
T.C.
Uludağ Üniversitesi
Fen Bilimleri Enstitüsü
KONUŞMACI TANIMADA MAP UYARLAMALI SINIFLANDIRICILAR
Cemal HANİLÇİ
Doktora Tezi

KONUŞMACI TANIMADA MAP UYARLAMALI SINIFLANDIRICILAR

Cemal HANİLÇİ

T.C.
ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KONUŞMACI TANIMADA MAP UYARLAMALI SINIFLANDIRICILAR

Cemal HANİLÇİ

Yrd. Doç. Dr. Figen ERTAŞ
(Danışman)

DOKTORA TEZİ
ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

BURSA–2013

TEZ ONAYI

CEMAL HANİLÇİ tarafından hazırlanan "KONUŞMACI TANIMADA MAP UYARLAMALI SINIFLANDIRICILAR" adlı tez çalışması aşağıdaki jüri tarafından oy birliği ile Uludağ Üniversitesi Fen Bilimleri Enstitüsü Elektronik Mühendisliği Anabilim Dalı'nda **DOKTORA TEZİ** olarak kabul edilmiştir.

Danışman : Yrd. Doç. Dr. Figen ERTAŞ

Başkan:	Yrd. Doç. Dr. Figen ERTAŞ Uludağ Üniversittesi, Mühendislik-Mimarlık Fakültesi, Elektronik Mühendisliği Anabilim Dalı	İmza
Üye:	Prof.Dr. Tuncay ERTAŞ Uludağ Üniversittesi, Mühendislik-Mimarlık Fakültesi, Elektronik Mühendisliği Anabilim Dalı	İmza
Üye:	Prof.Dr. Erdoğan DİLAVEROĞLU Uludağ Üniversittesi, Mühendislik-Mimarlık Fakültesi, Elektronik Mühendisliği Anabilim Dalı	İmza
Üye:	Prof.Dr. İsmail AVCIBAŞ Turgut Özal Üniversitesi, Mühendislik Fakültesi, Elektrik-Elektronik Mühendisliği Anabilim Dalı	İmza
Üye:	Prof.Dr. Ahmet Hamdi KAYRAN İstanbul Teknik Üniversitesi, Elektrik-Elektronik Mühendisliği Fakültesi, Elektronik ve Haberleşme Mühendisliği Anabilim Dalı	İmza

Yukarıdaki sonucu onaylarım / ONAY

Prof. Dr. Ali Osman DEMİR
Enstitü Müdürü
… / ... / 2013

U.Ü. Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

07/05/2013

Cemal HANİLÇİ

ÖZET

Doktora Tezi

KONUŞMACI TANIMADA MAP UYARLAMALI SINIFLANDIRICILAR

Cemal HANİLÇİ

Uludağ Üniversitesi
Fen Bilimleri Enstitüsü
Elektronik Mühendisliği Anabilim Dalı

Danışman: Yrd. Doç. Dr. Figen ERTAŞ

Konuşmacı tanıma, üzerinde uzun zamandır çalışılan ancak henüz arzu edilen başarım oranlarına erişilememiş zorlayıcı bir örüntü tanıma uygulamasıdır. Güvenlik sistemleri, adli uygulamalar, telefon bankacılığı ve erişim kontrolü gibi birçok alanda kullanılan konuşmacı tanıma sistemlerinin başarımı, veri miktarı, kayıtlardaki gürültü, kayıtlar arasındaki oturum farklılıkları, kullanılan öznitelik vektörleri ve sınıflandırıcı algoritmalar gibi birçok bileşenden kolayca etkilenmektedir. Bu tezde bu bileşenlerin metinden bağımsız konuşmacı tanıma performansına etkileri, güncel ve başarılı sınıflandırma yöntemleri kullanılarak incelenmiştir.

Sınıflandırıcı algoritmalar olarak Gauss karışım modeli (GMM), vektör nicemleme (VQ), en büyük ardıl olasılık (MAP) tabanlı (genel arkaplan modeli-UBM) GMM ve VQ (GMM-UBM ve VQ-UBM), Destek Vektör Makineleri (SVM) ve GMM süpervektör (GMM-SV) yöntemleri kullanılmıştır. UBM yönteminin konuşmacı tanımaya etkisi öncelikli olarak incelenmiş olup GMM, VQ, GMM-UBM ve VQ-UBM yöntemleri karşılaştırılmıştır. Konuşmacı tanımada eğitim ve test veri miktarlarının performansa etkileri detaylı olarak incelenmiş olup, GMM-UBM, VQ-UBM, SVM ve GMM-SV yöntemlerinin karşılaştırılması yapılmıştır. En çok kullanılan konuşmacıyı karakterize eden öznitelik vektörlerinden olan Mel-frekansı kepstrum katsayılarının (MFCC) toplamsal gürültü durumunda konuşmacı tanıma performansı incelenmiş olup, toplamsal gürültü nedeniyle tanıma başarımında meydana gelen düşüşü iyileştirmek amacı ile öznitelik çıkarımında değişik yaklaşımlar önerilmiştir. Ayrıca diğer bir popüler öznitelik çıkarma yöntemi olan doğrusal öngörü kepstrum katsayıları (LPCC) ile doğrusal öngörü neticesinde ortaya çıkan hata işaretinin konuşmacının kimliği ile ilgili bilgi içerip içermediği incelenmiştir. Oturum farklılıklarından dolayı skor ve öznitelik seviyesinde meydana gelen negatif etkileri azaltmak amacıyla sırası ile test normalizasyonu (TNorm) ve NAP yöntemleri önerilmiştir. Öznitelik vektör boyutunu azaltarak konuşmacı tanıma performansını iyileştirmek amacı ile temel bileşen analizi (PCA) tabanlı bir öneri sunulmuştur.

Anahtar Kelimeler: Konuşmacı tanıma, öznitelik çıkarımı, sınıflandırma algoritmaları.

2013, xii+103 sayfa

ABSTRACT

PhD Thesis

MAP ADAPTED CLASSIFICATION TECHNIQUES FOR SPEAKER RECOGNITION

Cemal HANİLÇİ

Uludağ University
Graduate School of Natural and Applied Sciences
Department of Electronic Engineering

Supervisor: Yrd. Doç. Dr. Figen ERTAŞ

Speaker recognition is a difficult pattern recognition task which has been focused for decades and the performance is not at the desired levels yet. Speaker recognition, which is used in security systems, forensic application, telephone banking and access control, and its performance depend on various parameters such as data duration, additive noise, session variability, features and modeling technique. In this thesis, the effects of these parameters on text-independent speaker recognition performance are analyzed by utilizing the most recent speaker modeling techniques.

Gaussian mixture models (GMM), vector quantization (VQ), maximum a *posteriori* (MAP) adapted GMM and VQ (also known as universal background model - UBM) (GMM-UBM and VQ-UBM), support vector machines (SVM) and GMM supervector (GMM-SV) are the classification techniques used. First, the effect of MAP adaptation on the speaker recognition performance is analyzed and GMM, VQ, GMM-UBM and VQ-UBM methods are compared in terms of recognition accuracy. The effect of training and test data duration on the recognition performance are investigated and the performance comparison of the four modeling algorithm is considered. Mel-frequency cepstral coefficients (MFCC), the most popular feature extraction method, which parameterizes the speaker characteristics is analyzed for the speaker recognition under additive noise contamination and different approaches are proposed against the performance degradation due to additive noise. Another well-known feature extraction method, linear prediction cepstral coefficients (LPCC) are compared with the features extracted from the linear prediction residual error signal. It is shown that residual error contains information about the speaker which is commonly not thought so. To reduce the effect of the channel variability test normalization (TNorm) and nuisance attribute projection (NAP) methods are proposed for the VQ-UBM and SVM based speaker recognition on the score and feature levels, respectively. Finally, local principal component analysis (PCA) based method is proposed for VQ-UBM based speaker recognition which yields better recognition accuracy than the baseline method by reducing the feature dimension.

Keywords: Speaker recognition, feature extraction, classification techniques.

2013, xii+103 pages

TEŞEKKÜR

Çalışmalarım boyunca sağladığı katkılarıyla ve sabırla bana yol gösteren tez danışmanım Yrd. Doç. Dr. Figen ERTAŞ'a teşekkür ederim. Engin bilgi ve tecrübelerini benimle paylaşan, bilimsel anlamda bana yol gösteren değerli hocam Prof. Dr. Tuncay ERTAŞ'a sonsuz teşekkürlerimi sunarım. Ayrıca bana Doğu Finlandiya Üniversitesi (University of Eastern Finland) ses ve görüntü işleme grubunun (SIPU) bir parçası olma şansını veren Dr. Tomi Henrik Kinnunen ve Prof. Dr. Pasi Fränti'ye ve orada bulunduğum süre zarfında desteklerini esirgemeyen Dr. Ville Hautamäki ve Dr. Rahim Saeidi'ye teşekkür ederim.

Son olarak bugüne kadar sabırları ve anlayışlarıyla bana destek olan aileme, eski ve yeni tüm arkadaşlarıma sonsuz teşekkürler.

Cemal HANİLÇİ

İÇINDEKILER

Sayfa

SİMGELER DİZİNİ

α	Doğrusal öngörü katsayıları
$\mathbf{b}$	Polinomsal açılım vektörü
$b(\mathbf{x})$	Çok boyutlu Gauss dağılımı
$\mathbf{c}$	Kod vektör
$\mathbf{C}$	Kod kitabı
D	Öznitelik vektör boyutu
$e(n)$	Doğrusal öngörü hata işareti
K	GMM ve VQ model boyutu
$K(\mathbf{X}_i, \mathbf{X}_j)$	SVM çekirdek fonksiyonu
m	Polinomsal açılım derecesi
N	Bir çerçevedeki örnek sayısı
p	Doğrusal öngörü derecesi
$p(\mathbf{X}\|\lambda)$	$\mathbf{X}$ öznitelik vektör kümesinin şartlı olasılık işlevi
$R(k)$	Özilinti katsayıları
$\mathbf{r}$	Özilinti vektörü
$\mathbf{r}_{\text{wlp}}$	Ağrılıklandırılmış özilinti vektörü
$\mathbf{R}$	Özilinti matrisi
$\mathbf{R}_{\text{wlp}}$	Ağrılıklandırılmış özilinti matrisi
w_i	GMM ağırlık katsayıları
$w(n)$	Pencere fonksiyonu
$\mathbf{x}$	D-boyutlu bir öznitelik vektörü
$\mathbf{X}$	Öznitelik vektörleri kümesi
μ_i	GMM ortalama vektörleri
$\mathbf{\Sigma}_i$	GMM ortak değişinti matrisleri
λ	RLP düzenleştirme sabiti
$\Phi(\mathbf{a})$	RLP penaltı fonksiyonu
$\Lambda(\mathbf{X})$	Logaritmik olabilirlik oranı
Δ	Öznitelik vektörlerinin birinci türevi
$\Delta\Delta$	Öznitelik vektörlerinin ikinci türevi

Şekil Listesi

Tablo Listesi

KISALTMALAR DİZİNİ

CMVN	Kepstral ortalama ve varyans normalizasyonu
DAC	İkili özilinti katsayıları
DCT	Ayrık kosinüs dönüşümü
DET	Sezim hata ödünleşimi
EER	Eşit hata oranı
EM	Beklentinin maksimumlaştırılması
FFT	Hızlı Fourier dönüşümü
GLDS	Genelleştirilmiş doğrusal ayırtaç dizisi
GMM	Gauss karışım modeli
HMM	Saklı Markov modeli
ICS	Iteratif kepstral yumuşatma
LLR	Logaritmik olabilirlik oranı
LP	Doğrusal öngörü
LPCC	Doğrusal öngörü kepstrum katsayıları
MAP	En büyük ardıl olasılık
MFCC	Mel-frekansı kepstrum katsayıları
MinDCF	En küçük karar bedel fonksiyonu
ML	En büyük olabilirlik
MSE	Ortalama karesel hata
MVDR	Minimum varyans bozunumsuz cevap
NAP	Nuisance attribute projection
RLP	Düzenleştirilmiş doğrusal öngörü
SV	Süper vektör
SVM	Destek vektör makineleri
SWLP	Kararlı ağırlıklandırılmış doğrusal öngörü
TNorm	Test normalizasyonu
UBM	Genel arka plan modeli
VQ	Vektör nicemleme
WLP	Ağırlıklandırılmış doğrusal öngörü

1. GİRİŞ

Gelişen teknoloji ile birlikte biyometrik sistemlerin günlük yaşamda kullanımı her geçen gün artmaktadır. Biyometrik, bir kişinin fiziksel veya davranışsal özelliklerinin kullanılarak kimliğinin tespit edilmesi ile ilgilenen bilim dalıdır. Biyometrik sistemlerin yaygınlaşması, telefon bankacılığı, güvenlik amacı ile giriş kontrolleri ve adli uygulamalar gibi birçok alanda kişi tanıma/doğrulama ihtiyacını doğurmuştur. Genel olarak herhangi bir kişi üç değişik şekilde tanınabilir (Prabhakar ve ark., 2003; Ortega-Garcia ve ark., 2004):

- Kişinin sahip olduğu bir nesne ile (kimlik kartı veya bir anahtar)
- Kişinin sahip olduğu bir bilgi ile (şifre)
- Kişinin kendisinde var olan fiziksel bir özelliği ile (parmak izi, retina, yüz ve ses gibi)

İlk iki metot uzun yıllardır kişi tanıma için kullanılan geleneksel yöntemler olmakla birlikte bu yöntemler bazı sıkıntıları beraberinde getirmektedir. Kimlik kartı veya anahtarın kaybedilme veya çalınma ihtimali ve şifrelerin unutulması veya yanlış girilmesi bu yöntemlerin dezavantajlarındandır. Üçüncü yöntem, biyometrik kişi tanıma/doğrulama olarak adlandırılmakta olup ilk iki yöntem için bahsedilen problemler bu yöntemde bulunmamaktadır (Prabhakar ve ark., 2003; Kittler ve Nixon, 2003). Parmak izi, yüz, iris, retina, el geometrisi, imza ve ses en çok kullanılan biyometrik özellikler arasında yer almaktadır. Biyometrik tanıma sistemleri sistemin güvenilirliği, kullanım kolaylığı, kolay uygulanabilirliği ve düşük maliyet gibi kriterler çerçevesinde gerçek zamanlı uygulamalarda kullanılmaktadır. García ve ark. (2003) bilinen ve yaygın olarak kullanılan biyometrik tanıma sistemlerinin belirtilen kriterlere göre karşılaştırmalarını Çizelge 1.1. de gösterildiği şekilde incelemiştir. Çizelgeden de görüldüğü gibi ses birçok açıdan diğer yöntemlerin önüne geçmektedir. Bunların yanında ses, günümüzde mobil iletişimin yaygınlaşması neticesinde kullanılabilecek en doğal biyometrik bilgi kaynağıdır (Bonastre ve ark., 2003).

Çizelge 1.1: Biyometrik tanıma sistemlerinin karakteristikleri (García ve ark., 2003).

Biyometrik yöntem	Başarım	Kullanım Kolaylığı	Kullanıcı Tercihi	Uygulama Kolaylığı	Maliyet
Parmak izi	Yüksek	Orta	Düşük	Yüksek	Orta
Yüz	Düşük	Yüksek	Yüksek	Orta	Düşük
iris	Orta	Orta	Orta	Orta	Orta
Retina	Yüksek	Düşük	Düşük	Düşük	Yüksek
El geometrisi	Orta	Yüksek	Orta	Orta	Yüksek
İmza	Orta	Yüksek	Yüksek	Düşük	Orta
Ses	Orta	Yüksek	Yüksek	Yüksek	Düşük

Ses işareti birçok bilgi içermektedir. İçerdiği bilgilerden birinci seviyede algılanan ve yararlanılan kelimeler aracılığı ile iletilen mesajdır. İletilen mesajın yanında ses işareti konuşulan dil, konuşan kişinin cinsiyeti, ruh hali ve genel olarak konuşan kişinin kimliği hakkında bilgiler içermektedir. Konuşma tanıma ses işareti aracılığıyla söylenen kelimeleri tanımayı amaçlamaktadır. Bu tezde kısaca *ses işaretlerini kullanarak kişi tanıma* olarak tanımlayabileceğimiz *konuşmacı tanıma* problemi ele alınmaktadır. Konuşmacı tanıma, ses örneklerinin kullanılarak konuşan kişinin kimliğinin tespit edilmesi işlemidir (Kinnunen ve Li, 2010; Bimbot ve ark., 2004; Sturim ve ark., 2007). Kişi tanıma/doğrulama, adli uygulamalar ve konuşma tanıma gibi uygulamalar günümüzde konuşmacı tanıma sistemlerinin temel kullanım alanları arasında yer almaktadır. Adli uygulamalar, konuşmacı tanıma sistemlerinin en önemli uygulama alanlarından biridir. Örneğin bir suçun işlendiği sırada kaydedilmiş bir ses işareti delil olarak kullanılıyor ise, şüpheli kişinin sesi bu ses kaydı ile karşılaştırılarak iki ses arasındaki benzerlik tespit edilebilir (Rose, 2006). Konuşmacı tanıma işleminin adli uygulamaları Rose (2002) tarafından detaylı bir şekilde incelenmiştir.

Konuşmacı tanıma işleminin en önemli avantajı ses işaretinin doğal olarak üretilmesinden kaynaklanmaktadır. İnsanlar arasındaki iletişimin temeli ses işaretlerine dayanmaktadır. Bir diğer avantaj ise, herhangi bir özel, yüksek maliyetli bir cihaza gerek duymadan sadece bir mikrofon aracılığı ile ses işaretleri kayıt altına alınabildiğinden dolayı düşük maliyetli olmasıdır. Bu durum parmak izi ve retina için geçerli

değildir. Bunun nedeni, bu uygulamalarda özel bir tarayıcı cihaza ihtiyaç duyulmasıdır. Son yıllarda ses işaretinin diğer biyometrik tanıma yöntemleri ile birlikte kullanımının (*çok şekilli kişi tanıma*) kişi tanıma performansını artırdığı belirlenmiştir (Toh, 2003; Toh ve Yau, 2005; Brunelli ve Falavigna, 1995). 2003 yılında yapılan uluslararası ses ve görüntü temelli kişi tanıma konferansında (*Audio and Video Based Person Authentication-AVBPA*) ses işareti, yüz ve parmak izinden sonra üçüncü popüler biyometrik özellik olarak belirlenmiştir (Kittler ve Nixon, 2003). Yine bu konferansta dokuz değişik çok şekilli kişi tanıma sistemi tanıtılmış ve ses ile konuşmacı tanıma bu dokuz sistemin altı tanesinde yer almıştır.

Genel olarak konuşmacı tanımanın güvenilir bir tanıma sistemi olmadığı inancı oldukça yaygın bir görüştür. Bunun sebebi, kişinin sesinin, örneğin parmak izi gibi değişmez olmamasıdır. Parmak izi ve ses işareti arasındaki temel fark, parmak izinin kişinin vücudundan direkt olarak elde edilebilmesidir. Ses işareti ise daha çok kişinin vücut hareketlerinin neticesinde elde edilen davranışsal bir işarettir. Çünkü ses işareti, kişinin ses üretiminde kullandığı organların hareketlerinin sonucu ortaya çıkan bir işarettir. Örneğin, bir kişinin farklı zamanlarda söylemiş olduğu aynı içeriğe sahip (söylenen kelime veya cümle) iki ses işareti hiçbir zaman birbirinin aynısı olmayacaktır.

Bir konuşmacı tanıma sistemi *işitsel*, *yarı-otomatik* veya *otomatik* olmak üzere üç farklı şekilde tasarlanabilir. İşitsel konuşmacı tanımada ses ve kullanılan dil alanında uzman bir kişi ses işaretlerini dinleyerek konuşmacının kimliği hakkında bir yargıda bulunur. Yarı otomatik sistemlerde yine konusunda uzman bir kişi ses işaretinin spektrogram ve dalga şekli gibi farklı özelliklerini görsel olarak inceleyerek konuşmacının kimliği hakkında fikir beyan eder. Otomatik sistemlerde ise tanıma işleminin bütün evreleri herhangi bir uzman kişinin müdahalesi olmadan bilgisayar aracılığıyla gerçekleştirilmektedir. Bu tezde *otomatik konuşmacı tanıma* sistemleri incelenmektedir.

Konuşmacı tanıma işlemi, konuşmacı belirleme ve konuşmacı doğrulama olmak üzere iki gruba ayrılır (Reynolds, 2002; Petrovska-Delacrétaz ve ark., 2007). Konuşmacı

belirleme, bilinmeyen konuşmacıdan elde edilen ses örneğinin daha önce veritabanına kaydedilmiş N adet bilinen konuşmacıdan hangisine ait olduğunun belirlenmesi işlemidir. Konuşmacı belirlemede ses örneği veritabanındaki N adet konuşmacının her biri ile karşılaştırılır ve en yüksek benzerliği sağlayan konuşmacı, bilinmeyen ses işaretinin sahibi olarak karar verilir. Konuşmacı doğrulama, verilen bir ses örneğinin iddia edilen kişiye ait olup olmamasına karar verilmesi işlemidir. Konuşmacı doğrulamada ses örneği ile iddia edilen konuşmacı modeli karşılaştırılır. Hesaplanan benzerlik derecesi önceden belirlenmiş bir eşik değerinden yüksek ise iddia kabul edilir. Benzerlik derecesi eşik değerden düşük ise reddedilir. Konuşmacı belirleme genellikle daha zor bir problem olarak düşünülmektedir. Bunun nedeni, konuşmacı belirlemede veritabanına kaydedilen konuşmacı sayısı arttıkça sistemin yanlış karar verme ihtimalinin de artmasıdır (Doddington, 1985). Konuşmacı doğrulamada ise böyle bir durum söz konusu değildir. Sadece iki konuşmacı arasında (bilinmeyen ses işareti ve iddia edilen kişi) bir karşılaştırma işlemi yapıldığından dolayı sistemin performansı veritabanındaki konuşmacı sayısından etkilenmemektedir.

Konuşmacı belirleme, *Açık Küme* ve *Kapalı Küme* olmak üzere iki farklı şekilde yapılabilir (Furui, 1997). Bilinmeyen ses işaretinin veritabanındaki kişilerden mutlaka birine ait olduğu varsayımı yapılıyor ise bu işlem *kapalı küme konuşmacı belirleme* olarak adlandırılır. Bilinmeyen ses işaretinin veritabanına kayıtlı kişilerin dışında ve sistem tarafından bilinmeyen bir konuşmacıya ait olabileceği ihtimali de göz önünde bulundurularak bir belirleme işlemi yapılıyor ise bu işleme *açık küme konuşmacı belirleme* adı verilmektedir. Açık küme konuşmacı belirleme, kapalı küme belirleme işlemine göre daha zordur. Bunun nedeni, kapalı küme belirlemede sistem, benzerlik derecesi ne kadar küçük olursa olsun bilinen konuşmacılardan birine karar vermeye zorlanırken, açık kümede sistem daha önce belirlenmiş bir tolerans seviyesine sahip olmalıdır ki bilinmeyen ses örneği ile konuşmacı arasındaki benzerlik derecesi bu tolerans seviyesi sınırlarında ise karar verilmesidir. Bu açıdan bakıldığında aslında konuşmacı doğrulama işleminin, veritabanında sadece bir konuşmacının bulunduğu ($N = 1$) açık küme belirleme işleminin özel bir halinin olduğu aşikârdır.

Bunlara ilaveten konuşmacı tanıma işlemi, *metinden bağımsız* ve *metine bağımlı* olmak üzere iki gruba daha ayrılır. Metine bağımlı sistemlerde kişinin hangi kelimeyi veya cümleyi söyleyeceği önceden bilinmektedir. Metinden bağımsız sistemlerde ise kişi ne söyleyeceğine kendisi karar verir ve bu konuda herhangi bir sınırlama yoktur.

Bir konuşmacı tanıma sistemi temel olarak iki adımdan oluşmaktadır. Bu adımlardan birincisi, bilinen kişilerin sisteme tanıtıldığı ve bu kişilerin veritabanına kaydedildiği adım olan *eğitim* aşaması, ikincisi ise bilinmeyen ses örneğinin sistemde kayıtlı kişiler ile karşılaştırılarak karar verme işleminin gerçekleştirildiği *sınıflandırma (test)* aşamasıdır. Bu aşamalar ve detayları bu tezin sonraki bölümlerinde ayrıntılı olarak verilecektir.

Konuşmacı tanıma performansını olumsuz anlamda etkileyen birtakım negatif faktörler bulunmaktadır. Bu faktörlerden bazıları konuşmacının kendisinden kaynaklanmakta olup bazıları da teknik ve çevresel etkilerden kaynaklanmaktadır. Sağlık problemleri (örneğin soğuk algınlığı), konuşmacının ruhsal durumu (depresyon, sinir, mutsuzluk v.b.), yaşlanma ve kilo değişimi gibi nedenlerden dolayı ses işaretindeki değişiklikler konuşmacının kendisinden kaynaklanan etkilerdir (Kinnunen, 2003). Bir konuşmacıdan aynı gün içerisinde aynı teknik şartlarda alınan iki ses örneğinin birbiriyle tam olarak örtüşmediği bilinen bir gerçektir. Bu nedenle bir konuşmacı sisteme tanıtılırken kullanılan ses işaretlerinin farklı zamanlarda kaydedilmiş sesler olması önemli bir avantaj sağlamaktadır. Kayıt ortamından kaynaklanan toplamsal gürültü (arka plan gürültüsü, yankı v.b.), eğer telefon hattından kayıt yapılıyor ise iletim kanalından kaynaklanan gürültü, iletim için sesin kodlanması ile oluşan kayıp ve mikrofondan kaynaklanan bozulmalar gibi etkiler ise teknik ve çevresel nedenler ile ortaya çıkan negatif etkilerdir. Kayıt sırasında ortamdaki gürültüler (bilgisayar fan sesi, motor sesi, trafik sesi v.b.) orijinal ses işaretine eklenir ve bu işarette bozulmalara neden olur. Ayrıca kayıt, yankıya sebep olan bir ortamda yapılıyor ise orijinal sesin kendisi belirli bir gecikme ile kayıt sesine eklenir (Castellano, 1996). Düşük kaliteli mikrofonlar da ses işaretinde bozulmalara neden olmaktadır. Örneğin, Quatieri ve ark. (2000) aynı ses işaretini farklı kalitedeki mikrofonlarla kaydederek incelemiş ve düşük kaliteli mikrofonların işaretin spektrumunda bozulmalara neden

olduğunu göstermiştir. Ayrıca ses isaretine uygulanan kodlama yönteminin türü de konuşmacı tanıma başarımını etkilemektedir (Besacier ve ark., 2000). Bütün bu negatif etkiler konuşmacı tanıma performansını düşürmektedir. Ancak performanstaki bu düşüş eğitim ve test sesleri arasında uyuşmazlık olduğunda ortaya çıkmaktadır (Reynolds, 2002; Mammone ve ark., 1996). Eğitim ve test sesleri arasindaki uyuşmazlık ile anlatılmak istenen, eğitim ve test aşamalarında kullanılan ses örneklerinin kayıtları sırasında ortaya çıkan etkilerin aynı olmamasıdır. Örneğin, eğitim sesleri gürültüsüz ortamda (laboratuvar), test sesleri ise telefon hattından veya gürültülü bir ortamda elde edildiğinde bu farklılık performansta düşüşe sebep olmaktadır. Gürültülü de olsa her ikisi de aynı ortamda kaydedilen eğitim ve test sesleri kullanıldığında performansta ciddi anlamda bir düşüş olmadığı yapılan çalışmalarda ortaya konulmuştur.

Bu tezde metinden bağımsız kapalı küme konuşmacı tanıma sistemi ele alınmıştır. Yapılan araştırmaların büyük bölümünde konuşmacı doğrulama sistemi incelenmiştir. Bu tezin amaçları şu şekilde özetlenebilir:

- Literatürde bilinen ve yaygın bir şekilde kullanılan başarılı sınıflandırıcı algoritmaların zayıf ve güçlü noktalarını ortaya koyarken yukarıda bahsedilen negatif etkilerden kaynaklanan performans düşüşlerine çözümler getirmek.
- Konuşmacı tanıma sistemleri için alternatif öznitelik çıkarma yöntemleri geliştirmek.
- Telefon hattından kaynaklanan bozulmaları öznitelik seviyesinde azaltmak için kanal dengeleme yöntemleri ortaya koymak.
- Konuşmacı tanıma performansını iyileştirmek amacıyla skor seviyesinde çözümler ortaya koymak.
- Toplamsal gürültü durumunda konuşmacı tanıma performansını iyileştirici yöntemler geliştirmek.
- Bilinen en hızlı ve basit sınıflandırıcı olan VQ yöntemine alternatif olarak TBA temelli bir sınıflandırıcı yöntemi önermek.

1.1. Tez Çalışması İle Elde Edilen Bilimsel Yenilikler

Bu tez sırasında yapılan çalışmalar neticesinde ortaya koyulan ve yayınlanan bilimsel katkılar ise şu şekilde özetlenebilir:

1. Temel bileşen analizi (PCA) tabanlı bir konuşmacı belirleme sistemi önerilerek gürültüsüz laboratuvar ortamında kayıt yapılan TIMIT ve kablolu sabit telefon hattı aracılığı ile kayıt yapılan NTIMIT veritabanları üzerinde başarımları bilinen basit fakat başarılı bir sınıflandırma yöntemi olan vektör nicemleme (VQ) sınıflandırıcısı ile karşılaştırılmıştır. Yapılan deneyler neticesinde önerilen yöntemin TIMIT veritabanı ile VQ yöntemine oldukça yakın performans gösterirken, NTIMIT veritabanında VQ yöntemine nazaran daha düşük tanıma oranı elde edildiği gösterilmiştir. Ancak skor birleştirme yöntemi ile önerilen yöntem ve VQ skorları bilerştirilerek NTIMIT veritabanı için literatürde daha önce yapılan çalışmalar içerisinde en iyi başarım elde edilmiştir (Hanilçi ve Ertaş, 2009).

2. Son yıllarda önerilen en büyük ardıl (*maximum a Posteriori-MAP*) uyarlamalı VQ yöntemi VQ-UBM için eğitim aşamasında oluşturulan konuşmacı modellerinin temsil kabiliyetini artırmak amacıyla bölgesel TBA metodu geliştirilmiştir. Cep telefonları ile yapılan konuşmalardan oluşan NIST 2001 veritabanı ile yapılan konuşmacı doğrulama deneylerinde bölgesel PCA yönteminin VQ-UBM yöntemine uyarlanması neticesinde daha az sayıda işlem yükü ile daha iyi konuşmacı doğrulama başarımı elde edilmiştir (Hanilçi ve Ertaş, 2011d).

3. VQ-UBM yöntemi ile konuşmacı doğrulama sisteminin performansını iyileştirmek amacı ile skor dengeleme yöntemlerinden Test Normalizasyonu (Tnorm) VQ-UBM yöntemine uyarlanmış ve NIST 2001 veritabanı ile yapılan deneylerde Tnorm yöntemi ile başarım iyileştirilmiştir (Hanilçi ve Ertaş, 2011c).

4. Doğrusal öngörü (*linear prediction*) yönteminin ses işaretine uygulanması ile elde edilen öngörücü katsayılarının kişinin ses üretim mekanizmasını modellediği bilinmektedir. Genellikle doğrusal öngörü neticesinde ortaya çıkan hata

işaretinin konuşmacının kimliği hakkında bilgi içermediği, daha çok konuşulan metin hakkında bilgi içerdiği iddia edilmektedir. Hata sinyali kullanılarak konuşmacı doğrulama sistemi gerçekleştirilerek, öznitelik vektörlerinin hata işaretinden çıkarıldığı ve orijinal işaret kullanılarak çıkarıldığı durumların karşılaştırmaları yapılmıştır. NIST 2001 veritabanı ve MAP uyarlamalı Gauss karışım modeli (GMM-UBM) sınıflandırıcısı ile yapılan deneylerde, hata işaretinin de konuşmacı kimliği hakkında bilgi içerdiği gösterilmiştir (Hanilçi ve Ertaş, 2011b).

5. Standart VQ ve GMM yöntemleri ile MAP uyarlamalı VQ-UBM ve GMM-UBM yöntemleri konuşmacı doğrulama ve belirleme sistemleri için karşılaştırılmıştır. Ayrıca VQ yönteminde kod kitaplarının oluşturulmasında ve GMM yönteminde model ortalama vektörlerinin başlangıç atamalarında LBG ve K-means yöntemlerinin kullanılmasının performansa etkisi incelenmiştir. TIMIT, NTIMIT ve NIST 2001 veritabanları ile yapılan konuşmacı doğrulama ve belirleme deneylerinde MAP uyarlamalı sistemlerin standart sınıflandırıcılara göre daha iyi performans gösterdiği ve LBG algoritmasının K-ortalama algoritmasına nazaran daha iyi başarım verdiği elde edilmiştir (Hanilçi ve Ertaş, 2011a).

6. Destek vektör makineleri (SVM) ile konuşmacı doğrulama sisteminde konuşmacıların ses işaretlerinin farklı zamanlarda kaydedilmesi sebebiyle (*oturum farkı*) işarette telefon hattındaki farklılıklardan kaynaklanan bozulmaları gidermek amacıyla öznitelik vektörleri üzerinde kanal dengelemesi işlemi uygulanmıştır. NIST 2002 veritabanı ile yapılan deneylerde önerilen kanal dengeleme yönteminin konuşmacı doğrulama performansını iyileştirdiği gösterilerek *MAP* uyarlamalı GMM-UBM yöntemi ve SVM yöntemlerinin karşılaştırmalı analizleri yapılmıştır (Hanilçi ve Ertaş, 2012).

7. Konuşmacı doğrulamada toplamsal gürültünün performansta ortaya çıkardığı olumsuz etkiyi azaltmak amacıyla öznitelik vektörlerini çıkarırken kullanılan spektrum hesaplama yöntemine değişik alternatif öneriler getirilmiştir. Yapılan çalışmalarda spektrum hesaplamanın konuşmacı doğrulama performansında oldukça büyük etkisinin olduğu ortaya koyulmuş olup yaygın olarak kullanı-

lan hızlı Fourier dönüşümü (HFD) ile 11 farklı spektrum hesaplama yöntemi karşılaştırılmıştır (Hanilçi ve ark., 2012c).

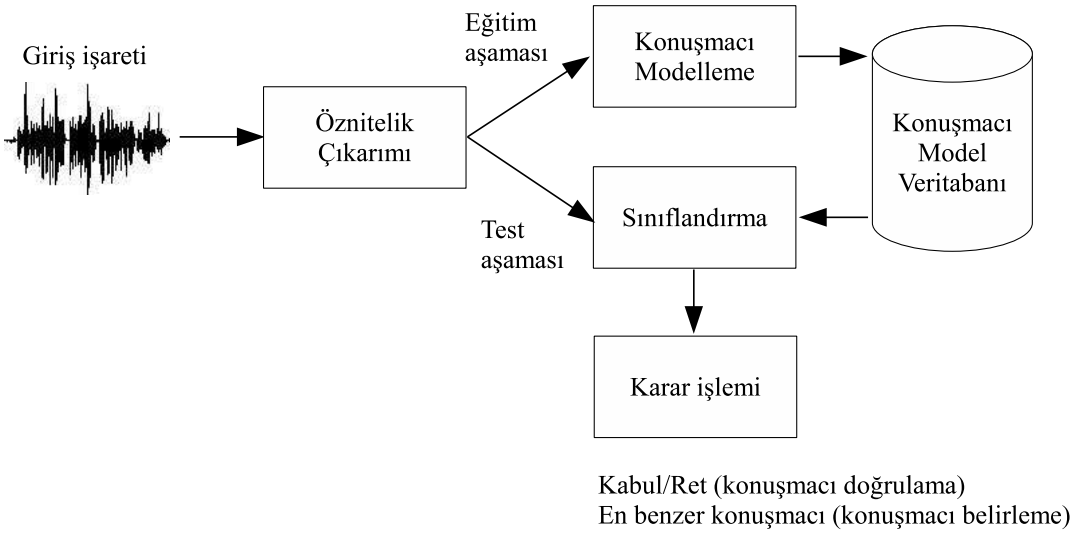
8. Yakın zamanda ses kodlama uygulamaları için önerilen düzenleştirilmiş doğrusal öngörü (RLP) yöntemi konuşmacı tanımada öznitelik çıkarımı için kullanılmıştır ve performansta önemli bir artış elde edilmiştir. Ayrıca RLP yönteminde özilinti ortamında kullanılan pencere fonksiyonu için alternatif bir yöntem önerilmiş ve orijinal yöntemden daha iyi performans elde edilmiştir (Hanilçi ve ark., 2012b,a).

9. Konuşmacı doğrulamada eğitim ve test işaretlerinin sürelerinin etkileri günümüzde kullanılan en popüler sınıflandırıcı algoritmalar ile karşılaştırmalı olarak incelenmiştir. NIST 2002 ve NIST 2005 veritabanları ile yapılan deneylerde günümüzde en iyi sınıflandırıcı yöntem olarak bilinen Gauss karışım modeli-süpervektör (GMM-SV) sınıflandırıcısının kısa süreli ses işaretleri kullanıldığında başarısız olduğu gösterilmiştir. SVM ve GMM-UBM yöntemleri kısa süreli işaretlerde daha yüksek performans gösterirken, uzun ses işaretleri kullanıldığında GMM-SV yönteminin en iyi başarımı verdiği gösterilmiştir (Hanilçi ve Ertaş, 2013b).

