalar $\mathbf{w}.\mathbf{x}+b = 0$ eşitliğini sağlamaktadır. Şekil 3.8'den de görüldüğü gibi, **w** hiperdüzleme dik olan bir doğruyu temsil etmektedir.



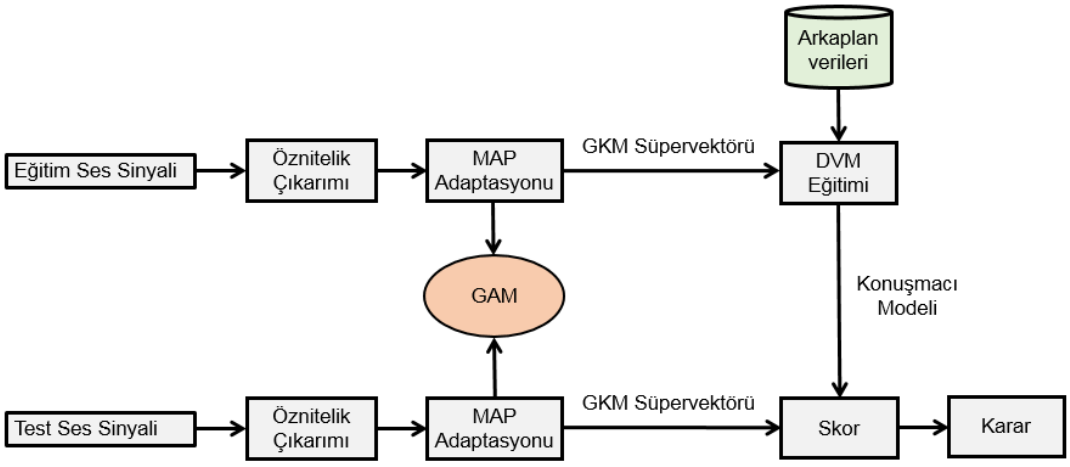
Şekil 3.8 : DVM yapısı

Şekil 3.8'de, $\{x_i, y_i\}$ doğrusal olarak ayrılabilir veriler için iki sınıflı bir durum göz önüne alındığında her bir örnek -1 (sınıf -1) olarak negatif sınıf yada +1 (sınıf +1) olarak pozitif sınıf ile etiketlendirilerek optimum ayırıcı hiperdüzlem bulunmaya çalışılır. Ayırıcı hiperdüzlemi bulmak için sınıfların ait olduğu sınırları belirlenmesi gerekir.

Her sınıfın sınırlarının belirlendiği doğrular şu şekilde ifade edilir;

$$y_i(\mathbf{w}.\mathbf{x} + b) \geq 1 \tag{3.22}$$

Denklem 3.22'de bahsedilen sınır doğruları, Şekil 3.8'de gösterilen ayırıcı hiperdüzlemin her iki tarafında bulunan kesikli çizgilerle belirtilen doğrulardır. Sınıf sınırlarının belirlendiği bu doğrular üzerinde bulunan ve Denklem 3.22 şartını sağlayan vektörlere *destek vektörleri* adı verilmektedir. Şekil 3.8'de görüldüğü gibi doğrusal olarak ayrılabilen veriler üzerinde sınıflandırma yapmak ve maksimum aralığın bulunması işlemi kolaydır. Fakat, doğrusal olarak ayrılamayan verilerin sınıflandırılması için doğrusal olarak ayrılabilecekleri farklı bir uzaya taşınması gerekir. Bu sebeple, doğrusal olmayan bir fonsiyona ihtiyaç duyulur. Bu fonksiyon, *çekirdek fonsiyonu (kernel function)* olarak bilinmektedir. Çekirdek fonksiyonu verilen öznitelik vektörleri, doğrusal olmayan bir fonksiyon aracılığıyla daha yüksek boyutlu bir uzaya taşımak amacıyla kullanılır. Böylelikle giriş uzayında doğrusal olarak sınıflandırılamayan öznitelikler daha yüksek boyutlu uzayda doğrusal olarak sınıflandırılabilir hale getirilir [44]. Çekirdek fonksiyonu Denklem 2.8'de daha önce belirtildiği gibi K(:,:) şeklinde ifade edilir. Çekirdek fonksiyonu olarak kullanılan farklı eşitlikler vardır. Bu tezde, doğrusal çekirdek fonksiyonu kullanılmıştır. DVM'in eğitim aşamasında temel amaç, sınıflar arasındaki sınırı modellemektir [43]. Konuşmacı doğrulamada, GKM'nin konuşmacıyı temsil etme kapasitesiyle DVM'nin ayırt edici özelliklerini birleştiren GKM – DVM modelini ilk olarak Campbell ve arkadaşları kullanmıştır [56]. GKM – DVM yöntemi ile oluşturulan konuşmacı tanıma sistemi Şekil 3.9'da gösterilmiştir.

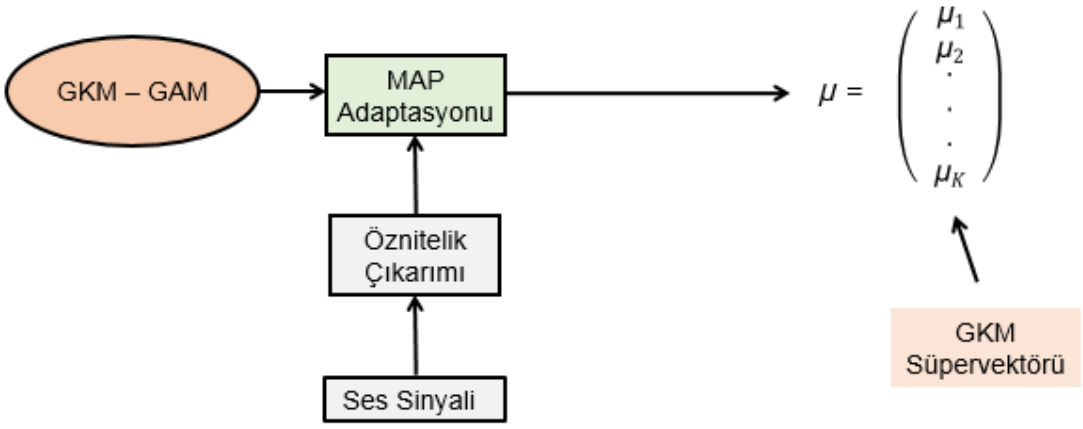


Şekil 3.9 : GKM – DVM yönteminin işlem adımları.

Şekil 3.9'da görüldüğü gibi ses sinyalleri öznitelik vektörlerine dönüştürülür. Daha sonra bu öznitelik vektörlerinin GAM modeli ile MAP adaptasyonu sonucunda oluşturulan GKM modelindeki ortalama vektörlerinin (μ) uç uca eklenmesiyle bir süpervektör oluşturulur [56,57]. Böylelikle öznitelik vektör kümesi tek bir yüksek boyutlu vektöre dönüştürülmüş olur ve bu vektöre GKM süpervektörü adı verilir. GKM – GAM modeli şu şekildedir;

$$p(x) = \sum_{k=1}^{K} \omega_k \, N(x; \mu_k, \sum_k) \tag{3.23}$$

Denklem 3.23'te ifade edilen ω_k, μ_k, $\sum_k$, sırasıyla karışım ağırlıklarını, ortalama vektörlerini ve kovaryans matrislerini ifade etmektedir [56]. GKM süpervektörü, $X=\{x_1,..., x_k\}$ konuşmacıya ait öznitelikleri temsil etmek üzere öznitelik vektörlerinin GAM modelinden uyarlanarak oluşturulan GKM parametrelerinden birisi olan ortalama vektörlerinin uç uca eklenmesiyle oluşturulur. GKM süpervektörünün yapısı Şekil 3.10'da gösterilmiştir.



Şekil 3.10 : GKM Süpervektörü.

GKM süpervektörü elde edildikten sonra DVM eğitilir. Daha öncede belirtildiği gibi DVM, iki sınıflı bir sınıflandırıcı olduğundan eğitim aşamasında, eğitilecek konuşmacının sesine (pozitif sınıf) ve arkaplan verilerine (negatif sınıf) ihtiyaç duyar. Burada arkaplan verisini ifade eden negatif sınıf, hedef konuşmacının dışında çok fazla sayıda konuşmacıdan alınan ses örneklerinden elde edilen öznitelik vektörleridir. Negatif sınıf, eğitilecek konuşmacının genel konuşmacı uzayında ayırt edici olmasını sağlar. DVM eğitimi tamamlandıktan sonra konuşmacı modeli oluşturularak konuşmacı tanıma sisteminin eğitim aşaması tamamlanarak test aşamasına geçilir.

