

**T.C.
ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

İSTATİKSEL MODELLEME İLE KONUŞMACI TANIMA

Ömer ESKİDERE

**DOKTORA TEZİ
ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI**

BURSA 2007

**T.C.
ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

İSTATİKSEL MODELLEME İLE KONUŞMACI TANIMA

Ömer ESKİDERE

**DOKTORA TEZİ
ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI**

Bu tez 05/11/2007 tarihinde aşağıdaki jüri tarafından oybirliği/oy çokluğu ile kabul edilmiştir.

Yrd. Doç. Dr. Figen ERTAŞ
Danışman

Prof. Dr. Atalay BARKANA

Prof. Dr. Erdoğan DİLAVEROĞLU

Prof. Dr. Osman KOPMAZ

Yrd. Doç. Dr. Rifat EDİZKAN

ÖZET

Kişilerin konuşmalarından kim olduklarının belirlenebilmesi önemi giderek artan bir ilgi alanı haline gelmiştir. Uzun yıllardır kullanılan parmak izi ve retina gibi kişiye has, kişinin kimliğini tanımlayıcı biyometrik özelliklere son yıllarda ses de eklenmiştir. Konuşma örneğinden kişinin kimliğinin belirlenebilmesinin günümüzde özellikle güvenlik, giriş ve/veya erişim kontrolü, telefon bankacılığı gibi önemli uygulama alanları mevcuttur. Bu tip gerçek zamanlı sistemlerde en büyük sorun seslerin kaydedildiği ortamın gürültülü olması ya da konuşmaların iletildiği kanalların (özellikle telefon hattı) bozucu etkisidir. Dolayısıyla, son yıllarda amaç, sistem başarımını olumsuz etkileyen bu tip etkileri en aza indirmek ve/veya bu şartlarda çalışacak dayanıklı sistemler geliştirmektir. Bu tezde Gauss Karışım Modeli (GKM) temeline dayanan, telefon hattı etkilerine karşı dayanıklı, bir konuşmacı tanıma sistemi oluşturulmuştur. Sistem eğitim ve test olmak üzere iki aşamalıdır. Kişinin sesinden kimliğini en iyi temsil eden öznelikler olarak da MFCC kullanılmış ve model parametreleri beklentinin maksimumlaştırılması algoritması ile kestirilmiştir. Test aşamasında aday konuşmacıya ait öznelikler, eğitim aşamasında oluşturulan her bir konuşmacı modele uygulanmakta ve maksimum olasılığı veren model konuşmacıyı belirlenmektedir.

Konuşmacı tanıma sistemi, temiz konuşma (TIMIT) ve telefon konuşması (NTIMIT) içeren iki veritabanı ile denenmiştir. Her iki veritabanı için, eğitim ve test aşamalarında, konuşmacı tanıma sistemine etkisi olan tüm parametreler incelenmiş ve parametrelerin optimum değerleri belirlenmiştir. Ayrıca formant frekansları, perde frekansı ve enerji gibi sesin bürünsel özellikleri tek başına ve MFCC öznelikleri ile birlikte kullanılarak konuşmacı tanıma performansı ölçülmüş, perde frekansının, telefon ortamında ortalama 8.34 puan tanıma artışı sağladığı görülmüştür. Özneliklerin oluşturulmasında kestrum katsayılarının kümelenecek ağırlıklandırılması ve konuşmacı frekans bandı parçalara ayrılıp, bu parçalara F-oranına bağlı olarak süzgeçler yerleştirilmesi önerilmiş olup, bu iki yöntem ile konuşmacı tanıma oranında 10 puana varan artış sağlanmıştır.

ANAHTAR KELİMELER: Konuşmacı tanıma, Gauss Karışım Modeli, MFCC, Öznelik vektörleri, TIMIT/NTIMIT veritabanları

ABSTRACT

Identifying speakers from their voices has been an area of interest that received ever increasing attention. In recent years, voice has also been added to the individual-specific biometric features representing the identity of individuals such as commonly employed finger print and retina, and the identification of speakers from their voice samples has recently found place particularly in security, access control, and telephone banking applications. The problem in such real time systems is the noise and/or distortion induced by the environments where the speech samples are taken and the media (particularly telephone lines) through which the speech samples are transmitted, respectively. In recent years, efforts have been made to minimize the impact of such factors that severely damage the identification performance, or to develop systems that are robust to such disturbances.

In this thesis, a speaker identification system based on Gaussian Mixture Model (GMM) has been developed that is robust to telephone line distortion, employing mel frequency cepstrum coefficients (MFCC) as speaker specific features, which are known to best represent speakers' identity, along with the Expectation Maximization algorithm for the estimation of speaker model parameters. The system consists of two stages, namely, training and testing. In the training session, a model is produced for each speaker to represent their identity, and the input speaker is identified in the test session by deciding on the model that provides the highest probability. The system has been tested on both clean speech (TIMIT) and telephone speech (NTIMIT) databases. From feature extraction to model training and testing, various parameters that affect the system performance have been investigated and optimized using both speech databases. Identification performance of the system has been determined for cases where prosodic features of speech such as formant frequency, pitch frequency, and energy are employed on their own and in combination with MFCC. It has been found that pitch frequency provides 8.34 point increase in identification performance on telephone speech when used in combination with MFCC. Weighted clustering of cepstral coefficients and adaptive filtering have been introduced in extracting discriminatory features. Up to 10 point increase in identification performance has been obtained by each technique.