2. kısımda kuramsal olarak konuşmacı doğrulamanın temelleri anlatılmıştır. Ayrıca yine bu bölümde literatür özeti ve konuşmacı tanımanın tarihsel gelişimi kısaca anlatılmıştır. 3. bölümde bu tezde kullanılan öznitelik çıkarma ve sınıflandırma algoritmalarının temelleri verilmiş olup, kullanılan konuşmacı tanıma veritabanları tanıtılmıştır. 4. bölümde elde edilen deneysel sonuçlar verilmiştir ve son olarak 5. bölümde deneysel sonuçlara ilişki tartışma ve gelecekte yapılabilecek çalışmalara ilişkin öneriler yer almaktadır.

2. KAYNAK ÖZETLERİ

Genel bir konuşmacı tanıma sistemi eğitim ve test aşaması olmak üzere iki aşamadan oluşmaktadır (Şekil 2.1). Eğitim aşamasında giriş işaretinden ilgili konuşmacıyı temsil edecek akustik bir model oluşturulur ve test aşamasında kullanılmak üzere veritabanına kaydedilir. Test aşamasında ise bilinmeyen konuşmacıya ait ses örneği veritabanındaki modeller ile karşılaştırılarak karar verme işlemi gerçekleştirilir. Konuşmacı doğrulamada, bilinmeyen konuşmacıya ait ses örneği ile sadece iddia edilen konuşmacı modeli karşılaştırılırken, konuşmacı belirlemede sistemdeki bütün modeller karşılaştırılır.



Şekil 2.1: Genel konuşmacı tanıma sistemi.

2.1. Öznitelik Çıkarımı

Konuşmacı belirleme ve konuşmacı doğrulama sistemleri aynı bileşenlerden oluşmaktadır (Şekil 2.1). Eğitim ve test aşamalarının her ikisinde de ortak olan işlem öznitelik çıkarımıdır. Öznitelik çıkarımı, verilen bir ses işaretini konuşmacıya özgü parametrelerin vurgulandığı, konuşmacı hakkında bilgi içermeyen bölümlerin atıldığı vektörlere dönüştürme işlemidir.

Öznitelik çıkarımı, matematiksel olarak orijinal yüksek boyutlu vektörlerin daha düşük boyutlu vektörlere dönüştürülme süreci olarak tanımlanabilir. Yani öznitelik çıkarımı, $f : \mathcal{R}^N \rightarrow \mathcal{R}^d$, $d \ll N$ şeklinde bir dönüşümdür. Öznitelik çıkarma işleminin yapılmasının iki önemli nedeni vardır. Birincisi, eğitim aşamasında oluşturulan konuşmacı modellerinin ilgili konuşmacıyı daha iyi temsil edebilmesi için eğitim için kullanılan vektör sayısının orijinal işaretin boyutundan daha yüksek olması gerekmektedir. Gerekli olan vektör sayısı, orijinal işaretin boyutuyla üstel olarak artmaktadır. Bu durum literatürde *boyutun laneti* (*curse of dimensionality*) olarak bilinmektedir (Jain ve ark., 2000; Jain ve Zongker, 1997). İkinci neden ise, öznitelik çıkarımı ile hesaplama ve işlem yükünün azaltılmasıdır.

Konuşmacı tanıma amacıyla kullanılacak ideal özniteliklerin taşıması gereken özellikler şu şekilde belirtilmektedir (Rose, 2002):

- Konuşma sırasında sıklıkla ve doğal olarak ortaya çıkmalıdır.
- Kolay hesaplanabilir olmalıdır.
- Taklitlere, bozucu etkilere ve gürültüye karşı dayanıklı olmalıdır.
- Diğer özniteliklerle arasında ilinti olmamalıdır.

Konuşma sırasında sıklıkla ve doğal olarak ortaya çıkmalıdır. Bu sayede öznitelikler kısa süreli konuşma işaretlerinden de kolaylıkla elde edilebilecektir. Ayrıca kolay ölçülebilir olmalıdır çünkü otomatik sistemlerde konunun uzmanı bir kişinin yardımı olmadan da bu öznitelikler hesaplanabilecektir. Diğer öznitelikler ile arasında ilinti olmamalıdır. Bunun nedeni, eğer iki öznitelik birbiri ile ilintili ise herhangi bir kazanç elde edilmeyecektir hatta belki de performansta düşüşe neden olacaktır. Birbirinden ilintisiz öznitelikler birbirinden farklı bilgiler içerecektir.

Bilinen hiçbir öznitelik türü tüm bu özelliklerin tamamını aynı anda sağlamamaktadır. Uygulamanın amacı ve türüne göre olması gereken özellikler arasında tercih yapılmalıdır. Genellikle konuşmacı tanımada ses işaretinin spektrumundan elde edilen öznitelikler kullanılmaktadır. Bu nedenle literatürde spektrumdan elde edilen

özniteliklere spektral öznitelikler adı verilmektedir. Konuşmacı tanımada kullanılan öznitelikler kronolojik olarak incelendiginde, 1960'lı yıllarda gerçek kepstrum katsayıları (*real cepstrum coefficients*) (Oppenheim, 1969), 1970'li yıllarda doğrusal öngörü katsayıları (*linear prediction coefficients-LPC*) (Atal ve Hanauer, 1971) ve doğrusal öngörü kepstrum katsayıları (*linear prediction cepstrum coefficients-LPCC*) (Atal, 1974) ve 1980'li yıllarda mel-frekansı kepstrum katsayıları (MFCC) (Davis ve Mermelstein, 1980) karşımıza çıkmaktadır. Sonraki yıllarda algısal doğrusal öngörü (*perceptual linear prediction-PLP*) (Hermansky, 1990), adaptif ağırlıklandırılmış bileşen katsayıları (*adaptive component weighting coefficients*) (Assaleh ve Mammone, 1994b,a) ve dalgacık dönüşümü tabanlı birçok öznitelik önerilmiştir. Fakat önerilen yöntemlerin hiç biri MFCC öznitelikleri kadar iyi performans göstermemiş olmakla birlikte bir kısmınında hesaplanabilirliği işlem yükü açısından MFCC kadar kolay olmadığından tercih edilmemiştir. Konuşmacı tanımada en çok kullanılan öznitelikler mel-frekansı kepstrum katsayıları (MFCC) ve doğrusal öngörü kepstrum katsayılarıdır (LPCC) (Kinnunen ve Li, 2010; Bimbot ve ark., 2004). Bu iki öznitelik türü konuşmacı tanımada yaygın olarak kullanılmakta olup, yüksek başarım göstermektedirler. Ayrıca dinamik özniteliklerde MFCC ve LPCC öznitelikleri ile birlikte yaygın bir şekilde kullanılmaktadır. Dinamik öznitelikler, MFCC ve/veya LPCC özniteliklerinin zamanla değişimini temsil etmektedir ve performansı iyileştirdiği yapılan birçok çalışmada belirtilmiştir (Kinnunen ve Li, 2010).

Spektral ve dinamik özniteliklerden farklı olarak, pitch frekansı, formant frekansları ve enerji gibi bürünsel (*prosodic*) özniteliklerde konuşmacı tanıma uygulamalarında kullanılmaktadır (Kinnunen ve Li, 2010). Bürünsel özellikler konuşmacının konuşma tarzı ile ilgili bilgiler içerdiğinden konuşmacı tanımada çoğu zaman başarımı artırmaktadır. Enerji hesabı oldukça kolay olmakla beraber, pitch frekansı ve formant frekanslarının hesaplanması oldukça zor ve dikkat gerektiren işlemlerdir. Çünkü bu öznitelikler çevresel faktörlerden dolayı (gürültü v.b.) işarette meydana gelen değişikliklerden çok kolay etkilenmektedir ve doğru olarak hesaplanması oldukça güçtür. Pitch ve formant frekanslarının hesaplanması ile ilgili önerilen yöntemlerin çoğu laboratuvar ortamında kaydedilen ses işaretleri için oldukça iyi çalışmakla beraber

gürültülü sesler için doğru sonuçlar vermemektedir (Ganchev, 2005).

Daha önce yapılan çalışmalarda, farklı alanlarda kullanılan genel amaçlı öznitelik çıkarma veya dönüştürme yöntemleri konuşmacı tanıma için de kullanılmıştır. Temel bileşen analizi (*principal component analysis-PCA*) (Hanilçi ve Ertaş, 2009, 2011d), doğrusal ayırtaç analizi (*linear discriminant analysis-LDA*) ve bağımsız bileşen analizi (*independent component analysis-ICA*) sıklıkla kullanılan yöntemlerden bazılarıdır (Duda ve ark., 2001; Fukunaga, 1990). Bu yöntemlerin konuşmacı tanımada kullanılmasındaki temel amaç, orijinal özniteliklerin bu yöntemler aracılığı ile ayırt edicilik özelliklerinin fazla ve birbirleri ile olan ilintilerinin daha az olduğu yeni bir uzaya taşınmasıdır.

2.2. Konuşmacı Modelleme ve Sınıflandırma

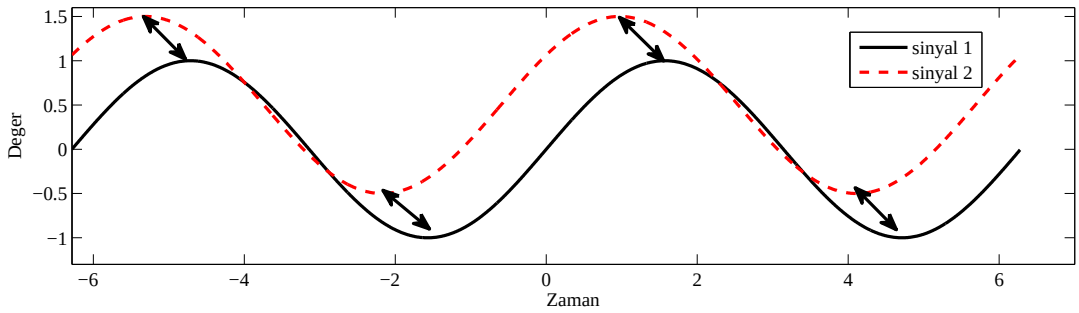
Konuşmacı tanıma bir örüntü tanıma problemidir. Verilen bir ses işaretinin hangi konuşmacıya ait olduğu sınıflandırıcı algoritma tarafından tespit edilir. Bunu yapabilmek için öncelikle yine sınıflandırıcı algoritma tarafından eğitim verileri kullanılarak her bir konuşmacıya ait akustik bir model oluşturulur. Test aşamasında test işaretinden elde edilen öznitelik vektörleri ile konuşmacı modelleri arasındaki benzerlik hesaplanarak karar verilir.

Konuşmacı tanımada değişik sınıflandırma yöntemleri kullanılmaktadır. Genel olarak sınıflandırma yöntemleri, parametrik ve parametrik-olmayan yöntemler olarak iki gruba ayrılır (Duda ve ark., 2001). Parametrik yöntemlerde öznitelik vektörlerinin sabit fakat bilinmeyen bir dağılıma sahip olduğu kabul edilir ve bilinmeyen dağılıma ait parametreler eğitim verileri kullanılarak tahmin edilir. Test aşamasında ise bilinmeyen konuşmacının ses işaretinden elde edilen öznitelik vektörleri ile konuşmacı modeli arasındaki şartlı olasılık hesaplanarak karar verilir. Parametrik olmayan yöntemlerde ise özniteliklerin dağılımı ile ilgili herhangi bir ön kabul yapılmamakla birlikte, test öznitelikleri ile eğitim özniteliklerinin birbirleri ile karşılaştırılmaları sonucunda karar verilir. Parametrik ve parametrik olmayan sınıflandırıcılara ek olarak

ayırt-edici (*discriminative*) sınıflandırıcılarda bulunmaktadır (Mitchell, 1997). Ayırt-edici sınıflandırıcılar ise vektör uzayına dağıtılmış farklı konuşmacılara ait eğitim özniteliklerini birbirinden ayıracak ayırıcı düzlemi bulma prensibine dayanmaktadır. Dinamik zaman eğirme (*dynamic time warping-DTW*), vektör nicemleme (VQ) ve en yakın komşuluk (*K-nearest neighbor-KNN*) yöntemleri konuşmacı tanımada kullanılan parametrik olmayan sınıflandırıcı algoritmalardır. Saklı Markov modelleri (*hidden Markov models-HMM*) ve Gauss karışım modeli (GMM) en popüler parametrik modellerdir. Polinomsal sınıflandırıcılar ve destek vektör makineleri (SVM) ise konuşmacı tanımada kullanılan ayırt-edici sınıflandırıcılardır. Bu bölümde bu sınıflandırıcıların kısa detayları verilecektir.

2.2.1. Dinamik Zaman Eğirme

Dinamik zaman eğirme (DTW), iki işaret arasındaki optimum uyuşmayı bulmaya yarayan bir sınıflandırma yöntemidir (Bonifas ve ark., 1995). İlk olarak 1960'lı yıllarda konuşma tanıma için önerilmiş olup, günümüzde el yazısı tanıma ve online imza eşleştirmesi gibi uygulamalarda kullanılmaktadır (Shanker ve Rajagopalan, 2007). Temel olarak DZE algoritması iki işaretin benzer noktalarının en küçük uzaklık prensibine göre zaman ortamında eşleştirilmesidir (Şekil 2.2.). Konuşmacı tanımada DTW yöntemi metine bağlı uygulamalarda kullanılmıştır (Bonifas ve ark., 1995). Çünkü tanımından da anlaşılacağı üzere benzer işaretler arasındaki örtüşme prensibine dayalı olduğundan farklı içeriklere sahip işaretler arasında yapılamamaktadır.



Şekil 2.2: DZE yöntemi ile iki işaretin karşılaştırılması. Örtüşen noktalar oklarla belirtilmiştir.

2.2.2. Vektör Nicemleme

DTW yöntemi metine bağımlı konuşmacı tanıma uygulamalarında kullanıldığından dolayı amaç metinden bağımsız konuşmacı tanıma gerçekleştirmek ise muhtemel yöntemlerden birisi konuşmacıya ait eğitim işaretlerinden elde edilen öznitelik vektörlerinin tamamını kullanmaktır.

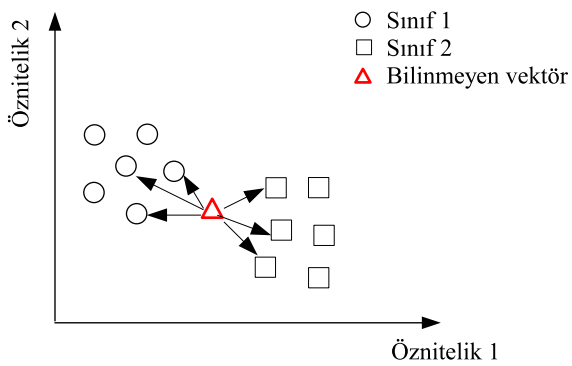
VQ yöntemi başlangıçta bir sıkıştırma algoritması olarak önerilmiş olup daha sonra çeşitli örüntü tanıma problemlerinde kullanılmıştır. 1980'li yıllardan itibaren ise konuşma ve konuşmacı tanıma uygulamalarında kullanılmaya başlanmıştır (Soong ve ark., 1985). VQ yöntemi, eğitim aşamasında eğitim özniteliklerini kullanarak, bu vektörleri temsil eden daha az sayıda vektörden oluşan kişiye özgü kod kitabı oluşturan bir algoritmadır. Her konuşmacı bir kod kitabı ile temsil edilir ve test aşamasında bilinmeyen konuşmacıya ait ses işaretinden elde edilen öznitelikler ile bilinen konuşmacıların kod kitapları arasındaki ortalama uzaklık hesaplanarak, en küçük uzaklığı veren konuşmacı, bilinmeyen konuşmacı olarak tanımlanır (Kinnunen ve ark., 2006). VQ yönteminin matematiksel olarak detayları bir sonraki bölümde verilecektir.

2.2.3. En Yakın Komşuluk

KNN yöntemi, öznitelik uzayında en küçük uzaklık prensibine göre nesneleri sınıflandırmak amacıyla kullanılan ve örüntü tanıma uygulamalarında kullanılan en basit sınıflandırıcılardan biridir.

KNN yönteminde eğitim aşaması eğitilecek konuşmacıya ait öznitelik vektörlerinin saklanmasından ibarettir. Test aşamasında test işaretinden elde edilen öznitelik vektörlerinin eğitim vektörlerine olan uzaklıkları hesaplanır ve k adet en yakın vektöre olan ortalama uzaklıkları elde edilerek karar verilir. Örneğin Şekil 2.3. ile verilen örnekte bilinmeyen test vektörünün 1 en yakın komşuluk durumunda ($k = 1$) Sınıf 1'e, 3 en yakın komşuluk durumunda ise ($k = 3$) Sınıf 2'ye ait olduğuna karar

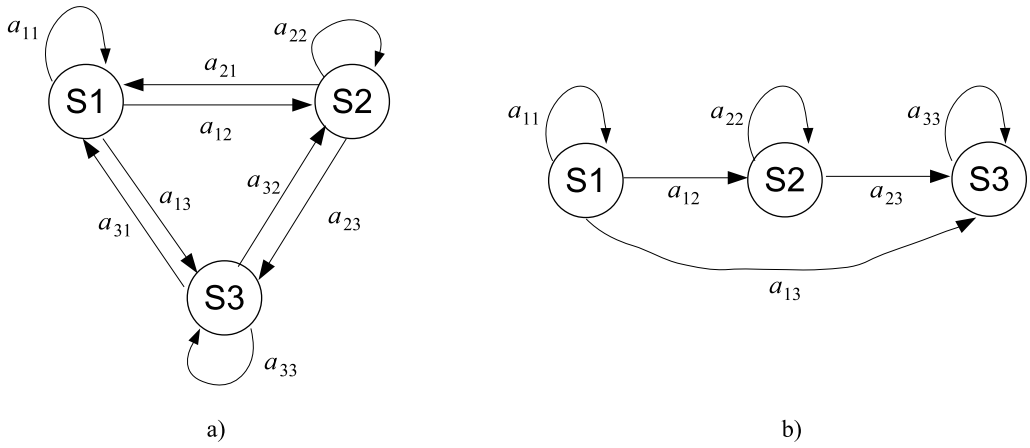
verilecektir.



Şekil 2.3: EYK yöntemi ile sınıflandırma işlemi.

2.2.4. Saklı Markov Modelleri

Saklı Markov modelleri (HMM) dizilerin modellenmesinde kullanılan istatistiksel bir model türüdür (Rabiner, 1989). HMM, her bir gözlem vektörünün (öznitelik vektörü) bir durumun istatistiksel bir fonksiyonu olduğu bir süreçtir. Bu istatistiksel fonksiyon direkt olarak gözlenemez fakat başka bir istatistiksel süreç tarafından gözlenebilmektedir. Bu nedenle Saklı Markov Modelleri adını almaktadır (Rabiner, 1989). HMM sonlu sayıda durumdan oluşan ve her durumun öznitelik vektörüne ait olasılık yoğunluk fonksiyonunu içerdiği bir modeldir. HMM'de durumlar birbirlerine bir durum geçiş işlevi ile bağlıdır (Şekil 2.3.) ve durum geçiş olasılıkları, a_{ij}, bir durumdan diğer bir duruma geçiş olasılıklarını belirtir. HMM genellikle konuşma tanıma uygulamalarında kullanılmakta olup, metine bağımlı konuşmacı tanıma uygulamalarında da yüksek performans göstermektedir (Furui, 1997). 1990'li yıllarda metinden bağımsız konuşmacı tanıma çalışmalarında da yaygın olarak kullanılmıştır (Matsui ve Furui, 1994).



Şekil 2.4: SMM modeli (a) Ergodik SMM, (b) Soldan Sağa SMM.

2.2.5. Gauss Karışım Modeli

Gauss Karışım Modeli (GMM), metinden bağımsız konuşmacı tanıma uygulamalarında kullanılan en başarılı ve önemli yöntemlerden biridir. İlk olarak 1992 yılında (Reynolds, 1992; Reynolds ve Rose, 1995) tarafından önerilmiş olup, bu yöntem konuşmacı tanıma probleminde bir dönüm noktası olarak görülmektedir. Günümüzde konuşmacı tanımada kullanılan sınıflandırıcıların çok büyük bir bölümünün temelini GMM yöntemi oluşturmaktadır. GMM yöntemi, bir öznitelik vektörünün olasılığını M adet çok boyutlu Gauss olasılık yoğunluk fonksiyonunun ağırlıklandırılmış toplamı şeklinde ifade etmektedir.

$$p(\mathbf{x}|\lambda_s) = \sum_{i=1}^{M} w_i b_i(\mathbf{x}). \tag{2.1}$$

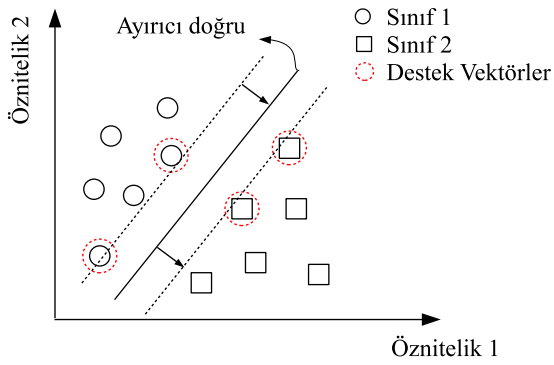
Bu ifadede w_i, $\sum_{i=1}^{M} w_i = 1$, ağırlık katsayılarını ve $b_i(\mathbf{x})$, ortalaması μ_i ve ortak değişinti matrisi Σ_i olan çok boyutlu Gauss olasılık yoğunluk fonksiyonunu göstermektedir. GMM yönteminin eğitim aşamasında amaç, verilen eğitim özniteliklerini kullanarak model parametrelerini (ağırlık katsayıları, ortalama vektörleri ve ortak değişinti matrisleri) tahmin etmektir. Bu nedenle eğitim aşamasında beklentinin maksimumlaştırılması (*Expectation Maximization - EM*) algoritması kullanılır (Reynolds ve Rose, 1995).

GMM yöntemi, tek durumlu sürekli ergodik HMM yöntemine karşılık gelmektedir (Reynolds, 1992). GMM yönteminden sonra önerilen ve günümüzde yaygın olarak kullanılan başarılı bütün istatistiksel yöntemler, Faktör Analizi (FA) ve i-vektör yöntemleri, öznitelik çıkarımı işleminden sonra ilk adım olarak GMM eğitim aşamasını kullanarak işlemlerine devam etmektedir (Kenny ve ark., 2007b).

2.2.6. Destek Vektör Makineleri

Destek vektör makineleri (SVM) günümüzde kullanılan en başarılı ve popüler sınıflandırma yöntemlerinden biridir. SVM ikili bir sınıflandırıcı olup birçok örüntü tanıma probleminde başarılı performans göstermektedir (Burges, 1998). Sınıflandırma işleminin haricinde SVM aynı zamanda başarılı bir bilgi birleştirme (*information fusion*) yöntemi olarak da kullanılmaktadır.

SVM, iki sınıfa ait öznitelik vektörlerini birbirinden en iyi şekilde ayıran doğruyu (*separating hyperplane*) bulmaya çalışan bir sınıflandırıcıdır. Bu doğruyu bulurken, doğrunun her sınıfa ait en yakın öznitelik vektörüne olan uzaklığını da en büyük yapmaya çalışmaktadır (Şekil 2.5). Ancak, çoğu zaman öznitelik vektörleri doğrusal



Şekil 2.5: SVM ile sınıflandırma.

olarak birbirinden ayrılamamaktadır. Bu durumda doğrusal olmayan bir fonksiyona ihtiyaç duyulur. SVM, çekirdek fonksiyonu (*kernel function*) adı verilen fonksiyonlar kullanarak bu işlemi gerçekleştirir. Çekirdek fonksiyonu, verilen öznitelik vektörlerini doğrusal olmayan bir fonksiyon aracılığı ile daha yüksek boyutlu başka bir uzaya ta-

şıma işlemini gerçekleştirir. Taşınan yüksek boyutlu yeni uzayda öznitelik vektörleri doğrusal olarak birbirlerinden ayırılabilir hale getirilmiş olur. Bu durumda ayırıcı doğru şu şekilde tanımlanır:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i t_i K(\mathbf{x}, \mathbf{x}_i) + d \tag{2.2}$$

bu ifadede $t_i \in \{-1, 1\}$ ideal sınıf etiketlerini, N toplam destek vektör sayısını, α_i, $\sum_{i=1}^{N} \alpha_i t_i = 0$ ve $\alpha_i > 0$ olmak üzere destek vektör katsayılarını ve $\mathbf{x}_i$ destek vektörleri belirtmektedir. $K(:,:)$ seçilen çekirdek fonksiyonunu temsil etmektedir. SVM ile sınıflandırma yönteminin eğitim aşamasında amaç, verilen eğitim özniteliklerini kullanarak α_i ve $\mathbf{x}_i$ parametrelerini elde etmektir. Test aşamasında ise bilinmeyen konuşmacıya ait öznitelik vektörleri, $\mathbf{x}$, kullanılarak eşitlik (2.2) ile verilen değer hesaplanır ve sonuç belirli bir eşik seviyesinden yüksek ise konuşmacı kabul edilir. Eşik seviyenin altında ise reddedilir.

1990'lı yıllarda GMM yönteminin ve 2000'li yıllarda SVM yönteminin konuşmacı tanıma probleminde kullanılmaya başlanmasıyla konuşmacı tanıma yeni bir boyut ve ivme kazanmıştır. Bu yıllara kadar az sayıda konuşmacıdan oluşan ve laboratuvar ortamında kaydedilen gürültüsüz seslerin bulunduğu veritabanları konuşmacı tanıma probleminde kullanılırken, GMM ve SVM yöntemleri ile elde edilen başarılı sonuçlardan sonra gürültülü ve telefon hattından geçirilmiş kayıtların bulunduğu veritabanları kullanılmaya başlanmıştır. 2000'li yıllara kadar yüksek sayıda konuşmacıların bulunduğu veritabanları ile en büyük başarım elde etmek temel amaç iken günümüzde konuşmacı tanıma konusunda çalışan araştırmacıların büyük çoğunluğunun ortak amacı yüksek sayıda konuşmacı kümeleri ile mobil telefon kanallarından kaynaklanan işaretteki bozulmalardan ve kayıtlar arasındaki oturum farklılıklarından doğan performans düşüşlerine çözümler aramaktır. Bu aşamada konuşmacı tanıma problemi için önerilen *MAP* uyarlamalı GMM sınıflandırıcısı, *GMM Genel Arkaplan Modeli* (GMM-UBM), konuşmacı tanımada bir dönüm noktası olmuştur (luc Gauvain ve ark., 1994; Reynolds ve ark., 2000). *MAP* uyarlamalı GMM ile 2000'li yıllara kadar elde edilmiş olan konuşmacı tanıma performansları oldukça düşük kal-

mıştır. GMM-UBM yönteminin ortaya çıkmasıyla birlikte önce VQ yönteminin *MAP* uyarlaması önerilmiş ve oldukça yüksek başarımlar elde edilmiş (Hautamäki, 2008; Hautamäki ve ark., 2008; Kinnunen ve ark., 2009) daha sonra da SVM ile GMM yöntemlerinin birleştirildiği *Gauss karışım modeli - süpervektör* (GMM-SV) yöntemi (Campbell ve ark., 2006b) literatüre girmiştir. Bu yöntemlerin detayları bir sonraki bölümde anlatılacaktır.

3. YÖNTEM

3.1. Veritabanı

Bir konuşmacı tanıma sisteminin eğitim aşamasında bilinen konuşmacıları sisteme tanıtmak ve test aşamasında ise konuşmacının kimliği belirtilmeden sistemi test etmek için bir veritabanına ihtiyaç vardır. Kullanılcak veritabanını seçerken; herkesin erişebileceği, yaygın olarak kullanılan ve en önemlisi dünyada konuşmacı tanıma araştırmacıları tarafından kabul edilen bir veritabanı olmasına dikkat edilmelidir. Ancak bu gibi özelliklere sahip veritabanları ile yapılan çalışmalar birbirleriyle kıyaslanabilir. Bu nedenlerden dolayı bu tezdeki çalışmalarda teknik özellikleri Çizelge 3.1 de verilen TIMIT, NTIMIT, NIST 2001, NIST 2002 ve NIST 2005 veritabanları kullanılmıştır.

Çizelge 3.1: Konuşmacı tanıma veritabanları ve teknik özellikleri

	TIMIT	NTIMIT	NIST 2001	NIST 2002	NIST 2005
Dil	İngilizce	İngilizce	İngilizce	İngilizce	İngilizce
Kişi sayısı	630	168	174	330	646
Test Sayısı	-	-	2038 doğru+ 20380 yanlış	2982 doğru+ 36277 yanlış	2775 doğru+ 28643 yanlış
Ses türü	Okuma	Okuma	Konuşma	Konuşma	Konuşma
Kayıt Ortamı	Laboratuvar	Laboratuvar	Telefon	Telefon	Telefon
Örnekleme Hızı	16 kHz	16 kHz	8 kHz	8 kHz	8 kHz
Oturum Farkı	Yok	Yok	Var	Var	Var
Kanal Farkı	Yok	Yok	Var	Var	Var
Eğitim süresi	21 s	21 s	2 dk	2 dk	5 dk
Test süresi	3 s	3 s	5 s - 1 dk	15 s - 45 s	5 dk

TIMIT veritabanı (Reynolds, 1995a,b), Amerikan İngilizcesinin 8 ana lehçesine sahip bölgelerden seçilmiş 438 erkek ve 192 kadın olmak üzere toplam 630 konuşmacıdan oluşmaktadır. Veritabanındaki ses işaretleri laboratuvar ortamında (gürültüsüz) karbon mikrofon kullanılarak kaydedilmiştir. Bu nedenle veritabanındaki sesler gürültü içermemektedir. TIMIT veritabanında her konuşmacıya ait her biri yaklaşık 3 saniye

uzunluğunda 10 adet ses kaydı mevcuttur. TIMIT veritabanı ile yapılan deneylerde her konuşmacının 10 adet cümlesinden 7 tanesi konuşmacıyı eğitmek, kalan 3 tanesi ise test için kullanılmıştır. NTIMIT veritabanı (Reynolds, 1995a,b), TIMIT veritabanındaki ses işaretlerinin kablolu sabit telefon hattı üzerinden geçirilmesi sonucunda elde edilmiş kayıtlardan oluşmaktadır. Telefon hattı kullanıldığından dolayı NTIMIT veritabanı gürültülü işaretlerden oluşan bir veritabanıdır. NTIMIT veritabanı ile yapılan deneylerde, veritabanının test klasöründe bulundan 168 konuşmacı kullanılmıştır.

TIMIT ve NTIMIT veritabanlarının laboratuvar ortamında kayıt edilmesinden ve oturum/kanal farklılıklarının olmamasından dolayı konuşmacı tanıma çalışmalarında 1990'lı yıllardan itibaren fazla tercih edilmemişlerdir. Mobil iletişimin gelişmesi ve kullanımının yaygınlaşması nedeniyle cep telefonları üzerinden günlük konuşmaların kaydedilmesi ile oluşturulan veritabanlarına ihtiyaç duyulmaya başlanmıştır. Bu doğrultuda, Amerikan Ulusal Standartlar ve Teknoloji Enstitüsü (*National Instutute of Standards and Technology - NIST*) NIST tarafından ilki 1997 yılında olmak üzere her yıl konuşmacı tanıma değerlendirme (*Speaker Recognition Evaluation*) organizasyonları düzenlenmeye başlanmıştır(Doddington ve ark., 2000)[1]. Bu organizasyonların en önemli avantajı, organizasyona katılan araştırmacılara kullanacakları veritabanı, NIST tarafından gönderilerek her katılımcının aynı veritabanı üzerinden sistemini geliştirmesine olanak sağlamasıdır. Bu sayede sonuçların karşılaştırılması ve en iyi sistemin seçilmesi adil bir ortamda yapılmış olmaktadır. Bu tezde NIST tarafından 2001, 2002 ve 2005 yıllarında yapılan organizasyonlar için oluşturulan veritabanları kullanılmıştır. NIST 2001, NIST 2002 ve NIST 2005 veritabanlarının tamamı cep telefonu kayıtlarından oluşmakta olup, kayıtlar farklı zamanlarda alındığından dolayı konuşma işaretlerinde iletim hattı (kanal) farklılıkları ve oturum farkları da bulunmaktadır. Bu nedenle, bu veritabanları kullanılarak TIMIT ve NTIMIT veritabanlarına nazaran gerçek zaman uygulamalarına daha uygun sonuçlar elde edilebilmektedir.

[1] `http://www.itl.nist.gov/iad/mig/tests/sre/`

3.2. Performans Değerlendirmesi ve Başarım Ölçütü

Önerilen yeni bir öznitelik çıkarma veya sınıflandırma yönteminin başarılı olup olmadığını belirlemek amacı ile konuşmacı tanıma sistemlerinde standart bir değerlendirme yapılması ve performans ölçütünün kullanılması gerekmektedir. 1980'li yıllar ve öncesinde yapılan konuşmacı tanıma çalışmalarında bir kaç veya en fazla bir kaç düzine konuşmacıdan oluşan veritabanları kullanılmakta idi. Ancak günümüzde, konuşmacı tanıma veritabanları yüzlerce konuşmacı içerdiğinden değerlendirmenin ve performans ölçütünün standartlaştırılması amacıyla yoğun çalışmalar yapılmaktadır.

Bir önceki bölümde bahsedilen NIST veritabanlarında yer alan her konuşmacıya ait ses işaretleri eğitim ve test işaretleri olarak etiketlenmiştir. Bu sayede, bir konuşmacının ses örneklerinden hangilerinin eğitim ve hangilerinin test aşamasında kullanılacağı sabittir. Böylece NIST veritabanlarını kullanan bütün araştırmacılar eğitim ve test aşamalarında aynı örnekleri kullanmaktadır ve dolayısı ile test aşamasındaki toplam deneme (sınama) sayısı aynı olmaktadır. Daha önce de belirtildiği gibi, konuşmacı belirleme ve konuşmacı doğrulama sistemleri arasındaki temel fark test aşamasında ortaya çıkmaktadır. Konuşmacı belirleme sisteminin test aşamasında bir ses örneği sistemin girişine uygulanır ve bu sesin veritabanındaki hangi konuşmacıya ait olduğunun belirlenmesi istenmektedir. Konuşmacı doğrulamada ise, test aşamasında kullanılan ses örneği *doğru sınama (target trial)* ve *yanlış sınama (impostor trial)* olmak üzere iki farklı türde olabilir. Doğru sınama, verilen ses örneğinin gerçekte iddia ettiği kişiye ait olması durumudur. Yanlış sınama ise, verilen ses örneğinin iddia edilen kişiye ait olmayıp sistemin yanıltılmaya çalışılması durumudur.