Test aşamasında da eğitim aşamasındaki gibi test örnekleri kullanılarak aynı şekilde GKM süpervektörü oluşturulur (Şekil 3.). Eğitilmiş konuşmacı modeli ile test süpervektörü kıyaslanarak, test süpervektörünün DVM ayırıcı hiperdüzlemine olan uzaklığı hesaplanarak skor hesabı yapılır ve karar verilerek GKM – DVM modeli ile konuşmacı doğrulama sisteminin süreci tamamlanır [58].

3.5.3 Birleşik etmen analizi (JFA)

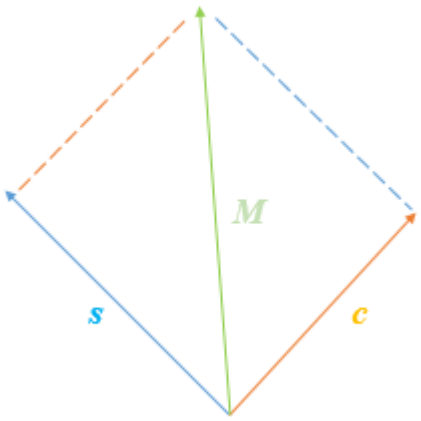
JFA, son yıllarda konuşmacı doğrulama alanında kullanılan oldukça popüler hale gelen bir yaklaşım olmuştur [48]. JFA yöntemi, kanal değişkenliğine ve konuşmacıya bağımlı bir GKM modeli olarak görülebilir [59].

Aslında JFA modeli, GKM modelinde kanal değişikliklerinde ve oturum farklılıklarında ortaya çıkan problemleri iyileştirmek için geliştirilmiş bir yöntemdir [50]. Daha önce belirtildiği gibi, GKM modelinde her bir konuşmacı, öznitelik uzayında tanımlanan, ortalama vektörleri (μ), karışım ağırlıkları (ω) ve kovaryans matrisleri (∑) olan çok değişkenli GKM bileşenleriyle temsil edilir. GKM modelinde konuşmacı modeli (target speaker), çok fazla sayıda ses kullanılarak eğitilmiş GAM modelinden MAP adaptasyonu yapılarak oluşturulur. MAP adaptasyonunda, GAM'dan ortalama vektörler uyarlanırken, ağırlıklar ve kovaryanslar tüm konuşmacılar arasında paylaşılır. Böylece bir konuşmacı modeli, ortalama vektörlerinin birleşmesiyle elde edilerek süpervektör adı verilen tek bir vektörle temsil edilir. Bir konuşmacı için farklı eğitim ses örneklerinden hesaplanan süpervektörler, özellikle eğititim ses örnekleri farklı cihazlardan alındığında aynı olmayabilir. Bu nedenle farklı kanallardan elde edilen test verilerinin konuşmacı modeline uygun şekilde hesaplanabilmesi için kanal değişkenliğinden kaynaklanan problemin telafisi gerekir. Kanal telafisinin yapılması amacıyla kanal değişkenliğinin modellenmesi gerekir. Bu amaçla JFA tekniği önerilmiştir.

JFA modeli, GKM süpervetörünün değişkenliğini konuşmacı ve kanal değişkenliğinin doğrusal bir kombinasyonu olarak kabul eder. Daha önce Denklem 2.9'da belirtildiği gibi, bir eğitim ses örneği verildiğinde, konuşmacıya bağlı ve kanala bağlı iki süpervektörü temsil eden $\boldsymbol{M}$, istatistiksel olarak bağımsız iki farklı bileşene ayrılır ve şu şekilde ifade edilir;

$$\boldsymbol{M} = \boldsymbol{s} + \boldsymbol{c} \tag{3.24}$$

Denklem 3.24'te görülen ***s*** ve ***c*** süpervektörleri sırasıyla, konuşmacı ve kanal süpervektörleri olarak adlandırılır. Şekil 3.11'de ***M*** süpervektörü gösterilmiştir.



Şekil 3.11 : *M* Süpervektörü.

d, ses öznitelik vektörlerinin boyutu olarak ve *K*, GAM'daki karışım sayısı olarak varsayıldığında, ***M, s, c*** süpervektörlerinin *Kd*- boyutlu parametre uzayında olduğu kabul edilir. Daha önce Denklem 2.11'de bahsedildiği gibi, kanal değişkenliği şu şekilde modellenmiştir;

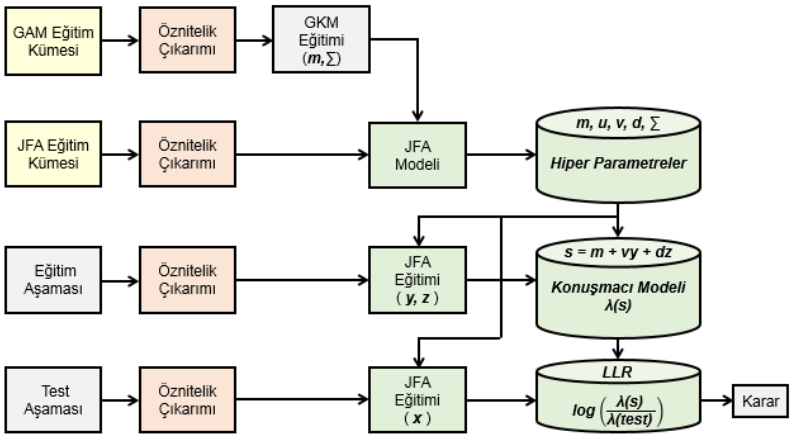
$$\boldsymbol{c} = \boldsymbol{Ux} \tag{3.25}$$

Denklem 3.25'te görülen ***U***, özkanalı (eigenchannel) temsil eden dikdörtgen bir matrisi, ***x***, verilen bir ses örneğinden hesaplanan kanal etmenlerini (channel factors) ifade eder. ***U*** matrisinin sütunları, konuşmacı veri kümesi için hesaplanan öz kanallardır (eigenchannels). Daha önce Denklem 2.10'da ifade edildiği gibi ***s*** konuşmacı süpervektörü şu şekilde hesaplanır;

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{vy} + \boldsymbol{dz} \tag{3.26}$$

Denklem 3.26'da ifade edilen ***m***, GAM'dan uyarlanan ortalama vektörlerinin birleşmesiyle oluşan konuşmacı ve kanaldan bağımsız süpervektörü temsil eder. ***v***, sütunlarının her biri özses (eigenvoice) olarak adlandırılan dikdörtgen bir matrisi, ***y***, konuşmacı modeline ait konuşmacı etmenlerini (speaker factors), ***d***, köşegen bir matrisi, ***z***, *Kd x 1* boyutlu sütun vektörünü temsil eder. JFA yönteminde, ***U***, ***v***, ***d*** matrisleri hiperparametreler olarak bilinmektedir [8]. JFA yönteminin eğitim aşamasında, büyük bir eğitim setinde sırasıyla ***v, U***, ***d*** hiper parametreleri tahmin edilir, daha sonra ***x*** , ***y*** , ***z*** konuşmacı ve kanal etmenleri birleşik olarak tahmin edilir.

Son olarak ***c*** kanal süpervektörü atılarak , ***M – Ux*** konuşmacı modeli oluşturulur ve elde edilen konuşmacı modeli ile test ses öznitelik vektörleri arasında olabilirlik oranı hesaplanarak karar verilir [50]. Şekil 3.12'de JFA yöntemi ile genel bir konuşmacı doğrulama sistemi gösterilmiştir.



Şekil 3.12 : JFA yöntemi ile konuşmacı doğrulama sistemi.

JFA yöntemini kısaca özetlemek gerekirse, GAM'dan MAP adaptasyonu ile elde edilen GKM bileşenlerinden ortalama vektörleri (m) ve kovaryans matrisleri (∑) hiperparametreler olarak JFA modelinin eğitilmesi için kullanılır. JFA eğitiminde temel amaç hiper parametreleri eğitmektir. Hiper parametreler MLE algoritmasıyla hesaplanır [59]. Eğitim aşamasında konuşmacı modeli, ***u,*** ∑, hiper parametreleri sabit tutularak sadece konuşmacıya bağlı hiper parametreler (***m, v, d***) ve her bir hedef konuşmacı için konuşmacı etmenleri (***y, z***) hesaplanarak elde edilir. Test aşamasında ise test ses örneklerinin öznitelikleri çıkarılarak, konuşmacıdan bağımsız hiper parametre (***u***) kullanılarak ve her bir sahte konuşmacı için ***x*** kanal etmeni hesaplanarak sahte konuşmacı modeli (non – target speaker) oluşturulur. Kanala bağlı bileşenler, yalnızca test ses örneklerinden doğrudan hesaplanarak model oluşturulur. Daha sonra iki model arasındaki logaritmik olabilirlik oranına göre karar verilir (Şekil 3.12).