Keywords: Speaker Identification, Gaussian Mixture Models, MFCC, Feature vectors, TIMIT/ NTIMIT databases

İÇİNDEKİLER

ÖZET.....	i
ABSTRACT.....	ii
SİMGELER DİZİNİ.....	vii
KISALTMALAR DİZİNİ.....	ix
ŞEKİLLER DİZİNİ.....	x
ÇİZELGELER DİZİNİ.....	xvi
1. GİRİŞ.....	1
1.1 Tezin Katkısı.....	3
1.2 Tez İçeriği.....	3
2. KAYNAK ARAŞTIRMASI.....	6
2.1 Konuşmacı Tanımda Kullanılan Algısal İpuçları.....	6
2.2 Konuşmacı Tanıma Süreci.....	7
2.3 Öznitelik Vektörleri.....	9
2.3.1 İdeal öznitelikler.....	10
2.3.2 Mel frekansı kestrum katsayıları.....	11
2.3.3 Doğrusal öngörü katsayıları.....	12
2.3.4 Doğrusal algı öngörü yöntemi.....	14
2.3.5 Göreceli spektra yöntemi.....	15
2.3.6 Formant frekansları.....	15
2.3.7 Temel frekans.....	16
2.3.8 Yoğunluk.....	16
2.3.9 Öznitelik seçimi.....	17
2.4 Sınıflandırma Teknikleri	18
2.4.1 Şablon temelli yaklaşım.....	18
2.4.1.1 Dinamik zaman eğirme.....	19
2.4.1.2 Vektör nicemeleme.....	19
2.4.2 İstatistiksel Yaklaşım.....	20
2.4.2.1 Gauss karışım modeli.....	20
2.4.2.2 Saklı markov model.....	21

2.4.3 Yapay sinir ağıları.....	23
2.4.4 Destek vektör makinesi.....	25
3. MATERYAL ve YÖNTEM.....	27
3.1 Gauss Karışım Modeli.....	28
3.1.1 Model tanımı.....	28
3.1.2 Akustik sınıf modelleme.....	29
3.1.3 Maksimum benzerlik sınıflandırıcı.....	31
3.1.4 Maksimum benzerlik kestirimi.....	33
3.1.4.1 Beklentinin maksimumlaştırılması.....	35
3.1.5 Tezde kullanılan veritabanları.....	36
3.2 Konuşmacı Tanıma için GKM'nin Deneysel Değerlendirilmesi.....	37
3.2.1 Model eğitimi aşamasında yapılan düzenlemeler.....	39
3.2.1.1 Beklenti Maksimumlaştırılması algoritmasının özyineleme sayısı	39
3.2.1.2 Model başlangıç değerleri	40
3.2.1.3 Ortak değişinti matrisi seçimi.....	41
3.2.1.4 Değişinti sınırlandırılması	42
3.2.2 Karışım bileşen sayısı ve eğitilen veri miktarı konuşmacı tanımaya etkisi.	43
3.2.2.1 İdeal karışım bileşen sayısının bulunması.....	44
3.2.2.2 Eğitim ve test süresi değişimi.....	47
3.2.2.3 Konuşmacı sayısı değişimi.....	53
3.3 Öznitelik Vektörü Çıkartma ve Parametre Kestirimi	55
3.3.1 Mel ölçek kepstrum katsayıları.....	56
3.3.1.1 Çerçeveleme.....	58
3.3.1.2 Pencereleme.....	60
3.3.1.3 Hızlı fourier dönüşümü.....	63
3.3.1.4 Ön vurgulama.....	66
3.3.1.5 Mel ölçekte dizilmiş dizileri.....	71
3.3.1.6 Logaritma alma.....	78
3.3.1.7 Ayrık kosinüs dönüşümü.....	85
3.3.1.8 Sıfırıncı kepstrum katsayısı.....	88
3.3.2 Kepstrum katsayı değişimlerinin konuşmacı tanımaya etkisi.....	89
3.3.3 Süzgeç dizileri frekans ölçekleri.....	94