Eğitim ve test aşamalarının gerçekleştirilme şekli aynı olmasına rağmen, konuşmacı belirleme ve konuşmacı doğrulama sistemlerinde kullanılan performans ölçütleri farklılık göstermektedir. Bunun nedeni, bu iki sistemde ortaya çıkacak muhtemel hataların farklı olmasıdır. Konuşmacı belirlemede ortaya çıkabilecek tek hata yanlış belirleme hatasıdır. Bu durum sistemin, bir ses örneğinin yanlış bir kişi tarafından üretildiğine karar vermesidir. Bu nedenle bir konuşmacı belirleme sisteminde perfor-

mans ölçütü olarak *doğru belirleme oranı* (identification rate) kullanılmaktadır:

$$\text{Belirleme Oranı} = \frac{\text{Doğru olarak belirlenen sınama sayısı}}{\text{Toplam sınama sayısı}} \times 100. \tag{3.1}$$

Konuşmacı doğrulamada ise iki muhtemel hata söz konusudur: *Yanlış Kabul (False Alarm)* ve *Yanlış Ret (Miss Detection)*. Yanlış kabul, yanlış sınama olan bir test işaretinin iddia edilen kişiye ait olduğu şeklinde karar verilmesidir. Yanlış ret ise bir doğru sınama işaretinin iddia edilen kişiye ait olmadığına karar verilmesidir. Yanlış kabul (P_{fa})ve yanlış ret (P_{miss}) oranları şu şekilde hsaplanmaktadır:

$$P_{\text{fa}} = \frac{\text{Kabul edilen yanlış sınama sayısı}}{\text{Toplam yanlış sınama sayısı}} \times 100, \tag{3.2}$$

$$P_{\text{miss}} = \frac{\text{Reddedilen doğru sınama sayısı}}{\text{Toplam doğru sınama sayısı}} \times 100. \tag{3.3}$$

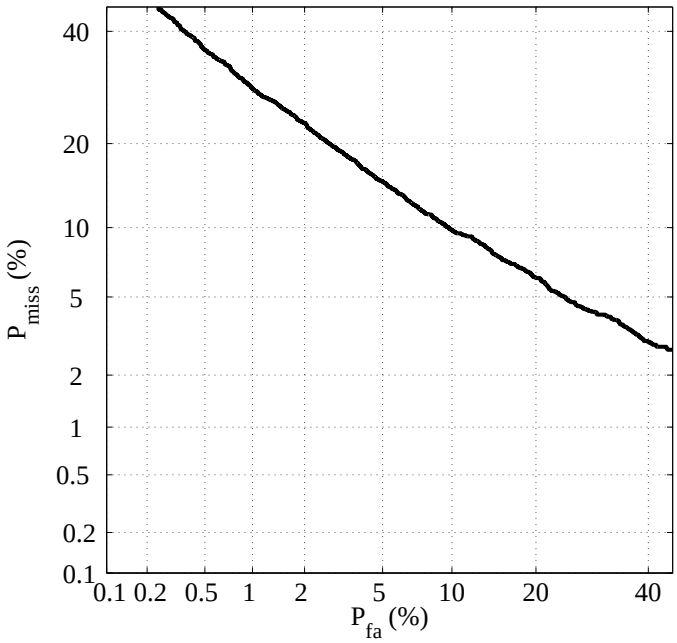
Bu iki tür hata konuşmacı doğrulama sisteminin karar aşamasında kullanılan eşik değere bağlıdır. Bu nedenle uygulamanın türüne göre eşik değer seçilmelidir. Dolayısı ile uygulamalarda bu iki tür hata arasında bir tercih yapılması gerekmektedir. Örneğin, yanlış kabul oranının (P_{fa}) minimum seviyede olması isteniyorsa bunun neticesinde yanlış ret oranında (P_{miss}) artış olacaktır. Akademik çalışmalarda konuşmacı doğrulama performansları sunulurken genel olarak P_{fa} ve P_{miss} oranlarının eşit olduğu eşik değer kullanılır ve bu noktadaki hata oranı *eşit hata oranı (Equal Error Rate - EER)* olarak adlandırılmaktadır (Kinnunen ve Li, 2010; Reynolds, 2002; Bimbot ve ark., 2004). EER kriterinden farklı olarak konuşmacı doğrulama çalışmalarında kullanılan bir diğer performans ölçütü ise *en küçük karar bedel fonksiyonu (Minimum Decision Cost Function - MinDCF)* olup şu şekilde tanımlanmaktadır:

$$MinDCF = \min(0.1 \times P_{\text{miss}} + 0.99 \times P_{\text{fa}}). \tag{3.4}$$

Burada en küçük hesaplama işlemi eşik değer üzerinden yapılmaktadır.

Konuşmacı doğrulama sisteminin performansını grafiksel olarak gösterebilmek ve her iki hata oranının birbirine göre değişimlerini göstermek amacı ile *sezim hata ödün-*

leşimi (Detection Error Trade-off - DET) eğrileri kullanılmaktadır (Martin ve ark., 1997). Şekil 3.1 de örnek bir DET eğrisi gösterilmektedir. Gösterilen eğriden de görüleceği gibi P_{fa} ve P_{miss} değerlerinin eşit olduğu değer, $EER = 10$ dur.

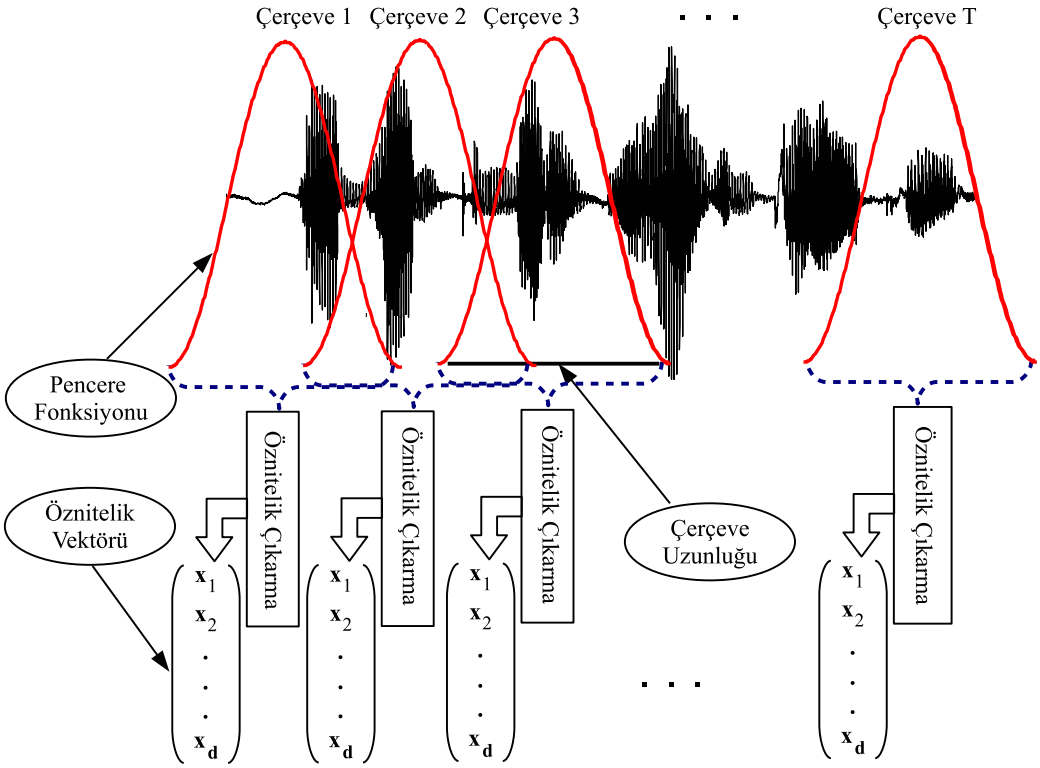


Şekil 3.1: Örnek bir DET eğrisi.

3.3. Öznitelik Çıkarımı

Ses işareti zamanla değişen ve durağan olmayan bir işaret olduğundan dolayı öznitelikler ses işaretinin kısa parçalarından çıkarılmalıdır. İşaret küçük parçalar halinde işlendiği zaman seçilen kısa aralık boyunca ses işaretinin özelliklerinin zamanla değişmediği varsayılmaktadır. Bu sayede işaret, seçilen kısa aralıklar boyunca kararlı özellikler göstermektedir (Deller ve ark., 2000). Ses işaretinin küçük parçalar halinde işlenmesine *kısa dönem analizi* adı verilmektedir. Şekil 3.2 de bir ses işaretinin kısa dönem analizinin adımları gösterilmektedir.

Şekil 3.2 den görüleceği gibi, ses işareti belirli kısımları örtüşen ardışık kısa süreli çerçevelere bölünür. Genellikle konuşmacı tanıma uygulamalarında 20-30 ms uzunluğunda ve 10-15 ms örtüşen çerçeveler kullanılmaktadır. Çerçeveleme işlemi aslında



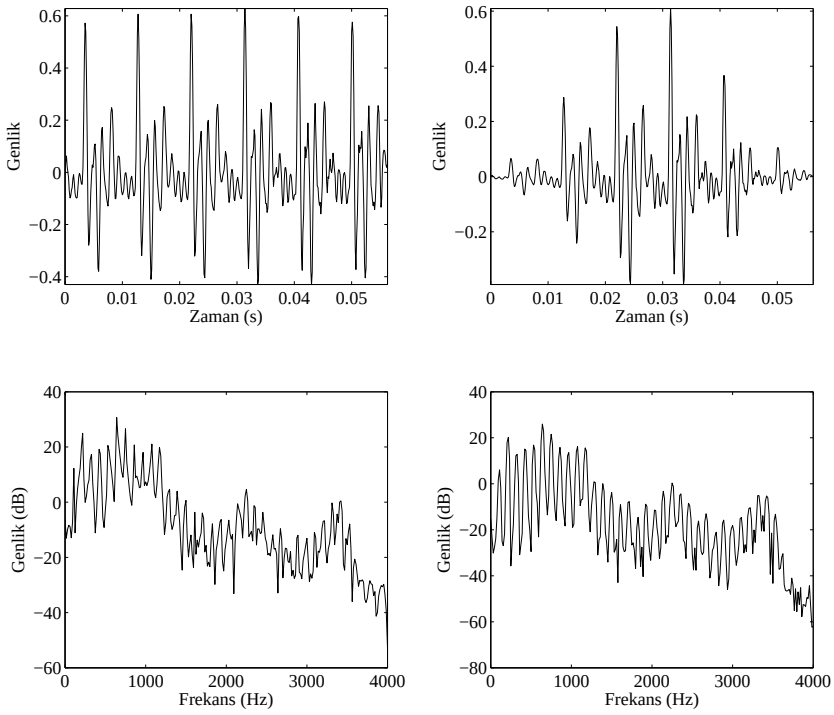
Şekil 3.2: Kısa Dönem Analizi.

ses işaretini çerçeve uzunluğundaki bir dikdörtgen pencere ile çarpmak demektir. Bu nedenle çerçeveleme işlemi ile işarette meydana gelen bozulmayı gidermek amacı ile her bir çerçeve bir pencere fonksiyonu ile çarpılır. Zaman ortamında pencereleme işleminin karşılığı, çerçeve ve pencere vektörlerinin elemanlarının çarpımıdır. Frekans ortamında ise bu işlem çerçevenin spektrumu ile pencere fonksiyonunun spektrumlarının konvolüsyonuna karşılık gelmektedir. İyi bir pencere fonksiyonun transfer fonksiyonunun ana lobunun (*main lobe*) dar, yan lobunun (*side lobe*) ise geniş olması istenmektedir.

En basit pencere fonksiyonu *dikdörtgen pencere* olup şu şekilde tanımlanmaktadır:

$$w[n] = \begin{cases} 1 & 0 \leq n \leq N-1. \\ 0 & \text{diğer} \end{cases} \tag{3.5}$$

Bir ses işaretinin Denklem (3.5) ile verilen dikdörtgen pencere ile pencerelenmesi,



Şekil 3.3: Sesli bir konuşma çerçevesi. Dikdörtgen pencere ile pencerelenmiş çerçeve (sol) ve Hamming pencere ile pencerelenmiş çerçeve (sağ).

aslında işarette hiç bir değişiklik yapmadan sadece pencere sınırları içerisinde kalan bölgenin seçilmesi anlamına gelmektedir. Ses işleme uygulamalarında kullanılan en yaygın pencere fonksiyonu Hamming penceresidir. Hamming pencerenin matematiksel ifadesi şu şekilde verilmektedir:

$$w[n] = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N}) & 0 \leq n \leq N-1 \\ 0 & \text{diğer} \end{cases} \tag{3.6}$$

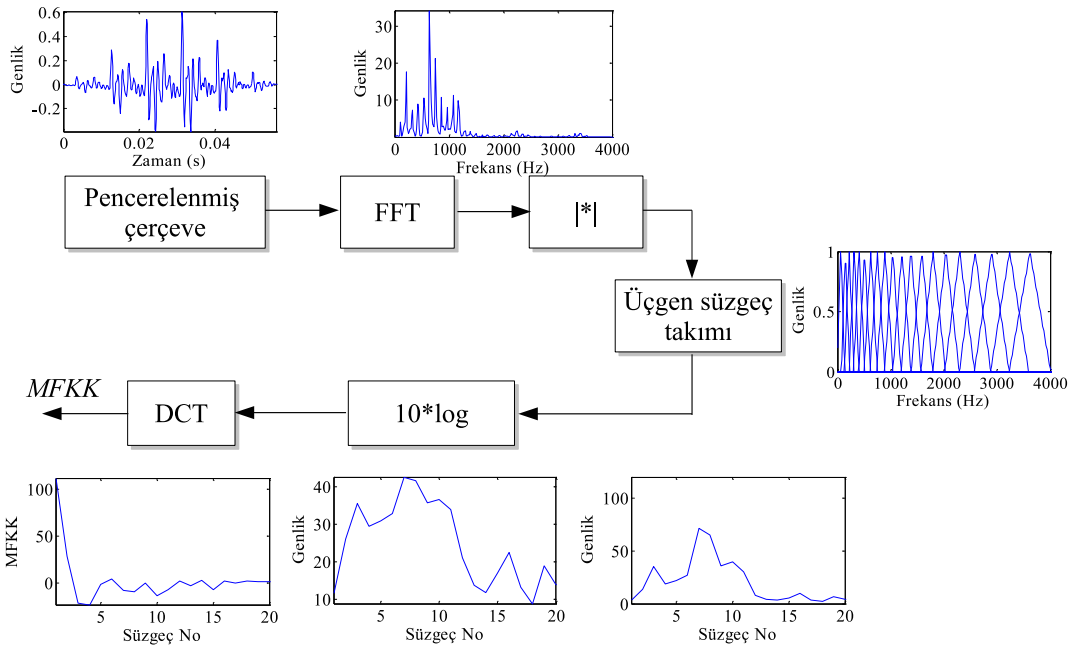
Bir ses işaretinden alınan bir çerçevenin Hamming pencere ile pencerelenmesi işaretin başlangıç ve bitiş noktalarında çerçevelemeden dolayı oluşan süreksizlikleri gidermektedir. Şekil 3.3 de bir çerçevenin dikdörtgen ve Hamming pencereler ile pencerelenmesi ile elde edilen işaretler ve genlik spektrumları gösterilmektedir.

Bütün öznitelik çıkarma yöntemleri için bahsedilen kısa dönem analizi işlemleri (çerçeveleme ve pencereleme) ön-işlem (*pre-processing*) olarak kullanılmaktadır. Bu ön işlemlerden sonra her bir çerçeve için istenen öznitelikler çıkarılır. Bu tezdeki deneysel çalışmalar sırasında konuşmacı tanıma için kullanılan en popüler özniteliklerden

olan MFCC ve LPCC öznitelikleri kullanılmıştır.

3.3.1. Mel-Frekansı Kepstrum Katsayıları

MFCC öznitelikleri ilk olarak 1980'li yılların başında konuşma tanıma için önerilmiş olup, daha sonra konuşmacı tanıma çalışmalarında da kullanılmaya başlanmıştır. Konuşmacı tanımada gösterdiği üstün başarıdan dolayı popüler hale gelmiştir. MFCC özniteliklerinin çıkarılmasında kullanılan adımlar Şekil 3.4 de gösterilmektedir. MFCC öznitelikleri elde edilirken ilk olarak, Hamming pencere ile pencerelenmiş



Şekil 3.4: MFCC özniteliklerinin çıkarılmasında kullanılan işlem adımları.

ses işaretinden elde edilen kısa süreli çerçeve işaretinin *Ayrık Fourier Dönüşümü* (AFD) alınarak işaret zaman tanım bölgesinden frekans tanım bölgesine taşınır. Uygulamalarda AFD işlemi için *Hızlı Fourier Dönüşümü* (*Fast Fourier Transform-FFT*) kullanılmaktadır. $\mathbf{s} = [\, s(0),\ s(1),\ \ldots,\ s(N-1)]^{\mathrm{T}}$ şeklinde ifade edilen bir çerçevenin genlik spektrumu şu şekilde tanımlanmaktadır:

$$S_{\mathrm{FFT}}(k) = \left| \sum_{n=0}^{N-1} s(n) e^{-j2\pi nk/N} \right|^2, 0 \leq k \leq N-1 \tag{3.7}$$

burada k ayrık frekans indisini, $s(n)$ ise pencerelenmiş çerçeveyi belirtmektedir. $s(n)$ sinyalinin $[0, N-1]$ aralığı dışında 0 olduğu varsayılmaktadır. Genlik spektrumu çok fazla detay içermektedir ve genellikle konuşmacı tanıma için bu detaylar bilgi taşımamaktadır (Bimbot ve ark., 2004). Bu nedenle işaretin genlik spektrumunun zarfının hesaplanması için spektrum bir süzgeç takımı ile çarpılır. Süzgeç takımı, ardışık olarak yerleştirilmiş band-geçiren üçgen süzgeçlerden oluşmaktadır. Süzgeç takımı kullanılan süzgeçlerin şekli ve frekanslarının (alt, üst ve merkez frekanslar) hesaplanma türüne göre adlandırılmaktadır. Standart olarak MFCC özniteliklerinin hesaplanmasında mel-ölçeğinde yerleştirilmiş üçgen süzgeçler kullanılmaktadır. Mel-ölçeği insan kulağının frekans ölçeğine benzer bir işitsel ölçek olup süzgeçlerin merkez frekansları şu şekilde hesaplanır:

$$f_{\mathrm{MEL}} = 1000 \frac{\log(1 + f_{\mathrm{Hz}}/1000)}{\log(2)} \tag{3.8}$$

Denklem (3.8) de belirtilen f_{Hz}, Hz biriminden mel birimine dönüştürülmek istenen frekans değerini belirtmektedir. Süzgeç takımındaki bir süzgecin alt kesim frekansı bir önceki süzgecin merkez frekansı, merkez frekansı ise kendisinden önceki süzgecin üst kesim frekansı olacak şekilde süzgeçler yerleştirilmektedir. Süzgeç takımında toplam L adet süzgecin bulunduğunu ve i. süzgecin genlik cevabının $H_i(k), i = 1, \ldots, L; k = 0, \ldots, N-1$ ile ifade edildiği varsayılırsa, i. süzgeç çıkışı, $Y(i)$, şu şekilde elde edilir:

$$Y(i) = \sum_{k=0}^{N-1} S(k) H_i(k) \tag{3.9}$$

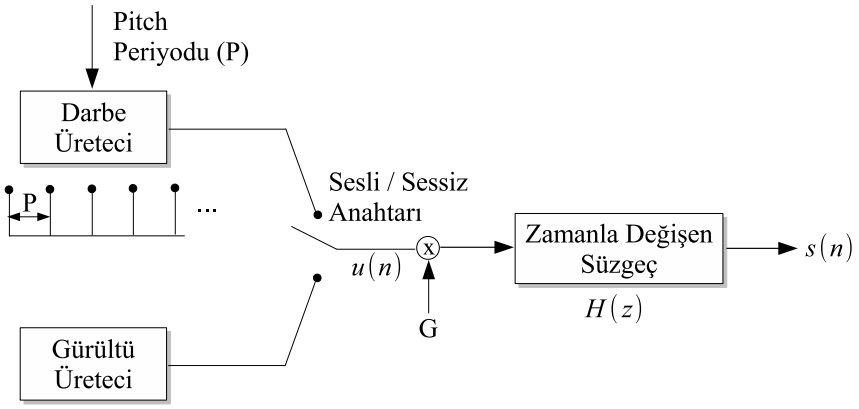
Yani i. süzgeç çıkışı, ses işaretinin spektrumunun süzgecin alt ve üst kesim frekanslarının sınırları arasında kalan bölgesinin süzgecin genlik spektrumu ile ağırlıklandırılmış toplamı şeklinde elde edilmektedir. Son aşamada ise süzgeç çıkışlarının logaritması alınarak dB türünden elde edilen çıkışlar *ayrık kosinüs dönüşümü* (*Discrete Cosine Transform-DCT*) ile MFCC katsayılarına dönüştürülür:

$$x_t = \sum_{i=1}^{L} 10 \log(Y(i)) \cos\left[t\left(i - \frac{1}{2}\right)\frac{\pi}{L}\right], t = 1, \ldots, d \tag{3.10}$$

Denklem 3.10 da belirtilen d hesaplamak istediğimiz MFCC öznitelik sayısını belirtmektedir. Bir ses işaretinden toplam T adet çerçeve elde edildiğinde sonuç olarak bu ses işaretinden her biri d boyutlu olan toplam T adet vektör, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T], \mathbf{x}_i = [x_{i1}, \ldots, x_{id}]^{\mathrm{T}}, i = 1, \ldots, T$ şeklinde MFCC öznitelikleri elde edilmiş olmaktadır.

3.3.2. Doğrusal Öngörü Kepstrum Katsayıları

İnsanın ses üretim mekanizması matematiksel olarak genellikle Şekil 3.5 de gösterildiği gibi modellenmektedir. Bu modelde insanın ses üretim mekanizması zamanla değişen bir süzgeç olarak tasarlanmaktadır. Bu süzgecin girişi periyodik darbe dizisi veya beyaz gürültü olabilir. Eğer sesli bir ses üretilmek isteniyorsa süzgeç girişine periyodik darbe dizisi uygulanır. Ses işaretinin üretiminin bu şekilde modellenmesi literatürde kaynak-süzgeç modeli (*source-filter model*) olarak bilinmekte olup ilk olarak 1960 yılında Fant (1960) tarafından önerilmiştir. Günümüzde hala bu model geçerliliğini korumaktadır. Şekil 3.5 de belirtilen $H(z)$ süzgecinin transfer fonksiyonu



Şekil 3.5: İnsan ses üretim mekanizmasının matematiksel modeli.

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.11}$$

şeklinde tanımlanmaktadır (Rabiner ve Schafer, 2010). Bu süzgecin girişine G kazancı ile çarpılan $u(n)$ işareti uygulanarak çıkışında $s(n)$ ses işareti elde edilir. Denklem 3.11 de görüleceği gibi, $H(z)$ süzgeci tamamen kutuplardan oluşmaktadır. Bu

nedenle literatürde bu süzgece tüm-kutup süzgeci (*all-pole filter*) adı verilmektedir. Bütün bu bilgiler ışığında, insan ses üretim mekanizmasında yer alan parametreler:

- Sesli/Sessiz sınıflandırması
- Pitch periyodu P
- Kazanç faktörü G
- $H(z)$ süzgeç parametreleri $\{a_k,\ k = 1,\ \ldots,\ p\}$

şeklinde sıralanabilir. Sesli/Sessiz ayrımı ve pitch periyodunun tahmini işlemleri uzun yıllardır ses işleme çalışmalarına konu olmuş araştırma alanlarındandır. Şekil 3.5 de gösterilen ses üretim modelinin en önemli avantajı kazanç faktörü G ve süzgeç katsayılarının, $\{a_k\}$, kolay ve başarılı bir şekilde doğrusal öngörü (*linear prediction-LP*) yöntemiyle tahmin edilmesidir.

$s(n)$ işareti, giriş işareti $u(n)$ ve süzgeç katsayıları $\{a_k\}$ ile şu şekilde tanımlanan fark denklemi ile ifade edilmektedir (Rabiner ve Schafer, 2010; Huang ve ark., 2001):

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{3.12}$$

Derecesi p olan bir doğrusal öngörü sistemi, bir işaretin geçmiş p adet örneğini kullanarak n. örneğini tahmin etmeyi amaçlamaktadır. Öngörü katsayıları α_k olan bir doğrusal öngörü sisteminin çıkışı şu şekildedir:

$$\hat{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{3.13}$$

Denklem 3.13 de belirtilen sistem çıkışı $\hat{s}(n)$, tahmin edilen işareti belirtmektedir. Öngörü hatası, $e(n)$, orijinal işaret ile tahmin edilen işaret arasındaki farktır ve

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{3.14}$$

şeklinde hesaplanır. Optimum öngörücü katsayıları, $\{\alpha_k\}$, ortalama karesel hatanın minimum yapılması prensibine göre hesaplanır. Ortalama karesel hata

$$\begin{aligned} E[e^2(n)] &= E\left[\left(s(n) - \sum_{k=1}^{p} \alpha_k s(n-k)\right)^2\right] \\ &= E[s^2(n)] - 2\sum_{k=1}^{p} \alpha_k E[s(n)s(n-k)] \\ &\quad + \sum_{k=1}^{p} \alpha_k \sum_{j=1}^{p} \alpha_j E[s(n-k)s(n-j)] \\ &= r(0) - 2\mathbf{r}^{\mathrm{T}}\mathbf{a} + \mathbf{a}^{\mathrm{T}}\mathbf{R}\mathbf{a} \end{aligned} \tag{3.15}$$

şeklinde hesaplanır (Rabiner ve Schafer, 2010; Makhoul, 1975). Burada, $\mathbf{a} = [\alpha_1, \ldots, \alpha_p]^{\mathrm{T}}$ öngörücü katsayılarını, $\mathbf{s} = [s(n-1) \ldots s(n-p)]^{\mathrm{T}}$ olarak tanımlandığında $\mathbf{R} = E[\mathbf{s}\mathbf{s}^{\mathrm{T}}]$ özilinti matrisini ve $\mathbf{r} = E[s(n)\mathbf{s}]$ özilinti vektörünü temsil etmektedir. Ortalama karesel hatanın α_k katsayılarına göre kısmi türevi alınıp 0'a eşitlenerek optimum öngörücü katsayıları elde edilir.

$$\frac{\partial}{\partial \mathbf{a}} E(e^2(n)) = -2\mathbf{r}^{\mathrm{T}} + 2\mathbf{a}^{\mathrm{T}}\mathbf{R} = 0 \tag{3.16}$$

Denklem (3.16) sonucunda optimum öngörücü katsayıları

$$\mathbf{a}_{\mathrm{opt}}^{\mathbf{LP}} = \mathbf{R}^{-1}\mathbf{r} \tag{3.17}$$

şeklinde hesaplanmaktadır. $\mathbf{R}$, Toeplitz bir matris olup (simetrik ve köşegen üzerindeki bütün elemanlar aynı değere sahip) $\mathbf{a}$, öngörücü katsayılarını içeren vektör ve $\mathbf{r}$ özilinti vektörünü temsil etmektedir. $\mathbf{R}$ ve $\mathbf{r}$, p adet özilinti katsayısı ile şu şekilde tanımlanmaktadır:

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & \cdots & R(p-2) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R(p-1) & R(p-2) & R(p-3) & \cdots & R(0) \end{bmatrix}^{-1} \times \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix} \tag{3.18}$$

Bir $s(n)$ ses işaretinin özilinti katsayıları

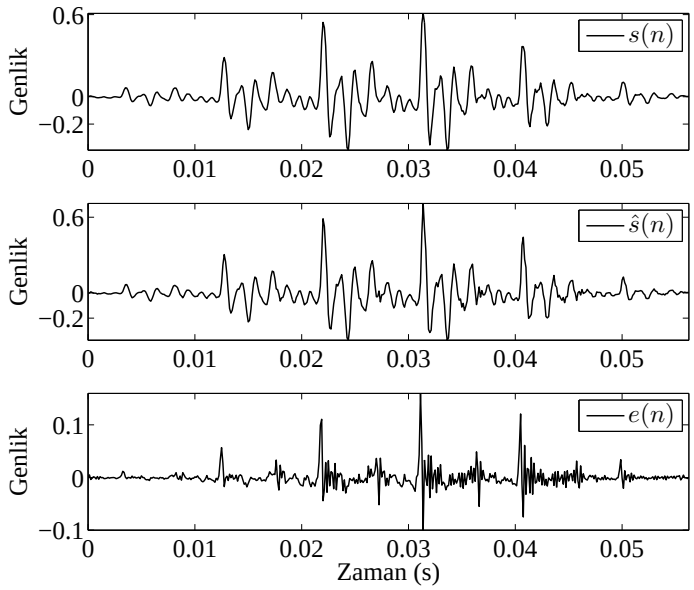
$$R(k) = \sum_{n=0}^{N-1-k} s(n)s(n-k),\ k = 0, \ldots,\ p \tag{3.19}$$

şeklinde hesaplanmaktadır (Rabiner ve Schafer, 2010; Makhoul, 1975). Denklem (3.19) da belirtilen N, ses işaretindeki toplam örnek sayısını belirtmektedir. Denklem 3.17 ve 3.18 de belirtilen $\mathbf{a}$ katsayılarının hesaplanmasında genellikle *Levinson-Durbin* algoritması kullanılmaktadır (Rabiner ve Schafer, 2010; Makhoul, 1975; Deller ve ark., 2000). Levinson-Durbin algoritması, giriş olarak özilinti katsayılarını alarak sonuçta optimum öngörücü katsayılarını hesaplamaktadır. Denklem (3.12) ve (3.14) incelendiğinde hata işaretinin, $Gu(n)$ işaretine eşit olduğu görülmektedir.

$$e(n) = s(n) - \hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) - \sum_{k=1}^{p} \alpha_k s(n-k) = Gu(n) \tag{3.20}$$

Genellikle konuşmacı tanıma uygulamalarında hata işareti, $e(n)$, ses üretim modelindeki zamanla değişen ve insan ses üretim mekanizmasını modelleyen süzgecin girişine uygulanan periyodik darbe dizisi veya gürültü olduğundan dolayı, bu işaretin konuşmacının kimliği hakkında bilgi içermediğine inanılmaktadır. Bu nedenle konuşmacı tanıma uygulamalarında $H(z)$ süzgecinin parametrelerinden türetilmiş öznitelikler kullanılmaktadır. Hanilçi ve Ertaş (2011b) hata işaretinin de konuşmacı kimliği hakkında bilgi içerdiği ve oldukça başarılı performans gösterdiğini ortaya koymuştur. Şekil 3.6 da bir ses işaretinden alınan ve Hamming pencere fonksiyonu ile pencerelenmiş bir çerçeve, $s(n)$, LP analizi ile tahmin edilen işaret, $\hat{s}(n)$, ve hata işareti $e(n)$ gösterilmiştir.

LP katsayıları, konuşma veya konuşmacı tanıma uygulamalarında öznitelik vektörü olarak nadiren kullanılmıştır (Deller ve ark., 2000; Rosenberg ve Sambur, 1975). Ancak yapılan çalışmalarda ardışık öngörü katsayılarının oldukça fazla ilintili oldukları belirtilmiştir (Rosenberg ve Sambur, 1975). Bu nedenle LPC katsayılarından türetilen ve daha az ilintili olan doğrusal öngörü kepstrum katsayıları (*linear predictive cepstral coefficients-LPCC*) yaygın olarak kullanılmaya başlanmıştır. Öngörücü kat-



Şekil 3.6: Bir ses işaretinden alınan bir çerçeve, LPC yöntemi ile tahmin edilmiş işaret ve hata işareti ($p = 12$).

sayıları α_k, $k = 1, \ldots, p$ kullanılarak LPCC öznitelikleri şu şekilde elde edilir (Huang ve ark., 2001):

$$x_n = \begin{cases} \alpha_n + \sum_{k=1}^{n-1} \frac{k}{n} x_k \alpha_{n-k}, & 1 \leq n \leq p \\ \sum_{k=1}^{n-1} \frac{k}{n} x_k \alpha_{n-k}, & n > p \end{cases} \tag{3.21}$$

Denklem (3.21), ilk olarak Atal (1974) tarafından konuşmacı tanıma için yeni bir öznitelik türü olarak önerilmiş olup, günümüzde hala yaygın ve başarılı bir şekilde hem konuşma hem de konuşmacı tanımada kullanılmaktadır. Atal (1974) LPCC özniteliklerini, LP katsayıları (α_k) ve özilinti katsayıları ($R(k)$) ile karşılaştırarak konuşmacı tanıma performanslarını incelemiş ve LPCC özniteliklerinin en iyi performansı gösterdiğini ortaya koymuştur.

3.3.3. Öznitelik Çıkarımında Farklı Spektrum Tahmin Yöntemleri

Literatürde konuşmacı tanıma için yapılan çalışmaların tamamında MFCC öznitelikleri hesaplanırken Şekil 3.4 de belirtilen ve Denklem (3.7) de verilen FFT yöntemi ile işaretin spektrumu hesaplanmaktadır (Quatieri, 2002). Ancak son yıllarda yapılan çalışmalarda spektrum hesaplama yönteminin konuşmacı tanıma performan-

sında önemli etkilerinin olduğu ortaya konulmuştur (Hanilçi ve ark., 2012c,b; Saeidi ve ark., 2010). Bu tezdeki deneyler sırasında değişik spektrum hesaplama yöntemlerinin konuşmacı doğrulama performansına etkileri incelenmiştir. Bu bölümde, kullanılan spektrum hesaplama yöntemleri kısaca anlatılacaktır.

Ses işleme uygulamalarında bir işaretin spektrumunu hesaplamak için FFT yönteminden sonra en çok kullanılan yöntem doğrusal öngörü (*linear prediction-LP*) metodudur (Makhoul, 1975). LP yöntemi genellikle pitch ve formant frekanslarının hesaplanmasında spektrum hesaplama yöntemi olarak kullanılmaktadır (Quatieri, 2002; Deller ve ark., 2000; Rabiner ve Schafer, 2010). Bölüm 3.3.2. de bahsedilen doğrusal öngörü yönteminde Denklem (3.12) gözönüne alındığında ses işaretinin z-dönüşümünün

$$S(z) = \frac{E(z)}{A(z)} \tag{3.22}$$

şeklinde hesaplanabileceği görülmektedir. Burada $E(z)$, hata işareti $e(n)$ nin z-dönüşümü ve $A(z)$, öngörücü katsayıları α_k, $k = 1, \ldots, p$, nin z-dönüşümünü belirtmektedir. Optimum öngörücü katsayılarının hatanın karesel ortalaması minimum olacak şekilde hesaplandığı bölüm 3.3.2. de anlatılmıştı. Zaman tanım bölgesinde hatanın karesel ortalamasının minimum yapılması, frekans tanım bölgesinde hata işaretinin spektrumunun maksimum derecede düzleştirilmesi (*flattening*) anlamına gelmektedir (Atal ve Hanauer, 1971). Bu nedenle $A(z)$, ses işleme literatüründe beyazlaştırıcı süzgeç ya da ters süzgeç (*inverse-filter*) olarak adlandırılmaktadır. Frekans tanım bölgesinde beyaz gürültü düzgün bir spektruma sahiptir ve bu nedenle $s(n)$ ses işareti kullanılarak ortalama karesel hata minimum olacak şekilde α_k, $k = 1, \ldots, p$, LP katsayıları tahmin edildiğinde $s(n)$ işaretinin spektrumu şu şekilde hesaplanabilmektedir:

$$S_{\mathrm{LP}}(f) = \frac{1}{\left|1 - \sum_{k=1}^{p} \alpha_k e^{-j2\pi fk}\right|^2}. \tag{3.23}$$

1993 yılında Ma ve ark. (1993) tarafından LP yöntemine alternatif olarak ağırlıklandırılmış doğrusal öngörü (*weighted linear prediction-WLP*) önerilmiştir. WLP yön-

teminde, LP yönteminden farklı olarak *ağırlıklandırılmış* hata işaretinin karesinin ortalaması minimum yapılacak şekilde öngörücü katsayıları hesaplanmaktadır.

$$E_{\text{wlp}} = \sum_{n} e^2(n)\Psi_n = \sum_{n} \left(s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \right)^2 \Psi_n. \tag{3.24}$$

Denklem (3.24) de belirtilen Ψ_n, zaman ortamında tanımlanmış bir ağırlık fonksiyonudur ve $\Psi_n = \sum_{i=1}^{M} s(n-i)$ şeklinde hesaplanmaktadır. (3.24) eşitliğini minimum yapacak optimum öngörücü katsayıları

$$\mathbf{a}_{\text{wlp}} = \mathbf{R}_{\text{wlp}}^{-1}\mathbf{r}_{\text{wlp}}, \tag{3.25}$$

şeklinde hesaplanır. Burada, $\mathbf{R}_{\text{wlp}}$ ve $\mathbf{r}_{\text{wlp}}$ sırasıyla ağırlıklandırılmış özilinti matrisi ve vektörünü belirtmekte olup, $\mathbf{s} = [s(n-1) \ \ldots s(n-p)]^{\text{T}}$ ses işaretini temsil etmek üzere, şu şekilde hesaplanmaktadır:

$$\mathbf{R}_{\text{wlp}} = \sum_{n} \mathbf{s}\mathbf{s}^{\text{T}}\Psi_n, \tag{3.26}$$

$$\mathbf{r}_{\text{wlp}} = \sum_{n} s(n)\mathbf{s}\Psi_n. \tag{3.27}$$

Optimum WLP katsayıları kullanılarak ses işaretinin spektrumu Denklem (3.23) ile hesaplanmaktadır. WLP yönteminin matematiksel detayları (Ma ve ark., 1993) çalışmasında yer almaktadır. Ses işleme uygulamalarında özellikle ses kodlama ve sentezleme çalışmalarında kullanılan süzgeçlerin durağan (*stable*) olması, süzgecin kutuplarının tamamının birim çember içerisinde yer alması, önemli bir özelliktir. Ayrıca Saeidi ve ark. (2010) süzgecin durağan olması durumunda konuşmacı doğrulama performansının da arttığını göstermiştir. Standart LP yöntemi parametreleri tahmin edilen süzgecin durağan olduğunu garanti etmektedir. Ancak aynı durum WLP yöntemi için geçerli değildir. Bu nedenle Magi ve ark. (2009) tarafından kararlılaştırılmış WLP yöntemi (*Stabilized Weighted Linear Prediction-SWLP*) önerilmiştir. SWLP yönteminde $\mathbf{R}$ matrisi ve $\mathbf{r}$ vektörü değiştirilerek hesaplanmakta olup bunun sonucunda elde edilen süzgecin durağan olduğu Magi ve ark. (2009) tarafından gösterilmiştir.