3.5.4 i – vektör yaklaşımı

Konuşmacı ve kanal etmenlerine dayanan klasik JFA yöntemi, özses matrisi (***v***) tarafından tanımlanan konuşmacı uzayı ve özkanal matrisi (***u***) tarafından tanımlanan kanal uzayı olmak üzere iki farklı uzaydan oluşur. i-vektör yaklaşımı, JFA modelinde kullanılan iki farklı uzay yerine tek bir uzay tanımlamaya dayanmaktadır. Bu yeni uzay, *toplam değişkenlik uzayı* (total variability space) olarak adlandırılmaktadır.

JFA modelinde olduğu gibi kanal ve konuşmacı değişkenliklerini içerir. i-vektör yaklaşımının ortaya atılmasındaki temel sebep, JFA'da sadece kanal etkilerini modelleyen kanal etmenlerinin konuşmacı hakkında da bilgi içermesidir [50]. Bu kanal uyumsuzluğu problemini çözmek için kullanılan i-vektör yaklaşımı, i-vektör özniteliklerinin elde edilmesi (i-vector extraction), değişkenlik telafisi (variability compensation), skor hesaplama (score computation) olmak üzere üç temel adımdan oluşmaktadır [60].

i-vektör yönteminde konuşmacıya ait ses sinyalinden çıkarılan öznitelik vektörleri sabit uzunluğa sahip tek bir vektöre dönüştürülür. GKM – GAM yöntemine benzer şekilde çok sayıda konuşmacının ses sinyalinden çıkarılarak elde edilen öznitelik vektörleriyle genel arkaplan modeli eğitilir.

i-vektör yaklaşımında, bir konuşmacıya ait ses sinyalinden çıkarılan öznitelik vektörlerinin genel arkaplan modeli ile MAP adaptasyonu sonucunda oluşturulan GKM modelinin ortalama vektörlerinin uç uca eklenmesiyle oluşan kanala ve konuşmacıya bağlı süpervektör şu şekildedir;

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{\omega} \tag{3.27}$$

Denklem 3.27'de görülen $\boldsymbol{M}$, konuşmacıya ait süpervektörü, $\boldsymbol{m}$, GAM modelinin ortalama vektörlerinin uç uca eklenmesiyle elde edilen konuşmacıdan ve kanaldan bağımsız süpervektörü, $\boldsymbol{T}$, düşük ranklı bir dikdörtgen matrisi, $\boldsymbol{\omega}$, i-vektör özniteliğini temsil etmektedir. Konuşmacı doğrulama sisteminin eğitim ve test kümesinde, konuşmacılara ait i-vektör öznitelik vektörleri çıkarıldıktan sonra konuşmacılara ait öznitelik vektörlerinin sınıflandırılmasına yada bu öznitelik vektörlerinin kullanılarak sistem için karar skorunun hesaplanmasına ihtiyaç vardır.

Deneysel çalışmalarda, karar skorunu hesaplamak amacıyla iki farklı yöntem kullanılarak performans başarı oranları incelenmiştir. Kullanılan bu yöntemlerden birisi basit fakat etkili bir yöntem olan kosinüs uzaklığı skorlaması yöntemidir (cosine distance scoring - CDS) [50]. CDS yönteminde, w_{hdf} ve w_{tst} şeklinde verilen bir eğitim ve test i-vektör ikilisi için kosinüs benzerliği şu şekilde hesaplanır [50];

$$cosine\,(w_{hdf}, w_{tst}) = \frac{w_{hdf}^{T} w_{tst}}{\|w_{hdf}\| \|w_{tst}\|} \tag{3.28}$$

CDS tekniği ile birlikte deneysel çalışmalarda olasılıksal doğrusal ayırt edici analizi (PLDA) tekniği de [61, 62] kullanılmıştır. w_{hdf} ve w_{tst} şeklinde verilen bir eğitim ve test i-vektör ikilisi için PLDA sınıflandırıcısı ile LLR skoru şu şekilde hesaplanmaktadır [62];

$$\text{LLR} = \log \frac{p(w_{hdf}, w_{tst} | H_s)}{p(w_{hdf} | H_d) p(w_{tst} | H_d)} \tag{3.29}$$

Denklem 3.29'da görülen H_s, her iki i-vektör özniteliğinin aynı konuşmacıya ait olma hipotezini ifade ederken H_d ise her iki i-vektör özniteliğinin farklı konuşmacılara ait olma hipotezini ifade etmektedir [62].

4. SONUÇLAR VE ÖNERİLER

Bu tezde MATLAB programı yardımıyla GKM – GAM, GKM – DVM, JFA, i-vektör yöntemleri kullanılarak KD sistemleri oluşturulmuştur. Bu yöntemlerle birlikte iki farklı öznitelik çıkarma metodu (MFKK, DGKK) kullanılarak sistemler üzerindeki başarı performansları karşılaştırmalı olarak ele alınmıştır ve ayrıca arkaplan seslerinin GKM – GAM modeliyle oluşturulan KD sistemleri üzerindeki etkisi incelenmiştir.

4.1 GKM – GAM Yönteminde GAM Modeli Türkçe Seslerle Eğitildiğinde Gauss Bileşen Sayısının Ve Öznitelik Katsayılarının Sistem Performansına Etkisi

Deneysel çalışmanın ilk aşamasında GKM – GAM modeli için en ideal sistem parametrelerini bulmak amacıyla deneyler gerçekleştirilmiştir. Yapılan bu deneylerde GAM eğitimi için kullanılan arkaplan sesleri Türkçe seslerden oluşmaktadır. Deneysel çalışmalarda ilk olarak, GKM – GAM modelinde önerilecek olan en ideal gauss bileşen sayısını bulmak hedeflenmiştir. Bu sebepten dolayı, öncelikle öznitelik çıkarma işlemi için MFKK yöntemi kullanılmıştır ve öznitelik boyutu 18 olarak sabit tutulmuştur. Gauss bileşenleri ise [8,512] aralığında olmak üzere 7 farklı değer için deneyler gerçekleştirilmiştir. Deneysel sonuçlar Çizelge 4.1'de verilmektedir.

Çizelge 4.1 : GAM eğitiminde Türkçe sesler kullanıldığında farklı sayıdaki gauss bileşenleri için EER değerleri. MFKK öznitelik boyutu 18 olarak alınmıştır.

GKM Bileşen Sayısı	EER (%)
8	9,26
16	6,65
32	5,68
64	5,20
128	5,14
256	5,26
512	6,80

Çizelge 4.1'de görüldüğü üzere, GKM bileşen sayısı [8,512] aralığında arttıkça, sistem performansını ifade eden EER değeri önemli miktarda azalmaktadır. Ancak 32 GKM bileşeninden sonra sistem performansında daha az oranda bir artış gözlenmektedir.

En düşük sistem hatasını %5,14 ile veren ideal GKM bileşen sayısı 128 olarak belirlendikten sonra (Çizelge 4.1), ideal MFKK öznitelik boyutunun tespit edilmesi amaçlanmıştır. Bu amaçla, [4,26] aralığında değişen farklı sayılardaki MFKK öznitelikleri kullanılarak elde edilen EER değerleri Çizelge 4.2'de verilmektedir.

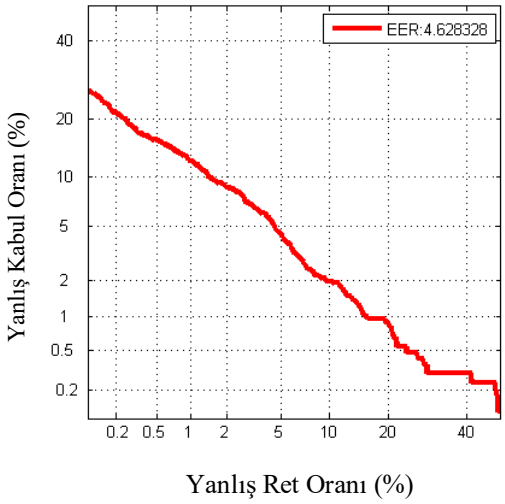
Çizelge 4.2 : GAM eğitiminde Türkçe sesler kullanıldığında farklı boyuttaki MFKK öznitelikleri için EER değerleri. Gauss bileşen sayısı 128 olarak alınmıştır.