3.3.3.1 Mel ölçek	94
3.3.3.2 Bark ölçek.....	94
3.3.3.3 ERB ölçek.....	95
3.3.3.4 Doğrusal ölçek	95
3.3.4 İnsan işitsel sistemi benzetiminin konuşmacı tanımaya uygulanması.....	102
3.3.4.1 İnsan kulağının yapısı ve işitme.....	102
3.3.4.2 Basilar membran ve gamaton süzgeçler.....	104
3.3.4.3 Gamaton süzgeçlerin konuşmacı tanımaya uygulanması.....	107
3.4 Telefon İletiminin Konuşmacı Tanıma Üzerine Etkilerinin Azaltılması.....	114
3.4.1 Spektral değişim kompanzasyonu.....	114
3.4.1.1 Ortalama normalizasyonu.....	114
3.4.1.2 Kepstrum fark katsayıları.....	115
3.4.1.3 Frekans eğirme.....	115
3.4.2 Öznitelik vektörlerinin kümelenerek ağırlıklandırılması.....	117
3.4.2.1 Spektral analiz.....	119
3.4.2.2 Kümeleme	119
3.4.2.3 Süzgeç dizileri.....	120
3.4.2.4 F-oranı analizi.....	120
3.4.2.5 Öznitelik vektörlerinin kümelenerek ağırlıklandırma.....	
deneysel sonuçları.....	122
3.4.3 Kepstrum Katsayıları ile F-oranı Analizi.....	124
3.4.3.1 Kepstrum katsayıları ile F-oranı analizinin deneysel sonuçları...126	
3.4.4 Öznitelik vektörleri oluşturulmasında F-oranına bağlı olarak süzgeç.....	
uygulanması.....	127
3.5 Bürünsel Özniteliklerin (Prosodic Features) Konuşmacı Tanımaya Etkisi.....	130
3.5.1 Temel frekans.....	131
3.5.1.1 Perde frekansı izlemenin zorlukları.....	133
3.5.1.2 Perde frekansı izleme aşamaları.....	135
3.5.1.3 Temel frekans deneysel değerlendirilmesi.....	138
3.5.2 Formant Frekansları.....	149
3.5.2.1 Formant frekansının etkisinin deneysel değerlendirilmesi.....	150
3.5.3 Enerji.....	151

3.5.3.1 Teager enerji operatörü.....	152
3.5.3.2 Enerji etkisinin deneysel değeriendirilmesi.....	155
3.5.4 Formant Genlik ve frekans modülasyonu parametreleri.....	155
3.5.4.1 Formant GM-FM öznitelik vektörü oluşturma yöntemi.....	159
3.5.4.2 Formant GM-FM parametrelerinin deneysel değeriendirilmesi..	164
3.5.5 Doğrusal olmayan öznitelik parametrelerinin eldesinde özilinti.....	
katsayılarının kullanılması ve polinom benzetimi.....	165
4. ARAŞTIRMA SONUÇLARI ve TARTIŞMA.....	172
4.1 Araştırma Sonuçları.....	172
4.2 Tartışma.....	176
4.3 Öneriler.....	180
KAYNAKLAR	182
EK 1 TERİMLER SÖZLÜĞÜ.....	191
EK 2 GKM PARAMETRE KESTİRİMİ.....	192
Teşekkür.....	197
Özgeçmiş.....	198

SİMGELER DİZİNİ

A_l	-	Süzgeçlerin bant genişliğine bağlı normalizasyon katsayısı	
$a(n)$	-	Anlık genlik kestirimi	
α	-	F-oranına bağlı olarak süzgeç oluşturulmasında kullanılan ağırlık katsayısı	
$\vec{\sigma}^2_i$	-	i. karışım bileşeninin değışinti vektörü	
B_i	-	Sınıflar arası değışinti	
$b_i(\vec{x})$	-	Bileşen yoğunlukları	
p_i	-	Karışım ağırlıkları	
C_l	-	Mel süzgecin merkez frekansı	(Hz)
D	-	Boyut	
E	-	Enerji	
F_1	-	Birinci formant frekansı	(Hz)
F_2	-	İkinci formant frekansı	(Hz)
F_3	-	Üçüncü formant frekansı	(Hz)
f	-	Frekans	(Hz)
f_0	-	Temel Frekans	(Hz)
f_c	-	Gabor bant geçiren süzgecin merkez frekansı	(Hz)
f_s	-	Örnekleme frekansı	(Hz)
$G(f)$	-	Gırtlak kaynak karakteristiği	
$H(w)$	-	Gabor bant geçiren süzgecin frekans cevabı	
\vec{h}	-	Telefon hattının süzgeç etkisi	
mfb	-	Mel süzgeç dizisi	
$\vec{\mu}_i$	-	Ortalama vektör	
k	-	Kepstrum katsayıları	
l	-	Mel süzgeç indisi	
L	-	Merkez frekanslar arası uzaklık	(Hz)
L_l	-	Mel süzgecin alt kesim frekansı	(Hz)
M	-	Karışım bileşen sayısı	
N	-	FFT örnek sayısı	
U_l	-	Mel süzgecin üst kesim frekansı	(Hz)