Ses işaretinin yüksek pitch frekansına sahip (darbe dizisinin periyodunun düşük olması durumu) bir konuşmacı tarafından üretildiği durumlarda, LP, WLP ve SWLP yöntemleri ile hesaplanan ses işaretlerinin spektrumlarında keskin tepelerin oluşmasına neden olmaktadır. Bu durum konuşma kodlama uygulamalarında problemlere sebep olmaktadır ve bu problemin giderilmesi için yüksek pitch frekansına sahip konuşmacılar için alternatif spektrum hesaplama yöntemleri önerilmiştir. Bu yöntemlerden en popüler olanları düzenleştirilimiş doğrusal öngörü (*Regularized Linear Prediction-RLP*) (Ekman ve ark., 2008; Murthi ve Kleijn, 2000) ve minimum varyans bozunumsuz cevap (*Minimum Variance Distortionless Response-MVDR*) (Murthi ve Rao, 2000; Dharanipragada ve ark., 2007) yöntemleridir.

Standart LP yönteminde hatanın karesel ortalaması minimum yapılırken, RLP yönteminde hata işaretine bir kontrol parametresi ve fonksiyonu toplamsal olarak eklenir:

$$E_{\mathrm{RLP}} = \sum_{n} \left(s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \right)^2 + \lambda \phi(\mathbf{a}) \tag{3.28}$$

Denklem (3.28) de belirtilen $\phi(\mathbf{a})$, bilinmeyen öngörücü katsayıları α_k, $k = 1, \ldots, p$, nın bir fonksiyonu olan penaltı ölçütü (*penalty measure*) ve $\lambda > 0$ ise düzenleştirme sabiti olup, hesaplanan spektrum zarfının yumuşaklığını kontrol etmektedir. Denklem (3.28) den görüleceği gibi $\lambda \to 0$ durumunda RLP yöntemi standart doğrusal öngörü, LP, yöntemine karşılık gelmektedir. Ekman ve ark. (2008) ile Murthi ve Kleijn (2000) çalışmalarında penaltı ölçütünü şu şekilde seçmişlerdir:

$$\phi(\mathbf{a}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{A'(e^{j\omega})}{W(\omega)} \right|^2 d\omega \tag{3.29}$$

Burada $1/|W(\omega)|^2$, yaklaşık olarak seçilmiş spektral zarfı ve $A'(e^{j\omega})$, RLP ters süzgeci, $A(e^{j\omega}) = \sum_{k=0}^{p} c_k e^{-j\omega k}$, nin frekansa göre türevini temsil etmektedir. Penaltı fonksiyonunun bu şekilde seçilmesinin nedenleri, seçilen bu fonksiyonun kapalı form ve iteratif olmayan bir çözümünün mevcut olmasıyla beraber hesaplama açısından kolay olmasıdır. Ekman ve ark. (2008) ve Murthi ve Kleijn (2000) çalışmalarında

spektral zarf, $1/|W(\omega)|^2$, pencerelenmiş özilinti fonksiyonu ile türetilmiş ve matris formunda şu şekilde tanımlanmıştır:

$$\phi(\mathbf{a}) = \mathbf{a}^{\mathrm{T}}\mathbf{DFDa} \tag{3.30}$$

Denklem (3.30) da verilen $\mathbf{a} = [a_1, \ldots, a_p]^{\mathrm{T}}$ öngörücü katsayılarını, $\mathbf{D}$ köşegen elemanları ilgili satır veya sütun numarası olan köşegen bir matrisi ve $\mathbf{F}$ ise, $r(m) = \sum_{n=0}^{N-1} s(n)s(n-m)$, $m = 0, \ldots, p-1$ özilinti fonksiyonunu ve $v(m)$ bir pencere fonksiyonunu belirtmek üzere, pencerelenmiş özilinti fonksiyonu, $f(m) = r(m)v(m)$, ile elde edilen Toeplitz bir matrisi belirtmektedir. $\mathbf{F}$ matrisi denklem (3.29) de paydada belirtilen $W(\omega)$ terimini temsil etmektedir ve $v(m)$ pencere fonksiyonu dikdörtgen pencere olarak seçildiğinde $\mathbf{F}$ matrisi standart LP yönteminde kullanılan ve denklem (3.17) de verilen $\mathbf{R}$ matrisine eşit olmaktadır. RLP yöntemi ile optimum öngörücü katsayıları ise

$$\mathbf{a}_{\mathrm{opt}}^{\mathrm{rlp}} = -(\mathbf{R}_{\mathrm{lp}} + \lambda\mathbf{DFD})^{-1}\mathbf{r}_{\mathrm{lp}} \tag{3.31}$$

şeklinde hesaplanır. Hesaplanan öngörücü katsayıları ile işaretin spektrumu denklem (3.23) ile elde edilir.

MVDR yöntemi (Murthi ve Rao, 2000; Dharanipragada ve ark., 2007), aynı zamanda Capon yöntemi olarak da bilinmekte olup, yüksek pitch frekansına sahip konuşmacılar tarafından üretilen ses işaretlerinde spektrum hesaplama yöntemi olarak önerilmiştir. Bu yöntem aynı zamanda sesli ve sessiz bölgelerin spektrumlarının hesaplanmasında başarılı bir yöntem olarak bilinmektedir. Dharanipragada ve ark. (2007) bu yöntemin konuşma tanımada başarılı sonuçlar verdiğini göstermiştir. MVDR yöntemi ile spektrum hesaplamada LP katsayıları kullanılmaktadır. Derecesi m olan bir MVDR spektrumu şu şekilde hesaplanmaktadır:

$$S_{\mathrm{MVDR}}(f) = \frac{1}{|\sum_{k=-m}^{m} \mu(k)e^{-j2\pi fk}|^2} \tag{3.32}$$

Denklem (3.32) de verilen $\mu(k)$ katsayıları, LP katsayıları, α_k, $k = 1, \ldots, p$ kullanı-

larak şu şekilde elde edilir:

$$\mu(k) = \begin{cases} \sum_{i=0}^{m-k}(m+1-k-2i)a_i a_{i+k}, & k = 0, 1, ..., m \\ \mu(-k), & k = -m,\ -m+1,\ \ldots,\ -1. \end{cases} \tag{3.33}$$

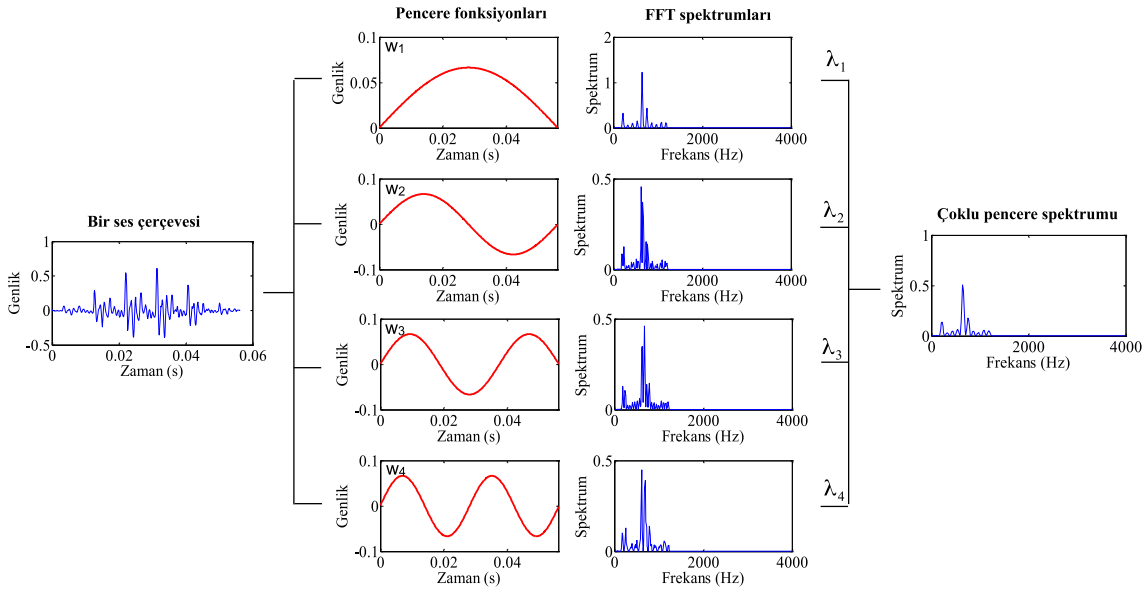
Doğrusal öngörü analizine dayanan yöntemler dışında ses işleme uygulamalarında önerilen ve başarılı sonuçlar veren diğer spektrum hesaplama yöntemleri de mevcuttur. Bu yöntemlerden en dikkat çekici olanlar çoklu pencere (*multitaper*) (Sandberg ve ark., 2010; Kinnunen ve ark., 2010, 2012; Alam ve ark., 2013) ve iteratif kepstral yumuşatma (*iterative cepstral smoothing-ICS*) (Robel ve ark., 2007) yöntemleridir.

Çoklu pencere yöntemi, işaretin birden fazla pencere fonksiyonu kullanılarak hesaplanan FFT spektrumlarının ağırlıklandırılmış toplamıdır. Bir $s(n), n = 0, \ldots, N-1$ ses işaretinin güç spektrumu çoklu pencere yöntemi ile

$$S_{\mathrm{MT}}(f) = \sum_{k=1}^{K} \lambda_k \left| \sum_{n=0}^{N-1} w_k(n) s(n) e^{-j2\pi n f/N} \right|^2, \tag{3.34}$$

şeklinde hesaplanmaktadır. Burada, K kullanılan pencere sayısını, $w_k(n)$, $k = 1, \ldots, K$ pencere fonksiyonlarını ve λ_k her bir pencere fonksiyonunun ağırlık katsayısını belirtmektedir. $w_k(n)$ pencere fonksiyonları uygulamanın türüne göre değişiklik gösterebilmektedir ancak yapılan çalışmalarda konuşmacı tanıma uygulamalarında *Thomson, Multipeak* ve *SWCE* fonksiyonlarının en iyi performansı gösterdiği ortaya koyulmuştur. Bu pencere fonksiyonlarına ilişkin detaylar Sandberg ve ark. (2010),Kinnunen ve ark. (2010), Kinnunen ve ark. (2012) ve Alam ve ark. (2013) tarafından yapılan çalışmalarda verilmiştir. Şekil 3.7 de $K = 4$ adet *SWCE* pencere fonksiyonu kullanılarak çoklu pencere yöntemi ile ses işaretinin spektrumunun elde edilmesi gösterilmektedir.

ICS yöntemi ise işaretin FFT spektrumunun iteratif olarak yumuşatılması prensibine dayanan ve ses kodlamada başarılı bir şekilde kullanılan bir yöntemdir (Robel ve ark., 2007). Bu yöntemin başlangıç adımında ses işaretinden alınan ve incelenen çerçevenin FFT spektrumu, $S(f)$, (3.7) eşitliğinde belirtildiği şekilde hesaplanır. i.



Şekil 3.7: Çoklu pencere yöntemi ile spektrum hesaplama.

iterasyonda spektrumun zarfı, $A_i(f)$, orijinal spektrumun ve o andaki zarfın, $C_{i-1}(f)$, maksimumu olacak şekilde güncellenir:

$$A_i(f) = \max(\log|S(f)|, C_{i-1}(f)). \tag{3.35}$$

Burada $C_i(f)$, i. adımda kepstral olarak yumuşatılmış spektrum olup $C_i(f) = \mathrm{DCT}(\log|A_{i-1}(f)|)$ şeklinde hesaplanmaktadır. Başlangıç aşamasında $C_0(f)$, $A_0(f) = \log|S(f)|$ spektrumu kullanılarak hesaplanmaktadır.

Tüm bu yöntemlerin toplamsal gürültü durumunda konuşmacı doğrulama sistemi için karşılaştırmalı analizleri Hanilçi ve ark. (2012c) ve Hanilçi ve ark. (2012b) çalışmalarında yapılmıştır ve elde edilen bulgular Bölüm 4.5. de detaylı bir şekilde verilecektir.

3.3.4. Öznitelik Vektörlerinin Türevleri

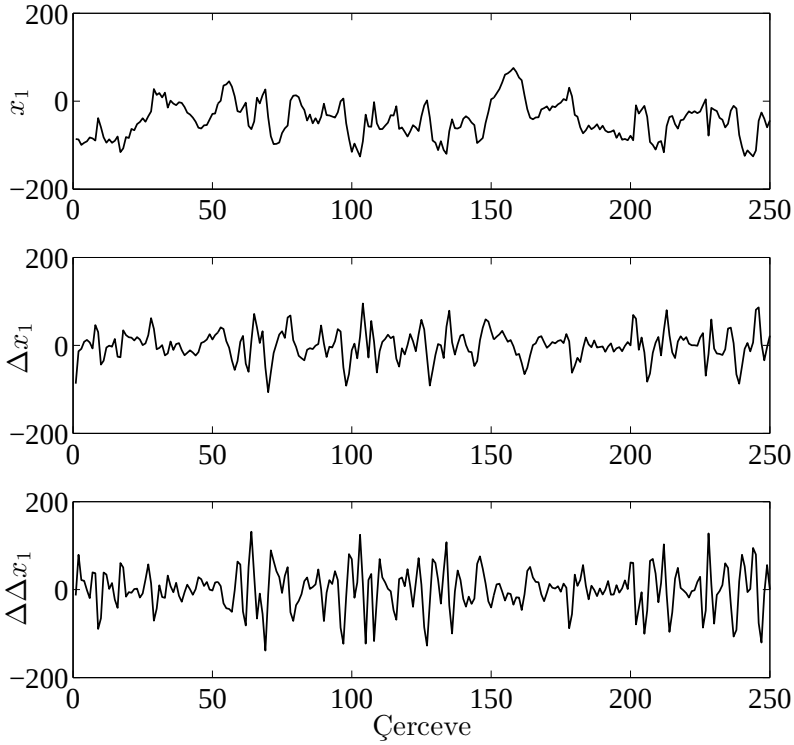
Konuşmacı tanıma ile ilgili yapılan bir çok çalışmada öznitelik vektörlerinin türevleri de orijinal özniteliklere eklenerek kullanılmaktadır (Huang ve ark., 2001). Öznitelik

vektörlerinin türevleri ses işaretindeki zamanla değişimlerden kaynaklanan dinamik bilgileri içermektedir. Özellikle bir konuşmacıya ait ses işaretlerinin farklı zamanlarda kayıt altına alındığı uygulamalarda bu bilgiler konuşmacı tanıma performansını önemli ölçüde düşürmektedir (Kinnunen ve Li, 2010). Yapılan bazı çalışmalarda öznitelik vektörlerinin türevlerinin orijinal öznitelik vektörlerine eklendiğinde boyut arttığından dolayı daha fazla eğitim verisinin gerekli olduğu belirtilmiştir (Kinnunen ve ark., 2003). Örneğin bir konuşma çerçevesinden elde edilen 12 adet MFCC katsayılarına bu katsayıların birinci türevleri eklendiğinde yeni öznitelik vektörünün boyutu $12 + 12 = 24$ olmaktadır. Soong ve Rosenberg (1988) hem orijinal öznitelik vektörlerini kullanarak hem de öznitelik vektörlerinin türevlerini kullanarak elde ettikleri skorları birleştirerek konuşmacı tanıma performansının arttırılabileceğini göstermiştir.

Literatürde öznitelik vektörlerinin birinci ve ikinci türevleri *delta* (Δ) ve *delta-delta* ($\Delta\Delta$) parametreleri olarak adlandırılmaktadır (Bimbot ve ark., 2004). Yapılan bir çok çalışmada Δ ve $\Delta\Delta$ parametreleri değişik birçok öznitelik vektörü türü ile kullanılmıştır (Furui, 1981; Soong ve Rosenberg, 1988). Hangi tür öznitelik vektörü elde etme yönteminin kullanıldığından bağımsız olarak işaretin kısa dönemli çerçevelerinin spektrumundan elde edilen öznitelikler için birinci ve ikinci türev parametreleri spektral değişimlerin zamanla değişimini temsil etmektedir (Kinnunen, 2003). $x_k(i)$ k. çerçeveden elde edilen i. özniteliği temsil etmek üzere $x_k(i)$ özniteliğinin birinci türevi

$$\Delta x_k(i) = x_{k+M}(i) - x_{k-M}(i) \tag{3.36}$$

şeklinde hesaplanmaktadır. Burada M, komşu çerçeveleri temsil etmektedir ve genellikle $1 - 2$ olarak seçilmektedir. Türev alma işlemi her öznitelik için ayrı ayrı hesaplanarak bütün öznitelik vektörünün türevi elde edilir. $\Delta\Delta$ parametreleri ise Δ parametreleri kullanılarak (3.36) eşitliği ile hesaplanır. Şekil 3.8 de bir ses işaretinden elde edilen 1. öznitelik katsayıları, x_1, ile bu katsayıların birinci ve ikinci türevleri gösterilmektedir.



Şekil 3.8: x_1 öznitelik katsayıları ile Δx_1 ve $\Delta\Delta x_1$ katsayıları.

3.3.5. Öznitelik Normalizasyonu

Konuşma ve konuşmacı tanıma uygulamalarında karşılaşılan en büyük problemlerden bir tanesi ses işaretine ortamdan veya iletim kanalından kaynaklanan gürültünün etki etmesidir. Ayrıca farklı zamanlarda kaydedilmiş ses işaretleri de performansı olumsuz yönde etkilemektedir. Bu gibi etkileri öznitelik bazında azaltmak amacıyla öznitelik vektörlerine değişik işlemler uygulanmaktadır. Bu işlemlerin amacı, ortam veya kanal gürültüsünün elde edilen öznitelik vektörleri üzerindeki olumsuz etkisini minimize etmektir (Alam ve ark., 2011; Viikki ve Laurila, 1998). Bu tür iyileştirme işlemlerinin bir çoğunda yapılan ön kabul, yapılan kayıt boyunca kanal veya ortam etkisinin sabit olmasıdır. Konuşmacı tanımada yaygın ve başarılı bir şekilde kullanılan yöntemlerin en önemlisi kepstral ortalama ve varyans normalizasyonudur (*Cepstral mean and variance normalization-CMVN*). CMVN yönteminin amacı, öznitelik vektörlerinin doğrusal olarak aynı segmental istatistiklere sahip bir düzleme taşınmasıdır. $x_t(i)$ t. çerçeveden elde edilmiş i. öznitelik katsayısını belirtmek üzere

CMVN işlemi şu şekilde gerçekleştirilmektedir:

$$\hat{x}_t(i) = \frac{x_t(i) - \mu(i)}{\sigma_t(i)} \tag{3.37}$$

Denklem (3.37) de $\mu(i)$, $x_t(i)$, $t = 1, \ldots, T$ $i.$ öznitelik katsayısının bütün çerçeveler boyunca ortalaması ve $\sigma_t(i)$ ise standart sapmasını temsil etmektedir ve şu şekilde hesaplanmaktadırlar:

$$\mu(i) = \frac{1}{T} \sum_{t=1}^{T} x_t(i) \tag{3.38}$$

$$\sigma_t(i) = \left(\frac{1}{T} \sum_{t=1}^{T} (x_t(i) - \mu(i))^2 \right)^{1/2} \tag{3.39}$$

3.4. Sınıflandırma Yöntemleri

Şekil 2.1 de gösterildiği üzere, bir konuşmacı tanıma sisteminde ilk olarak ses işaretinden kullanılacak öznitelik vektörleri elde edilmektedir. Bölüm 3.3. de bu tez çalışmasında kullanılan öznitelik vektörlerinin elde edilme yöntemleri anlatılmıştır. Sonraki adım olan eğitim aşamasında ise öznitelik vektörleri kullanılarak konuşmacıyı temsil eden bir model oluşturulur. Test aşamasında ise bilinmeyen konuşmacıya ait ses işaretinden elde edilen öznitelik vektörleri konuşmacı modeliyle karşılaştırılarak karar verme işlemi gerçekleştirilir. Bu bölümde eğitim ve test aşamalarının gerçekleştirildiği sınıflandırma yöntemleri detayları ile anlatılacaktır.

3.4.1. Vektör Nicemleme

2. bölümde kısaca anlatıldığı gibi vektör nicemleme (VQ) algoritması aslında bir veri sıkıştırma yöntemi olarak literatürde yer almaktadır (Linde ve ark., 1980; Kanungo ve ark., 2002). Konuşmacı tanımada ilk defa Soong ve ark. (1985) tarafından kullanılmış olup, günümüze kadar yaygın olarak kullanılagelmiştir. VQ yönteminin en önemli avantajı, basit bir yöntem olmasına karşın konuşmacı tanımada yüksek performans göstermesidir.

VQ yöntemi ile konuşmacı tanıma sisteminde bir konuşmacı modeli, konuşmacıya ait toplam T adet konuşma çerçevesinden elde edilen ve her biri D boyutlu olan öznitelik vektörlerinin, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, K adet birbiri ile örtüşmeyen gruba ayrılması ile oluşturulur. Her bir grup, ilgili gruba ait öznitelik vektörlerinin ortalaması olan bir kod vektörü (*code vector*), $\mathbf{c}_i$, $i = 1, \ldots, K$ ile temsil edilir. K adet kod vektörden oluşan küme, $\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$, kod kitabı (*codebook*) olarak ifade edilir ve konuşmacının modelini temsil eder. Kod vektör sayısı, K, model boyutu olarak adlandırılmakta olup, toplam öznitelik vektörü sayısından çok küçüktür, $K \ll T$. Kod vektörlerinin dağılımı öznitelik vektörleri ile aynı dağılıma sahiptir. Bu nedenle, VQ yöntemi toplam veri miktarını düşürmekle beraber orijinal öznitelik vektörlerinin içermiş olduğu bilgileri korumaktadır (Kinnunen, 2003).

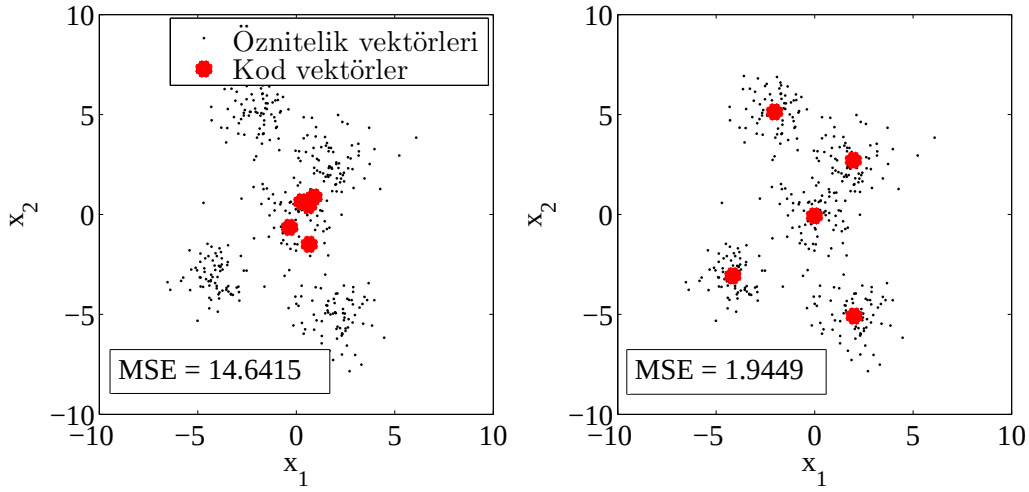
VQ yöntemi ile verilen $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ öznitelik vektör kümesini kullanarak kod kitabı, $\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$, oluştururken ulaşılmak istenen hedef, amaç fonksiyonunu minimum yapacak $\mathbf{C}$ kod kitabının elde edilmesidir. Amaç fonksiyonu olarak *ortalama karesel hata* (*Mean Square Error-MSE*) kullanılmaktadır:

$$MSE(\mathbf{X}, \mathbf{C}) = \frac{1}{T} \sum_{t=1}^{T} \min_{1 \leq k \leq K} \|\mathbf{x}_t - \mathbf{c}_k\|^2 \tag{3.40}$$

Burada $\|\mathbf{x}_t - \mathbf{c}_k\|^2$, D boyutlu iki vektör arasındaki karesel Öklit mesafesini belirtmekte olup şu şekilde hesaplanmaktadır:

$$\|\mathbf{x}_t - \mathbf{c}_k\|^2 = \sum_{i=1}^{D} (x_i - c_i)^2 \tag{3.41}$$

VQ yönteminin uygulanması sırasında belirlenmesi gereken iki temel unsur bulunmaktadır: (1) kullanılacak olan kod kitabı oluşturma yöntemi ve (2) model boyutu (K). VQ yöntemi kullanılan konuşmacı tanıma uygulamalarında genel olarak model boyutu arttıkça konuşmacı tanıma performansının da arttığı gözlenmiştir (Hanilçi ve Ertaş, 2011a, 2009; Kinnunen ve ark., 2006; Kinnunen, 2003). Kod kitabı oluşturma yöntemi olarak bilinen en popüler iki yöntem ise LGB algoritması (Linde ve ark., 1980) ve K-ortalama (*K-means*) (Kanungo ve ark., 2002) algoritmasıdır. İki yön-



Şekil 3.9: K-ortalama algoritması ile kod vektörlerin elde edilmesi.

temi birbirinden ayıran temel fark başlangıç aşamalarıdır. LBG algoritması, bütün öznitelik vektörlerinin ortalaması olan vektörü başlangıç kod vektörü olarak belirler ve her adımda kod vektörleri ikiye bölerek arzu edilen sayıda kod vektör elde edilinceye kadar işlemi sürdürür. K-ortalama algoritması ise başlangıçta K adet kod vektörü rasgele belirleyerek belirli sayıda iterasyon ile öznitelikler ile kod vektörler arasındaki MSE değerine göre kod vektörleri günceller. LBG ve K-ortalama algoritmalarının işlem akışları sırasıyla Algoritma 1 ve Algoritma 2'de verilmiştir. Algoritma 1'den görüleceği gibi LBG yöntemi tek bir kod vektör ile işleme başlamakta ve kod vektör sayısını her adımda 2 katına çıkarmaktadır. Bu nedenle LBG yöntemi ile oluşturulmak istenen kod vektör sayısı 2'nin kuvvetlerinden seçilmelidir. K-ortalama algoritmasında böyle bir sınırlama bulunmamaktadır. Ancak Algoritma 2'den de görüldüğü gibi başlangıç kod vektörleri rasgele seçildiğinden K-ortalama yöntemi başlangıç koşullarına yüksek derecede bağımlıdır. Başlangıç kod vektörleri iyi bir şekilde seçildiğinde yöntem hızlı bir şekilde çalışacak, aksi taktirde yöntemin sonuç üretme süresi artacaktır. Şekil 3.9 da K-ortalama algoritması kullanılarak $D = 2$ boyutlu öznitelik vektörleri ile kod vektörlerin elde edilmesi ve başlangıçta ve son aşamada hesaplanan MSE değerleri gösterilmektedir. Şekilden de görüleceği gibi başlangıçta (sol şekil) kod vektörler rasgele seçildiğinden yüksek MSE değeri hesaplanmış olup, $I = 20$ iterasyon sonunda ise (sağ şekil) kod vektörler güncellendiğinde düşük MSE değeri elde edilmiştir.

Algoritma 1 LBG algoritması

Giriş: $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ ve K ▷ Öznitelik vektörleri ve Kod vektör sayısı
Çıkış: $\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$ ▷ Kod kitabı
Prosedür $\mathbf{C}$ =LBG$(\mathbf{X}, K)$
    $\mathbf{c} = \frac{1}{T}\sum_t \mathbf{x}_t$ ▷ Başlangıç kod vektörü
    $k = 1$
    repeat
        $\mathbf{c}_k = (1+\epsilon)\mathbf{c}$
        $\mathbf{c}_{k+1} = (1-\epsilon)\mathbf{c}$
        $k = 2 \times k$
        $D = 0, D' = MSE(\mathbf{X}, \mathbf{C})$
        while $D - D' \neq 0$ **do** ▷ kod vektör güncellemesi durana dek
            $D = D'$
            $q_t = \text{argmin}_k \left\|\mathbf{x}_t - \mathbf{c}_k\right\|^2, t = 1, \ldots, T$
            $S_k = \{\mathbf{x}_t \in \mathbf{X} | q_t = k\}$ ▷ Öznitelik vektörlerini grupla
            $\mathbf{c}_k = \frac{1}{|S_k|}\sum_{\mathbf{x}_t \in S_k} \mathbf{x}_t$ ▷ Kod vektörlerini güncelle
            $D' = MSE(\mathbf{X}, \mathbf{C})$
        end while
    until $k == K$
end Prosedür

Algoritma 2 K-ortalama algoritması

Giriş: $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ ve K ▷ Öznitelik vektörleri ve Kod vektör sayısı
Çıkış: $\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$ ▷ Kod kitabı
Prosedür $\mathbf{C}$ =KMEANS$(\mathbf{X}, K)$
    Başlangıç kod vektörlerini rasgele belirle $\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$
    $D_0 = MSE(\mathbf{X}, \mathbf{C})$
    for i=1:I **do**
        for $k = 1 : K$ **do** ▷ Her bir kod vektörü için
            $q_t = \text{argmin}_{1 \leq k \leq K} \left\|\mathbf{x}_t - \mathbf{c}_k\right\|^2, t = 1, \ldots, T$
            $S_k = \{\mathbf{x}_t \in \mathbf{X} | q_t = k\}$ ▷ Öznitelik vektörlerini grupla
            $\hat{\mathbf{c}}_k = \frac{1}{|S_k|}\sum_{\mathbf{x}_t \in S_k} \mathbf{x}_t$ ▷ Yeni kod vektörü hesapla
            $\mathbf{c}_k = \hat{\mathbf{c}}_k$ ▷ Kod vektörünü güncelle
        end for
        $D_i = MSE(\mathbf{X}, \mathbf{C})$
        if $D_i - D_{i-1} == 0$ **then**
            Return $\mathbf{C}$ ▷ Eğer güncelleme işlemi bittiyse sonlandır
        end if
    end for
end Prosedür

VQ yöntemi kullanılan bir konuşmacı tanıma probleminin eğitim aşamasında LBG veya K-ortalama algoritmalarından herhangi biri ile her konuşmacı için eğitim öznitelik vektörleri kullanılarak kod kitapları oluşturulur ($\mathbf{C}_1, \ldots, \mathbf{C}_N$, N adet konuşmacı). Test aşamasında ise bilinmeyen konuşmacıya ait ses işaretinden elde edilen öznitelik vektörleri, $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_J\}$, ile her konuşmacı modeli (kod kitabı) arasındaki benzerlik hesaplanır ve en yüksek benzerliği veren kod kitabı hangi konuşmacıya ait ise bilinmeyen ses örneğinin o konuşmacıdan geldiğine karar verilir. VQ yönteminde genellikle benzerlik ölçütü olarak Denklem (3.40) ile verilen MSE kriteri kullanılır ve karar verme işlemi minimum MSE kriterine göre yapılır:

$$\delta = \arg\max_{1 \leq i \leq N} MSE(\mathbf{Y}, \mathbf{C}_i). \tag{3.42}$$

Daha öncede bahsedildiği gibi MSE kriterinde karesel Öklit metriği kullanılmaktadır. Ancak D-boyutlu iki vektör arasındaki uzaklığı hesaplamada genel olarak Minkowski metriği olarak da bilinen l_p-norm kullanılmaktadır:

$$l_p(\mathbf{x}, \mathbf{c}) = \left(\sum_{i=1}^{D} |x_i - c_i|^p \right)^{1/p} = \|\mathbf{x} - \mathbf{c}\|_p \tag{3.43}$$

En yaygın olarak kullanılan Minkowski metrikleri l_1, l_2 ve l_∞ durumlarıdır. Denklem (3.43) de görüldüğü gibi l_2 metriği Öklit mesafesine karşılık gelmektedir. VQ yöntemi ile konuşmacı tanıma sisteminde karar aşamasında bu üç metriğin kullanılmasının tanıma performansına etkileri (Hanilçi ve Ertaş, 2011a) tarafından detaylı bir şekilde incelenmiştir.

3.4.2. Gauss Karışım Modeli

GMM yöntemi, Denklem (2.1) de belirtildiği gibi K adet çok boyutlu Gauss yoğunluk fonksiyonunun ağırlıklandırılmış toplamı şeklinde ifade edilmektedir. Bir $\mathbf{x}$ öznitelik vektörünün olasılık yoğunluk fonksiyonu K adet çok boyutlu Gauss yoğun-

luk fonksiyonunun ağırlıklandırılmış toplamı olarak şu şekilde tanımlanmaktadır:

$$p(\mathbf{x}|\lambda) = \sum_{k=1}^{K} w_k b_k(\mathbf{x}) \tag{3.44}$$

Denklem (3.44) de $\mathbf{x}$, D-boyutlu öznitelik vektörü, $b_k(\mathbf{x})$, $i = 1, \ldots, K$ Gauss bileşenlerini ve $w_i, i = 1, \ldots, K$ ise karışım ağırlıklarını temsil etmektedir. Her bir karışım bileşeni, ortalaması μ_i ve ortak değişinti matrisi $\mathbf{\Sigma}_k$ olan D-boyutlu Gauss dağılımı olarak şu şekilde tanımlanmaktadır:

$$b_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x} - \mu_k)^{\mathrm{T}} \mathbf{\Sigma}_k^{-1}(\mathbf{x} - \mu_k)\}. \tag{3.45}$$

Karışım ağırlıkları $\sum_{k=1}^{K} = 1$ koşulunu sağlamaktadır. Bu koşul karışım modelinin doğru bir olasılık yoğunluk fonksiyonu olmasını sağlamaktadır. Böylece, bir karışım modeli karışım ağırlıkları w_k, ortalama vektörleri μ_k ve ortak değişinti matrisleri $\mathbf{\Sigma}_k$ ile temsil edilmekte ve şu şekilde tanımlanmaktadır:

$$\lambda = \{w_k, \mu_k, \mathbf{\Sigma}_k\}, k = 1, \ \ldots, \ K \tag{3.46}$$

GMM yönteminde her bir konuşmacı, bir λ modeli ile temsil edilmektedir ve eğitim aşamasında eğitilecek konuşmacının ses işaretinden elde edilen öznitelik vektörleri, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \ \mathbf{x}_T\}$, kullanılarak model parametreleri hesaplanmaktadır. GMM yöntemi ile model parametreleri hesaplanırken amaç fonksiyonunu maksimum yapacak parametrelerin hesaplanması hedeflenmektedir. Amaç fonksiyonu olarak logaritmik olabilirlik fonksiyonu (*log-likelihood function*) kullanılmaktadır:

$$\log p(\mathbf{X}|\lambda) = \frac{1}{T} \sum_{t=1}^{T} \log \sum_{k=1}^{K} w_k b_k(\mathbf{x_t}|\mu_k, \mathbf{\Sigma}_k). \tag{3.47}$$

Model parametreleri tahmin edilirken iteratif beklentinin maksimumlaştırılması (*expectation maximization-EM*) algoritması kullanılmaktadır. EM algoritması 4 adımdan oluşmaktadır:

1. Başlangıç model parametreleri, $\lambda^0 = \{w_k^0,\ \mu_k^0, \Sigma_k^0\}_{k=1}^K$, rasgele belirlenerek Denklem (3.47) ile verilen logaritmik olabilirlik değeri hesaplanır.

2. **E-adımı:** m. adımdaki $\lambda^{\mathrm{m}} = \{w_k^{\mathrm{m}},\ \mu_k^{\mathrm{m}}, \Sigma_k^{\mathrm{m}}\}_{k=1}^K$ parametrelerini kullanarak sonsal (*a posteriori*) olasılıkları hesaplanır:

$$\gamma(k,t) = \frac{w_k b_k(\mathbf{x}_t)}{\sum_{j=1}^K w_j b_j(\mathbf{x}_t)},\ k = 1,\ \ldots,\ K, \text{ve } t = 1,\ \ldots,\ T. \tag{3.48}$$

3. **M-adımı:** $N_k = \sum_{t=1}^T \gamma(k,t)$ olmak üzere, model parametreleri güncellenir:

$$\mu_k = \frac{1}{N_k} \sum_{t=1}^T \gamma(k,t)\mathbf{x}_t, \tag{3.49}$$

$$\mathbf{\Sigma}_k = \frac{1}{N_k} \sum_{t=1}^T \gamma(k,t)(\mathbf{x}_t - \mu_k)(\mathbf{x}_t - \mu_k)^{\mathrm{T}}, \tag{3.50}$$

$$w_k = \frac{N_k}{T} \tag{3.51}$$

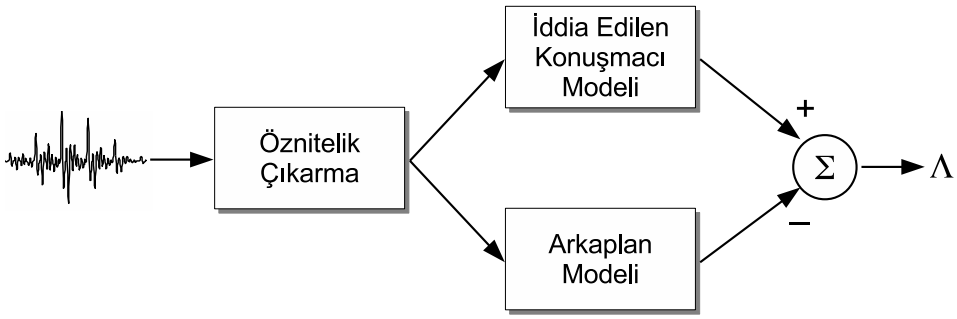
   $\lambda^{\mathrm{m+1}} = \{w_k^{\mathrm{m+1}}, \mu_k^{\mathrm{m+1}}, \mathbf{\Sigma}_k^{\mathrm{m+1}}\} \leftarrow \{w_k, \mu_k, \mathbf{\Sigma}_k\}$ şeklinde parametreler değiştirilir.