Öznitelik Boyutu	EER (%)
4	16,45
8	10,71
12	7,41
16	5,52
24	5,38
26	4,62

Çizelge 4.2'de görüldüğü üzere MFKK öznitelik boyutunun artması, konuşmacı doğrulama sisteminin performansı üzerinde çok önemli bir etkiye neden olmuştur. Kullanılan öznitelik boyutu arttıkça EER değerinde önemli ölçüde azalma olmaktadır. Sadece 4 adet MFKK boyutu için EER değeri %16,45 olurken, MFKK öznitelik boyutu 8 olarak arttırıldığında EER değerinde yaklaşık %35 değerinde azalma olmuş ve bu değer %10,71'e düşmüştür. Konuşmacı doğrulama sisteminin performansındaki bu değişim oranı öznitelik boyutu 26'ya arttırılıncaya kadar devam etmektedir ve 26 öznitelik boyutunda sistem en düşük hata oranını vermektedir. Bundan dolayı ideal MFKK öznitelik boyutu 26 olarak elde edilmiştir.

Farklı sayılardaki GKM bileşenleri (Çizelge 4.1) ile MFKK öznitelik boyutlarının (Çizelge 4.2) sistem performansı üzerindeki etkisi karşılaştırıldığında MFKK öznitelik boyutunun etkisinin, GKM bileşen sayısına göre nispeten daha fazla olduğu gözlenmektedir. Farklı sayılardaki GKM bileşenleri için elde edilen EER değerleri %9,26 ve %5,14 arasında elde edilirken (dinamik aralık 4.12) farklı MFKK öznitelik boyutu için elde edilen bu dinamik aralık miktarı 11,83 olarak elde edilmiştir. Buradan MFKK boyutunun Gauss bileşen sayısından daha etkili olduğu anlaşılmaktadır.

Konuşmacı doğrulama sistemlerinde çoğunlukla Sezim Hata Ödünleşimi (DET) eğrisi kullanılarak sistemin çalışma noktaları gösterilir [63]. Bundan dolayı sistemin performansı DET eğrisi kullanılarak gösterilmiştir. Gauss bileşen sayısı 128 olarak alındığında ve MFKK öznitelik boyutu 26 adet kullanıldığında elde edilen DET eğrisi Şekil 4.1'de gösterilmiştir.



Şekil 4.1 : Türkçe seslerle eğitilen arkaplan modeli ve MFKK öznitelikleri kullanılarak oluşturulan GKM – GAM modeli için elde edilen DET eğrisi.

Deneysel çalışmanın ikinci aşamasında yine GAM eğitimi için Türkçe sesler kulanılmıştır. Ancak aynı deneyler bu kez DGKK öznitelik çıkarma yöntemi kullanılarak, DGKK katsayılarının ve gauss bileşenlerinin sistem performansı üzerindeki etkisi incelenmiştir.

Deneysel çalışmalarda öncelikle DGKK öznitelik boyutu 18 olarak alınmış ve ideal GKM bileşen sayısını bulmak için gauss bileşen sayısı [8,512] aralığında değiştirilerek deneyler gerçekleştirilmiştir. Deneysel sonuçlar Çizelge 4.3'te verilmektedir.

Çizelge 4.3 : GAM eğitiminde Türkçe sesler kullanıldığında farklı sayıdaki gauss bileşenleri için EER değerleri. DGKK öznitelik boyutu 18 olarak alınmıştır.

GKM Bileşen Sayısı	EER (%)
8	11,12
16	9,70
32	7,92
64	7,75
128	7,69
256	7,21
512	7,11

Çizelge 4.3'de görüldüğü üzere, % 7,11 ile en düşük sistem hatasını veren ideal gauss bileşen sayısını 512 olarak belirledikten sonra (Çizelge 4.3), en iyi konuşmacı doğrulama sisteminin başarı performansını sağlayan DGKK öznitelik boyutunun belirlenmesi amaçlanmıştır. Bu amaçla, [4,24] arasında değişen 6 farklı öznitelik boyutu için elde edilen EER değerleri Çizelge 4.4'de verilmektedir.

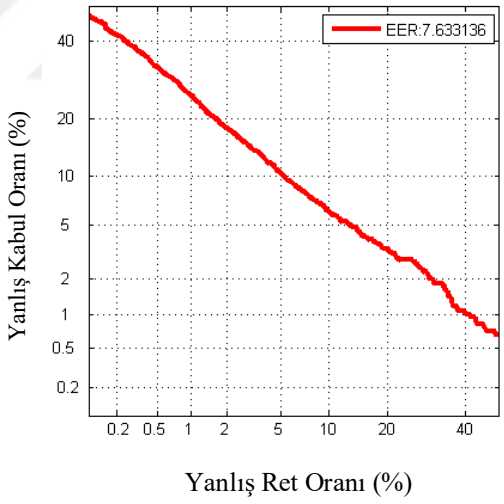
Çizelge 4.4 : GAM eğitiminde Türkçe sesler kullanıldığında farklı boyuttaki DGKK öznitelikleri için EER değerleri. Gauss bileşen sayısı 512 olarak alınmıştır.

Öznitelik Boyutu	EER (%)
4	17,15
8	10,65
12	8,46
16	8,63
20	8,04
24	7,63

Çizelge 4.4'de görüldüğü üzere, kullanılan öznitelik boyutu arttıkça EER değeri önemli oranda düşmektedir. Sadece 4 adet DGKK boyutu kullanıldığında sistemin EER değeri %17,15 olurken, DGKK öznitelik boyutu 8 olarak arttırıldığında EER değeri yaklaşık %38 azalarak %10,65'e düşmüştür.

En iyi sistem başarısı veren boyut sayısı %7,63'lük hata oranı veren 24 sayısıdır. Bundan dolayı ideal DGKK öznitelik boyutu 24 olarak elde edilmiştir.

GKM bileşen sayısı 512 ve DGKK öznitelik boyutunun 24 olarak kullanıldığında elde edilen DET eğrisi Şekil 4.2'de gösterilmiştir.



Şekil 4.2 : Türkçe seslerle eğitilen arkaplan modeli ve DGKK öznitelikleri kullanılarak oluşturulan GKM – GAM modeli için elde edilen DET eğrisi.

Türkçe sesler kullanılarak eğitilen GAM modeli için, GKM bileşen sayısı (Çizelge 4.3) ile DGKK öznitelik boyutlarının (Çizelge 4.4) sistem performansı üzerindeki etkisi karşılaştırılarak ele alındığında, DGKK öznitelik boyutunun sistem üzerindeki etkisinin Gauss bileşen sayısına göre nispeten daha fazla olduğu gözlenmektedir.

4.2 GKM – GAM Yönteminde GAM Modeli İngilizce Seslerle Eğitildiğinde Gauss Bileşen Sayısının Ve Öznitelik Katsayılarının Sistem Performansına Etkisi

GKM – GAM deneylerinin üçüncü aşamasında GAM eğitimi için İngilizce sesler kullanılmıştır. Arkaplan modelinin eğitilmesi amacıyla kullanılan seslerin diliyle konuşmacı doğrulama deneylerinde kullanılan dilin farklı olduğu durumda, konuşmacı doğrulama performansı üzerindeki etkisi incelenmiştir.

İngilizce seslerin kullanılarak eğitildiği GKM – GAM modelinde sırasıyla MFKK ve DGKK öznitelikleri kullanılarak konuşmacı doğrulama üzerinde performansları karşılaştırmalı olarak ele alınmıştır.

Deneysel çalışmalarda öncelikle, MFKK boyutu 18 olarak alınmış ve GKM bileşen sayısı [8,512] arasında değiştirilmiştir. Daha önce bahsedildiği gibi GAM modeli İngilizce sesler kullanılarak eğitilmiştir. Deneysel sonuçlar Çizelge 4.5'te verilmektedir.

Çizelge 4.5 : GAM eğitiminde İngilizce sesler kullanıldığında farklı sayıdaki gauss bileşenleri için EER değerleri. MFKK öznitelik boyutu 18 olarak alınmıştır.