- $P_r(t)$ - Konuşma işareti
 S - Konuşma frekans bandının ayrıldığı parça sayısı
 $s(t)$ - Boğaz kaynak işareti
 $R_s(k)$ - Otokorelasyon katsayısı
 $R(f)$ - Yayılım karakteristiği
 $T(f)$ - Ses yolu transfer fonksiyonu
 Σ_i - Ortak deęişinti matrisi
 $x[n]$ - Örnekleilmiş konuşma işareti
 \bar{x} - Öznitelik vektörleri
 Δk - Birinci dereceden dinamik katsayılar
 $\Delta \bar{z}_t$ - Konuşmacının t . çerçevesinin fark katsayıları
 $\Delta \Delta k$ - İkinci derece dinamik katsayılar
 $\Psi [\cdot]$ - Teager enerji operatörü
 Ω - Anlık frekans
 W_i - i . özelliğın sınıf içi deęişintisi
 λ - Karışım ağırlık, ortalama ve ortak deęişintilerini ifade eden model
 \bar{z} - Gözlenen kepstrum vektörü

KISALTMALAR DİZİNİ

- AEA - Ayrık enerji ayırma
AKD - Ayrık kosinüs dönüşümü
BM - Beklentinin maksimumlaştırılması
DÖK - Doğrusal öngörü katsayıları
DVM - Destek vektör makinesi
DZE - Dinamik zaman eğirme
ERB - Eşdeğer dikdörtgenel bant genişliği
FFT - Hızlı fourier transformu
FIR - Sınırlı uyartı cevaplı süzgeç
FS - Süzgeç sayısı
FM - Frekans modülasyonu
MFCC - Mel frekansı keprum katsayıları
GKM - Gauss karışım modeli
GM - Genlik modülasyonu
MLP - Çok katmanlı algılayıcı
NTIMIT - Nynetex tarafından TIMIT veritabanının telefonda kaydedilmiş hali
Pdf - Olasılık yoğunluk fonksiyonu
SMM - Saklı markov model
TEO - Teager enerji operatörü
TIMIT - Texas Instruments ve Massachusetts teknoloji enstitüsü tarafından hazırlanan veritabanı
VN - Vektör nicemleme
YSA - Yapay sinir ağı

ŞEKİLLER DİZİNİ

<u>Şekil</u>	<u>sayfa</u>
2.1 Konuşmanın taşıdığı bilgi seviyeleri ve ipuçları	7
2.2 Bir konuşmacı tanıma sisteminde hedef.....	8
2.3 Otomatik konuşmacı tanıma sistemi.....	9
2.4 MFCC işlemi blok diyagramı.....	12
2.5 VN temelli bir konuşmacı tanıma sisteminin blok diyagramı.....	20
2.6 GKM ile konuşmacı tanıma sistemi.....	21
2.7 Gözlem vektörlerinin her biri bir durum tarafından üretilen soldan sağa üçlü bir SMM.....	22
2.8 Bir Yapay Nöron.....	23
2.9 Genel YSA Modeli.....	24
2.10 (a) İki sınıflı veriyi ayıran bir altdüzlem, (b) en iyi altdüzlem.....	26
2.11 Düzgün dağılımlı olmayan örneklerin çekirdek fonksiyonları düzenlenmesi.....	26
3.1 M bileşenli Gauss karışım yoğunluğunun gösterimi.....	29
3.2 Gizli akustik sınıflardan elde edilen gözlem vektörleri.....	30
3.3 GKM'nin modelleme kabiliyeti örneği.....	31
3.4 Konuşmacı tanıma için kullanılan maksimum benzerlik sınıflandırıcı blok diyagramı.....	33
3.5 GKM konuşmacı modeli için BM algoritması adımları.....	35
3.6 GKM eğitim için BM algoritmasının benzerlik fonksiyonunun (a) karışım sayısı 32 (b) karışım sayısı 16 için değişimi.....	39
3.7 TIMIT veritabanı için karışım bileşen sayısına bağlı olarak konuşmacı tanıma oranları (%).....	45
3.8 NTIMIT veritabanı için karışım bileşen sayısına bağlı olarak konuşmacı tanıma oranları (%).....	47
3.9 TIMIT için elde edilen üç farklı eğitim süresine bağlı olarak 1 saniye uzunluğunda test ifadesi için konuşmacı tanıma oranları (%).....	49
3.10 TIMIT için elde edilen üç farklı eğitim süresine bağlı olarak 3 saniye uzunluğunda test ifadesi için konuşmacı tanıma oranları (%).....	50
3.11 TIMIT için elde edilen üç farklı eğitim süresine bağlı olarak 6 saniye uzunluğunda test ifadesi için konuşmacı tanıma oranları.....	50