4. **Tekrar:** Yeni parametreler ile logaritmik olabilirlik değeri hesaplanır ve 2. ve 3. adımları olabilirlik fonksiyonundaki artış durana kadar tekrarlanır.

3.5. Konuşmacı Doğrulama ve Olabilirlik Oran Testi

Daha önceki bölümlerde belirtildiği gibi konuşmacı doğrulama, verilen bir ses işaretinin iddia edilen kişiye ait olup olmadığına karar verilmesi işlemidir. Verilen bir ses işareti x, ve iddia edilen konuşmacı S için, konuşmacı doğrulama iki hipotezden oluşan basit bir hipotez testi şeklinde tanımlanabilir:

- H_0: x ses işareti, S konuşmacısına aittir.
- H_1: x ses işareti, S konuşmacısına ait değildir.



Şekil 3.10: Olabilirlik oran testi ile konuşmacı doğrulama sistemi.

Bu iki hipotez arasından bir karar vermek amacıyla uygulanabilecek en iyi yöntem olabilirlik oran testidir (*likelihood ratio test*):

$$\frac{p(x|H_0)}{p(x|H_1)} \begin{cases} \geq \Theta & H_0 \text{ hipotezi doğru} \\ < \Theta & H_1 \text{ hipotezi doğru} \end{cases} \tag{3.52}$$

Burada, $p(x|H_i)$, $i = 0, 1$, H_i hipotezinin ses işareti x için hesaplanan olasılık yoğunluk fonksiyonunu temsil etmektedir. Bu aynı zamanda verilen ses işareti için H_i hipotezinin olabilirliği olarak da bilinmektedir. Θ ise karar verme işleminde kullanılan eşik değerdir. Bir konuşmacı doğrulama sisteminin en temel amacı, ses işaretini kullanarak $p(x|H_0)$ ve $p(x|H_1)$ olabilirlik değerlerini hesaplamaktır.

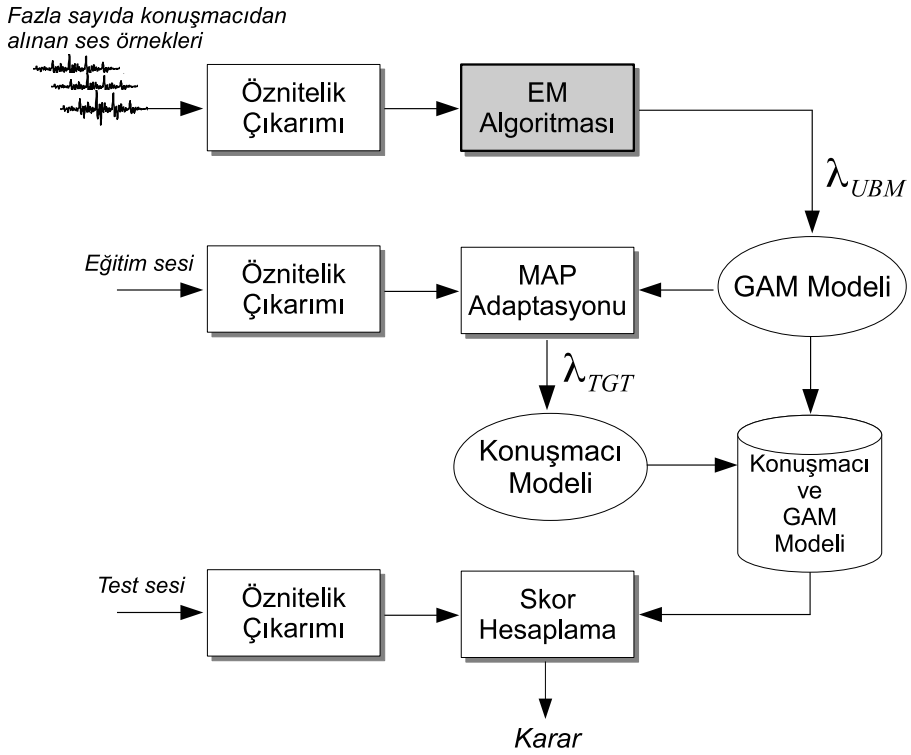
Şekil 3.10 da olabilirlik oran testi ile konuşmacı doğrulama sisteminin temel bileşenleri gösterilmektedir. Şekilden de görüleceği üzere ilk önce ses işaretinden öznitelik vektörleri elde edilmektedir. Öznitelik çıkarma, Bölüm 3.3.'de anlatıldığı gibi ses işaretinin kullanılarak konuşmacıyı temsil eden parametrelere dönüştürülmesi işlemidir. Bu adım neticesinde ses işaretini ve konuşmacıyı temsil eden ve her biri D-boyutlu olan öznitelik vektör kümesi, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, elde edilir. Elde edilen öznitelik vektörleri daha sonra H_0 ve H_1 hipotezlerinin olabilirlik değerlerini, $p(x|H_0)$ ve $p(x|H_1)$, hesaplamada kullanılır. Matematiksel olarak H_0 hipotezi, λ_{hyp} şeklinde ifade edilen bir model ile temsil edilmekte olup, bu model öznitelik uzayında S konuşmacısını karakterize etmektedir. Gauss karışım modeli, H_0 hipotezi için öznitelik vektörlerini temsil edecek en uygun yöntemdir ve bu nedenle λ_{hyp} ortalama vektörleri, kovaryans matrisleri ve karışım ağırlıklarından oluşan bir GMM modelini temsil etmektedir,

$\lambda_{\text{hyp}} = \{w_k, \mu_k, \Sigma_k\}_{k=1}^{K}$. Alternatif hipotez H_1 ise $\lambda_{\overline{\text{hyp}}}$ modeli ile temsil edilmektedir. Son aşamada ise olabilirlik oranı $p(\mathbf{X}|\lambda_{\text{hyp}})/p(\mathbf{X}|\lambda_{\overline{\text{hyp}}})$ değeri hesaplanır. Genellikle logaritmik olabilirlik oranı kullanılmakta olup şu şekilde hesaplanmaktadır:

$$\Lambda(\mathbf{X}) = \log p(\mathbf{X}|\lambda_{\text{hyp}}) - \log p(\mathbf{X}|\lambda_{\overline{\text{hyp}}}) \tag{3.53}$$

H_0 hipotezine ait λ_{hyp} modeli, iyi tanımlanmış olup, S konuşmacısına ait eğitim ses işaretleri kullanılarak bu modele ait parametreler tahmin edilebilir. Ancak $\lambda_{\overline{\text{hyp}}}$ modelinin S konuşmacısı dışındaki bütün muhtemel alternatif konuşmacıları temsil etmesi gerektiğinden dikkatle hesaplanmalıdır. Alternatif $\lambda_{\overline{\text{hyp}}}$ modelini hesaplamada iki değişik çözüm mevcuttur. Bunlardan ilki, çok fazla sayıda başka konuşmacı modelleri kullanmaktır. Çok fazla sayıdaki konuşmacı modelleri literatürde olabilirlik oranı kümeleri (Higgins ve ark., 1991), işbirlikçiler (*cohorts*) (Rosenberg ve ark., 1992) veya arkaplan konuşmacıları (Reynolds, 1995b) olarak adlandırılmaktadır. İkinci ve en yaygın yöntem ise S konuşmacısından farklı olmak koşulu ile, çok fazla sayıda konuşmacıdan alınan ses işaretleri ve bu işaretlerden elde edilen öznitelik vektörlerinin kullanılarak alternatif hipotez, $\lambda_{\overline{\text{hyp}}}$ modelini eğitmektir. Bu yöntem genel arka plan modeli (*universal background model-UBM*) olarak bilinmektedir (Reynolds, 1997). UBM yönteminde, farklı konuşmacılardan alınan çok sayıda ses örneğinden elde edilen öznitelik vektörleri kullanılarak alternatif hipotez H_1 i temsil edecek tek bir genel arkaplan modeli, λ_{UBM}, eğitilir. Daha sonra S konuşmacısına ait eğitim ses işaretlerinden elde edilen öznitelik vektörleri kullanılarak H_0 hipotezine ait konuşmacı modeli, λ_{TGT}, bu öznitelik vektörleri ile UBM modeli, λ_{UBM}, nin en büyük sonsal olasılık (*maximum a posteriori-MAP*) adaptasyonu ile elde edilir. GMM-UBM yöntemi ile konuşmacı modeline ait parametreler, λ_{TGT}, tahmin edilirken, bir önceki bölümde anlatılan EM yöntemi ile parametrelerin tahmin edilme yönteminin aksine, model parametreleri çok geniş bir öznitelik kümesi kullanılarak iyi bir şekilde tahmin edilmiş bir genel modelden (arkaplan modeli-UBM) uyarlanarak elde edilir.

Şekil 3.11 de UBM yöntemi ile konuşmacı doğrulama sisteminin temel adımları gösterilmektedir. Şekilden de görüleceği gibi öncelikli olarak, çok sayıda kişiden alınan



Şekil 3.11: UBM yöntemi ile konuşmacı doğrulama sistemi.

ses örnekleri (arkaplan işaretleri) kullanılarak Bölüm 3.4.2.'de anlatıldığı gibi enbüyük olabilirlik (*Maximum Likelihood-ML*) kriterine göre EM algoritması ile K adet Gauss bileşeninden oluşan genel arka plan modeli, λ_{UBM}, elde edilir. Daha sonra eğitim aşamasında konuşmacı modeli, λ_{TGT}, eğitim sesinden çıkarılan öznitelik vektörleri kullanılarak λ_{UBM} modelinden türetilir.

Verilen bir genel arka plan modeli, λ_{UBM}, ve eğitim öznitelikleri $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ için GMM-UBM yöntemi ile konuşmacı modeli parametreleri elde edilirken ilk olarak öznitelik vektörlerinin UBM modelindeki her bir Gauss bileşenine istatistiksel olarak aidiyetleri (Şekil 3.12a) hesaplanır:

$$Pr(i|\mathbf{x}_t) = \frac{w_i b_i(\mathbf{x}_t)}{\sum_{k=1}^{K} w_k b_k(\mathbf{x}_t)} \tag{3.54}$$

Burada $b_i(\mathbf{x}_t)$, UBM modelindeki Denklem (3.45) ile belirtilen D-boyutlu i. Gauss bileşenini temsil etmektedir. Daha sonra, $Pr(i|\mathbf{x}_t)$ ve $\mathbf{x}_t$ kullanılarak yeterli istatistikler (*sufficient statistics*) hesaplanır. Bunlar, her bir Gauss bileşenine atanan

vektör sayısı, birinci ve ikinci moment değerleridir ve şu şekilde hesaplanır:

$$n_i = \sum_{t=1}^{T} Pr(i|\mathbf{x}_t) \tag{3.55}$$

$$E_i(\mathbf{x}) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|\mathbf{x}_t)\mathbf{x}_t \tag{3.56}$$

$$E_i(\mathbf{x}^2) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|\mathbf{x}_t)\mathbf{x}_t^2 \tag{3.57}$$

Son olarak bu istatistikler kullanılarak konuşmacı modelindeki (λ_{TGT}) i. Gauss bileşeninin uyarlanmış parametreleri hesaplanır (Şekil 3.12b):

$$\hat{w}_i = [\alpha_i^w n_i/T + (1 - \alpha_i^w)w_i]\,\gamma \tag{3.58}$$

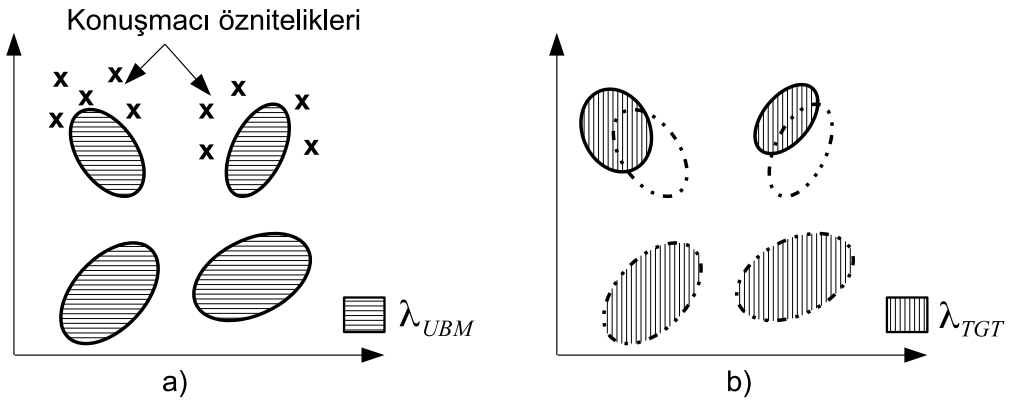
$$\hat{\mu}_i = \alpha_i^m E_i(\mathbf{x}) + (1 - \alpha_i^m)\mu_i \tag{3.59}$$

$$\hat{\sigma}_i^2 = \alpha_i^v E_i(\mathbf{x}^2) + (1 - \alpha_i^v)(\sigma^2 + \mu_i^2) - \hat{\mu}_i^2 \tag{3.60}$$

Burada, w_i, μ_i ve σ_i^2 sırasıyla λ_{UBM} modelindeki i. Gauss bileşenine ait ağırlık, ortalama ve ortak değişinti parametrelerini temsil etmekte olup, $\hat{w}_i$, $\hat{\mu}_i$ ve $\hat{\sigma}_i^2$ ise MAP uyarlaması ile hesaplanmış konuşmacı modeli, λ_{TGT}, nin i. Gauss bileşeni parametrelerini belirtmektedir. α_i^w, α_i^m ve α_i^v sırasıyla ağırlık, ortalama ve ortak değişinti parametrelerine ait uyarlama parametreleridir.

$$\alpha_i^\rho = \frac{n_i}{n_i + r}, \tag{3.61}$$

Burada, r sabit bir ilgi parametresi (*relevance factor*) olup UBM modelindeki kaç adet vektör ile kaç adet eğitim öznitelik vektörünün uyarlanacağının dengelenmesini sağlamaktadır. Şekil 3.12 de örnek olarak bir konuşmacı modelinin uyarlanması işlemi temsili olarak gösterilmektedir. Konuşmacı öznitelik vektörleri UBM modelindeki 4 adet Gauss bileşeninden sadece iki tanesine atandığından dolayı (Şekil 3.12a), uyarlanmış konuşmacı modelinde sadece ilgili iki Gauss bileşeninin parametreleri değişmektedir. Geriye kalan iki bileşen ise UBM modeli ile aynı parametrelere sahip



Şekil 3.12: UBM yöntemi ile konuşmacı modelinin uyarlanması.

olacak şekilde herhangi bir değişikliğe uğramamaktadır.

GMM-UBM yöntemi ile konuşmacı doğrulama sisteminin test aşamasında, bilinmeyen konuşmacıya ait öznitelik vektörleri, $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$, arkaplan modeli, λ_{UBM}, ve iddia edilen konuşmacı modeli, λ_{TGT}, kullanılarak logaritmik olabilirlik skoru şu şekilde hesaplanarak, karar verme işlemi Denklem (3.52) ile belirtildiği gibi gerçekleştirilir. Logaritmik olabilirlik skoru şu şekilde hesaplanmaktadır:

$$\Lambda(\mathbf{Y}) = \log p(\mathbf{Y}|\lambda_{\text{TGT}}) - \log p(\mathbf{Y}|\lambda_{\text{UBM}}), \tag{3.62}$$

Burada $\log p(\mathbf{Y}|\lambda)$ Denklem (3.47) de belirtildiği şekilde hesaplanır.

GMM yönteminin MAP uyarlaması (GMM-UBM) ile konuşmacı doğrulama performansındaki büyük artıştan hareketle, Hautamäki ve ark. (2008) tarafından VQ yönteminin MAP uyarlaması (VQ-UBM) önerilmiş olup, konuşmacı tanımada klasik VQ yöntemine göre oldukça yüksek başarım gösterdiği ortaya koyulmuştur (Kinnunen ve ark., 2009; Hanilçi ve Ertaş, 2011a). VQ-UBM yönteminde ilk olarak, GMM-UBM yönteminde olduğu gibi, çok fazla sayıda konuşmacıdan alınan ses işaretleri kullanılarak klasik VQ metodu (LBG veya K-ortalama) ile genel arkaplan modelini temsil edecek K adet kod vektörden oluşan bir kod kitabı, $\mathbf{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$, oluşturulur. Daha sonra konuşmacı kod kitabı $\mathbf{C}_{\text{TGT}} = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$, eğitim öznitelikleri, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, kullanılarak UBM modeli, $\mathbf{U}$, nin MAP uyarlaması ile elde edilir. VQ yönteminin MAP uyarlamasında ilk olarak başlangıç konuşmacı modeli

$\mathbf{C}_{\text{TGT}} = \mathbf{U}$ olarak seçilir. Daha sonra her bir kod vektöre en yakın eğitim vektörleri gruplandırılarak kod vektörleri güncellenir:

$$q_n = \arg \min_{1 \leq k \leq K} \|\mathbf{x}_t - \mathbf{c}_k\|^2 \tag{3.63}$$

$$S_k = \{\mathbf{x}_t \in \mathbf{X} | q_n = k\} \tag{3.64}$$

$$\mu_k = \frac{1}{\|S_k\|} \sum_{\mathbf{x}_t \in S_k} \mathbf{x}_t \tag{3.65}$$

$$\hat{\mathbf{c}}_k = w_k \mu_k + (1 - w_k)\mathbf{u}_k \tag{3.66}$$

Burada,

$$w_k = \frac{\|S_k\|}{\|S_k\| + r} \tag{3.67}$$

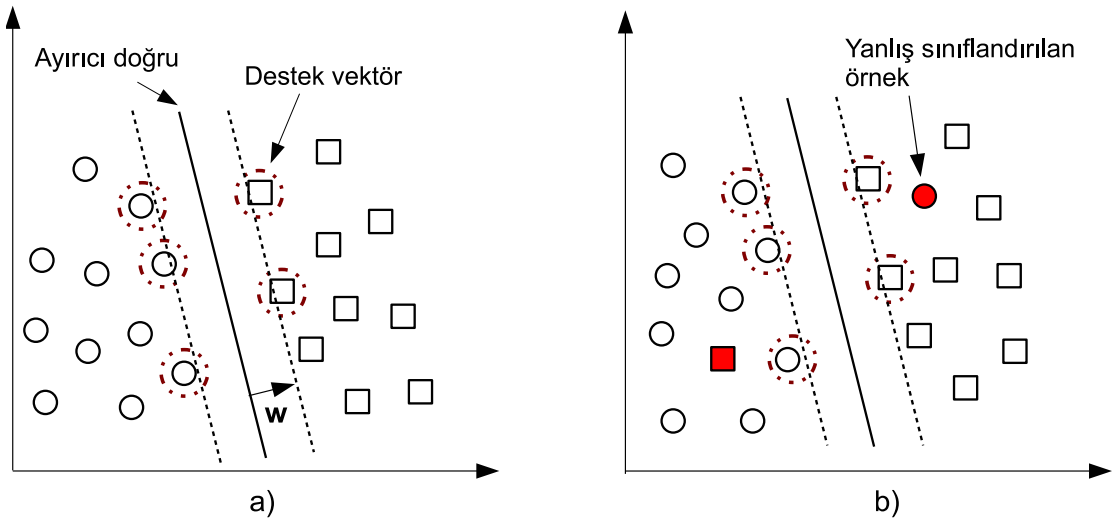
olup, r, GMM-UBM yönteminde olduğu gibi ilgi faktörüdür. Denklemler (3.63) - (3.66) ile belirtilen işlemler iteratif olarak tekrarlanarak (genellikle 4 ya da 5 tekrar) konuşmacı kod kitabı elde edilmiş olur. VQ-UBM yönteminin detayları (Hautamäki ve ark., 2008; Hanilçi ve Ertaş, 2011a; Kinnunen ve ark., 2009) çalışmalarında yer almaktadır.

VQ-UBM yöntemi ile konuşmacı doğrulama sisteminin test aşamasında, bilinmeyen konuşmacıya ait öznitelik vektörleri, $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$, arkaplan kod kitabı, $\mathbf{U}$, ve iddia edilen konuşmacı modeli, $\mathbf{C}_{\text{TGT}}$, kullanılarak karar verme skoru şu şekilde hesaplanır:

$$\Lambda(\mathbf{Y}) = MSE(\mathbf{Y}, U) - MSE(\mathbf{Y}, \mathbf{C}_{\text{TGT}}), \tag{3.68}$$

3.6. Destek Vektör Makineleri

Önceki bölümlerde detayları anlatılan VQ ve GMM yöntemleri öncelikli olarak konuşmacıyı karakterize edecek model parametrelerinin tahmin edilmesini amaçlayan modelleme yöntemleridir. Bu yöntemlerin test aşamasında ise istatistiksel veya uzaklığa dayalı kriterler gözönünde bulundurularak karar verme işlemi gerçekleştirilmektedir.



Şekil 3.13: SVM ile doğrusal olarak (a) ayrılabilen durum ve (b) ayrılamayan durum.

Son yıllarda ise makine öğrenme (*machine learning*) teknikleri örüntü tanıma problemlerine uyarlanmıştır. Bu tip modelleme yöntemleri, iki sınıfın örneklerinden yola çıkarak bu iki sınıfı birbirinden ayırt edecek şekilde eğitilmektedir. Bu tarz tekniklere ayırt edici (*discriminative*) modelleme teknikleri adı verilmektedir. Literatürde, örüntü tanıma problemlerinde oldukça popüler hale gelen en yaygın makine öğrenme tekniği destek vektör makineleri (*support vector machines-SVM*) yöntemidir (Burges, 1998). SVM, el yazısı dijit tanıma (Burges, 1998; Schölkopf ve ark., 1995), nesne tanıma (Blanz ve ark., 1996) ve yüz tanıma (Osuna ve ark., 1997) gibi uygulamalarda yaygın olarak kullanılmakta ve oldukça başarılı sonuçlar vermektedir.

SVM, son yıllarda konuşmacı belirleme ve doğrulama problemlerinde de ilgi çeken yöntemlerden biri olmuştur (Campbell, 2002; Campbell ve Richardson, 2007; Wan ve Renals, 2002; Campbell ve ark., 2006b,a)

SVM yönteminin çalışma prensibini tanımlayabilmek için, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$ şeklindeki her biri iki boyutlu olan toplam l adet eğitim vektör kümesinden örnekler alan iki sınıflı bir durumu gözönünde bulunduralım. Her bir örnek, $\mathbf{x}_i$, pozitif ya da negatif sınıftan hangisine ait olduğunu belirtmek üzere, $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, l$ ve $y_i \in \{-1, 1\}$, olacak şekilde temsil edilir. İki sınıfın birbirinden doğrusal bir doğru ile ayırt edilebildiği durum Şekil 3.13(a) da gösterilmiştir. Pozitif ve negatif sınıf-

ların birbirinden ayrılması işlemi şekilde de belirtilen *ayırıcı bir doğru* (*separating hyperplane*) ile gerçekleştirilir. Ayırıcı doğru üzerinde bulunan noktalar $\mathbf{w}.\mathbf{x}+b=0$ eşitliğini sağlamaktadır. Bu ifadede $\mathbf{w}$, ayırıcı doğruya dik bir vektörü belirtmektedir.

Ayırıcı doğrunun her iki tarafında yer alan ve Şekil 3.13 de kesikli çizgilerle belirtilen her sınıfın sınırlarının belirlendiği doğrular şu şekilde tanımlanmaktadır:

$$y_i(\mathbf{w}.\mathbf{x}+b) \geq 1. \tag{3.69}$$

Sınıf sınırlarının belirlendiği doğrular üzerinde bulunan ve Denklem (3.69) şartını sağlayan eğitim vektörleri destek vektörler olarak adlandırılmaktadır (Şekil 3.13). İsminden de anlaşılacağı gibi *destek vektörler*, ayırıcı doğruyu destekleyen veya tanımlayan eğitim vektörleridir. Eğitim vektörlerinden destek vektör olarak seçilmeyen vektörler SVM'nin eğitilmesinde herhangi bir katkı sağlamamaktadır. Yani bu vektörler eğitim kümesinden çıkarılsa dahi SVM'nin eğitilmesi neticesinde yine aynı ayırıcı doğru elde edilecektir.

SVM tekniğinin eğitilmesinde temel amaç, eğitim vektörleri kullanılarak ayırıcı doğrunun bulunması işlemidir. SVM'nin eğitilmesi genellikle Lagrange çarpanları, α_i, ile temsil edilir. Ayırıcı doğrunun optimum pozisyonu amaç fonksiyonu belirli kısıtlara göre maksimum yapılacak şekilde hesaplanır. Amaç fonksiyonu:

$$\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j \mathbf{x}_i.\mathbf{x}_j \tag{3.70}$$

şeklinde tanımlanmaktadır. Amaç fonksiyonu maksimum yapılırken dikkat edilecek kısıtlar :

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i, \tag{3.71}$$

$$\sum_i \alpha_i y_i = 0. \tag{3.72}$$

Bu şekilde eğitilen SVM tekniğinde, $\alpha_i > 0$ şartını sağlayan eğitim vektörleri destek vektörler olarak belirlenmektedir. Bu nedenle, α_i Lagrange çarpanları genellikle

destek vektör katsayıları olarak ifade edilmektedir.

SVM'nin test aşamasında ise, test vektörü, $\mathbf{x}$, direkt olarak sınıflandırma işleminde kullanılır. Karar aşamasında

$$f(\mathbf{x}) = sgn\left(\sum_i \alpha_i y_i \mathbf{x}.\mathbf{x}_i + d\right) \tag{3.73}$$

fonksiyonu kullanılır ve

$$f(\mathbf{x}) \begin{cases} > 0 & \text{Sınıf } +1 \\ < 0 & \text{Sınıf -1} \end{cases}. \tag{3.74}$$

olacak şekilde karar verme işlemi gerçekleştirilir. Denklem (3.71) de yapılan tanımlamaya göre karar fonksiyonu $f(\mathbf{x}) = sgn(\mathbf{w}.\mathbf{x} + b)$ şeklinde basitleştirebilir.

Eğitim vektörlerinin doğrusal bir doğru ile ayrılamayacağı durumda (Şekil 3.13(b)), yanlış sınıflandırılan eğitim vektörleri Lagrange çarpanlarının oldukça yüksek olacak şekilde hesaplanmasına neden olacağından dolayı bu durum SVM'nin eğitiminde uygun bir çözümün bulunmasına engel olmaktadır. Bu problemin çözümü için SVM tekniğinin temeli çekirdek fonksiyonuna (*kernel function*) dayanmaktadır. Çekirdek fonksiyonunun amacı, giriş eğitim ve test vektörlerini daha yüksek boyutlu uzaya taşımaktır. Bu işlem yapılırken temel amaç, yüksek boyutlu uzaya taşınan vektörlerin doğrusal olarak ayırt edilebilmesini sağlamaktır. Açıkça belirtilmemiş olsa da örneğin Denklem (3.73) de kullanılan çekirdek fonksiyonu, $K(\mathbf{x}, \mathbf{x}_i) = \mathbf{x}.\mathbf{x}_i$ şeklinde bir fonksiyondur ve bu *lineer çekirdek fonksiyonu* olarak bilinmektedir. Çekirdek fonksiyonu ile Denklem (3.73)

$$f(\mathbf{x}) = sgn\left(\sum_i \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + d\right) \tag{3.75}$$

şeklinde yazılabilir. Burada $\mathbf{x}$ giriş vektörünü, $\mathbf{x}_i$ eğitim öznitelik vektörlerini ve y_i eğitim vektörünün ait olduğu sınıf etiketini (± 1) ve α_i destek vektör katsayılarını belirtmektedir. Çekirdek fonksiyonu genellikle şu şekilde ifade edilir:

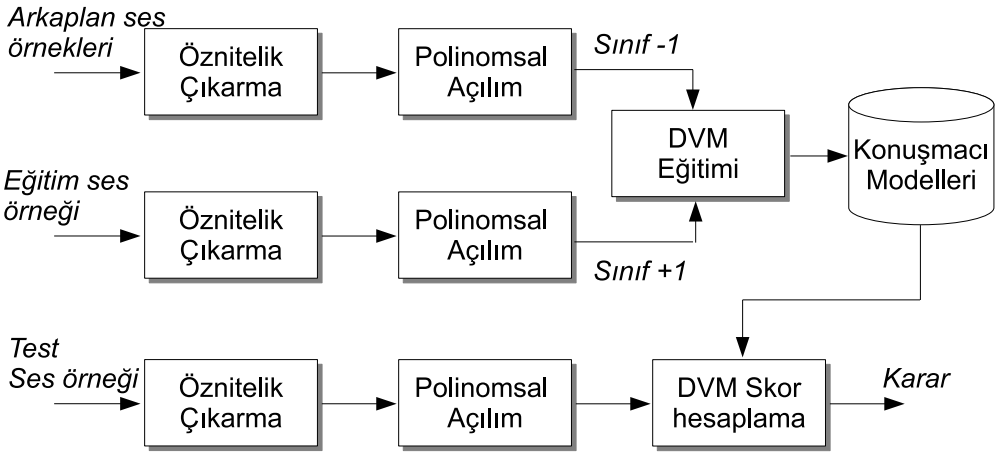
$$K(\mathbf{x}, \mathbf{x}_i) = \Phi(\mathbf{x}).\Phi(\mathbf{x}_i). \tag{3.76}$$

Burada $\Phi(\mathbf{x})$, $\mathbf{x}$ vektörünü istenen yüksek boyutlu uzaya taşımak için kullanılan fonksiyonu belirtmektedir. Φ fonksiyonunun seçimi uygulamanın türüne göre değişmektedir.

Daha önce de belirtildiği gibi, SVM iki sınıflı bir sınıflandırıcı olduğundan eğitim aşamasında iki sınıfa ait eğitim vektörlerinin kullanılması gerekmektedir. Bu sınıflardan birinin (Pozitif Sınıf) eğitilecek konuşmacının ses işaretinden elde edilen öznitelik vektörlerinin olduğu aşikardır. Negatif sınıf ise çok fazla sayıda farklı konuşmacıdan alınan ses örneklerinden elde edilen öznitelik vektörleri tarafından temsil edilmelidir. Negatif sınıf konuşmacılarına arkaplan konuşmacıları (*background speakers*) adı verilmektedir. Negatif sınıfın bu şekilde seçilmesi, eğitim aşamasında eğitilecek konuşmacının genel konuşmacı uzayında ayırt edici özelliklerini karakterize etmeye yardımcı olmaktadır.

Son yıllarda SVM ile konuşmacı doğrulamada dizi çekirdek (*sequence kernel*) fonksiyonları popüler hale gelmiş ve bu şekilde seçilen çekirdek fonksiyonları ile oldukça yüksek başarımlar elde edilmektedir (Lee ve ark., 2008; Bimbot ve ark., 2004; Fauve ve ark., 2007b). Dizi çekirdek fonksiyonları, değişen sayıda vektörden oluşan özniteliklerin tek bir karakteristik vektör ile temsil edilmesine olanak sağlamaktadır. Bu sayede, her ses işaretinden elde edilen öznitelik vektörleri değişik uzunlukta olmasına rağmen dizi çekirdek fonksiyonları ile her işaretin öznitelikleri aynı boyuttaki tek bir vektör ile temsil edilmektedir. Bu yaklaşımın bir diğer avantajı ise hafıza gereksinimini azaltmasıdır. Konuşmacı doğrulama işleminde test sesinden elde edilen her biri D-boyutlu olan öznitelik vektörleri kümesi, $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$, kullanılarak bir skor hesaplanır. Skor hesaplama fonksiyonu, $f(\mathbf{Y}) = \mathbf{w}^T\mathbf{b}(\mathbf{Y})$ şeklinde bir fonksiyon olup, burada $\mathbf{w}$, sınıflandırıcı parametrelerinden oluşan konuşmacı modeli ve $\mathbf{Y}$ ise öznitelik vektörlerinin yüksek boyutlu uzaya taşınması ile elde edilmiş bir vektördür (Bishop, 2006, 1995).

Genelleştirilmiş doğrusal ayırtaç dizisi çekirdek fonksiyonu (*generalized linear discriminant sequence kernel - GLDS*) (Campbell ve ark., 2006a; Campbell, 2002) ve GMM süpervektör (*GMM supervector - GMM-SV*) (Campbell ve ark., 2006b)



Şekil 3.14: GLDS-SVM yönteminin işlem adımları.

yöntemleri konuşmacı doğrulamada en sık kullanılan iki dizi çekirdek fonksiyonlarıdır. GLDS yönteminde (Şekil 3.14), bir ses işaretinden elde edilen öznitelik vektörleri polinomsal açılım ile daha yüksek boyutlu uzaya taşınırlar. Örneğin, $\mathbf{x} = [x_1\ x_2]^{\mathrm{T}}$ şeklinde verilen 2 boyutlu bir vektörün ikinci dereceden polinomsal açılımı $\mathbf{b}(\mathbf{x}) = [1\ x_1\ x_2\ x_1^2\ x_1x_2\ x_2^2]^{\mathrm{T}}$ şeklinde hesaplanır. Konuşmacı doğrulamada ise bir ses işaretinden elde edilen $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ şeklindeki her biri D-boyutlu toplam T adet öznitelik vektörlerinin ortalama polinomsal açılım vektörü hesaplanır ve ilgili konuşmacı böylece tek bir vektör ile temsil edilir:

$$\mathbf{b}(\mathbf{X}) = \sum_{t=1}^{T} \mathbf{b}(\mathbf{x}_t). \tag{3.77}$$

Eğitim aşamasında, eğitilecek konuşmacıya ait N adet eğitim ses örneklerinden elde edilen öznitelik vektörleri ile şu şekilde bir konuşmacı matrisi oluşturulur:

$$\mathbf{M}_{\mathrm{SPK}} = \begin{bmatrix} \mathbf{b}(\mathbf{X}_1)^t \\ \mathbf{b}(\mathbf{X}_2)^t \\ \vdots \\ \mathbf{b}(\mathbf{X}_N)^t \end{bmatrix}. \tag{3.78}$$

Benzer şekilde, arkaplan konuşacılarına ait ses örneklerinden elde edilen öznitelik

vektörleri ile $\mathbf{M}_{\text{IMP}}$ matrisi oluşturulur ve bütün eğitim vektörleri

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{\text{SPK}} \\ \mathbf{M}_{\text{IMP}} \end{bmatrix}, \tag{3.79}$$

şeklinde bir matris ile temsil edilir. Son olarak GLDS çekirdek fonksiyonu şu şekilde tanımlanır:

$$K(\mathbf{X}_i, \mathbf{X}_j) = \mathbf{b}(\mathbf{X}_i)\mathbf{R}^{-1}\mathbf{b}(\mathbf{X}_j). \tag{3.80}$$

Burada, $\mathbf{R} = (1/N_{imp})(\mathbf{M}^t\mathbf{M})$ ve N_{imp} negatif sınıfı (arkaplan kümesi) temsil etmek için kullanılan ses örneği sayısını belirtmektedir. GLDS çekirdek fonksiyonu ile SVM sınıflandırıcısının karar aşamasında kullanılan ve Denklem (3.81) ile belirtilen sonuç fonksiyonu

$$f(\mathbf{x}) = sgn\left(\sum_i \alpha_i y_i \mathbf{b}_i \mathbf{R}^{-1}\mathbf{b}_x + d\right) \tag{3.81}$$

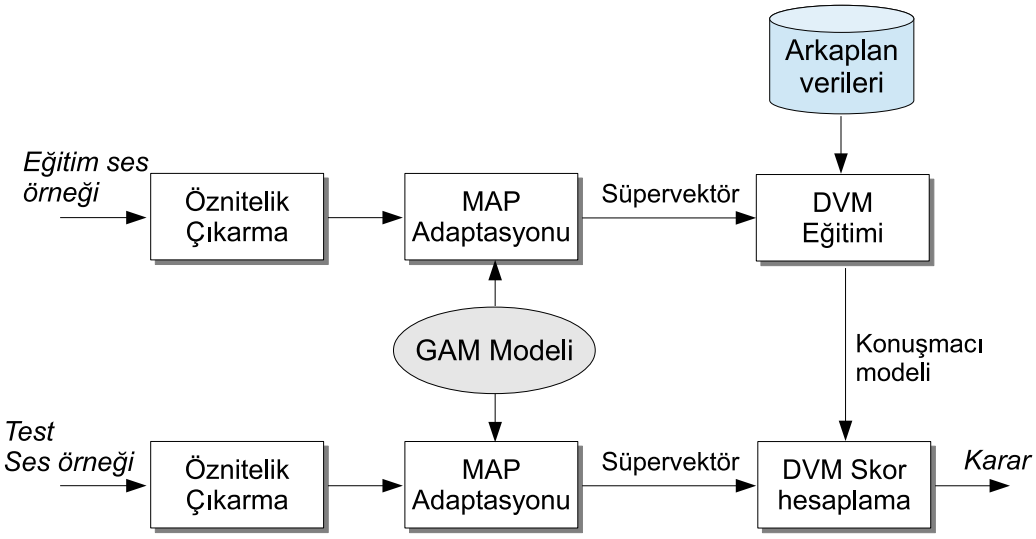
şeklinde yazılabilir. Burada $\mathbf{b}_i$ destek vektörleri, $\mathbf{b}_x$ ise test ses işaretinden elde edilen ortalama polinomsal açılım vektörünü temsil etmektedir. SVM yönteminin eğitilmesi neticesinde bütün destek vektörler

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{R}^{-1}\mathbf{b}_i + \mathbf{d} \tag{3.82}$$

şeklinde tek bir vektöre indirgenebilir ve bu eğitilen konuşmacının modeli olarak ifade edilmektedir. Test aşamasında ise skor şu şekilde hesaplanır:

$$Skor = \mathbf{w}^t\mathbf{b}_x. \tag{3.83}$$

Diğer bir popüler çekirdek fonksiyonu tekniği olan GMM-SV yöntemi (Şekil 3.15) ise öznitelik vektörlerinin *MAP* uyarlanması ile elde edilen GMM modelinin ortalama vektörlerinin konuşmacıyı temsil etme kapasitesi ile SVM yönteminin ayırt edicilik özelliklerinin birleştirilmesi prensibine dayanmaktadır. GMM-SV yöntemi, öznitelik vektör kümesinin tek bir yüksek boyutlu vektöre dönüştürülmesini sağlayan oldukça



Şekil 3.15: GMM-süpervektör yönteminin işlem adımları.

kolay bir yöntemdir. Daha öncede belirtildiği gibi GMM modeli,

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x}; \mu_k, \mathbf{\Sigma}_k). \tag{3.84}$$

şeklinde tanımlanmaktadır. Burada, w_k, μ_k ve $\mathbf{\Sigma}_k$ sırası ile karışım ağırlıklarını, ortalama vektörlerini ve ortak değişinti matrislerini temsil etmektedir. GMM-SV yönteminde bahsedilen süpervektör, $\mathbf{X}$ öznitelik vektörlerinin UBM modeli ile *MAP* uyarlaması netiesinde elde edilen GMM modelinin ortalama vektörlerinin arka arkaya eklenmesi ile oluşturulan ve $\mu = [\mu_1^t \dots \mu_K^t]$ şeklinde tanımlanan $K \times D$ boyutunda bir vektördür. GMM-SV yönteminde çekirdek fonksiyonu

$$K(\mathbf{X}_i, \mathbf{X}_j) = \mu_i(\mathbf{X}_i)\mathbf{R}^{-1}\mu(\mathbf{X}_j). \tag{3.85}$$

şeklinde tanımlanmaktadır. Burada $\mathbf{R}$, UBM modelinden elde edilen köşegen bir matris olup, köşegen elemanları UBM modelinin ağırlık katsayıları ve varyans parametrelerinden şu şekilde hesaplanır:

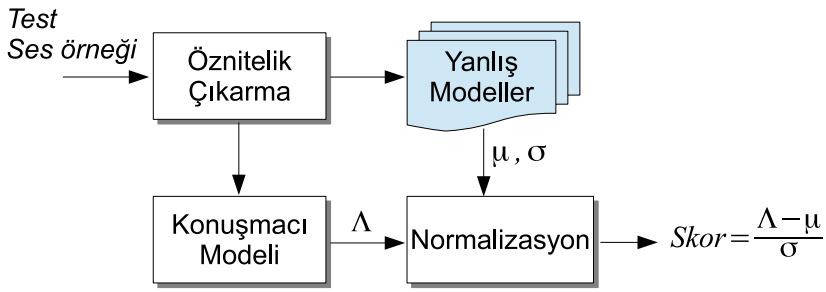
$$\begin{bmatrix} \sqrt{w}_1 \mathbf{\Sigma}_1^{-1/2} \\ \vdots \\ \sqrt{w}_K \mathbf{\Sigma}_K^{-1/2} \end{bmatrix}. \tag{3.86}$$

Çekirdek fonksiyonu bu şekilde tanımlandıktan sonra diğer aşamalar GLDS yöntemi ile aynı şekilde gerçekleştirilmektedir.