GKM Bileşen Sayısı	EER (%)
8	10,2
16	8,20
32	6,62
64	6,33
128	6,03
256	6,56
512	6,71

Çizelge 4.5'de görüldüğü üzere, GKM bileşen sayısı arttıkça EER değeri azalmaktadır. Ancak 32 gauss bileşeninden sonra sistem performansında daha düşük oranda artışlar görülmektedir.

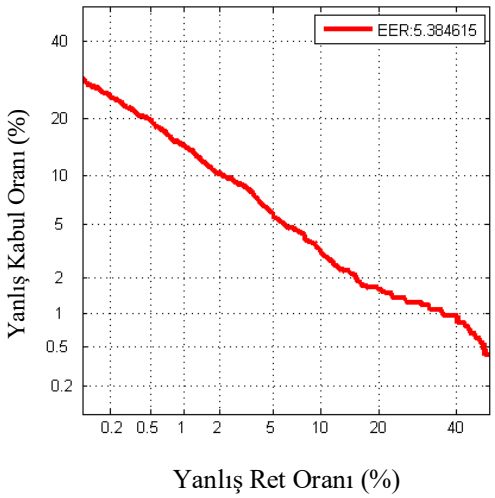
GKM bileşen sayısı 16'dan 32'ye arttırıldığında konuşmacı doğrulama sistemi performansında yaklaşık olarak % 20 oranında bir artış olmuştur.

İdeal gauss bileşen sayısı 128 olarak elde edildikten sonra (Çizelge 4.5), en iyi sistem başarı performansını veren MFKK öznitelik boyutunun belirlenmesi amaçlanmıştır. Bu amaçla, [4,24] arasında değişen farklı boyutlardaki MFKK öznitelikleri için bulunan EER değerleri Çizelge 4.6'da verilmektedir.

Çizelge 4.6 : GAM eğitiminde İngilizce sesler kullanıldığında farklı boyuttaki MFKK öznitelikleri için EER değerleri. Gauss bileşen sayısı 128 olarak alınmıştır.

Öznitelik Boyutu	EER (%)
4	14,4
8	10,4
12	9,28
16	7,04
20	5,73
22	5,38
24	6,13

Çizelge 4.6'da görüldüğü üzere, kullanılan öznitelik boyutunun artması EER değerini azaltmaktadır. Sadece 4 adet MFKK ile %14,4 EER değerine ulaşılırken, MFKK öznitelik boyutu 8 olduğunda EER %27 oranında düşerek %10,4'e azalmıştır. Sistem için en iyi başarı performansı %5,38 EER ile 22 öznitelik boyutu olmuştur. Bundan dolayı ideal MFKK öznitelik boyutu 22 olarak bulunmuştur. GKM bileşen sayısı (Çizelge 4.5) ile MFKK öznitelik boyutlarının (Çizelge 4.6) sistem performansı üzerindeki etkileri ele alındığında, MFKK öznitelik boyutunun sistem üzerindeki etkisinin Gauss bileşen sayısına nispeten daha fazla olduğu gözlenmiştir. Farklı sayıdaki GKM bileşenleri için elde edilen EER değerleri %10,2 ve %6,03 arasında değişirken (dinamik aralık 4.17) farklı boyutlardaki MFKK öznitelikleri için bu aralık 9,02 olarak gözlenmiştir. Buradan, MFKK boyutunun Gauss bileşen sayısından daha etkili olduğu anlaşılmaktadır. GKM bileşen sayısı 128 ve MFKK öznitelik boyutu 22 olarak kullanıldığında elde edilen DET eğrisi Şekil 4.3'te gösterilmiştir.



Şekil 4.3 : İngilizce seslerle eğitilen arkaplan modeli ve MFKK öznitelikleri kullanılarak oluşturulan GKM – GAM modeli için elde edilen DET eğrisi.

İkinci olarak, DGKK öznitelik boyutu 18 olarak alınmış ve GKM bileşen sayısı [8,512] arasında değiştirilmiştir. Deneysel sonuçlar Çizelge 4.7'te verilmektedir.

Çizelge 4.7 : GAM eğitiminde İngilizce sesler kullanıldığında farklı sayıdaki gauss bileşenleri için EER değerleri. DGKK öznitelik boyutu 18 olarak alınmıştır.

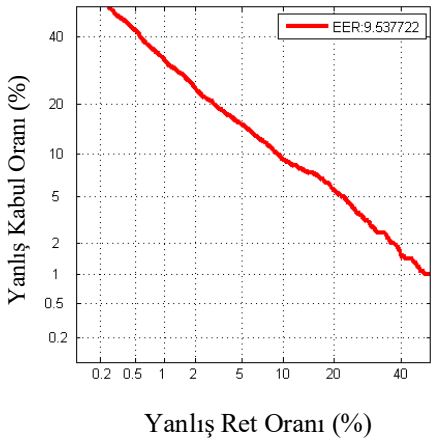
GKM Bileşen Sayısı	EER (%)
8	15,14
16	13,10
32	12,18
64	11,17
128	9,94
256	9,64
512	9,82

En düşük hata oranını veren ideal Gauss bileşen sayısı 256 olarak belirlendikten sonra (Çizelge 4.7), DGKK boyutları [4,24] aralığında seçilerek deneyler yapılmıştır ve elde edilen deneysel sonuçlar Çizelge 4.8'de gösterilmiştir.

Çizelge 4.8 : GAM eğitiminde İngilizce sesler kullanıldığında farklı boyuttaki DGKK öznitelikleri için EER değerleri. Gauss bileşen sayısı 256 olarak alınmıştır.

Öznitelik Boyutu	EER (%)
4	17,40
8	11,82
12	11,36
16	9,53
20	9,64
24	10,2

İdeal DGKK katsayısı 16 olarak belirlenmiştir ve 256 adet gauss bileşeni ile elde edilen DET eğrisi Şekil 4.4'te gösterilmiştir.



Şekil 4.4 : İngilizce seslerle eğitilen arkaplan modeli ve DGKK öznitelikleri kullanılarak oluşturulan GKM – GAM modeli için elde edilen DET eğrisi.

GKM – GAM modeli kullanılarak geliştirilen KD sistemlerinde sistem performansı üzerinde, kullanılan öznitelik katsayılarının gauss bileşen sayısına nazaran daha etkili olduğu görülmüştür. Aynı zamanda Türkçe sesler ve İngilizce sesler kullanılarak eğitilen GAM modeli ile gerçekleştirilen KD sistemlerinde, arkaplan verisi ile gerçekleştirme verisi arasında konuşulan dil anlamında uyumsuzluk olduğunda konuşmacı doğrulama performansı düşmektedir. Arkaplan verileri Türkçe seslerden oluşan MFKK öznitelikleri ile yapılan GKM – GAM modelinde EER değeri %4,62 iken, GAM İngilizce seslerle eğitildiğinde KD performans yaklaşık %17 düşerek EER değeri %5,38 olmuştur. Arkaplan verileri Türkçe seslerden oluşan DGKK öznitelikleri ile yapılan GKM-GAM modelinde EER değeri %7,63 iken, GAM İngilizce seslerle eğitildiğinde KD performansı yaklaşık % 25 düşerek EER değeri %9,53 olmuştur.

GKM – GAM modeliyle oluşturulan KD deney sonuçlarından anlaşılan bir diğer sonuç ise MFKK öznitelikleri, öznitelik sayısından bağımsız olarak Türkçe Metne Bağlı Konuşmacı Doğrulama Sistemi için DGKK özniteliklerinden daha üstündür. Elde edilen optimum değerlere göre Türkçe GAM için, MFKK öznitelikleriyle EER değeri % 4,62 iken, DGKK öznitelikleriyle EER değeri 7,63'tür. İngilizce GAM için, MFKK öznitelikleriyle EER değeri %5,38 iken, DGKK öznitelikleriyle elde edilen EER değeri %9,53'tür. GKM – GAM modeliyle oluşturulan bir KD sistemi için MFKK öznitelikleri önerilmiştir.

4.3 GKM – DVM Yönteminde Gauss Bileşen Sayısının Ve Öznitelik Katsayılarının KD Sistem Performansına Etkisi

GKM – DVM deneylerinde ilk olarak, MFKK öznitelikleri ile 18'e sabitlenerek yapılan deney sonuçları Çizelge 4.9'da gösterilmiştir.

Çizelge 4.9 : GKM – DVM modeli kullanılarak farklı sayıdaki gauss bileşenleri için elde edilen EER değerleri.