<u>Şekil</u>	<u>sayfa</u>
3.12 Eğitim sürelerinin değişimine bağlı olarak (a) test süresi 1 saniye (b) test süresi 3 saniye (c) test süresi 6 saniye için konuşmacı tanıma oranları.....	52
3.13 Konuşmacı sayısına bağlı olarak test kümesi için konuşmacı tanıma oranları.....	54
3.14 MFCC çıkarılma işleminin blok diyagramı.....	57
3.15 Bir konuşma ve ortalaması alınmış hali.....	57
3.16 Yirmi beş ms uzunluğunda konuşma parçası.....	58
3.17 Pencereleme fonksiyonları.....	61
3.18 Konuşma parçası ve Hamming pencereden geçirilmiş hali.....	62
3.19 Pencerelenen konuşma parçasının $ FFT ^2$ ve $ FFT $ alınmış hali.....	64
3.20 Yirmi ms uzunluğunda (a) ünsüz bir konuşma parçası (b) bu konuşma parçasının $ FFT $ (c) $ FFT ^2$ alınmış hali.....	65
3.21 Yirmi ms uzunluğunda (a) ünlü bir konuşma parçası (b) bu konuşma parçasının $ FFT $ (c) $ FFT ^2$ alınmış hali.....	66
3.22 Ön vurgulama süzgecinin değişik α değerleri için frekans cevabı.....	67
3.23 Bir cümleye çerçevelemeden önce ön vurgulama uygulanması.....	68
3.24 (a) Yirmi ms uzunluğunda bir konuşma parçası (b) bu konuşma parçasının ön vurgulamadan önce genlik spektrumu (c) ön vurgulama uygulandıktan sonra genlik spektrumu.....	68
3.25 (a) Yirmi ms uzunluğunda bir konuşma parçası (b) bu konuşma parçasının $ FFT ^2$ spektrumu (c) spektrumu alınmış işaretin ön vurgulanmış hali.....	69
3.26 Konuşma parçasının güç spektrumu alındıktan sonra, ön vurgulama yapılmadan önce ve sonraki halleri.....	69
3.27 Bir cümlenin birinci dereceden süzgeçten ($\alpha = 0.95$) (a) geçirilmeden (b) geçirildikten sonra zaman-frekans değişimi.....	71
3.28 Mel ölçek.....	71
3.29 Mel ölçekte dizilmiş üçgen süzgeç dizileri (Davis ve Mermelstein 1980).....	73
3.30 Mel ölçekte dizilmiş üçgen süzgeç dizileri (Slaney 1998).....	75
3.31 İşaretin süzgeç dizisinden geçirildikten sonraki durumu.....	77
3.32 Konuşma parçasına denklem 3.38 uygulanması durumunda elde edilen kepsrum.. katsayıları.....	79

<u>Şekil</u>	<u>sayfa</u>
3.33 Logaritmik ölçekte kök ve logaritma fonksiyonlarının değişimi.....	81
3.34 İşaretin süzgeç çıkışı ve logaritmali hali.....	82
3.35 Temiz (kırmızı) ve gürültülü (mavi) konuşmalar için logaritması alınmış Mel süzgeç dizilerinin enerjileri.....	84
3.36 Temiz (kırmızı) ve gürültülü (mavi) konuşmalar için logaritmasının karesi alınmış Mel süzgeç dizilerinin enerjileri.....	84
3.37 c_0 (kırmızı) ve c_1 (mavi) fonksiyonları.....	86
3.38 c_2 (kırmızı) ve c_3 (mavi) fonksiyonları.....	87
3.39 Ayrık kosinüs dönüşümü.....	87
3.40 (a) c_0 çıkartılmadan elde edilen kepstrum katsayıları (b) c_0 çıkartıldıktan sonra elde edilen kepstrum katsayı eğrileri.....	88
3.41 10-13. pencereler arası kepstrum katsayıları değişimi.....	90
3.42 Çerçeve sayısına bağlı olarak kepstrum katsayıları değişimi.....	90
3.43 Frekans ölçekleri karşılaştırması.....	95
3.44 Değişik frekans ölçeklerinin kepstrum katsayıları değişimlerine bağlı olarak karşılaştırılması (0-8000 Hz).....	99
3.45 Değişik frekans ölçeklerinin kepstrum katsayı değişimlerine bağlı olarak karşılaştırılması (0-4000 Hz).....	99
3.46 NTIMIT veritabanı test dizini (168 konuşmacı) için süzgeçlerin yerleştirildiği frekans bandı F-oranı.....	101
3.47 Kulağın yapısı.....	103
3.48 (a) Basilar membranın yapısı ve dalgaların hareket yönleri (b) basilar membranın duyarlı olduğu frekans bölgeleri (c) basilar membran boyunca ses dalgası hareketi	104
3.49 Salyangoz yapı boyunca basilar membranın, duyarlı olduğu frekans bölgeleri ve bant geçiren süzgeç özelliği.....	105
3.50 Gamaton fonksiyonunun dürtü cevabı.....	106
3.51 Yirmi adet gamaton süzgeç dizisi.....	107
3.52 Gamaton süzgeç dizisi genlik spektrumu (dB).....	108
3.53 Otuz iki adet gamaton süzgecin genliği bant genişliğine göre düzenlenmiş genlik spektrumu.....	109