3.7. Skor Normalizasyonu

Gerçek zamanlı uygulamalarda kullanılan konuşmacı doğrulama sistemleri ciddi performans kayıplarına sebep olan birçok faktöre maruz kalmaktadır. Toplamsal gürültü, iletim kanalından kaynaklanan ses işaretinde meydana gelen bozulmalar, kayıt cihazları arasındaki farklılıklar ve konuşmacının yaş ve sağlık problemleri bu faktörlerden bazılarıdır. Bu negatif etkileri ortadan kaldırmak veya en azından azaltmak için değişik normalizasyon teknikleri kullanılmaktadır. Konuşmacı doğrulamada kullanılan normalizasyon tekniklerinin amacı, öznitelik vektörlerinden, konuşmacı modellerinden ve/veya hesaplanan skorlardan bu negatif faktörlerin etkilerini azaltmaktır. Öznitelik vektörleri üzerinde uygulanan normalizasyon yöntemleri bölüm 3.3.5.'de anlatılmıştır. Model bazında normalizasyon ise genel arka plan modeli yönteminde anlatıldığı gibi, eğitilecek konuşmacı modellerinin geniş bir konuşmacı uzayındaki modelden (UBM modeli) türetilmesidir.

Skor normalizasyonu, skor dağılımının bir yanlış test kümesinden (*impostor trials*) elde edilen istatistiksel parametrelerin kullanılarak dengelenmesi işlemidir (Rosenberg ve ark., 1992; Finan ve ark., 1997; Ramos-Castro ve ark., 2007; Auckenthaler ve ark., 2000; Hanilçi ve Ertaş, 2011c). Bu yanlış testlerin uygulanmasında, kullanılan veri kümesi dışından seçilen bir konuşmacı kümesi kullanılmaktadır. Konuşmacı doğrulama uygulamalarında kullanılan en yaygın skor normalizasyonu tekniği *Test Normalizasyonu*dur (TNorm) (Auckenthaler ve ark., 2000) (Şekil 3.16). TNorm yönteminin temel amacı test işaretlerinin elde edildiği koşulların (ortam, kanal farklılıkları gibi) farklı olmasından dolayı skor seviyesinde meydana gelen negatif etkilerin azaltılmasıdır. TNorm işleminde test sesinden elde edilen öznitelik vektörleri, $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$, öncelikle bir dizi yanlış konuşmacı modeli ile test edilerek, elde edilen skorların ortalama ve standart sapma değerleri hesaplanır. Daha sonra test işareti iddia edilen konuşmacı modeli ile test edilerek gerçek skor değeri, $\Lambda(\mathbf{Y})$, he-



Şekil 3.16: TNorm skor normalizasyonu.

saplanır. Son aşamada ise normalize edilmiş ve karar aşamasında kullanılacak skor değeri şu şekilde elde edilir:

$$Skor = \frac{\Lambda(\mathbf{Y}) - \mu}{\sigma} \tag{3.87}$$

3.8. Kanal Etkilerinin Dengelenmesi

Bir konuşmacı doğrulama sisteminde kullanılan ses örnekleri telefon hattından elde edilmiş ve kayıtlar farklı zamanlarda alınmış ise iletim hattı farklılıklarından kaynaklanan negatif etki, konuşmacı doğrulama performansının düşmesine neden olmaktadır. Bu nedenle son yıllarda konuşmacı doğrulama uygulamalarında iletim hattının olumsuz etkisini azaltacak yöntemler üzerinde çalışmalar yapılmaktadır. NAP (*nuisance attribute projection*), GMM-SV yöntemi için kanal etkilerinin azaltılması amacıyla önerilmiş bir teknik olup konuşmacı tanıma performansında sağladığı artıştan dolayı oldukça popüler hale gelmiştir (A. Solomonof ve Boardman, 2005; A. Solomonof ve Quillen, 2004). NAP yöntemi, SVM tekniğinde yüksek boyutlu uzaya taşınan ve konuşmacıyı karakterize eden öznitelik vektöründen gereksiz ve bozucu (*nuisance*) etkileri azaltmayı amaçlamaktadır.

NAP yönteminde bozucu etkileri azaltmak amacıyla bir ses veritabanı kullanılarak oturum farklılıklarını temsil eden düşük boyutlu bir alt uzay matrisi hesaplanır. Daha sonra bu alt uzay matrisi kullanılarak eğitilecek veya test edilecek yüksek boyutlu uzaya taşınmış karakteristik vektörden (süpervektör) değişimin yüksek olduğu bileşenler atılır. NAP tekniği ile bir öznitelik vektöründen kanal bileşenlerinin

giderilmesi şu şekilde gerçekleştirilir:

Toplam N_s adet konuşmacıdan alınan N tane ses örneğinden elde edilmiş ve yüksek boyutlu uzaya taşınmış öznitelik vektörlerinden oluşan bir küme şu şekilde verilsin.

$$\{\mathbf{b}_{(1,s_1)} \ldots \mathbf{b}_{(h_1,s_1)} \ldots \mathbf{b}_{(1,s_{N_s})} \ldots \mathbf{b}_{(h_{N_s},s_{N_s})}\} \tag{3.88}$$

burada h_i, $i = 1, \ldots, N_s$ ve s_i konuşmacısının ses örneklerinin alındığı oturum indisleridir. Her konuşmacının vektörlerinden o konuşmacının tüm vektörlerinin ortalamaları çıkarılarak bir matris şu şekilde oluşturulur:

$$\mathbf{M} = [\tilde{\mathbf{b}}_{(1,s_1)} \ldots \tilde{\mathbf{b}}_{(h_1,s_1)} \ldots \tilde{\mathbf{b}}_{(1,s_{N_s})} \ldots \tilde{\mathbf{b}}_{(h_{N_s},s_{N_s})}] \tag{3.89}$$

$$\tilde{\mathbf{b}}_{(l,s_i)} = \mathbf{b}_{(l,s_i)} - \bar{\mathbf{b}}_{s_i} \tag{3.90}$$

$\mathbf{M}$ matrisi $E \times N$ boyutundadır. E, öznitelik vektörlerinin yüksek boyutlu uzaya taşınması neticesinde elde edilen karakteristik vektör boyutu ve $N = h_1 + \ldots + h_{N_s}$ bütün konuşmacılardan elde edilen toplam ses örneği sayısına karşılık gelmektedir. $\mathbf{M}$ matrisi bu sayede sadece kanal değişimlerini temsil eden bir matris haline gelmiştir ve konuşmacı bilgisi içermemektedir. Bu nedenle bu matristen elde edilen, değişimin en yüksek olduğu K adet özvektör ile tanımlanan alt uzay kanal değişim alt uzayını temsil edecektir ve konuşmacı verilerinden bu alt uzay bilgileri elenerek, konuşmacı vektörlerindeki kanal değişim etkileri azaltılmış olacaktır.

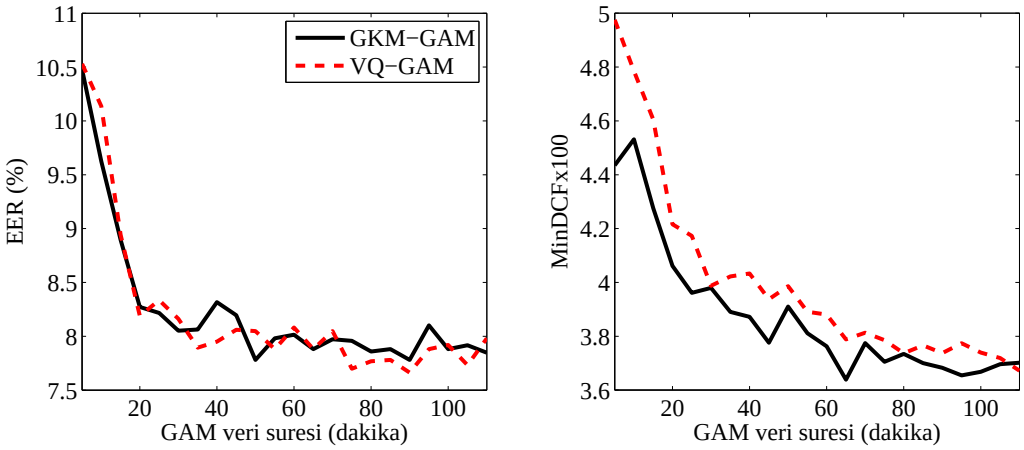
4. Deneysel Sonuçlar

Bu bölümde tez çalışmaları neticesinde elde edilerek değişik uluslararası/ulusal dergi ve konferanslarda yayınlanan konuşmacı tanıma sonuçları anlatılacaktır.

4.1. Arkaplan Veri Süresinin Konuşmacı Doğrulama Perfromansına Etkisi

MAP uyarlamalı sınıflandırıcılar kullanılırken genellikle UBM modeli çok fazla sayıda ses işareti kullanılarak (genellikle yüzlerce saatlik konuşma işaretleri) oluşturulmaktadır. Bu konuda yaygın olarak kabul edilen görüş, daha uzun ve fazla sayıda veriler kullanıldığında UBM modeli alternatif hipotezi daha iyi bir şekilde temsil ettiği ve dolayısıyla konuşmacı doğrulama performansının arttığıdır. Bir diğer görüş ise, arkaplan modeli oluşturulurken kullanılan ses işaretlerinin fazla sayıda farklı konuşmacılardan elde edilmesi durumunda UBM modelinin alternatif hipotezi temsil etme kabiliyetinin daha yüksek olacağıdır. Hasan ve Hansen (2011) tarafından yapılan çalışmada, UBM verilerinin en az 40 farklı konuşmacıdan elde edilmesi gerektiği deneysel olarak gösterilmiştir. Fakat UBM modelinin eğitiminde kullanılacak veri miktarının belirsizliğine ilişkin problem devam etmektedir. Çok uzun süreli verilerin kullanılmasının en büyük dezavantajı, UBM modeli parametrelerinin tahmin edilme zamanının veri miktarına bağlı olarak artmasıdır. Bu nedenle az miktarda veri kullanarak temsil kabiliyeti yüksek UBM modellerinin oluşturulması arzu edilmektedir.

Bu çalışmada, NIST 2002 veritabanı kullanılarak UBM veri süresinin konuşmacı doğrulama performansına etkisi deneysel olarak incelenmiştir (Hanilçi ve Ertaş, 2013a). Deneysel çalışmada, GMM-UBM ve VQ-UBM sınıflandırıcıları için değişik uzunlukta arka plan verisi kullanılarak performans karşılaştırması yapılmıştır. NIST 2001 veritabanından seçilen ve 5 dakika ile 110 dakika arasında değişen arka plan ses işaretleri UBM modelinin eğitilmesinde kullanılmıştır. Model boyutu olarak (GMM-UBM yönteminde Gauss bileşen sayısı ve VQ-UBM yönteminde kod kitabı sayısı)

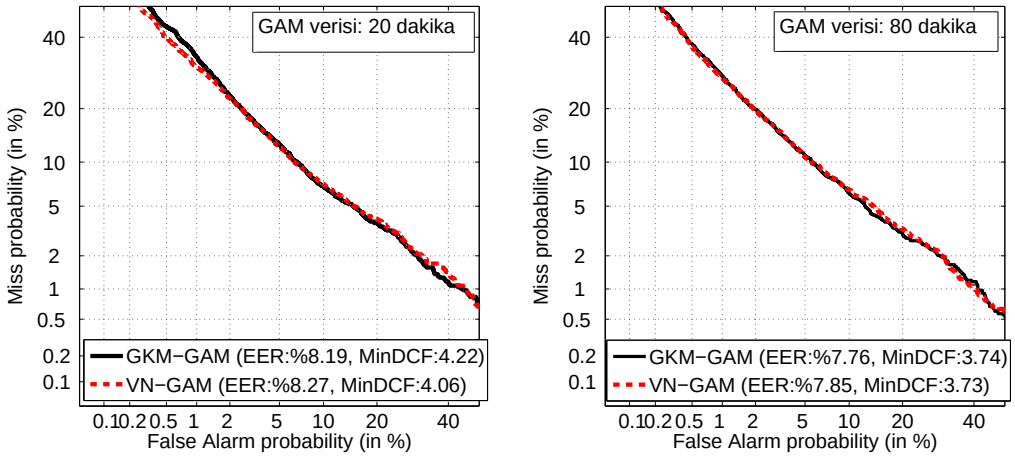


Şekil 4.1: Farklı arkaplan veri süreleri için elde edilen EER ve MinDCF değerleri.

$K = 512$ kullanılmıştır. 12 adet MFCC öznitelikleri ve bunların birinci ve ikinci türevleri (Δ ve $\Delta\Delta$) olmak üzere toplam 36 boyutlu öznitelik vektörleri kullanılmıştır. Sınıflandırıcı algoritmanın dizaynı ve özniteliklerle ilgili detaylı bilgiler (Hanilçi ve Ertaş, 2013a) çalışmasında verilmiştir.

Şekil 4.1 de arka plan veri süresinin konuşmacı doğrulama performansına etkisi gösterilmektedir. MinDCF değerleri karşılaştırmanın kolay yapılması amacıyla 100 ile çarpılarak verilmektedir. Şekilden görüldüğü gibi arka plan verisi 20 dakikadan daha kısa olduğunda her iki sınıflandırma yöntemi de oldukça düşük performans göstermektedir. Ancak 20 dakikadan daha uzun veri kullanıldığında performansta çok büyük değişiklik görülmemektedir (EER değeri %8.2 ile %7.6 arasında değişmektedir). Başarım ölçütü olarak MinDCF kriteri kullanıldığında ise 20 dakikadan uzun verilerde performansta daha küçük değişimler gözlenmektedir (3.9 ile 3.7 değerleri arasında değişen MinDCF değerleri elde edilmiştir). Şekil 4.2 ise 20 ve 80 dakikalık UBM verileri kullanıldığında elde edilen DET eğrilerini göstermektedir. Görüldüğü gibi iki durumda EER değerleri arasında yaklaşık 0.4 gibi bir fark ortaya çıkmaktadır ki bu UBM modelinin eğitim zamanı göz önüne alındığında kabul edilebilir bir farktır.

Sonuç olarak, yapılan deneysel çalışmalar neticesinde elde edilen bulgular UBM modelinin eğitiminde kullanılan veri süresinin yüzlerce dakika veya saat uzunluğunda olmasının gerekmediğini ortaya koymuştur. Bu nedenle, daha az miktarda veri kulla-



Şekil 4.2: Farklı arkaplan veri süreleri için elde edilen DET eğrileri.

nılarak daha hızlı bir şekilde konuşmacı doğrulama işlemi gerçekleştirilebilmektedir. Bu çalışma ile ilgili daha detaylı bilgiler Hanilçi ve Ertaş (2013a) çalışmasında yer almaktadır.

4.2. Veri Süresinin Konuşmacı Doğrulama Performansına Etkisi

Bu çalışmada, konuşmacı doğrulamada eğitim ve test veri sürelerinin performansa etkisi detaylı bir şekilde incelenmiş olup, GMM-UBM, VQ-UBM, SVM-GLDS ve GMM-SV yöntemlerinin karşılaştırılması ele alınmıştır (Hanilçi ve Ertaş, 2013b).

Eğitim ve test veri süreleri konuşmacı doğrulama performansına önemli şekilde etki etmektedir. Son yıllarda, özellikle NIST konuşmacı tanıma ve değerlendirme organizasyonları neticesinde konuşmacı doğrulamada uzun eğitim ve test verilerinin kullanımı yaygınlaşmıştır. Bunun en temel nedeni uzun süreli veriler kullanıldığında konuşmacıyı karakterize eden özniteliklerin daha kolay elde edilmesidir. Fakat, gerçek zamanlı uygulamalarda uzun süreli verilerin kullanılması pratik değildir. Çünkü gerçek zamanlı uygulamalarda kullanıcıyı uzun süreli ses üretmeye zorlamak uygulamayı zorlaştırmaktadır. Literatürde değişik veri sürelerinin kullanıldığı konuşmacı doğrulama çalışmaları mevcuttur (Mak ve ark., 2006; Vogt ve Sridharan, 2006; Fauve ve ark., 2007a; Vogt ve ark., 2008a,b, 2009; Pelecanos ve ark., 2004; McLaren

ve ark., 2010). Örneğin Mak ve ark. (2006), değişik model uyarlama yöntemleri kısa süreli eğitim verileri için karşılaştırmış ve MAP uyarlamalı GMM yönteminin (GMM-UBM) uzun süreli eğitim verileri kullanıldığında daha iyi performans verdiğini göstermiştir.

Genellikle konuşmacı doğrulama uygulamalarında bir sınıflandırıcı iyi performans gösteriyorsa, araştırmacıların çoğu o sınıflandırıcının her koşulda iyi performans sergileyeceğini düşünmektedir. Bu nedenle, bu çalışmada değişik eğitim ve test süreleri ile bilinen en yaygın dört sınıflandırma yöntemleri olan GMM-UBM, VQ-UBM, SVM-GLDS ve GMM-süpervektör yöntemlerinin karşılaştırılması yapılmıştır. Deneysel çalışmalarda NIST 2002 ve NIST 2005 veritabanları kullanılmıştır. Her iki veritabanı da telefon konuşmalarından oluşan ses işaretlerinden oluşturulmuştur. NIST 2002 veritabanında 191 bayan ve 139 erkek olmak üzere toplam 330 konuşmacı mevcut olup, 2982 doğru (1232 erkek ve 1750 bayan) and 36277 yanlış/geçersiz (14630 erkek ve 21647 bayan) sınama verisi bulunmaktadır. Test verileri ise ortalama 45 saniye uzunluğundadır. NIST 2005 veritabanı ise dört farklı eğitim-test veri uzunluğu kombinasyonuna sahiptir:

- Bir konuşma - Bir konuşma (1conv-1conv): Her konuşmacı 5 dakika uzunluğunda eğitim ve test verisine sahiptir.
- Bir konuşma - 10 saniye (1conv-10sec): Konuşmacıların eğitim verileri 5 dakika uzunluğunda olup test verileri ise 10 saniye uzunluğundadır.
- 10 saniye - 10 saniye (10sec-10sec): Eğitim ve test süreleri her konuşmacı için 10 saniyedir.
- 10 saniye - Bir konuşma (10sec-1conv): Bir konuşmacı 10 saniyelik veri ile eğitilip, 5 dakikalık veri ile test yapılmaktadır.

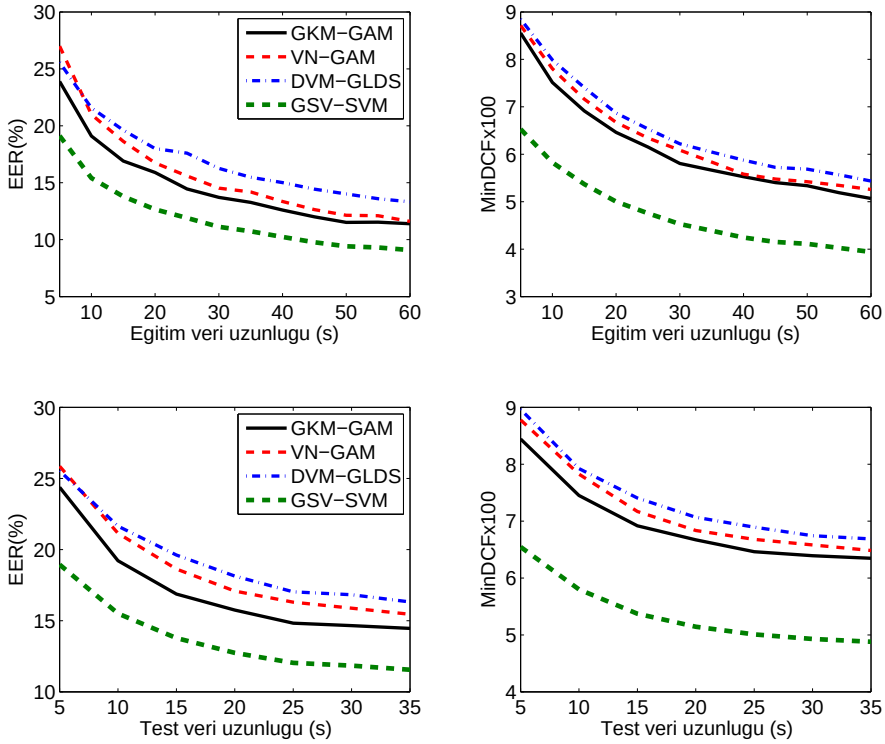
NIST 2005 veritabanında her eğitim/test kombinasyonu farklı sayıda konuşmacıya ve sınama sayısına sahiptir. Her bir kombinasyona ait konuşmacı, doğru ve geçersiz sınama sayıları erkek ve bayan konuşmacılar için Çizelge 4.1 de özetlenmiştir.

Çizelge 4.1: NIST 2005 veritabanındaki her bir eğitim/test veri süresi kombinasyonundaki konuşmacı ve sınama sayıları.

Eğitim-Test Koşulu	Bayan Konuşmacılar		
	Konuşmacı sayısı	Doğru sınama sayısı	Geçersiz sınama sayısı
1conv-1conv	372	1547	16247
1conv-10sec	372	1706	17856
10sec-10sec	378	1752	18114
10sec-1conv	378	1594	16410
Eğitim-Test Koşulu	Erkek Konuşmacılar		
	Konuşmacı sayısı	Doğru sınama sayısı	Geçersiz sınama sayısı
1conv-1conv	274	1228	12396
1conv-10sec	274	1391	13708
10sec-10sec	274	1391	13887
10sec-1conv	274	1227	12219

NIST 2002 ile yapılan deneylerde, GMM ve VQ yöntemleri için UBM modeli oluşturulmasında NIST 2001 veritabanından seçilen 56 bayan ve 77 erkek konuşmacıya ait ses örnekleri kullanılmıştır. NIST 2005 deneylerinde ise UBM modeli NIST 2004 veritabanından seçilen 246 erkek ve 370 bayan konuşmacıya ait ses örnekleri kullanılmıştır. GMM-UBM, VQ-UBM ve GMM-süpervektör (GSV-SVM) yöntemlerinde model derecesi (Gauss bileşeni sayısı ve kodvektör sayısı) $K = 512$ olarak seçilmiştir. SVM-GLDS yönteminde, 3. dereceden polinomsal açılım fonksiyonu kullanılmıştır.

Şekil 4.3 de NIST 2002 veritabanı ile farklı uzunluklardaki eğitim ve test verileri ile elde edilen konuşmacı doğrulama performansları dört farklı yöntem için gösterilmiştir. Şekilde birinci satır, eğitim veri uzunluğunun etkisini göstermekte olup bu durumda test verilerinin uzunluğu 15 saniye olacak şekilde sabitlenmiştir. Aynı şekilde ikinci satırdaki deneylerde ise, eğitim verilerinin uzunluğu 15 saniye olarak sabitlendiğinde test verilerinin uzunluğunun performansa etkisi gösterilmektedir. Şekilden görüleceği üzere, eğitim ve test verilerinin süresi arttıkça performans da sistematik bir şekilde artmaktadır. Eğitim verisinin uzunluğunun tanıma performansına etkisi test veri süresinin etkisinden daha yüksektir. Örneğin GMM-UBM yöntemi ile 5 saniyelik eğitim verileri kullanıldığında %23.2 EER değeri elde edilirken, 5 saniyelik test verisi kullanıldığında %24.37 EER değeri elde edilmiştir. GSV-SVM yöntemi her durumda diğer üç yöntemden daha iyi performans göstermektedir.



Şekil 4.3: Farklı veri süreleri için NIST 2002 veritabanı ile hesaplanan EER ve MinDCF değerleri.

Şekiller 4.4 ve 4.5'de sırası ile NIST 2005 veritabanının 1conv eğitim ve 10sec eğitim durumlarında değişik test koşullarında elde edilen EER ve MinDCF değerlerinin değişimleri gösterilmektedir. Deneysel çalışmalar neticesinde elde edilen bulgular şu şekilde özetlenebilir:

GMM-UBM yöntemi için,

- 5 dakika eğitim durumunda (1conv), test süresi 10 saniye olduğunda konuşmacı doğrulama performansı, test süresinin 1conv (5 dakika) olduğu duruma göre yaklaşık %60 oranında düşmektedir (EER değeri %10.33 den %16.47 ye yükselmektedir).

- 10 saniye (10sec) eğitim ve 5 dakikalık test durumunda, doğrulama performansı 1conv (5 dakika)-1conv (5 dakika) eğitim-test durumuna göre %72 daha düşüktür.

- 10sec-10sec durumunda, doğrulama performansı 10sec-1conv durumuna göre

%56 oranında azalmaktadır.

VQ-UBM yöntemi için,

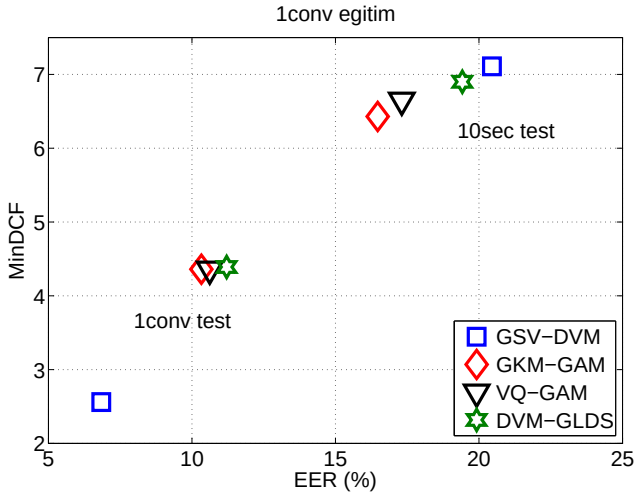
- 1conv eğitim durumunda, test süresi 10 saniye olduğunda %17.62 EER değeri elde edilirken, test süresi 1conv (5 dakika) olduğunda %10.62 EER değeri elde edilmiştir. Bu da performansta yaklaşık %40'lık bir artışa karşılık gelmektedir.
- 10sec-1conv koşulunda, EER değeri %19.49 a yükselmiştir ki bu, 1conv-1conv durumu ile karşılaştırıldığında performansta %83 lük bir düşüşe karşılık gelmektedir.
- 10sec-1conv durumunda (%19.49 EER), 10sec-10sec durumuna göre (%30.96 EER) performansta %59 luk bir iyileşme gözlenmiştir.

GMM-SV sistemi için,

- 1conv-1conv durumunda, GMM-SV diğer yöntemlere göre en düşük EER ve MinDCF değerlerini vermektedir.
- Eğitim ve test veri uzunluğu düşük olduğunda GMM-SV yöntemi diğer yöntemlerden daha düşük performans göstermektedir. Buna benzer bir bulgu yakın zamanda McLaren ve ark. (2010) tarafından, GMM-UBM ve GMM-SV yöntemlerinin karşılaştırıldığı bir çalışmada belirtilmiştir.

SVM-GLDS yöntemi için ise en ilginç bulgu, 1conv-10sec ve 10sec-1conv durumları arasında çok az performans farkı olmasıdır. Bunun dışındaki durumlarda elde edilen sonuçlar GMM-UBM ve VQ-UBM yöntemleri ile benzerlik göstermektedir.

Bu çalışma neticesinde görülmüştür ki eğitim ve/veya test veri uzunlukları konuşmacı tanıma performansına önemli derecede etki etmektedir. Ayrıca karşılaştırılan dört yöntemden GMM-GSV yönteminin literatürdeki bir çok çalışmada diğer yöntemlerden oldukça iyi performans sergilediği belirtilse de bunun ancak uzun eğitim



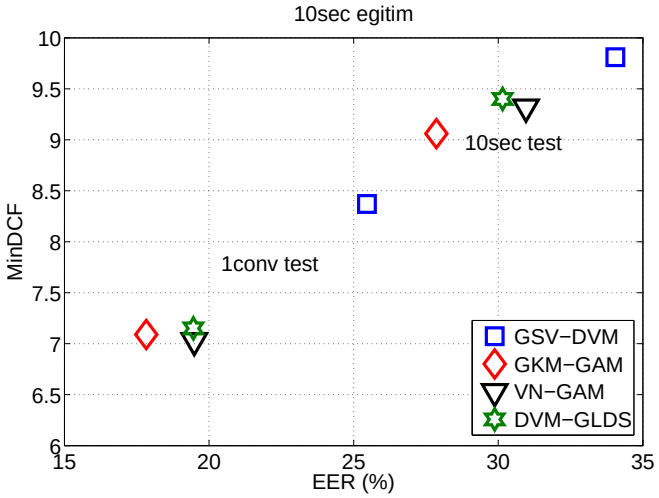
Şekil 4.4: Farklı veri süreleri için NIST 2002 veritabanı ile hesaplanan EER ve MinDCF değerleri.

ve test verileri durumunda geçerli olduğu gözlenmiştir. Ayrıca en önemli bulgulardan bir tanesi de, eğitim verisinin uzunluğu konuşmacı tanıma performansına etkisi test verisinin etkisinden daha yüksektir. Bu çalışma ile ilgili detaylar (Hanilçi ve Ertaş, 2013b) çalışmasında bulunabilir.

4.3. VQ-UBM Sınıflandırıcı için Skor Normalizasyonu

Bu çalışmada, TNorm skor normalizasyonu tekniğinin VQ-UBM sınıflandırıcısına uyarlanması ve SVM-GLDS yöntemi ile karşılaştırılması ele alınmıştır (Hanilçi ve Ertaş, 2011c). Literatürde skor normalizasyonu yöntemleri sadece GMM-UBM ve SVM-GLDS sınıflandırıcıları için incelendiğinden, VQ-UBM modelleme tekniği için performansa yapacağı etkiler incelenmediğinden dolayı bu tür bir çalışmanın önemli olacağı aşikardır.

Deneysel çalışmalarda, NIST 2002 konuşmacı doğrulama veritabanı kullanılmış olup, model boyutu (kod kitabı boyutu) olarak $K \in \{64, 512, 1024\}$ şeklinde üç farklı değer seçilmiştir. SVM-GLDS yönteminde ise $m = \{2, 3\}$ polinomsal açılım dereceleri kullanılmıştır. Öznitelik vektörleri olarak daha önceki çalışmalarda da olduğu gibi, 12 MFCC ve bunların birinci ve ikinci türevlerinden oluşan toplam 36 boyutlu vektörler

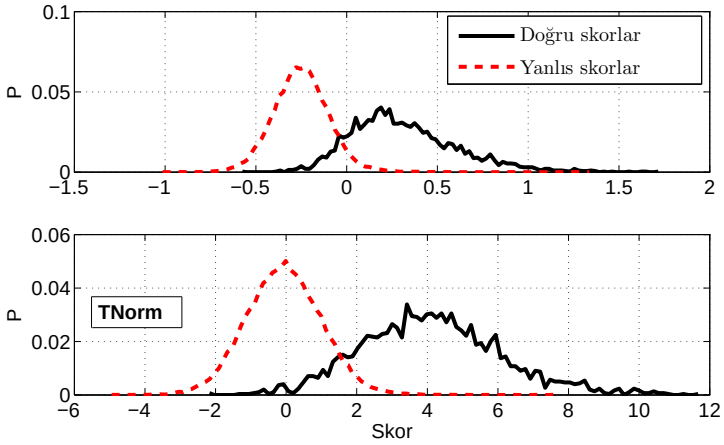


Şekil 4.5: Farklı veri süreleri için NIST 2002 veritabanı ile hesaplanan EER ve MinDCF değerleri.

kullanılmıştır. Öznitelik vektörlerine bölüm 3.3.5.'de anlatılan CMVN normalizasyonu uygulanmıştır. TNorm skor normalizasyonu için kullanılacak olan konuşmacı kümesi NIST 2001 veritabanından seçilmiştir.

İlk olarak skor normalizasyonu işleminin skor dağılımları üzerindeki etkisi incelenmiştir. Normalizasyon yapılmadan ve normalizasyon işlemi uygulandıktan sonraki doğru (*target trials*) ve yanlış sınama (*impostor trials*) skorlarının dağılımları Şekil 4.6 ile gösterilmiştir. Görüldüğü gibi normalizasyon işlemi sonrası yanlış sınama skorlarının ortalaması 0 noktasına çekilmiştir. Bununla beraber doğru sınama skorları ise 4 noktasında ortalanmıştır. Bu sayede sistemin karar aşamasında yapacağı olası hatalar azaltılmış olmaktadır.

Çizelge 4.2 TNorm normalizasyonu öncesi ve sonrasında elde edilen EER ve MinDCF değerlerini her iki sınıflandırıcı yöntemi ve değişik model dereceleri için göstermektedir. Çizelgeden de görüleceği gibi TNorm, her iki sınıflandırıcı için hem EER hem de MinDCF kriterleri göz önüne alındığında performansı artırmaktadır. Ancak SVM-GLDS yönteminde $m = 3$ polinomsal açılım derecesi kullanıldığında EER değeri az da olsa yükselmektedir. Fakat bu durum, TNorm yönteminin SVM-GLDS sınıflandırıcısının performansını düşürdüğü anlamına gelmemektedir. Bunu göstermek amacı ile normalizasyon öncesi ve sonrasında DET eğrileri Şekil 4.7 ile gösterilmektedir.