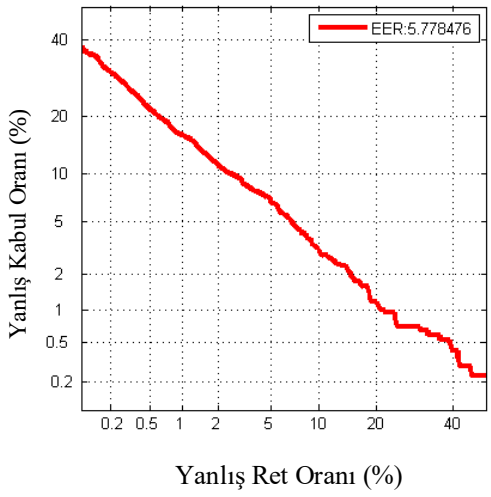
GKM Bileşen Sayısı	EER (%)
8	10,28
16	9,34
32	8,18
64	7,80
128	7,68
256	6,98
512	7,92

Çizelge 4.9'da görüldüğü gibi en düşük hata oranı 256 gauss bileşeninde elde edilmiştir (%6,98). Ancak yapılan deneylerde 128 gauss bileşen sayısı ve MFKK öznitelik boyutu 26 olarak alındığında GKM – DVM modeli için en ideal KD sistemi elde edilmiştir. Sonuçlar Çizelge 4.10'da gösterilmiştir.

Çizelge 4.10 : GKM – DVM modeli kullanılarak farklı sayıdaki gauss bileşenleri ve MFKK öznitelik boyutları için elde edilen EER değerleri.

GKM Bileşen Sayısı	Öznitelik Boyutu	EER (%)
256	8	9,89
256	12	10,88
256	16	10,88
256	20	8,22
256	24	6,64
128	12	8,52
128	16	8,5
128	22	7,86
128	26	5,77
64	20	7,7

Çizelge 4.10'da görüldüğü gibi GKM – DVM modeli için en ideal parametreleri bulmak amacıyla çeşitli deneyler yapılmıştır. GKM – DVM modeli kullanılarak oluşturulan bir KD sistemi için, EER değeri %5,77 ile en iyi KD performansı gösteren optimum değerler, Gauss bileşen sayısı için 128, MFKK öznitelik boyutu için 26 bulunmuştur. Sistemin DET eğrisi Şekil 4.5'te gösterilmiştir.



Şekil 4.5 : GKM – DVM modeli ve MFKK öznitelikleriyle oluşturulan KD sistemi DET eğrisi.

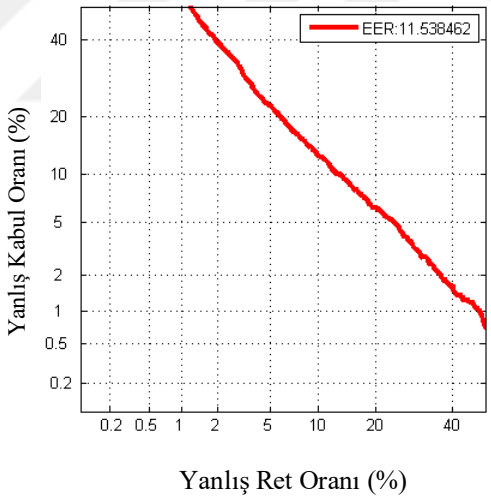
GKM – DVM deneylerinde ikinci olarak, farklı DGKK öznitelikleri ile ve farklı bileşen sayıları ile yapılan deney sonuçları Çizelge 4.11'de gösterilmiştir.

Çizelge 4.11 : GKM – DVM modeli kullanılarak farklı gauss bileşen sayıları ve DGKK öznitelik katsayıları için elde edilen EER değerleri.

GKM Bileşen Sayısı	Öznitelik Boyutu	EER (%)
8	18	19,58
16	18	16,44
32	18	13,37
64	18	13,14
128	12	11,53
128	18	12,61
128	22	12,68
128	26	13,47
256	18	12,48
256	20	12,06
512	18	12,89

Çizelge 4.11'de görüldüğü gibi, DGKK öznitelikleriyle yapılan GKM – DVM deneylerinde optimum parametreler, DGKK öznitelik sayısı için 12 ve gauss bileşen sayısı için 128 olarak elde edilmiştir.

Optimum parametreler için elde edilen KD sistemi DET eğrisi Şekil 4.6'da gösterilmiştir.



Şekil 4.6 : GKM – DVM modeli ve DGKK öznitelikleriyle oluşturulan KD sistemi DET eğrisi.

GKM – DVM modeli için, MFKK öznitelikleriyle yapılan deneylerde en düşük hata oranı %5,77 (Çizelge 4.10) iken DGKK öznitelikleriyle yapılan deneylerde bu oran %11,53'tür (Çizelge 4.11).

GKM – DVM modeli için gauss bileşeni 128 ve MFKK öznitelik boyutu 26 olarak önerilmektedir.

4.4 JFA Yönteminde Gauss Bileşen Sayısının, Öznitelik Katsayılarının, Özkanal Sayısının, Özses sayısının KD Sistem Performansına Etkisi

JFA modeli kullanılarak oluşturulan KD sisteminde gauss bileşen sayısı, öznitelik türü ve boyutu, özkanal sayısı, özses sayısı gibi farklı parametreler değiştirilerek yapılan deneylerin sonucunda KD sistemi için optimum değerlerin elde edilmesi amaçlanmıştır. Bu amaçla yapılan deneylerde ilk olarak MFKK öznitelik boyutu 18, Gauss bileşen sayısı 128 alınmıştır ve özkanal sayısı 5 olarak sabitlenmiştir. Deneysel sonuçlar Çizelge 4.12'de gösterilmiştir.

Çizelge 4.12 : MFKK öznitelikleriyle farklı konuşmacı etmenleri için EER değerleri.

Özkanal Sayısı	Özses Sayısı	EER (%)
5	5	6,92
5	10	9,84
5	15	9,95
5	20	9,88
5	25	10
5	30	9,57
5	35	9,96
5	40	9,76
5	45	9,74
5	50	9,94

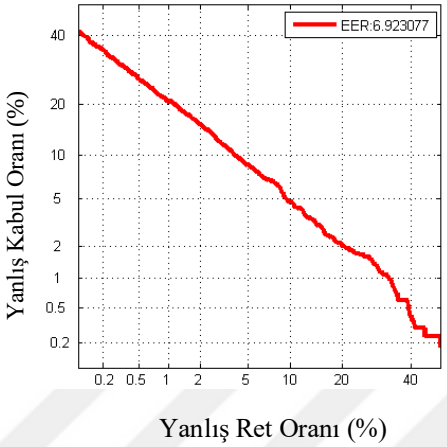
Özses sayısı 5 olarak sabitlendiğinde elde edilen deneysel sonuçlar Çizelge 4.13'te gösterilmiştir.

Çizelge 4.13 : MFKK öznitelikleriyle farklı kanal etmenleri için EER değerleri.

Özkanal Sayısı	Özses Sayısı	EER (%)
50	5	47,77
45	5	46,98
40	5	47,45
35	5	46,74
30	5	46,5
25	5	31,18
20	5	16,92
15	5	12,05
10	5	8,34

Çizelge 4.12'de ve Çizelge 4.13'te görüldüğü gibi, özkanal sayısı 5 olarak sabitlendiğinde özses sayısının [10,50] aralığında arttırılması sistem performansı üzerinde çok önemli bir değişikliğe neden olmazken (yaklaşık EER %9), özkanal sayısının arttırılması KD sistemini olumsuz yönde etkilemektedir. Örneğin özkanal sayısı 20'den 25'e çıkarıldığında performansta yaklaşık olarak %46 değerinde bir

düşüş meydana gelmektedir (Çizelge 4.13). MFKK öznitelikleri kullanılarak yapılan JFA deneyilerinde en düşük hata oranı % 6,92'dir. JFA modelinde konuşmacı ve kanal etmenleri için bulunan optimum değerler ise 5'tir. Sistemin DET eğrisi Şekil 4.7'de gösterilmiştir.



Şekil 4.7 : MFKK öznitelikleri ile elde edilen optimum JFA modeli.

JFA modeli ile yapılan deneylerde ikinci olarak DGKK öznitelik yöntemi kullanılmıştır. DGKK katsayıları MFKK'da olduğu gibi 18 olarak sabit tutulmuştur. Gauss bileşen sayısı 128 olarak sabit tutulmuştur. Daha önce MFKK öznitelikleriyle yapılan deneyler DGKK öznitelikleri kullanılarak tekrar yapılmıştır. Deneysel sonuçlar Çizelge 4.14'te ve Çizelge 4.15'te gösterilmiştir.