<u>Şekil</u>	<u>sayfa</u>
3.54 Sadece ERB bant genişliği içerisindeki süzgeç değerlerine genlik düzenlemesi uygulanırsa elde edilen süzgeç dizisi.....	110
3.55 Gamaton süzgeçlerin sınırlandırılmış ve sınırlandırılmamış halleri.....	110
3.56 ERB ölçek ve bant genişliğinde 32 adet üçgen süzgeç dizileri yerleştirilmesi.....	111
3.57 ERB ölçek ve bant genişliğinde dikdörtgen süzgeç dizileri yerleştirilmesi.....	112
3.58 Frekans eğirme örneği.....	116
3.59 Kümeleme ve ağırlıklandırma sonucu elde edilen öznelik vektörleri.....	118
3.60 Eşit aralıklarla dizilmiş süzgeç dizileri.....	120
3.61 Dört küme için süzgeç çıkışlarına göre F-oranı değeri.....	122
3.62 Kepstrum katsayılarına bağlı olarak F-oranı değerleri (küme sayısı 4).....	125
3.63 F-oranı değerinin kepstrum katsayılarına bağlı olarak değişimi (NTIMIT).....	128
3.64 Alfa ve parça genişliği değerlerine bağlı olarak konuşmacı tanıma oranları (TIMIT veritabanı).....	129
3.65 Alfa ve parça genişliği değerlerine bağlı olarak konuşmacı tanıma oranları (NTIMIT veritabanı).....	130
3.66 Ses tellerinin darbe üretici gibi davranması.....	132
3.67 NTIMIT veritabanından alınmış “She” sözcüğü.....	132
3.68 “She” sözcüğünün zaman-frekans-yoğunluk değişimi.....	132
3.69 Değişik sağlık koşullarında temel frekans değişimi.....	134
3.70 Merkez kırpması ile işaretin kırılması.....	135
3.71 Özilinti fonksiyonu ile elde edilen işaret.....	136
3.72 f_0 alt ve üst sınırları içerisindeki tepe değerinin bulunması.....	137
3.73 Bir konuşmacının f_0 değerleri.....	138
3.74 Dört farklı konuşmacının aynı cümleyi söylemesi ile elde edilen perde frekanslarının histogramları.....	139
3.75 (a) NTIMIT veritabanında bir konuşma işareti (b) $t=0.1$ için işaretten sessiz kısımların atılmış hali (c) $t=0.01$ için işaretten sessiz kısımların atılmış hali (d) $t=0.0025$ için işaretten sessiz kısımların atılmış hali.....	140
3.76 Özilinti Yöntemi.....	147
3.77 Kepstrum Yöntemi.....	148

<u>Şekil</u>	<u>sayfa</u>
3.78 Özilinti yöntemi ile elde edilen perde frekansın medyan süzgeç (a) öncesi (b) sonrası dağılımı (c) Kepstrum yöntemi ile elde edilen perde frekansın medyan süzgeç öncesi (d) sonrası dağılımı.....	148
3.79 Denklem 3.73'deki transfer fonksiyonlarının gösterimi.....	149
3.80 Bir konuşma örneği ve enerjisi alınmış hali.....	151
3.81 Ses yolunda girdap-hava akış etkileşimi.....	152
3.82 Bir sinüs işaretinin TEO ile genliğinin izlenmesi.....	154
3.83 Bir sinüsoidal işaretin frekans izlenmesi.....	154
3.84 (a) sönümlü sinüs işareti (b) sönümlü sinüs işaretinin genliği (c) sönümlü sinüs işaretinin frekansı (d) AEA-2 algoritması ile kestirilen sönümlü sinüs işaretinin genliği (e) AEA-2 algoritması ile kestirilen sönümlü sinüs işaretinin frekansı.....	158
3.85 Sönümlü sinüs işaretinden AEA-2 algoritması frekans ve genlik kestirimi sonucu oluşan hata oranları.....	158
3.86 Formant GM-FM parametrelerinin ölçümü için oluşturulan öznitelik vektörü oluşturma yönteminin blok diyagramı.....	159
3.87 NTIMIT veritabanından bir cümle.....	159
3.88 Bir cümle için formant frekansları (bir çerçeve 25 msn).....	160
3.89 Bir boyutlu gabor süzgeçlerin zaman ve frekans cevabı.....	161
3.90 (a) 25 msn uzunluğunda bir konuşma işareti parçası (b) gabor bant geçiren süzgeçten geçirilmiş konuşma işareti (c) Teager ayrıklaştırma ($\sqrt{\psi[x(n)]}$)	162
3.91 (a) 25 msn uzunluğunda bir konuşma işareti parçasının gabor bant geçiren süzgeçten geçirilmiş hali (b) AEA-2 kullanılarak genlik zarfının kestirimi (c) AEA-2 kullanılarak anlık frekans kestirimi.....	162
3.92 Bir cümle için formant GM-FM işlemi sonucu elde edilen kepstrum katsayıları.....	163
3.93 Ayrıklaştırma teager enerji olması durumunda tanıma oranları.....	164
3.94 Doğrusal olmayan konuşma özniteliklerinin analizi blok diyagramı.....	166
3.95 NTIMIT veritabanında bir konuşmacıya ait 25 msn lik çerçevede (a) orijinal konuşma (b) süzgeçlenmiş konuşma (c) TEO (d) AEA-1 ile (b) genlik kestirimi (e) AEA-1 algoritması ile frekans kestirimi	167