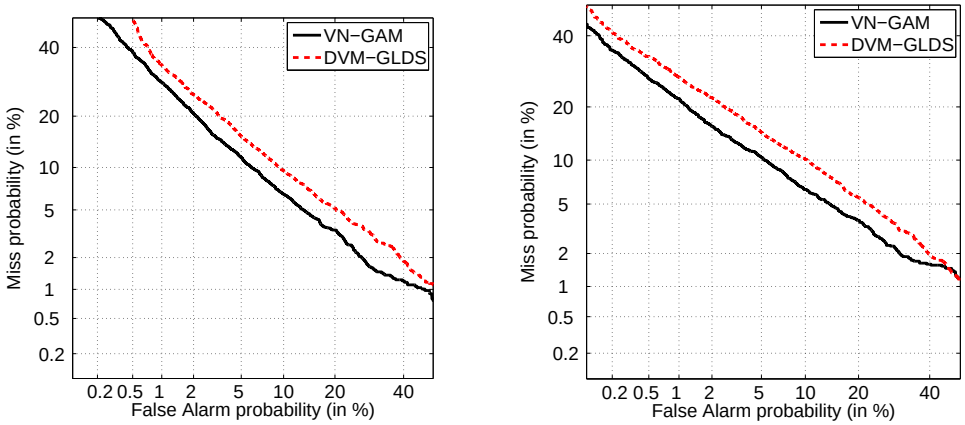
Şekil 4.6: Skor normalizasyonunun skor dağılımları üzerine etkisi.

Çizelge 4.2: Skor normalizasyonunun EER ve MinDCF değerleri üzerine etkisi.

Yöntem	Tnorm öncesi		Tnorm sonrası	
	EER (%)	MinDCFx100	EER (%)	MinDCFx100
VQ-UBM ($K = 64$)	9.27	4.28	8.91	3.49
VQ-UBM ($K = 512$)	7.98	3.86	7.69	3.13
VQ-UBM ($K = 1024$)	8.25	3.94	8.26	3.31
SVM-GLDS ($m = 2$)	12.97	6.35	12.45	5.22
SVM-GLDS ($m = 3$)	9.69	4.36	10.09	3.73

DET eğrilerinden de görüleceği gibi her iki sınıflandırıcı için TNorm uygulandığında düşük yanlış kabul bölgesinde yanlış ret oranları azalmaktadır (DET eğrilerinin sol üst köşesi). Bu nedenle EER değerinde değişiklik olmaması veya ufak bir artış olması aslında performansın düştüğünü göstermemektedir. Dolayısı ile doğrudan EER ve MinDCF değerlerine göre performans analizi yapmak yerine DET eğrileri ile analiz etmek daha yararlı olacaktır.

Sonuç olarak TNorm yönteminin VQ-UBM sınıflandırıcısına uygulanması ile konuşmacı doğrulama performansı model derecesinden bağımsız olarak artmaktadır. Ayrıca deneysel sonuçlar neticesinde elde edilen bulgular, VQ-UBM yönteminin SVM-GLDS sınıflandırıcısından skor normalizasyonu öncesinde ve sonrasında daha iyi başarım gösterdiğini ortaya koymuştur. Bu çalışma ile ilgili daha detaylı bilgiler (Hanilçi ve Ertaş, 2011c) çalışmasında yer almaktadır.



Şekil 4.7: Skor normalizasyonu öncesi (sol) ve sonrasında (sağ) elde edilen DET eğrileri.

4.4. Doğrusal Öngörü Hatası ile Konuşmacı Doğrulama

Bölüm 3.3.'de bahsedildiği gibi konuşmacı tanıma çalışmalarında en çok MFCC ve LPCC öznitelikleri kullanılmaktadır. Ayrıca yine LPCC özniteliklerinin hesaplanması anlatılırken de bahsedildiği gibi LP yöntemi ile ilgili yaygın olarak kabul edilen görüş, hata işaretinin konuşmacının kimliği hakkında herhangi bir bilgi içermediğidir. Bunun nedeni ses üretim mekanizmasındaki (Şekil 3.5) zamanla değişen süzgeç parametreleri LP katsayıları tarafından modellendiğinden ve giriş işareti $u(n)$ periyodik darbe dizisi veya rasgele gürültü olduğundan kişiyi temsil etmemektedir.

Bu çalışmada, konuşmacı doğrulamada en çok kullanılan iki farklı öznitelik çıkarma yöntemi olan MFCC ve LPCC yöntemleri, özniteliklerin doğrusal öngörü hata işareti ile çıkarıldığı durumlar ele alınmıştır (Hanilçi ve Ertaş, 2011b). Standart MFCC ve LPCC öznitelikleri ve doğrusal öngörü hatasından elde edilen kepstral katsayıların (*linear prediction residual cepstral coefficients-LPRC*) GMM-UBM sınıflandırıcısı kullanılarak NIST 2001 veritabanı ile konuşmacı doğrulama performansları karşılaştırılmıştır. Sınıflandırıcı için Gauss karışım sayısı $K = \{64, 256\}$ olarak iki farklı değer seçilmiştir. Çizelge 4.3 de MFCC, LPCC ve LPRC özniteliklerinin karşılaştırmaları ve öznitelik vektörlerinin birinci ve ikinci türevlerinin performansa etkileri gösterilmektedir. Çizelgeden da görüldüğü gibi her üç öznitelik vektör kümesi için Δ ve $\Delta\Delta$ öznitelikleri performansı artırmaktadır. Ayrıca bu çalışmanın da moti-

Çizelge 4.3: Değişik öznitelik kümeleri için elde edilen EER ve MinDCF değerleri.

Öznitelik	EER(%)		MinDCFx100	
	$K = 64$	$K = 256$	$K = 64$	$K = 256$
MFCC	18.06	17.39	7.79	7.22
MFCC + Δ	16.78	15.67	7.47	6.81
MFCC + Δ + $\Delta\Delta$	16.48	15.45	7.28	6.65
LPCC	18.49	18.15	7.48	7.13
LPCC + Δ	16.87	15.75	7.01	6.63
LPCC + Δ + $\Delta\Delta$	16.13	15.26	6.82	6.47
LPRC	17.90	17.3	7.61	6.98
LPRC + Δ	17.02	16.06	7.31	6.70
LPRC + Δ + $\Delta\Delta$	17.02	15.75	7.24	6.76

vasyonunu oluşturan, hata işaretinden öznitelik vektörleri elde edilmesi durumunda (LPRC öznitelikleri), elde edilen EER ve MinDCF değerleri MFCC ve LPCC öznitelikleri ile karşılaştırılabilir seviyededir. Buradan da anlaşılmaktadır ki hata işareti sanılanın aksine konuşmacı kimliği hakkında neredeyse orijinal işaret ve LP katsayıları kadar bilgi taşımaktadır. Bunun muhtemel nedeni ise, Şekil 3.5 de gösterilen insan ses üretim mekanizması modelinde giriş işareti $u(n)$ matematiksel olarak ve yapılan varsayımlara göre kişiden bağımsız olarak kabul edilmektedir. Aslında gerçekte giriş işaretini oluşturan mekanizma konuşmacının akciğerlerine çektiği hava ve organlarının etkileşimidir. Bu nedenle bu işaretin kişi hakkında bilgi taşıması kaçınılmazdır. Daha detaylı bilgiler ve deneysel çalışmalar (Hanilçi ve Ertaş, 2011b) çalışmasında yer almaktadır.

4.5. Spektrum Hesaplama Yönteminin Konuşmacı Doğrulama Performansına Etkisi

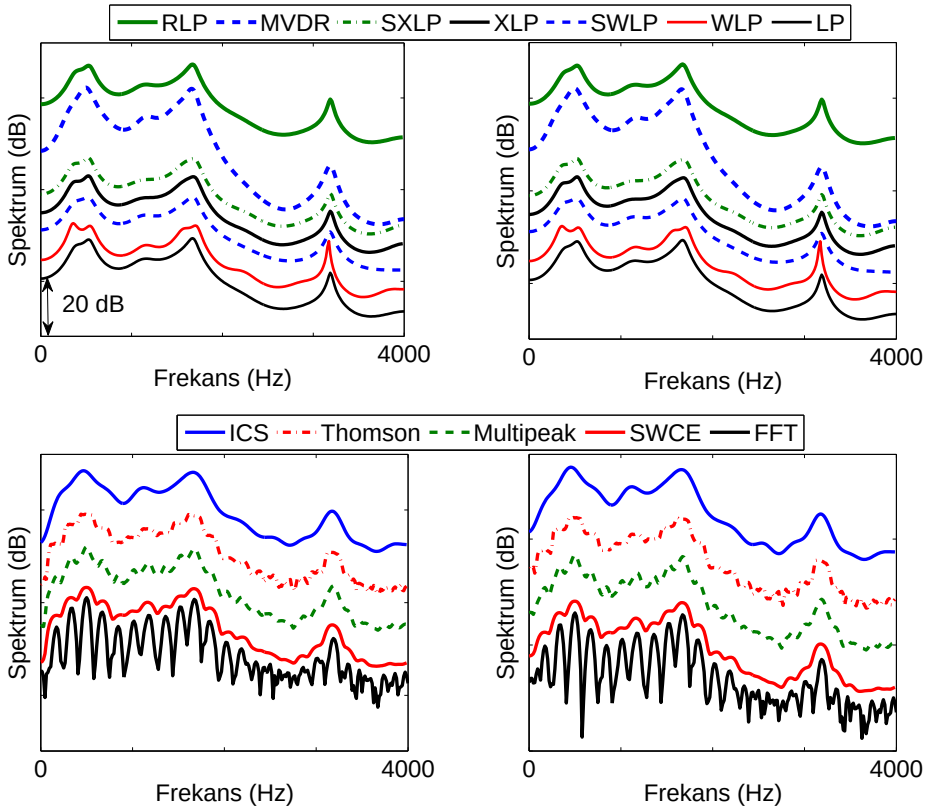
Bu çalışmada, 12 farklı spektrum hesaplama yönteminin toplamsal gürültü durumunda konuşmacı doğrulama performansına etkisi incelenmiştir (Hanilçi ve ark., 2012c). Çalışmada 2 farklı gürültü türü (factory ve babble gürültüler) ve 5'er farklı sinyal gürültü oranı (*Signal-to-Noise Ratio - SNR*) ele alınmıştır. NIST 2002 veritabanı ile yapılan deneylerde, 512 adet Gauss bileşeninden oluşan GMM-UBM

yöntemi kullanılmış olup TNorm skor normalizasyonu uygulanmıştır. Deneylerde 12 adet MFCC katsayısı, Δ ve $\Delta\Delta$ öznitelikleri ile toplam 36 boyutlu öznitelik vektörleri kullanılmıştır. Öznitelik vektörlerine bölüm 3.3.5.'de detayları verilen ortalama ve varyans normalizasyonu (CMVN) işlemi uygulanmıştır.

Bölüm 3.3.1.'de MFCC öznitelik vektörlerinin ses işaretinin FFT spektrumunun süzgeç takımından geçirilerek, süzgeç takımı çıkışlarının logaritmasının ayrık kosinus dönüşümü ile hesaplandığı belirtilmişti. Bu çalışmada FFT spektrum hesaplama adımı değiştirilerek farklı spektrum hesaplama yöntemleri kullanılmıştır. Kullanılan spektrum hesaplama yöntemlerinin detayları Bölüm 3.3.3.'de anlatılmıştır. Şekil 4.3 de bir ses işaretinden alınan bir çerçevenin ve aynı çerçeveye 0 dB SNR seviyesinde *factory* gürültü eklenmesi neticesinde elde edilen işaretin farklı spektrum hesaplama yöntemleri ile elde edilen spektrumları gösterilmektedir. Şekil 4.3(a) ve (b) doğrusal öngörü (LP) tabanlı spektrum hesaplama yöntemlerini, (c) ve (d) ise parametrik olmayan yöntemleri göstermektedir. Şekilde sol sütun orijinal işareti, sağ sütun ise gürültülü işareti göstermektedir. Şekilden görüleceği gibi, işarete gürültü eklenmesi spektrumda bozulmalara neden olmaktadır. LP ve WLP yöntemleri ile elde edilen spektrumlarda yüksek frekans değerlerinde keskin tepelerin oluştuğu görülmektedir. Ayrıca orijinal ve gürültülü işaretin FFT spektrumları karşılaştırıldığında, spektral dinamikler (Spektrumun maksimum ve minimum değerleri arasındaki fark) arasındaki farkın oldukça yüksek olduğu görülmektedir. Bu da performansta düşüşe neden olmaktadır.

Çizelge4.4 ve 4.5'de MFCC özniteliklerinin hesaplanmasında farklı spektrum hesaplama yöntemlerinin kullanılması durumunda elde edilen konuşmacı doğrulama performansları sırasıyla factory ve babble gürültüler için verilmektedir. Çizelgelerde, her alt-grup için en düşük EER değerleri altı çizili şekilde ve her satırdaki en düşük EER değeri kalın font ile belirtilmiştir. Ayrıca Şekil 4.9 de seçilen bazı yöntemler için elde edilen DET eğrileri verilmiştir.

Deneysel çalışmalar neticesinde elde edilen bulgular şu şekilde özetlenebilir:



Şekil 4.8: Bir ses işaretinin ve 0 dB SNR seviyesinde gürültü eklenmiş işaretin spektrumları.

Çizelge 4.4: Factory gürültü durumunda farklı spektrum hesaplama yöntemleri ile elde edilen EER (%) değerleri.

	Temel yöntemler		Ağırlıklandırılmış LP yöntemleri				Diğer yöntemler		
SNR (dB)	DFT	LP	WLP	SWLP	XLP	SXLP	ICS	MVDR	RLP
orijinal	7.65	7.44	7.48	7.81	7.94	7.78	8.01	7.62	7.57
20	8.08	7.83	**7.81**	8.22	8.04	7.98	8.45	8.30	7.81
10	9.32	**8.50**	8.79	9.11	8.85	8.85	9.55	9.12	8.75
0	10.46	**9.93**	10.34	10.06	10.01	9.99	10.88	10.36	10.29
-10	15.35	14.96	15.19	**14.35**	14.55	14.73	16.05	14.78	15.02

Çizelge 4.5: Babble gürültü durumunda farklı spektrum hesaplama yöntemleri ile elde edilen EER (%) değerleri.

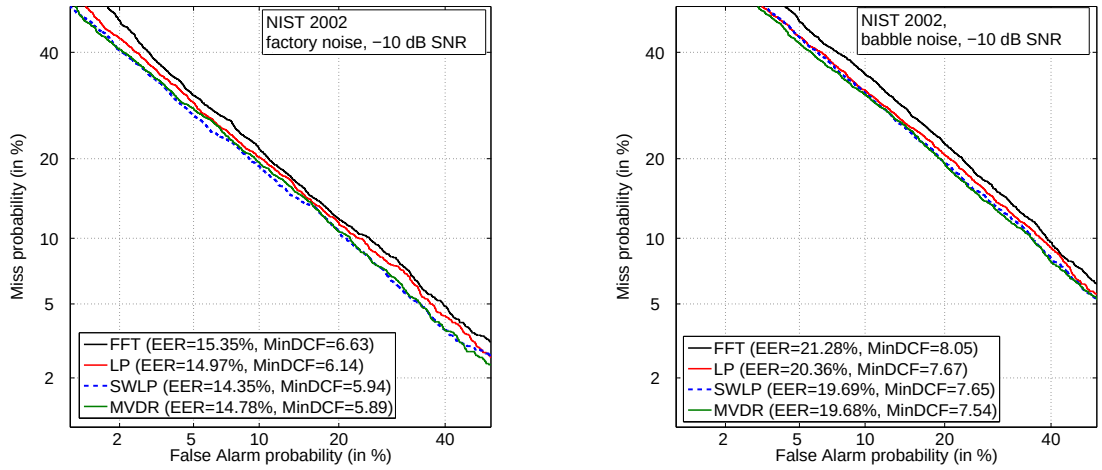
	Temel yöntemler		Ağırlıklandırılmış LP yöntemleri				Diğer yöntemler		
SNR (dB)	DFT	LP	WLP	SWLP	XLP	SXLP	ICS	MVDR	RLP
orijinal	7.65	7.44	7.48	7.81	7.94	7.78	8.01	7.62	7.57
20	7.83	7.78	**7.71**	8.11	7.94	7.93	8.28	8.19	7.81
10	8.85	**8.58**	8.70	8.78	8.68	8.85	9.56	9.19	8.92
0	11.62	11.23	11.47	10.93	**10.63**	10.83	11.91	11.70	10.94
-10	21.27	20.35	21.02	19.69	20.35	20.23	22.03	**19.68**	20.12

Orijinal ses işaretleri (gürültü eklenmediğinde) için;

- LP tabanlı yöntemler FFT tekniğine göre daha iyi performans göstermektedir (%7.65). XLP (%7.34) ve SWLP (%7.34) yöntemleri ile diğer LP tabanlı yöntemlerden daha iyi performans elde edilmiştir.
- ICS tekniği %8.01 EER ile en düşük performansı sergilemektedir.

Toplamsal gürültü durumunda:

- Standart LP metodu ile yüksek SNR seviyelerinde en düşük EER değeri elde edilmiştir (20 dB SNR seviyesinde WLP ve RLP yöntemlerinden daha yüksek).
- Gürültünün en fazla olduğu durumda (-10 dB SNR) SWLP (%14.35) outperforms yöntemi diğer ağırlıklandırılmış LP yöntemlerinden daha iyi performans göstermektedir (WLP, XLP ve SXLP yöntemleri için sırasıyla %15.19, %14.55 and %14.73 EER değerleri elde edilmiştir).
- Hemen hemen bütün SNR seviyelerinde ICS yöntemi en düşük performansı göstermektedir.
- babble gürültü durumunda ve en düşük SNR seviyesinde (-10 dB SNR) MVDR yöntemi en iyi performansı göstermektedir.

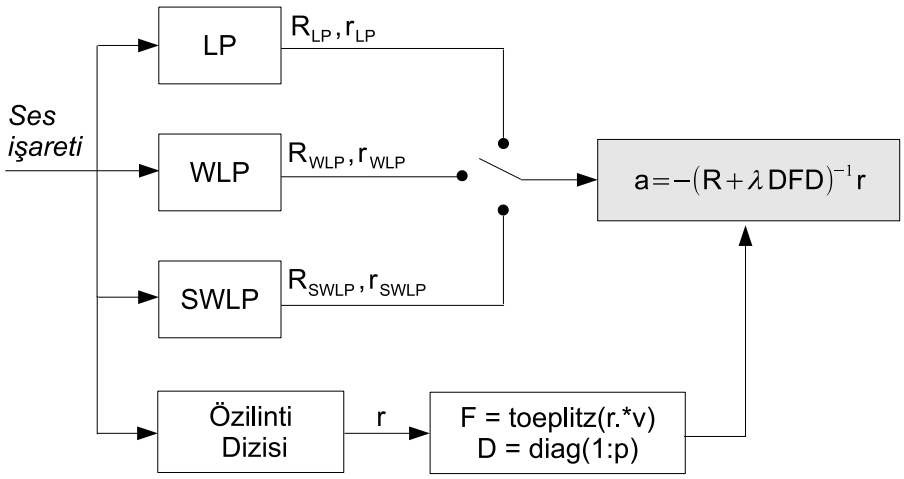


Şekil 4.9: Toplamsal factory (sol) ve babble (sağ) gürültü durumunda hesaplanan DET eğrileri (−10 dB SNR)

4.6. RLP Yöntemi ile Konuşmacı Doğrulama

Bölüm 4.5.'de öznitelik vektörlerinin hesaplanması sırasında işaretin spektrumunun genellikle FFT yöntemi ile hesaplandığı ve buna alternatif olarak değişik spektrum hesaplama yöntemleri kullanılabileceğinden bahsedilmişti. Bu çalışmada FFT yöntemi ile doğrusal öngörü (LP) tabanlı spektrum hesaplama yöntemleri toplamsal gürültü durumunda konuşmacı doğrulama problemi için karşılaştırılmıştır ve bölüm 3.3.3.'de anlatılan RLP yöntemi ile spektrum hesaplama yönteminde $\mathbf{F}$ matrisinin hesaplanması için alternatif pencere fonksiyonları ve ağırlıklandırılmış LP yöntemlerinin (WLP ve SWLP) de düzenleştirilmesi önerilmiştir (Hanilçi ve ark., 2012a,b).

Bir önceki deneysel çalışmalarda MFCC öznitelikleri elde edilirken kullanılan spektrum hesaplama yönteminin konuşmacı doğrulamada önemli etkisinin olduğu gösterilmişti (Hanilçi ve ark., 2012c). Bu çalışmada, yakın zamanda ses kodlama için önerilen ve Bölüm 3.3.3.'de detayları anlatılan düzenleştirilmiş doğrusal öngörü (RLP) (Ekman ve ark., 2008; Murthi ve Kleijn, 2000) ele alınmıştır. RLP yöntemi ile klasik FFT, LP ve ayrıca ağırlıklandırılmış LP yöntemleri (WLP ve SWLP) ile karşılaştırılmıştır. Deneysel çalışmalarda (Hanilçi ve ark., 2012c) çalışmasında olduğu gibi ve Bölüm 4.5.'de anlatıldığı gibi NIST 2002 veritabanı kullanılmıştır. İki farklı gürültü türü (factory ve babble) 5 farklı SNR seviyesinde ($SNR \in \{orijinal,\ 20,\ 10,\ 0,\ -10\}\ dB$) ses işaretlerine eklenmiştir.



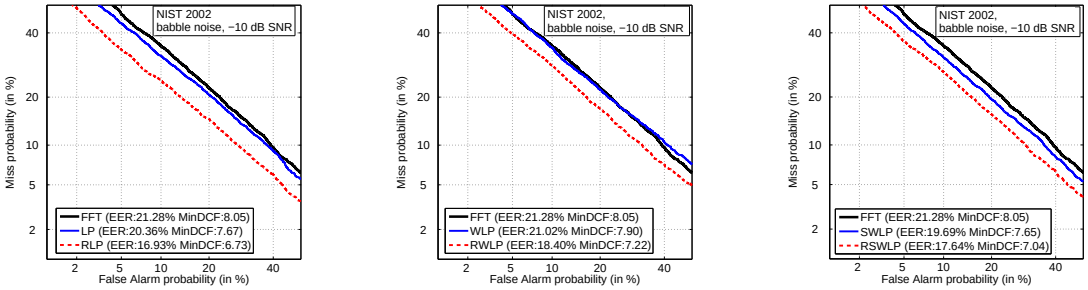
Şekil 4.10: RLP, RWLP ve RSWLP yöntemlerinin işlem adımları.

Bu çalışma ile önerilen yenilik ise, ağırlıklandırılmış WLP ve SWLP yöntemlerinin de düzenleştirilmesi işlemidir. Ekman ve ark. (2008) standart LP yöntemi için düzenleştirme (*regularization*) önermiş olup ses kodlamada RLP yönteminin iyi performans gösterdiğini belirtmiştir. LP, WLP ve SWLP yöntemlerinin düzenleştirilmesi işleminin adımları Şekil 4.10 ile gösterilmektedir. Şekilden de görüleceği gibi öncelikle ses işareti kullanılmak istenen yönteme göre (LP, WLP veya SWLP) işlenerek ilgili yöntemin özilinti matrisi ve özilinti vektörü hesaplanır. Bu yöntemlerin detayları bölüm 3.3.3.'de ve (Hanilçi ve ark., 2012b,a) çalışmalarında anlatılmıştır. Ayrıca işaretin özilinti vektörü (*autocorrelation*) bir pencere fonksiyonu ($\mathbf{v}$) ile çarpılarak düzenleştirme matrisi $\mathbf{F}$ ve köşegen $\mathbf{D}$ matrisi elde edilir. Son olarak düzenleştirilmiş öngörücü katsayıları Şekilde gri alanla gösterilen ifade yardımı ile elde edilir. Elde edilen öngörücü katsayıları kullanılarak Denklem (3.23) ile işaretin spektrumu hesaplanır ve hesaplanan spektrum, Şekil 3.4 de gösterilen MFCC özniteliklerinin çıkartılması adımlarında yer alan FFT bloğunun yerini alır.

Deneylerde öncelikle standart RLP yönteminde özilinti katsayılarının bir pencere fonksiyonu ile çarpılarak $\mathbf{F}$ matrisinin elde edildiği adım incelenmiştir. RLP yönteminin önerildiği çalışmalarda (Ekman ve ark., 2008; Murthi ve Kleijn, 2000) yazarlar, Blackman pencere fonksiyonunun kullanıldığını belirtmişlerdir. Ancak farklı pencere fonksiyonlarının kullanılması durumunda yöntemin başarımının nasıl değiştiğinin incelenmesi amacıyla bu çalışmada Blackman pencere fonksiyonunun yanında, Box-

Çizelge 4.6: RLP yönteminde özilinti ortamı pencere fonksiyonunun (**v**) etkisi.

	SNR (dB)	EER (%) Boxcar	Blackman	Hamming	DAC	MinDCFx100 Boxcar	Blackman	Hamming	DAC
	temiz	7.57	7.52	**7.37**	7.38	3.07	**3.02**	3.03	3.03
Factory	20	7.81	**7.78**	8.04	7.84	3.18	3.18	**3.16**	3.19
	10	8.75	8.85	8.85	**8.38**	3.57	3.55	3.57	**3.45**
	0	10.29	10.02	10.16	**9.41**	4.17	4.16	4.16	**3.81**
	-10	15.02	15.08	15.45	**13.61**	6.10	6.15	6.06	**5.81**
Babble	20	7.81	7.81	**7.78**	7.90	3.19	3.15	**3.14**	3.30
	10	8.92	8.51	8.68	**8.35**	3.44	3.41	**3.37**	3.46
	0	10.94	11.05	11.20	**9.61**	4.32	4.27	4.26	**3.96**
	-10	20.12	20.92	20.73	**16.93**	7.55	7.76	7.65	**6.63**



Şekil 4.11: -10 dB SNR seviyesinde DAC pencere fonksiyonu için elde edilen DET eğrileri.

car, Hamming ve ikinci özilinti fonksiyonu (*double autocorrelation - DAC*) (Mansour ve Juang, 1989; Shimamura ve Nguyen, 2010; Kobatake ve Matsunoo, 1994) pencere fonksiyonları kullanılmıştır. Çizelge 4.6 değişik pencere fonksiyonlarının toplamsal gürültü durumunda konuşmacı doğrulama performanslarını göstermektedir. Her SNR seviyesi için en iyi performans değerleri kalın font ile belirtilmiştir. Çizelgeden de görüleceği gibi DAC fonksiyonu (Mansour ve Juang, 1989; Shimamura ve Nguyen, 2010; Kobatake ve Matsunoo, 1994) en iyi başarımı göstermekte ve DAC fonksiyonu kullanıldığında diğer pencere fonksiyonlarına nazaran performansta önemli artış elde edilmektedir.

Deneylerin son aşamasında ise, DAC fonksiyonu ile WLP ve SWLP yöntemlerinin düzenleştirildiği RWLP ve RSWLP yöntemleri standart FFT ve LP yöntemleri ile karşılaştırılmıştır. Çizelge 4.7 de DAC fonksiyonunun kullanılması durumunda konuşmacı doğrulama başarımları verilmektedir. Ayrıca Şekil 4.11, -10 dB SNR seviyesinde elde edilen DET eğrilerini göstermektedir.

Çizelge 4.7: RLP, RWLP ve RSWLP yöntemleri ile toplamsal gürültü durumunda konuşmacı doğrulama başarımları.

	SNR	EER (%)						
	(dB)	FFT	LP	RLP	WLP	RWLP	SWLP	RSWLP
	orijinal	7.65	7.44	**7.38**	7.48	8.10	7.81	7.94
Factory	20	8.08	7.83	7.84	7.81	**7.75**	8.22	7.85
	10	9.32	8.50	8.38	8.79	**8.32**	9.11	8.50
	0	10.46	9.93	**9.41**	10.34	9.62	10.06	9.59
	-10	15.35	14.96	13.61	15.19	13.86	14.35	**13.32**
Babble	20	7.83	7.78	7.90	**7.71**	8.21	8.11	8.17
	10	8.85	8.58	**8.35**	8.70	8.48	8.78	8.65
	0	11.62	11.23	**9.61**	11.47	10.29	10.93	9.99
	-10	21.27	20.35	**16.93**	21.02	18.40	19.69	17.64
	SNR	MinDCFx100						
	(dB)	FFT	LP	RLP	WLP	RWLP	SWLP	RSWLP
	orijinal	3.07	3.05	3.03	**2.99**	3.33	3.08	3.41
Factory	20	3.25	3.22	3.19	**3.12**	3.14	3.21	3.24
	10	3.64	3.56	3.45	3.57	**3.32**	3.62	3.45
	0	4.13	4.21	**3.81**	4.19	3.92	4.17	3.92
	-10	6.63	6.14	**5.81**	6.19	6.03	5.94	5.87
Babble	20	3.14	3.12	3.30	**3.09**	3.35	3.19	3.44
	10	**3.44**	3.48	3.46	3.46	3.53	3.56	3.64
	0	4.53	4.34	**3.96**	4.49	4.35	4.38	4.27
	-10	8.05	7.67	**6.63**	7.90	7.22	7.65	7.04

4.7. Kanal Etkilerinin Dengelenmesi

Bu çalışmada, GMM-SV yöntemi için önerilen NAP (*nuisance attribute projection*) yönteminin SVM-GLDS metoduna uyarlanması ele alınmıştır (Hanilçi ve Ertaş, 2012). Deneyler sırasında NIST 2002 veritabanı ile konuşmacı doğrulama performansı, GMM-UBM ve SVM-GLDS yöntemleri için karşılaştırılmıştır. Literatürde yapılan çalışmalarda GMM-UBM yönteminin SVM-GLDS tekniğine göre daha iyi performans gösterdiği ortaya konulmuştur (Kinnunen ve ark., 2009). Bu çalışmada, SVM-GLDS yöntemine NAP kanal dengeleme yöntemi uyarlanarak GMM-UBM metodu ile karşılaştırılabilir seviyeye getirilmiştir.

Deneysel çalışmalarda NIST 2002 konuşmacı tanıma değerlendirme (Speaker Recognition Evaluation - SRE) veritabanı kullanılmıştır. NIST 2002, 139 erkek ve 191 kadın olmak üzere toplam 330 konuşmacıdan oluşmaktadır. Her konuşmacıya ait yaklaşık 2.5 dakika uzunluğunda eğitim verisi mevcut olup, test verileri 15 ile 45

saniye arasında değişmektedir. Veritabanında 2982 doğru (hedef/geçerli) ve 36277 adet yanlış (geçersiz/sahte) erişim olmak üzere toplam 39259 adet sınama verisi bulunmaktadır.

GMM-UBM yönteminde UBM modeli NIST 2001 veritabanından seçilen her biri 2 dakika uzunluğunda konuşma verisine sahip 38 erkek ve 22 bayan konuşmacı kullanılarak oluşturulmuştur. SVM-GDAD yönteminde ise aynı veriler negatif sınıfı oluşturmak için kullanılmıştır.

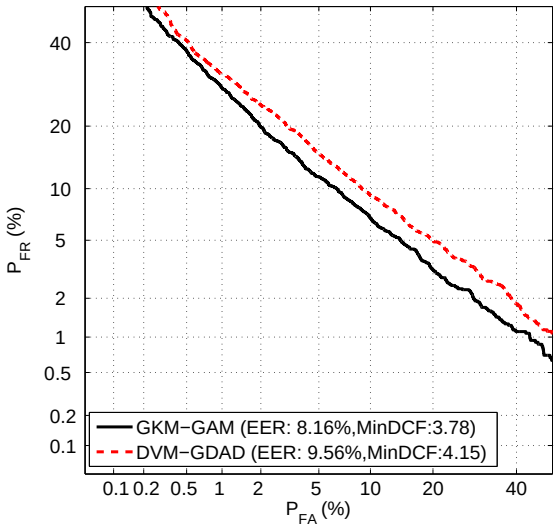
Kanal dengelemesi için NIST 2001 veritabanından seçilen toplam 1706 (666 erkek ve 1040 kadın) adet ses örneği kullanılmıştır.

Konuşmacıyı karakterize eden öznitelikler olarak MFCC kullanılmıştır. Ses işareti 15 ms'lik adımlar ile her biri 30 ms uzunluğunda olan çerçevelere bölünmüştür. Hamming pencere ile pencerelenen çerçevelerin genlik spektrumları 27 adet üçgen süzgeçten oluşan süzgeç takımından geçirilmiştir. Logaritması alınan süzgeç çıkışlarının ayrık kosinus dönüşümü alınarak MFCC öznitelik vektörleri elde edilmiştir. Deneysel çalışmalarda 12 MFCC katsayısı ve bunların birinci ve ikinci türevlerinden (Δ, $\Delta\Delta$) oluşan 36 boyutlu öznitelik vektörleri kullanılmıştır.

GMM-UBM yönteminde konuşmacıları modellemede 512 adet bileşenden oluşan GMM'ler kullanılmıştır. Konuşmacı modelleri oluşturulurken, sadece bileşen ortalamaları uyarlanmış olup bileşen ağırlıkları ve ortak değişinti matrisleri UBM modelinden aynen aktarılmıştır. İlgi parametresi (relevance factor) $r = 16$ alınmıştır. SVM-GLDS yönteminde ise konuşmacıları eğitirken 36 boyutlu öznitelik vektörlerinin 3. dereceden polinomsal açılımları kullanılmıştır. Bu sayede her bir konuşmacı 9139 boyutlu karakteristik vektörler ile temsil edilmiştir.

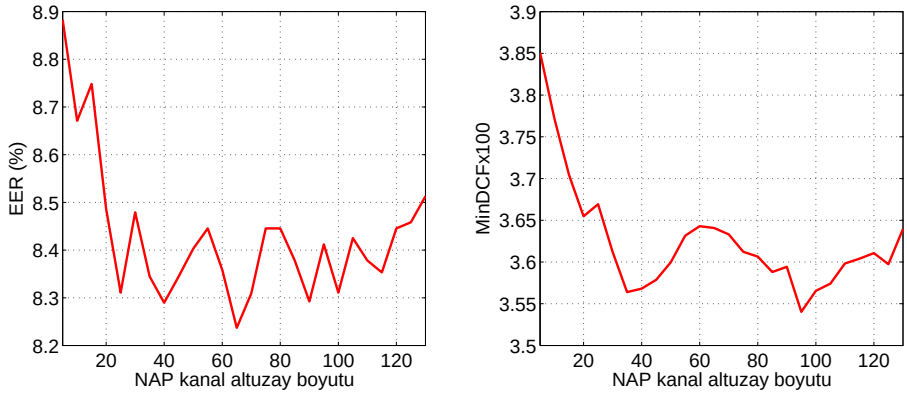
İlk olarak SVM-GDAD yöntemine kanal dengeleme uygulanmadan önce GMM-UBM yöntemi ile karşılaştırmak amacı ile referans sonuçlar elde edilmiştir. Şekil 4.12, GMM-UBM ve SVM-GLDS yöntemleri için elde edilen DET eğrilerini ve her bir yönteme ait EER ve MinDCF değerlerini göstermektedir. Şekilden de görüleceği üzere temel durumda GMM-UBM yöntemi ile SVM-GDAD yöntemine göre daha

yüksek başarım elde edilmiştir. Hem EER hem de MinDCF metriği açısından GMM-UBM yönteminin performansı SVM-GLDS sınıflandırıcısından daha yüksektir.



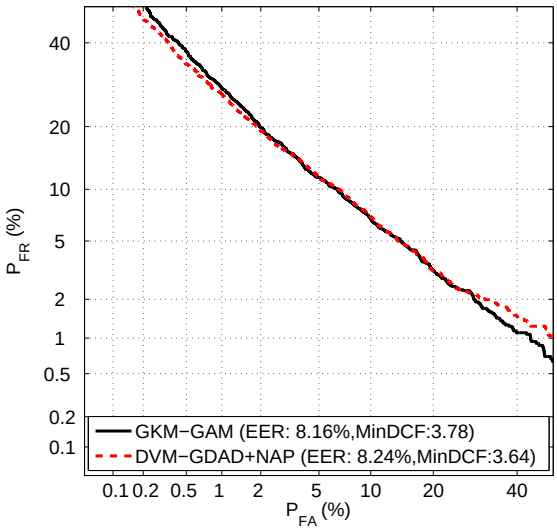
Şekil 4.12: GMM-UBM ve SVM-GLDS yöntemleri için DET eğrileri.

İkinci olarak, SVM-GLDS yöntemine kanal dengeleme uygulanırken optimum kanal alt uzay boyutunun belirlenmesi amaçlanmıştır. Bu nedenle öncelikle değişik kanal alt uzay boyutları için EER ve MinDCF değerleri elde edilmiştir. Şekil 4.13, değişik alt uzay boyutları için EER ve MinDCF değerlerinin değişimini göstermektedir. Şekilden de görüldüğü gibi kanal alt uzay boyutunun 65 seçilmesi durumunda en düşük EER elde edilmektedir. Bu nedenle sonraki deneylerde alt uzay boyutu 65 olarak sabitlenmiştir.