Çizelge 4.14 : DGKK öznitelikleriyle farklı konuşmacı etmenleri için EER değerleri.

Özkanal Sayısı	Özses Sayısı	EER (%)
5	5	18,64
5	10	17,76
5	15	17,89
5	20	17,69
5	25	17,86
5	30	17,74
5	35	17,62
5	40	17,69
5	45	17,71
5	50	17,53

Çizelge 4.14'te görüldüğü gibi, özkanal sayısı sabit tutulup, özses sayısı arttırıldığında EER değeri neredeyse sabit olup %17 civarındadır. Ancak özses sayısı sabit tutulup özkanal sayısı arttırıldığında sistem performansı ciddi bir şekilde düşmektedir. Bu durum Çizelge 4.15'te gösterilmiştir.

Çizelge 4.15 : DGKK öznitelikleriyle farklı kanal etmenleri için EER değerleri.

Özkanal Sayısı	Özses Sayısı	EER (%)
50	5	44,91
45	5	44,61
40	5	44,04
35	5	44,62
30	5	44,68
25	5	43,96
20	5	38,75
15	5	33,18
10	5	23,40

JFA modeli ile yapılan deneylerden elde edilen sonuçlara göre, özkanal sayısı arttıkça sistem performansı düşmektedir (Çizelge 4.15). DGKK öznitelikleriyle yapılan deneylerde optimum parametreler özkanal sayısı için 5, özses sayısı için 50 olarak elde edilmiştir ve EER değeri %17,53'tür.

JFA modeli ile oluşturulan KD sisteminde en düşük hata oranı MFKK öznitelikleriyle elde edilmiş olup %6,92'dir (Çizelge 4.12). Bu sebepten dolayı, JFA modeli için gauss bileşeni 128, MFKK öznitelik boyutu 18, özses sayısı ve özkanal sayısı 5 olarak önerilmektedir. Sistemin DET eğrisi Şekil 4.7'de gösterilmiştir.

4.5 i-Vektör Yaklaşımı İle KD Sistem Performansı

i-vektör öznitelikleri çıkarılırken, GKM – GAM sınıflandırıcısında kullanılmış olan genel arkaplan modeli kullanılmıştır. **T** matrisi, i-vektör çıkarıcısıdır ve GAM eğitimi için kullanılan ses örnekleri EM algoritması kullanılarak 5 iterasyon ile gerçekleştirilmiştir. Konuşmacılardan her biri için eğitim öznitelik vektörlerinden ve test öznitelik vektörlerinden 400 boyutlu i-vektör öznitelik vektörü elde edilmiştir. Elde edilen i-vektör özniteliklerinin birim uzunluğa sahip olmaları için uzunluk normalizasyonu uygulanmıştır [62].

PLDA sınıflandırıcısı, GAM modeli eğitiminde ve i-vektör çıkarıcısı matrisinin eğitiminde kullanılan ses sinyallerinden H_s ve H_d hipotezleri (Denklem 3.29) eğitilmiştir. Sonra, i-vektör yöntemi CDS ve PLDA sınıflandırıcıları kullanılarak konuşmacı doğrulama deneyleri gerçekleştirilmiştir.

Bayan konuşmacılar için elde edilen EER değerleri Çizelge 4.16'da, erkek konuşmacılar için elde edilen EER değerleri ise Çizelge 4.17'de gösterilmiştir.

Çizelge 4.16 : Bayan konuşmacılar için EER değerleri.

CDS	PLDA
% 9,64	% 21,99

Çizelge 4.17 : Erkek konuşmacılar için EER değerleri.

CDS	PLDA
% 9,98	% 24,41

Çizelge 4.16 ve Çizelge 4.17'den anlaşıldığı gibi, PLDA sınıflandırıcısı CDS sınıflandırıcısına nazaran daha düşük bir performans göstermiştir. Metinden bağımsız konuşmacı tanıma uygulamalarında PLDA sınıflandırıcısının genellikle CDS sınıflandırıcısına göre daha iyi performansa sahip olduğu bilinse de Çizelge 4.16 ve Çizelge 4.17'de görüldüğü gibi, Türkçe metne bağlı uygulamalarda CDS, PLDA sınıflandırıcısına göre yaklaşık iki kat daha iyi performans göstermiştir.

4.6 Tartışma

Bu tezde, Türkçe metne bağlı konuşmacı doğrulama sistemi geliştirilerek metinden bağımsız konuşmacı doğrulama uygulamalarında en çok kullanılan yöntemlerin, Türkçe metne bağlı konuşmacı doğrulama sistemindeki performansları analiz edilerek ele alınmıştır.

Deneysel çalışmalarda elde edilen hata oranı en düşük, performansı en iyi olan optimum modeller özet olarak Çizelge 4.18'de gösterilmiştir.

Çizelge 4.18 : Farklı KD sistemleri için elde edilen EER değerleri.

GKM – GAM	GKM – DVM	JFA	i-vektör
% 4,62	% 5,77	% 6,92	% 9,64 (CDS)

Bilinen klasik yöntemlerden GKM – GAM sınıflandırıcısı diğer sınıflandırıcılara nazaran daha iyi başarı performansı göstermiştir.

Ayrıca PLDA sınıflandırıcısı performansta herhangi bir artış sağlamamış tam tersi sistem performansını önemli ölçüde düşürmüştür.

MFKK öznitelik sayısının da, konuşmacı doğrulama sistem performansı üzerinde önemli bir etkisinin olduğu belirlenmiştir. MFKK öznitelikleriyle eğitilen sınıflandırıcıların performansı, DGKK özniteliklerine nazaran çok daha iyidir.

KAYNAKLAR

[1] Reynolds, D. A. (1992). *A Gaussian Mixture Modeling Approach to Text Independent Speaker Identification*. (Doktora tezi). Georgia Institude of Technology.

[2] Furui, S. (1997). Recent Advances in Speaker Recognition, *Pattern Recognition Letters*, 18(9), 859-872.

[3] Matsui, T. and Furui, S. (1995). Speaker Recognition Tecnology, *NNT Review*, 7(2), 40-48.

[4] Hanilçi, C. (2013). *Konuşmacı Tanımada Map Uyarlamalı Sınıflandırıcılar* (Doktora tezi). Uludağ Üniversitesi, Fen Bilimleri Enstitüsü, Bursa.

[5] Reynolds, D. A. (2002). An overview of automatic speaker recognition technology, *ICASSP'02*, (ss. 4072-4075). Orlando USA, Mayıs 13-17.

[6] Furui, S. (1995). Speech recognition-past, present, and future, *NTT review*, 7(2), 13.

[7] Furui, S. (2005). 50 years of progress in speech and speaker recognition research, *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, 1(2), 64-74.

[8] Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors, *Speech communication*, 52(1), 12-40.

[9] Doddington, G. R. (1985). Speaker Recognition – Identifying People by Their Voices, *Proceedings of the IEEE*, 73(11), 1651-1664.

[10] Büyük, O. (2015). Experiment on Fast Scoring for GMM Based Speaker Verification, *23.Sinyal İşleme ve İletişim Uygulamaları Kurultayı*, (ss.140-143). Malatya, Mayıs 16-19.

[11] Wang, L. (2002). Capture interspeaker information with a neural network for speaker identification, *IEEE Neural Networks*, 13(2), 436-445.

[12] Hanilçi, C. (2007). *Konuşmacı Tanıma Yöntemlerinin Karşılaştırmalı Analizi* (Yüksek lisans tezi). Uludağ Üniversitesi, Fen Bilimleri Enstitüsü, Bursa.

[13] Farah, S., & Shamim, A. (2013). Speaker Recognition System Using Mel - Frequency Cepstrum Coefficients, Linear Prediction Coding and Vector Quantization, *International Conference on Computer, Control & Communication*, (ss.1-5). Karachi: Pakistan, 25-26 Eylül.

[14] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE transactions on acoustics, speech, and signal processing*, 28(4), 357-366.

[15] Kinnunen, T. (2003). *Spectral Features for Automatic Speaker Recognition* (Doktora tezi). University of Joensuu, Finland.

[16] Campbell, J. P. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9), 1437-1462.

[17] Rose, P. (2002). *Forensic Speaker Identification*. New York, Taylor & Francis Forensic.

[18] Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of Speech Recognition*. New Jersey, Prentice Hall Englewood Cliffs.