<u>Şekil</u>	<u>sayfa</u>
3.96 AEA-1 genlik kestirimi uygulanmış 25 ms'nlik işaretin 21 nokta medyan süzgeç ve ortalama bileşenler atılmış hali.....	168
3.97 AEA-1 genlik kestirimi uygulanan işaretin özilintisi ve özilintisinin..... genlik zarfı alınmış şekil.....	169
3.98 Özilinti genlik zarfı işaret ve bu işarete ait (a) $N= 5$ için (b) $N= 19$ için polinomlara ait eğriler.....	169
4.1 Frekans ölçeklerinin karşılaştırılması (NTIMIT veritabanı).....	175
4.2 MFCC ve f_0 birlikte kullanıldığında konuşmacı tanıma oranları.....	179
4.3 Formant GM-FM parametrelerinin tanıma oranlarının karşılaştırılması.....	180

ÇİZELGELER DİZİNİ

<u>Çizelge</u>	<u>sayfa</u>
3.1 Test setinin tamamındaki konuşmacıların bölgelere göre dağılımı.....	37
3.2 TIMIT ve NTIMIT veritabanlarının karakteristikleri.....	37
3.3 Farklı model başlangıç metotları için konuşmacı tanıma oranları (%).....	41
3.4 Köşegen ve tam değişinti matrisleri için konuşmacı tanıma oranları (%).....	42
3.5 Farklı minimum değişinti değerleri için konuşmacı tanıma oranları (%).....	43
3.6 Karışım bileşen sayısına bağlı olarak konuşmacı tanıma oranı (%).....	44
3.7 Karışım bileşen sayısına bağlı olarak konuşmacı tanıma oranları (%).....	46
3.8 GKM'in 9 saniye eğitilmesi ile elde edilen konuşmacı tanıma oranları (%).....	47
3.9 GKM'in 15 saniye eğitilmesi ile elde edilen konuşmacı tanıma oranları (%).....	48
3.10 GKM'in 24 saniye eğitilmesi ile elde edilen konuşmacı tanıma oranları (%).....	48
3.11 Değişik eğitim süreleri için elde edilen konuşmacı tanıma oranları (%).....	51
3.12 TIMIT veritabanının tamamı için doğru konuşmacı tanıma oranı (%).....	53
3.13 Test süresi kullanılış biçimlerine göre tanıma oranı değişimi (%).....	55
3.14 Çerçeveleme sürelerinin konuşmacı tanımaya etkisi (%).....	59
3.15 Pencereleme fonksiyonlarına bağlı olarak konuşmacı tanıma oranları (%).....	62
3.16 FFT kuvvetlerinin konuşmacı tanıma üzerine etkisi (%).....	64
3.17 Ön vurgulamanın konuşmacı tanıma üzerine etkisi (%).....	70
3.18 İki farklı Mel ölçeğin merkez frekansları ve bant genişlikleri.....	76
3.19 Çizelge 3.18'de tanımlanan Mel ölçeklerin konuşmacı tanıma oranı (%).....	77
3.20 Süzgeç çıkışlarının logaritması ve kuvvetleri alınmasının tanımaya etkisi (%)..	82
3.21 Süzgeç çıkışlarının logaritması alınmasının tanımaya etkisi (%).....	85
3.22 AKD'nin konuşmacı tanımaya etkisi (%).....	88
3.23 Sıfırıncı kepstrum katsayısının konuşmacı tanımaya etkisi (%).....	89
3.24 Kepstrum katsayıları ve test süresi değişimlerinin konuşmacı tanımaya etkisi..	92
3.25 Kepstrum katsayıları değişimlerinin konuşmacı tanımaya etkisi.....	93
3.26 Kepstrum katsayıları değişimlerinin test süresine göre konuşmacı tanımaya etkisi (%).....	93
3.27 Değişik süzgeç ölçekleri için konuşmacı tanıma oranları (%).....	96
3.28 Karışım sayısına bağlı olarak değişik frekans ölçekleri için tanıma oranları ...	97