Şekil 4.13: NAP kanal alt uzay boyutunun konuşmacı doğrulama başarımına etkisi.

Son olarak, SVM-GLDS yöntemine NAP tekniğinin uygulanması neticesinde elde edilen konuşmacı doğrulama performansı, GMM-UBM yöntemi sonuçları ile karşılaştırılmıştır. Şekil 4.14, her iki yöntem için elde edilen DET eğrilerini ve EER ve MinDCF değerlerini göstermektedir. Görüldüğü gibi SVM-GLDS, NAP kanal dengeleme yöntemi sonucunda GMM-UBM yöntemine oldukça yakın başarım göstermektedir. Ayrıca MinDCF metriği açısından SVM-GLDS yöntemi az bir fark ile GMM-UBM yönteminden daha iyi başarım göstermektedir.



Şekil 4.14: GMM-UBM ve SVM-GLDS (NAP) yöntemleri için DET eğrileri.

Bu çalışmalar neticesinde, NAP tekniğinin SVM-GLDS yöntemine de uyarlanabileceği ve bu sayede konuşmacı doğrulama başarımının arttırılabileceği görülmüştür. Ayrıca temel olarak SVM-GLDS yönteminin, GMM-UBM yönteminden daha düşük başarım göstermesine rağmen, kanal dengeleme sonrasında GMM-UBM yöntemine EER metriği açısından oldukça yakın, MinDCF metriği açısından ise daha iyi başarım gösterdiği gözlemlenmiştir.

5. Tartışma ve Gelecek Çalışmalara İlişkin Öneriler

Bu tez çalışmasında, genel olarak metinden bağımsız konuşmacı tanıma problemi incelenmiş olup yapılan deneysel çalışmaların büyük bir kısmında konuşmacı doğrulama sistemi ele alınmıştır. Konuşmacı tanıma sisteminin temel üç aşaması olan *öznitelik çıkarımı*, *modelleme* ve *karar* aşamalarının her birine yönelik değişik öneriler getirilmiş olup bu öneriler neticesinde performansta iyileştirmeler sağlanmıştır. Her bir adımda geliştirilen önerileri özetlemek gerekir ise:

Öznitelik çıkarımı aşaması için

- Öncelikle literatürde ses işleme uygulamalarında yaygın bir şekilde kullanılan doğrusal öngörü analizi (LP) ile ortaya çıkan hata işareti $e(n)$ (Eşitlik 3.14 ve Şekil 3.6), insan ses üretim mekanizmasını modelleyen zamanla değişen süzgecin girişine (Şekil 3.5) uygulanan $u(n)$ işaretini temsil ettiğinden konuşmacı kimliği hakkında bilgi içermediği kabul edilmektedir. Genel görüş, konuşmacıyı asıl temsil eden bileşenin LP katsayıları tarafından modellenen zamanla değişen süzgeç olduğudur. Bu görüşün eksikliklerini ortaya koymak amacı ile öznitelik vektörlerinin orijinal işaretten değil de, hata işaretinden ($e(n)$) çıkarılması durumunda konuşmacı doğrulama performansı incelenmiştir (Hanilçi ve Ertaş, 2011b). Yapılan çalışmada gösterilmiştir ki, zamanla değişen süzgeç girişine uygulanan $u(n)$ giriş işareti en az orijinal işaret kadar konuşmacı kimliği hakkında bilgi taşımaktadır. Bu nedenle bu bilginin de göz önünde bulundurularak öznitelik çıkarma işleminin gerçekleştirilmesi gerektiği vurgulanmıştır.

- Çoğu konuşmacı tanıma uygulamalarında öznitelik olarak kullanılan MFCC vektörleri ses işaretinin spektrumundan elde edilmektedir. Bütün çalışmalarda ses işareti, zaman ortamından frekans ortamına hızlı Fourier dönüşümü (FFT) yöntemi ile dönüştürülmektedir. Ancak çok iyi bilinmektedir ki ses işleme uygulamalarında asıl önemli olan işaretin spektrumundaki detaylar yerine spektrumun zarfıdır. Çünkü arzu edilen çoğu bilgi (konuşmacının kimliği, konuşulan

metin ve dil v.s.) spektrumun detaylarında değilde harmonik yapıda bulunmaktadır. Çünkü formant ve pitch frekansları gibi önemli bilgiler bu harmonik bölgelerde bulunmaktadır. Bu nedenle spektrum hesaplanırken FFT yerine değişik yöntemler önerilmiştir (Hanilçi ve ark., 2012c,b,a). Spektrum hesaplama yönteminin konuşmacı doğrulama performansına önemli ölçüde etkisinin olduğu (özellikle toplamsal gürültü durumunda) yapılan çalışmalarda ortaya koyulmuştur. Bu nedenle literatürde yer alan ancak daha önce konuşma ve konuşmacı tanıma gibi uygulamalarda kullanılmamış olan spektrum hesaplama yöntemleri karşılaştırılmıştır ve kullanılan çoğu yöntemin FFT metodundan daha iyi performans gösterdiği belirtilmiştir. Örneğin Hanilçi ve ark. (2012c) 12 değişik spektrum hesaplama yöntemini, toplamsal gürültü durumunda, konuşmacı doğrulama problemi için karşılaştırmalı olarak incelemiş ve doğrusal öngörü (LP) ve ağırlıklandırılmış LP (WLP) yöntemlerinin oldukça başarılı performans verdiğini göstermiştir.

- Gelişen mobil iletişim teknolojilerinin günlük hayatta kullanımının yaygınlaştığı bilinen bir gerçektir. Ancak bu durum konuşmacı tanıma problemini zorlaştırmaktadır. Çünkü telefon hattı ses işaretlerinde bozucu etkiler meydana getirmektedir. Bu bozucu etkiler farklı zamanlarda kaydedilen ses işaretlerinin olması durumunda daha çok artmaktadır. Çünkü her aramada işaretin maruz kaldığı kanal etkisi bir öncekinden farklı olma ihtimali oldukça yüksektir. Dolayısı ile konuşmacı tanıma performansı büyük ölçüde düşmektedir. Bu nedenle kayıtlardaki iletim hattının olumsuz etkilerini azaltmak amacıyla öznitelik vektörleri üzerinden bu etkilerin dengelenmesi metodu önerilmiştir (Hanilçi ve Ertaş, 2012). Önerilen yöntem, çok fazla sayıda ve değişik zamanlarda gerçekleştirilen telefon aramaları sırasında kayıt altına alınmış ses işaretleri kullanarak genel bir iletim hattı profili oluşturmaktadır. Böylece bu model kullanılarak öznitelik vektörleri üzerinde iletim hattından kaynaklanan olumsuz bilgiler azaltılabilmektedir. Yapılan deneysel çalışmalarda önerilen yöntemin konuşmacı doğrulama performansını büyük ölçüde iyileştirdiği gösterilmiştir (Hanilçi ve Ertaş, 2012)

Modelleme aşaması için

- MAP uyarlamalı sınıflandırıcıların konuşmacı tanımada kullanılmaya başlaması bu alanda önemli gelişmelere temel oluşturmuştur. İlk olarak 2000 yılında Reynolds ve ark. (2000) tarafından GMM yönteminin MAP uyarlaması (GMM-UBM) konuşmacı doğrulama için önerilmiştir. GMM-UBM yöntemi ile çok büyük başarım elde edilmiş ve bu yöntem konuşmacı tanımada modelleme aşamasına farklı bir bakış açısı getirmiştir. Sonraki yıllarda bunu VQ algoritmasının MAP uyarlaması (VQ-UBM) (Hautamäki ve ark., 2008; Hautamäki, 2008) ve son olarak GMM-UBM yöntemi ile SVM sınıflandırıcılarının birlikte kullanıldığı GMM-süpervektör (GMM-SV) (Campbell ve ark., 2006b) izlemiştir. GMM-SV yönteminin GMM-UBM, VQ-UBM ve SVM-GLDS yöntemlerinden daha yüksek performans gösterdiği yapılan birçok çalışmada ortaya koyulmuştur (Vogt ve Sridharan, 2006; Kinnunen ve ark., 2009). Ancak tüm çalışmalarda, eğitim ve test aşamalarında en az 5 dakika uzunluğunda ses işaretleri kullanılmıştır. Bu nedenle bir sınıflandırıcı yönteminin her durumda başka bir yöntemden daha iyi performans gösterip göstermediğini belirtmek için değişik eğitim ve test veri süreleri ile bu dört sınıflandırma yönteminin karşılaştırmalı analizleri yapılmıştır (Hanilçi ve Ertaş, 2013b). Yapılan çalışmalar neticesinde, kısa süreli eğitim veya test işaretleri kullanıldığında GMM-SV yönteminin performansının diğer üç yöntemden daha düşük olduğu gösterilmiştir. Ayrıca yine aynı çalışmada gösterilen ilginç sonuçlardan biri de, bayan konuşmacıların erkek konuşmacılara nazaran daha zor tanındığıdır. Bunun nedeni de muhtemelen bayan konuşmacıların pitch frekanslarının erkek konuşmacılardan daha yüksek olmasıdır. Çünkü literatürde pitch frekansının yüksek olmasının konuşmacının kimliğinin tespit edilebilirliğini zorlaştığını belirten çalışmalar mevcuttur (Zilea ve ark., 2003).

- UBM yönteminin kullanıldığı bütün çalışmalarda, UBM modeli eğitilirken genellikle toplam olarak yüzlerce saate varan ses işaretleri kullanılmaktadır (Kinnunen ve ark., 2009, 2011; Kenny ve ark., 2007a). Bunun sebebi, çoğunlukla daha çok veri kullanmanın UBM yönteminde ifade edilen alternatif hipotez

için oluşturulan modelin temsil kabiliyetini artıracağına olan inançtır. Ancak bunun sonucunda UBM modelinin eğitimi çok uzun sürmektedir. Bu amaçla, UBM modelini oluşturmak için kullanılan veri miktarının konuşmacı doğrulama performansına etkisi incelenmiştir (Hanilçi ve Ertaş, 2013a). Deneysel çalışmalar neticesinde UBM modelinin eğitilmesi için kullanılan veri miktarının performansa etkisinin sanıldığı kadar büyük olmadığı ortaya koyulmuştur. Bu nedenle yüzlerce saatlik veri yerine bir kaç saatlik verinin de alternatif hipotezi yeterince iyi temsil ettiği gösterilmiştir.

Karar aşaması için

- Bir konuşmacı tanıma sisteminin eğitim ve test aşamasında kullanılan ses işaretlerinin farklı zamanlarda ve farklı ortamlarda kayıt yapılmış sesler olması durumunda, bunun karar aşamasında hesaplanan skorlar üzerinde bozucu etkiler (*nuisance*) meydana getirdiği daha önceki çalışmalarda tespit edilmiştir (Auckenthaler ve ark., 2000). Bu bozucu etkileri gidermek amacıyla yaygın olarak skor normalizasyonu yöntemleri kullanılmaktadır (Kinnunen ve Li, 2010; Apsingekar ve Leon, 2011; Ramos-Castro ve ark., 2007). Ancak literatürde yapılan çalışmalarda VQ-UBM yöntemi için skor normalizasyonu önerilmemiş ve kullanılmamıştır. Bu amaçla, VQ-UBM sınıflandırıcısı için test normalizasyonu (TNorm) işlemi önerilmiş ve elde edilen sonuçlar TNorm işleminin konuşmacı doğrulama performansında önemli artış sağladığını göstermiştir (Hanilçi ve Ertaş, 2011c).

Bu tezde elde edilen sonuçlar göstermiştir ki konuşmacı tanıma uygulamalarının performansı kullanılan verinin türü (telefon hattı veya mikrofon) ve miktarı, öznitelik vektörlerinin çıkarımı, sınıflandırma algoritma ve karar aşaması gibi birçok parametreden önemli ölçüde etkilenmektedir. Özellikle son yıllarda önerilen ve çok iyi performans gösterdiği belirtilen JFA (Kenny ve ark., 2007a) ve i-vector gibi yöntemler çok fazla veri miktarının kullanılmasını (en az 5 dakikalık eğitim ve test verileri) gerektirmektedir. Fakat bu durum gerçek zamanlı uygulamalarda oldukça önemli bir

problemdir. Bu nedenle kısa süreli ses işaretlerinin kullanılması durumunda yüksek performans elde etmek önemli bir ihtiyaçtır. Çünkü bu tezdeki deneysel sonuçlar göstermiştir ki uzun süreli veriler için yüksek başarım veren yöntemler kısa süreli veriler için bu performanslarını koruyamamaktadır.

Konuşmacı doğrulama için artık standart hale gelen NIST veritabanları telefon hattından veya mikrofon aracılığı ile kayıt altına alan seslerden oluşturulmaktadır. Bu nedenle çalışmalarda ortaya koyulan sonuçlar çoğunlukla bu iki tür kayıt ortamı için sunulmaktadır. Ancak bunların dışında konuşmacı doğrulama performansını direkt etkileyen, duygusal durum, hastalık, yaşlanma ve konuşma sırasındaki vurgulamalar gibi insan odaklı başka unsurlar da mevcuttur. Bu nedenle konuşmacı tanımada bu gibi insan odaklı faktörlere karşı dayanıklı öznitelikler ve sınıflandırıcı algoritmalara ihtiyaç duyulmaktadır.

Konuşmacı tanımanın gerçek zamanlı uygulamalarda kullanıldığı durumlarda sorgulanması gereken en önemli unsurlardan bir tanesi de, bir konuşmacının sesinin bir kayıt cihazı ile kaydedilmesi ve giriş veya erişim izni olmayan bir kişinin sistemi yanıltmak amacı ile bu kaydı kullanarak giriş yapmaya çalışması durumunda sistemin nasıl bir performans göstereceğidir. Bir diğer unsur da günümüzde mevcut gelişmiş yazılımlar kullanılarak bir kişinin sesi çok başarılı bir şekilde taklit edilebilmekte, değiştirilebilmekte ve hatta başka bir kişinin sesine dönüştürülebilmektedir. Bu gibi durumların ileriki zamanlarda yapılacak olan konuşmacı tanıma çalışmalarında ve özellikle konuşmacı tanıma probleminin adli uygulamaları açısından ele alınması oldukça önemli katkılar sağlayacaktır.

KAYNAKLAR

A. Solomonof, W.C., Boardman, I., 2005. Advances in channel compensation for SVM speaker recognition, *Proc. International Conference on Acoustics, Speech and Signal Processing, (ICASSP'05)*, 629-632.

A. Solomonof, W.C., Quillen, C., 2004. Channel compensation for SVM speaker recognition, *Proc. Odyssey The Speaker and Language Recognition Workshop*, 57-62.

Alam, M.J., Kinnunen, T., Kenny, P., Ouellet, P., O'Shaughnessy, D., 2013. Multitaper MFCC and PLP features for speaker verification using i-vectors, *Speech Communication,* 55(2): 237-251.

Alam, M.J., Ouellet, P., Kenny, P. and O'Shaughnessy, D., 2011. Comparative Evaluation of Feature Normalization Techniques for Speaker Verification, *Proc. 5th International Conference on Nonlinear Speech Processing, (NOLISP)*, 246-253.

Apsingekar, V.R., Leon, P.L.D., 2011. Speaker verification score normalization using speaker model clusters, *Speech Communication*, 53(1): 110-118.

Assaleh, K.T., Mammone, R.J., 1994a. New LP-derived features for speaker identification, *IEEE Transactions on Speech and Audio Processing*, 2(4):630-638.

Assaleh, K.T., Mammone, R.J., 1994b. Robust cepstral feature for speaker identification, *Proc. International Conference on Acoustics, Speech and Signal Processing, (ICASSP'94)*, 1:129-132.

Atal, B.S., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *The Journal of the Acoustical Society of America*, 55:1304-1312.

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*, 50:637-655.

Auckenthaler, R., Carey, M., Lloyd-Thomas, H., 2000. Score normalization for text-independent speaker verification systems, *Digital Signal Processing*, 10:42-54.

Besacier, L., Grassi, S., Dufaux, A., Ansorge, M., Pellandini, F., 2000. GSM speech coding and speaker recognition, *Proc. International Conference on Acoustics, Speech and Signal Processing, (ICASSP'00)*, 1085:1088.

Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau,

I., Meignier, S., Merlin, Té., Ortega-Garcia, J., Petrovska-Delacrétaz, D., Reynolds, D.A., 2004. A tutorial on text-independent speaker verification, *EURASIP Journal on Applied Signal Processing*, 2004(4): 430-451.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press Inc., New York, NY, USA, 504 pp.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag Inc., New York, USA, 738 pp.

Blanz, V., Schölkopf, B., Bülthoff, H.H., Burges, C., Vapnik, V., Vetter, T., 1996. Comparison of view-based object recognition algorithms using realistic 3D models, *Proc. International Conference on Artificial Neural Networks (ICANN-96)*, 251-256.

Bonastre, J.F., Bimbot, F., Boë, L.J., Campbell, J.P., Reynolds, D.A., Magrin-Chagnolleau, I., 2003. Person authentication by voice: A need for caution, *Proc. INTERSPEECH*, 33-36.

Bonifas, J.L., Rioja, I.H., Gonzalez, B.E., Saoudi, S., 1995. Text-dependent speaker verification using dynamic time warping and vector quantization of LSF, *Proc. Eurospeech.*

Brunelli, R., Falavigna, D., 1995. Person Identification Using Multiple Cues, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10): 955-966.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, 2: 121-167.

Campbell, W.M., 2002. Generalized linear discriminant sequence kernels for speaker recognition, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'02)*, 161-164.

Campbell, W., Sturim, D., Reynolds, D., 2006a. Support vector machines using GMM supervectors for speaker verification, *IEEE Signal Processing Letters*, 13(5): 308-311.

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

Campbell, W.M., Richardson, F.S., 2007. Discriminative keyword selection using support vector machines, *Proc. Advances in Neural Information Processing Systems, (NIPS).*

Castellano, P.J., 1996. Speaker recognition in reverberant enclosures, *Proc. IEEE*

International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'02), 117-120.

Davis, S.B., 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.

Deller, J., Hansen, J., Proakis, J., 2000. Discrete-time processing of speech signals. Wiley-Interscience, New York, 936 pp.

Dharanipragada, S., Yapanel, U.H., Rao, B.D., 2007. Robust feature extraction for continuous speech recognition using MVDR spectrum estimation method, *IEEE Transactions on Audio, Speech and Language Processing*, 15(1): 224-234.

Doddington, G., 1985. Speaker recognition - identifying people by their voices, *Proceedings of the IEEE*, 73(11): 1651-1664.

Doddington, G., Przybocki, M.A., Martin, A.F., Reynolds, D.A., 2000. The NIST speaker recognition evaluation-overview, methodology, systems, results, perspective, *Speech Communication*, 31(2-3): 225-254.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification. Wiley-Interscience, 680 pp.

Ekman, L.A., Kleijn, W.B., Murthi, M.N., 2008. Regularized linear prediction of speech, *IEEE Transactions on Audio, Speech and Language Processing*, 16(1): 65-73.

Fant, G., 1960. Acoustic Theory of Speech Production. Mouton.

Fauve, B.G.B., Evans, N.W.D., Pearson, N., Bonastre, J.F., Mason, J.S.D., 2007a. Influence of task duration in text-independent speaker verification, *Proc. Interspeech*, 794-797.

Fauve, B.G.B., Matrouf, D.; Scheffer, N., Bonastre, J.F., Mason, J.S.D., 2007b. State-of-the-art performance in text-independent speaker verification through open-dource software, *IEEE Transactions on Audio, Speech and Language Processing*, 15(7): 1960-1968.

Finan, R.A., Sapeluk, A.T., Damper, R.I., 1997. Impostor cohort selection for score normalisation in speaker verification, *Pattern Recognition Letters*, 18(9): 881-888.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press, 592 pp.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2): 254-272.

Furui, S., 1997. Recent advances in speaker recognition, *Pattern Recognition Letters*, 18(9): 859-872.

Ganchev, T.D., 2005. Speaker Recognition. *Ph.D. Thesis,* University of Petras.

García, R.D.L., Alberola-López, C., Aghzout, O., Ruiz-Alzola, J., 2003. Biometric identification systems, *Signal Processing*, 83(12): 2539-2557.

luc Gauvain, J., hui Lee, C., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, *IEEE Transactions on Speech and Audio Processing*, 2(2): 291-298.

Hanilçi, C., Ertaş, F., 2009. Principal Component Based Classification for Text-Independent Speaker Identification, *Proc. IEEE International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control (ICSCCW 2009)*, 1-4.

Hanilçi, C., Ertaş, F., 2011a. Comparison of the impact of some Minkowski metrics on VQ/GMM based speaker recognition, *Computers & Electrical Engineering*, 37(1): 41-56.

Hanilçi, C., Ertaş, F., 2011b. Impact of Voice Excitation Features on Speaker Verification , *Proc. International Conference on Electrical and Electronics Engineering (ELECO 2011)*, 137-140.

Hanilçi, C., Ertaş, F., 2011c. Score Normalization for VQ-UBM Based Text-Independent Speaker Verification, *Proc. International Conference on Electrical and Electronics Engineering (ELECO 2011)*, 132-136.

Hanilçi, C., Ertaş, F., 2012. Destek Vektör Makineleri ile Konuşmacı Tanımada Kanal Etkilerinin Dengelenmesi, *Proc. IEEE Conference on Signal Processing and Communication Applications (SIU 2012)*, 1-4.

Hanilçi, C., Ertaş, F., 2013a. Effects Of Background Data Duration On Speaker Verification Performance, *Uludağ University Journal of the Faculty of Engineering and Architecture*, 18(1): 111-119.

Hanilçi, C., Ertaş, F., 2013b. Investigation of The Effect of Data Duration and Speaker Gender on Text-Independent Speaker Recognition , *Computers & Electrical Engineering*, 39(3): 441-452.

Hanilçi, C., Kinnunen, T., Ertaş, F., Saeidi, R., Pohjalainen, J., Alku, P., 2012a. Regularized all-pole models for speaker verification under noisy environ-

ments, *IEEE Signal Processing Letters*, 19(3): 163-166.

Hanilçi, C., Kinnunen, T., Saeidi, R., Pohjalainen, J., Alku, P., Ertaş, F., 2012b. Regularization of all-pole models for speaker verification under additive noise, *Proc. Odyssey The Speaker and Language Recognition Workshop.*

Hanilçi, C., Kinnunen, T., Saeidi, R., Pohjalainen, J., Alku, P., Ertaş, F., Sandberg, J., Hansson-Sandsten, M., 2012c. Comparing spectrum estimators in speaker verification under additive noise degradation, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'12)*, 4769-4772.

Hasan, T., Hansen, J.H.L., 2011. A study on universal background model training in speaker verification, *IEEE Transactions on Audio, Speech and Language Processing*, 19(7): 1890-1899.

Hautamäki, V., 2008. Improving Pattern Recognition Methods for Speaker Recognition, *Ph.D. Thesis.* University of Joensuu.

Hautamäki, V., Kinnunen, T., Kärkkäinen, I., Saastamoinen, J., Tuononen, M., Fränti, P., 2008. Maximum a posteriori adaptation of the centroid model for speaker verification, *IEEE Signal Processing Letters*, 15: 162-165.

Hermansky, H., 1990. Perceptual linear predictive PLP analysis for speech, *Journal of the Acoustical Society of America*, 87: 1738-1752.

Higgins, A., Bahler, L., Porter, J., 1991. Speaker verification using randomized phrase prompting, *Digital Signal Processing*, 1(1991): 89-106.

Huang, X., Acero, A., Hon, H.W., 2001. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1008 pp.

Jain, A.K., Duin, R.P.W., Mao, J., 2000. Statistical pattern recognition: A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1): 4-37.

Jain, A.K., Zongker, D.E., 1997. Feature selection: Evaluation, application, and small sample performance, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2): 153-158.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y., 2002. An efficient K-Means clustering algorithm: Analysis and implementation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7): 881-892.

Kenny, P., Boulianne, G., Ouellet, P., Dumochel, P., 2007a. Joint factor

analysis versus eigenchannels in speaker recognition, *IEEE Transactions on Audio, Speech and Language Processing*, 15(4): 1435-1447.

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

Kinnunen, T., 2003. Spectral Features for Automatic Text-Independent Speaker Recognition. *Licentiate Thesis.* University of Joensuu.

Kinnunen, T., Hautamäki, V., Fränti, P., 2003. On the fusion of dissimilarity-based classifiers for speaker identification, *Proc. INTERSPEECH,* 2641-2644.

Kinnunen, T., Karpov, E., Fränti, P., 2006. Real time speaker identification and verification, *IEEE Transactions on Audio, Speech and Language Processing*, 14(1): 278-288.

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

Kinnunen, T., Saastamoinen, J., Hautamäki, V., Vinni, M., Fränti, P., 2009. Comparative evaluation of maximum a Posteriori vector quantization and gaussian mixture models in speaker verification, *Pattern Recognition Letters*, 30(4): 341-347.

Kinnunen, T., Saeidi, R., Sandberg, J., Hansson-Sandsten, M., 2010. What else is new than the Hamming window? Robust MFCCs for speaker recognition via multitapering, *Proc. INTERSPEECH*, 2734-3737.

Kinnunen, T., Saeidi, R., Sedlak, F., Lee, K.A., Sandberg, J., Hansson-Sandsten, M., Li., H., 2012. Low-variance multitaper MFCC features: A case study in robust speaker verification, *IEEE Transactions on Audio, Speech and Language Processing*, 20(7): 1990-2001.

Kinnunen, T.; Sidoroff, I.; Tuononen, M., Fränti, P., 2011. Comparison of clustering methods: A case study of text-independent speaker modeling, *Pattern Recognition Letters*, 32(1): 1604-1617.

Kittler, J., Nixon, M.S., 2003. Audio and Video Based Biometric Person Authentication, 4th International Conference, AVBPA 2003. Lecture Notes in Computer Science, Springer, Guildford, UK.

Kobatake, H., Matsunoo, Y., 1994. Degraded word recognition based on segmental signal-to-noise ratio weighting, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'94)*, 425-428.

Lee, K.A., You, C., Li, H., Kinnunen, T., Zhu, D., 2008. Characterizing speech utterances for speaker verification with sequence kernel SVM, *Proc. INTERSPEECH*, 1397-1400.

Linde, Y., Buzo, A., Gray, R., 1980. An algorithm for vector quantizer design, *IEEE Transactions on Communications*, 1: 84-95.

Ma, C.; Kamp, Y., Willems, L., 1993. Robust signal selection for linear prediction analysis of voiced speech, *Speech Communication*, 12(1): 69-81.

Magi, C., Pohjalainen, J., Bäckström, T., Alku, P., 2009. Stabilized weighted linear prediction, *Speech Communication*, 51(5): 401-411.

Mak, M.W., Hsiao, R., Mak, B., 2006. A comparison of various adaptation methods for speaker verification with limited enrollment data, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'06)*, 929-932.

Makhoul, J., 1975. Linear prediction: a tutorial review, *Proceedings of the IEEE*, 64(4): 561-580.

Mammone, R., Zhang, X., Ramachandran, R., 1996. Robust speaker recognition: a feature based approach, *IEEE Signal Processing Magazine*, 13(5): 50-62.

Mansour, D., Juang, B., 1989. The short-time modified coherence representation and noisy speech recognition, *IEEE Transactions on Acoustic and Signal Processing*, 37(6): 795-804.

Martin, A.F., Doddington, G.R., Kamm, T., Ordowski, M., Przybocki, M. A., 1997. The DET curve in assessment of detection task performance, *Proc. EUROSPEECH.*

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.

McLaren, M., Vogt, R., Baker, B., Sridharan, S., 2010. Experiments in SVM-based speaker verification using short utterances, *Proc. Odyssey The Speaker and Language Recognition Workshop*, 83-90.

Mitchell, T., 1997. Machine Learning. McGraw Hill, 432 pp.

Murthi, M.N., Kleijn, W.B., 2000. Regularized linear prediction all-pole models, *Proc. Speech Coding Workshop*, 96-98.

Murthi, M.N., Rao, B.D., 2000. All-pole modeling of speech based on the minimum variance distortionless response spectrum, *IEEE Transactions on Speech and*

Audio Processing, 8(3): 221-239.

Oppenheim, A.V., 1969. A speech analysis-synthesis system based on Homomorphic filtering *Journal of the Acoustical Society of America*, 45: 458-465.

Ortega-Garcia, J., Bigun, J., Reynolds, D., Gonzalez-Rodriguez, J., 2004. Authentication gets personal with biometrics *IEEE Signal Processing Magazine*, 21(2): 50-62.

Osuna, E., Freund, R., Girosi, F., 1997. Training support vector machines: An application to face detection, *Proc. Conference on Computer Vision and Pattern Recognition (CVPR '97)*, 130-136.

Pelecanos, J., Chaudhari, U., Ramaswamy, G., 2004. Compensation of utterance length for speaker verification, *Proc. Odyssey The Speaker and Language Recognition Workshop.*

Petrovska-Delacrétaz, D., Hannani, A.E., Chollet, G., 2007. Text-independent speaker verification: State of the art and challenges, *Lecture Notes in Computer Science*, 4391: 135-169.

Prabhakar, S., Pankanti, S., Jain, A.K., 2003. Biometric recognition: Security and privacy concerns, *IEEE Security & Privacy*, 1(2): 33-42.

Quatieri, T.F., 2002., Discrete Time Speech Signal Processing: Principles and Practice. Prentice Hall, 816 pp.

Quatieri, T.F., Reynolds, D.A., O'Leary, G.C., 2000. Estimation of handset nonlinearity with application to speaker recognition, *IEEE Transactions on Speech and Audio Processing*, 8(5): 567-584.

Rabiner, L., Schafer, R., 2010. Theory and Applications of Digital Speech Processing. Prentice Hall, Upper Saddle River, NJ, USA, 1056 pp.

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

Ramos-Castro, D., Fierrez-Aguilar, J., Gonzalez-Rodriguez, J., Ortega-Garcia, J., 2007. Speaker verification using speaker- and test-dependent fast score normalization, *Pattern Recognition Letters*, 28(1): 90-98.

Reynolds, D.A., 1992. A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification. *Ph.D. Thesis.* Georgia Institute of Technology.

Reynolds, D.A., 1995a. Large population speaker identification using clean and telephone speech, *IEEE Signal Processing Letters*, 2(3): 46-48.

Reynolds, D.A., 1995b. Speaker identification and verification using Gaussian mixture speaker models, *Speech Communication*, 17: 91-108.

Reynolds, D.A., 1997. Comparison of background normalization methods for text-independent speaker verification, *Proc. EUROSPEECH*, 963-966.

Reynolds, D.A., 2002. An overview of automatic speaker recognition technology, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'02)*, 4072-4075.

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.

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.

Robel, A., Villavicencio, F., Rodet, X., 2007. On cepstral and all-pole based spectral envelope modeling with unknown model order, *Pattern Recognition Letters*, 28(11): 1343-1350.

Rose, P., 2002. Forensic Speaker Identification. Taylor & Francis Forensic Science Series, 380 pp.

Rose, P., 2006. Technical forensic speaker recognition: Evaluation, types and testing of evidence, *Computer Speech & Language*, 20(2-3), 159-161.

Rosenberg, A., Sambur, M., 1975. New techniques for automatic speaker verification, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23(2): 169-176.

Rosenberg, A.E., DeLong, J., Lee, C.H., Juang, B.H., Soong, F.K., 1992. The use of cohort normalized scores for speaker verification, *Proc. International Conference on Spoken Language Processing (ICSLP)*, 599-602.

Saeidi, R., Pohjalainen, J., Kinnunen, T., Alku, P., 2010. Temporally weighted linear prediction features for tackling additive noise in speaker verification, *IEEE Signal Processing Letters*, 17(6): 599-602.

Sandberg, J., Hansson-Sandsten, M., Kinnunen, T., Saeidi, R., Flandrin, P., Brognat, P., 2010. Multitaper estimation of frequency-warped cepstra with application to speaker verification, *IEEE Signal Processing Letters*, 17(4): 343-346.

Schölkopf, B., Burges, C., Vapnik, V., 1995. Extracting support data for a given task, *Proc. International Conference on Knowledge Discovery and Data Mining (KDD-95)*, 252-257.

Shanker, A.P., Rajagopalan, A.N., 2007. Off-line signature verification using DTW, *Pattern Recognition Letters*, 28(12): 1407-1414.

Shimamura, T., Nguyen, N.D., 2010. Autocorrelation and double autocorrelation based spectral representations for a noisy word recognition systems, *Proc. INTERSPEECH*, 1712-1715.

Soong, F., Rosenberg, A., 1988. On the use of instantaneous and transitional spectral information in speaker recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(6): 871-879.

Soong, F., Rosenberg, A., Rabiner, L.R., Juang, B., 1985. A vector quantization approach to speaker recognition, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'85)*, 387-390.

Sturim, D.E., Campbell, W.M., Reynolds, D.A., 2007. Classification Methods for Speaker Recognition, *Speaker Classification*, 1: 278-297.

Toh, K.A., 2003. Fingerprint and speaker verification decisions fusion, *Proc. International Conference on Image Analysis and Processing (ICIAP)*, 626-631.

Toh, K.A., Yau, W.Y., 2005. Fingerprint and speaker verification decisions fusion using a functional link network, *IEEE Transactions on Systems, Man, and Cybernetics*, 35(3): 357-370.

Viikki, O., Laurila, K., 1998. Cepstral domain segmental feature vector normalization for noise robust speech recognition, *Speech Communication*, 25(1-3), 133-147.

Vogt, R., Baker, B., Sridharan, S., 2008a. Factor analysis subspace estimation for speaker verification with short utterances, *Proc. INTERSPEECH*, 853-856.

Vogt, R., Lustri, C., Sridharan, S., 2008b. Factor analysis modelling for speaker verification with short utterances, *Proc. Odyssey The Speaker and Language Recognition Workshop.*

Vogt, R., Pelecanos, J.W., Scheffer, N., Kajarekar, S.S., Sridharan, S., 2009. Within-session variability modelling for factor analysis speaker verification, *Proc. INTERSPEECH*, 1563-1566.

Vogt, R., Sridharan, S., 2006. Experiments in session variability modelling for speaker verification , *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'06)*, 897-900.

Wan, V., Renals, S., 2002. Evaluation of kernel methods for speaker verification and identification, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'02)*, 669-672.

Zilea, R.D., Navratil, J., Ramaswamy, G.N., 2003. Depitch and the role of fundamental frequency in speaker recognition, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'03)*, 81-84.

ÖZGEÇMİŞ

Adı Soyadı	: Cemal HANİLÇİ
Doğum Yeri ve Tarihi	: Malatya, 14.02.1983
Yabancı Dili	: İngilizce (80.0-ÜDS, Mart-2012)
Eğitim Durumu (Kurum ve Yıl)	
Lise	: Turgut Özal Anadolu Lisesi, 2001
Lisans	: Uludağ Üniversitesi, 2005
Yüksek Lisans	: Uludağ Üniversitesi, 2007
Çalıştığı Kurum ve Yıl	: Uludağ Üniversitesi, 2006-…
İletişim (e-posta)	: chanilci@uludag.edu.tr

Yayınları (SCI ve diğer) :

Hanilçi C., Ertaş F. 2009. Principal component based classification for text-independent speaker identification. 5th International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control (ICSCCW 2009).

Hanilçi C., Ertaş F. 2011. VQ-UBM based speaker verification through dimension reduction using local PCA. 19th European Signal Processing Conference (EUSIPCO 2011).

Hanilçi C., Ertaş F. 2011. Score normalization for VQ-UBM based text-independent speaker verification. 7th international conference on Electrical and Electronics Engineering, (ELECO 2011).

Hanilçi C., Ertaş F. 2011. Impact of voice excitation features on speaker verification. 7th international conference on Electrical and Electronics Engineering, (ELECO 2011).

Hanilçi C., Ertaş F. 2011. Comparison of the Impact of Some Minkowski Metrics on VQ/GMM Based Speaker Recognition. *International Journal of Computers and Electrical Engineering*, 37(1): 41-56.

Hanilçi C., Kinnunen T., Saeidi R., Pohjalainen J., Alku P., Ertaş F., Sandberg J., Sandsten M. 2012. Comparing spectrum estimators in speaker verification under additive noise degradation. IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP 2012).

Hanilçi C., Kinnunen T., Saeidi R., Pohjalainen J., Alku P. Ertaş F. 2012. Regularization of all-pole models for speaker verification under additive noise. Odyssey 2012: The Speaker and Language Recognition Workshop.

Hanilçi C., Ertaş F. 2012. Destek vektör makineleri ile konuşmacı doğrulamada kanal etkilerinin dengelenmesi. IEEE Sinyal işleme ve iletişim uygulamaları Kurultayı (SIU-2012).

Hanilçi C., Kinnunen T., Ertaş F., Saeidi R., Pohjalainen J., Alku P. 2012. Regularized All-Pole Models for Speaker Verification Under Noisy Environments. *IEEE Signal Processing Letters*, 19(3): 163-166.

Hanilçi C., Ertaş F. 2013. Investigation of The Effect of Data Duration And Speaker Gender on Text-Independent Speaker Recognition. *International Journal of Computers and Electrical Engineering*, 39: 441-452.