[19] Atal, B. S. (1974). Effectiveness of Linear Prediciton Characteristics of the Speech wave for Automatic Speaker Identification and Verification, *Journal of the Acoustical Society of America*, 55(6), 1304-1312.

[20] Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave, *The journal of the acoustical society of America*, 50(2B), 637-655.

[21] Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition, *The Journal of the Acoustical Society of America*, 51(6B), 2044-2056.

[22] Hermansky, H., & Morgan, N. (1994). RASTA Processing of Speech, *IEEE Transactions on Speech and Audio Processing*, 2(4), 578-589.

[23] Hegde, R. M., Murthy, H. A., & Gadde, V. R. R. (2007). Significance of the modified group delay feature in speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1), 190-202.

[24] Karpov, E. (2003). *Real-Time Speaker Identification* (Yüksek lisans tezi). University of Joensuu, Finlandiya.

[25] Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A review on speech recognition technique, *International Journal of Computer Applications*, 10(3), 16-24.

[26] Yu, K., Mason, J., & Oglesby, J. (1995). Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation, *IEE Proceedings-Vision, Image and Signal Processing*, 142(5), 313-318.

[27] Deng, J. & Hu, Q. (2003). Open Set Text-Independent Speaker Recognition Based on Set-Score Pattern Classification, *ICASSP'03*, (ss.73-77). Hong Kong, Nisan 6-10.

[28] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models, *Digital signal processing*, 10(1-3), 19-41.

[29] Gish, H., & Schmidt, M. (1994). Text-independent speaker identification, *IEEE signal processing magazine*, 11(4), 18-32.

[30] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, 77(2), 257-286.

[31] Tisby, N. Z. (1991). On the application of mixture AR hidden Markov models to text independent speaker recognition, *IEEE Transactions on Signal Processing*, 39(3), 563-570.

[32] Matsui, T., & Furui, S. (1994). Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's, *IEEE transactions on speech and audio processing*, 2(3), 456-459.

[33] Lippmann, R. (1987). An introduction to computing with neural nets, *IEEE Assp magazine*, 4(2), 4-22.

[34] Park, J., & Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks, *Neural computation*, 3(2), 246-257.

[35] Park, J., & Sandberg, I. W. (1993). Approximation and radial-basis-function networks, *Neural computation*, *5*(2), 305-316.

[36] Zhang, Q., & Benveniste, A. (1992). Wavelet networks, *IEEE transactions on Neural Networks*, 3(6), 889-898.

[37] Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks:: The state of the art, *International journal of forecasting*, 14(1), 35-62.

[38] Bennani, Y., & Gallinari, P. (1991). On the use of TDNN-extracted features information in talker identification, *ICASSP'91*, (ss. 385-388). Toronto, Kanada, Nisan.

[39] Farrell, K. R., Mammone, R. J., & Assaleh, K. T. (1994). Speaker recognition using neural networks and conventional classifiers, *IEEE Transactions on speech and audio processing*, 2(1), 194-205.

[40] Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models, *IEEE transactions on speech and audio processing*, 3(1), 72-83.

[41] Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech communication*, 17(1-2), 91-108.

[42] Vapnik, V. (1995). *The Nature of Statistical Leanzing Theory*. New York, Springer-Verlag.

[43] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., & Torres-Carrasquillo, P. A. (2006). Support vector machines for speaker and language recognition, *Computer Speech & Language*, 20(2-3), 210-229.

[44] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition, *Data mining and knowledge discovery*, 2(2), 121-167.

[45] Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Joint factor analysis versus eigenchannels in speaker recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4), 1435-1447.

[46] Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Speaker and session variability in GMM-based speaker verification, *IEEE Transactions on Audio Speech and Language Processing*, 15(4), 1448-1460.

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

[48] Glembek, O., Burget, L., Dehak, N., Brummer, N., & Kenny, P. (2009). Comparison of scoring methods used in speaker recognition with joint factor analysis, *ICASSP'09*, (ss. 4057-4060). Taipei, Taiwan, Nisan 19-24.

[49] Verma, P., & Das, P. K. (2015). i-Vectors in speech processing applications: a survey, *International Journal of Speech Technology*, 18(4), 529-546.

[50] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification, *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 788-798.

[51] Vergin, R., O'shaughnessy, D., & Farhat, A. (1999). Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition, *IEEE Transactions on speech and audio processing*, 7(5), 525-532.

[52] Cochran, W. T., Cooley, J. W., Favin, D. L., Helms, H. D., Kaenel, R. A., Lang, W. W., ... & Welch, P. D. (1967). What is the fast Fourier transform?, *Proceedings of the IEEE*, *55*(10), 1664-1674.

[53] Claudio, B., & Ricotti, L. P. (1999). *Speech Recognition Theory and C++ Implementation*. England, John WILEY&Sons.

[54] Murthy, H. A., & Gadde, V. (2003). The modified group delay function and its application to phoneme recognition, *ICASSP'03*, (ss.1-68). Hong Kong, Nisan 6-10.

[55] Hegde, R. M., Murthy, H. A., & Gadde, V. R. R. (2004). Continuous speech recognition using joint features derived from the modified group delay function and MFCC, *In 8th International Conference on Spoken Language Processing*, (ss.905-908). Korea, Ekim 4-8.

[56] Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification, *IEEE Signal Processing Letters*, 13(5), 308-311.

[57] Wan, V., & Campbell, W. M. (2000). Support vector machines for speaker verification and identification, *In Neural Networks for Signal Processing X*, (ss.775-784). Sydney, Australia, Aralık 11-13.

[58] McLaren, M., Vogt, R., Baker, B., & Sridharan, S. (2007). A comparison of session variability compensation techniques for SVM-based speaker recognition, *8th Annual Conference of the International Speech Communication Association*, (ss.790-793). Antwerp, Belgium, Ağustos 27-31.

[59] Kenny, P. (2005). Joint factor analysis of speaker and session variability: Theory and algorithms, *CRIM, Montreal,(Report) CRIM-06/08-13*, *14*, 28-29.

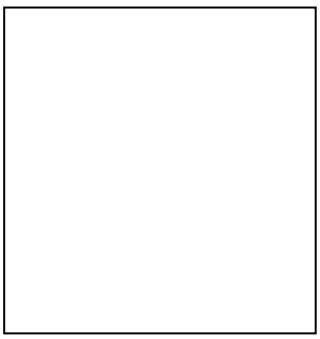
[60] Chen, L., & Yang, Y. (2013). Emotional speaker recognition based on i-vector through atom aligned sparse representation, *ICASSP'13*, (ss.7760-7764). Vancouver, BC, Canada, Mayıs 26-31.

[61] Prince, S. J., & Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity, *11th International Conference on Computer Vision*, (ss.1-8). Rio de Janeiro, Brazil, Ekim 14-20.

[62] Garcia-Romero, D. & Espy-Wilson C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems, *12th Annual Conference of the Intenational Speech Communication Association*, (ss.249-252). Florence, Italy, Ağustos 27-31.

[63] Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of detection task performance (Rapor No. 5). ABD : National Inst of Standards and Technology Gaithersburg MD.

ÖZGEÇMİŞ



Ad-Soyad **:** HAVVA ÇELİKTAŞ

Doğum Tarihi ve Yeri **:** 11/06/1991

E-posta **:** havvaceliktas@gmail.com

ÖĞRENİM DURUMU:

- **Lisans** **:** 2014, Uludağ Üniversitesi, Mühendislik Fakültesi, Elektronik Mühendisliği
- **Yüksek Lisans** **:** 2019, Bursa Teknik Üniversitesi, Elektrik Elektronik Mühendisliği A.B.D.

TEZDEN TÜRETİLEN ESERLER, SUNUMLAR VE PATENTLER:

- **Çeliktaş, H., & Hanilçi, C.** (2017). Impact of Background Language On Turkish Text Dependent Speaker Verification, *5th International Conference on Advanced Technology Sciences*, (ss.607-611). İstanbul, Mayıs 9-12.
- **Çeliktaş, H., & Hanilçi, C.** (2017). A study on Turkish text-dependent speaker recognition, *25.Sinyal İşleme ve İletişim Uygulamaları Kurultayı*, (ss.1-4). Antalya, Mayıs 15-18.
- **Hanilçi, C., & Çeliktaş, H.** (2018). Turkish text-dependent speaker verification using i-vector/PLDA approach, *26.Sinyal İşleme ve İletişim Uygulamaları Kurultayı,* (ss. 1-4). İzmir, Mayıs 2-5.