<u>Çizelge</u>	<u>sayfa</u>
3.29 Süzgeç aralığı 0-4 KHz için değişik süzgeç ölçekleri için konuşmacı tanıma oranları (%).....	97
3.30 Değişik frekans ölçekleri için konuşmacı tanıma oranları (%).....	98
3.31 Değişik frekans ölçekleri için konuşmacı tanıma oranları (%).....	100
3.32 Gamaton süzgeç sayısına bağlı konuşmacı tanıma oranları (%).....	109
3.33 Genliği bant genişliğine göre düzenlenmiş gamaton süzgeçler için konuşmacı tanıma oranları (%).....	109
3.34 Sadece ERB bant genişliği içerisindeki süzgeç değerleri alınırsa elde edilen konuşmacı tanıma oranları (%).....	111
3.35 Üçgen süzgeç dizileri ile konuşmacı tanıma oranları.....	111
3.36 Dikdörtgen süzgeç dizileri ile konuşmacı tanıma oranları.....	112
3.37 Spektral değişim kompanzasyonu yöntemlerinin tanımaya etkisi.....	117
3.38 Küme sayısına bağlı olarak tanıma oranları (%).....	123
3.39 Küme sayısına bağlı olarak tanıma oranları (%).....	123
3.40 Kümeleme ile konuşmacı tanıma oranları (%).....	124
3.41 Küme sayıları değişimlerine bağlı olarak tanıma oranları (%).....	126
3.42 Kümeleme ile konuşmacı tanıma oranları (%).....	127
3.43 Eşik parametresi t 'ye bağlı olarak konuşmacı tanıma oranları (%).....	141
3.44 Mel frekansı keprstrum katsayılarına f_0 eklenmesi ile elde edilen tanıma oranları (%).....	142
3.45 Çerçeveleme sürelerine bağlı olarak temel frekansın tanımaya etkisi (%).....	142
3.46 Ön vurgulamaya bağlı olarak tanıma oranları.....	143
3.47 Karışım sayısının konuşmacı tanımaya etkisi.....	143
3.48 Süzgeç dizileri frekans ölçeğine bağlı olarak tanıma oranları.....	144
3.49 Bant sınırlamalı durumda tanıma oranları.....	145
3.50 Örnekleme hızının düşürülmesinin tanımaya etkisi	145
3.51 Konuşmadan sessiz kısımların atılması	146
3.52 Özilinti ve keprstrum yöntemlerinin konuşmacı tanımaya etkisi (%).....	147
3.53 Formant frekansları için tanıma oranları (%).....	150
3.54 Enerjinin konuşmacı tanımaya etkisi.....	155
3.55 AEA-1 ve AEA-2 genlik kestirimi ile tanıma oranları (%).....	165

<u>Çizelge</u>	<u>sayfa</u>
3.56 I. ve II. yöntemler öznitelik vektörleri olarak kullanılması durumunda	
konuşmacı tanıma oranları.....	170
3.57 TIMIT ve NTIMIT veritabanları için çerçeve başına formant	
karşılaştırmaları.....	171
4.1 TIMIT ve NTIMIT veritabanı için ideal öznitelik parametreleri.....	173
4.2 TIMIT ve NTIMIT veritabanı için ideal eğitim parametreleri.....	173
4.3 TIMIT veritabanında literatür karşılaştırması.....	177
4.4 NTIMIT veritabanında literatür karşılaştırması.....	177
4.5 Küme sayısına bağlı olarak tanıma oranları.....	178
4.6 Küme sayısına bağlı olarak tanıma oranları.....	178

1. GİRİŞ

Konuşma işareti pek çok seviye bilgi taşır. Konuşma işareti, kelime veya konuşulan mesaj hakkında bilgi taşımakla birlikte ayrıca konuşanın kimliği hakkında bilgi taşır. Bilgisayarların kullanıldığı sesli iletişimde, konuşma tanıma, söylenen sözcüğün anlamı ile ilgilenilirken konuşmacı tanıma ise sözcüğü söyleyen kişinin kimliği ile ilgilenilir. Son zamanlarda ses araştırmacıları bu konu üzerinde yoğunlaşmaktadır.

Otomatik konuşmacı tanıma son on yıl içerisinde büyük ilerlemeler göstermiştir. Birkaç yıl öncesine kadar söylenen kelimeler arasında boşluk verilerek tanıma işlemi yapılabilirken, günümüzde sürekli konuşulan bir konuşma için bile konuşmacı tanımayı sağlayan sistemler ticari anlamda kullanılmaktadır (Matsui ve Furui 1995). Bilgisayar teknolojisindeki gelişmeler sonucu, günümüzde gerçek zamanlı konuşma ve konuşmacı tanıma gibi karmaşık uygulamalar gerçekleştirilmektedir.

Konuşmacı tanıma sistemi, genellikle gizli bir kaynağı (bina, fabrika, laboratuvar, gizli bilgilerin saklandığı bir oda vb.) koruyup giriş kontrolü yapmakta kullanılır. Giriş kontrolünde bir anahtar, bir şifre veya bir kart kullanılabilir. Bunların hepsi kolayca çalınabilir, kaybolabilir, taklit edilebilir. Bununla birlikte kişiye özel olup başka kimsede olmayan kişiye has biometrik özellikler vardır. Biometrik kişinin, kişisel özelliklerinin otomatik olarak ölçülmesi tekniği olup kişinin tanınması amacı ile kişinin karakteristik özelliklerinin bir veritabanı ile karşılaştırılıp kişi belirlenmeye çalışılmasıdır. Biometrik fiziksel özellik olarak parmak izi, el geometrisi ve retina yapısı; kişisel özellik olarak ise ses yapısı ve el yazısını kullanır (Woodward 1997). Bahsedilen pek çok biometrik teknikten ses karakteristikleri, ses tanıma ve konuşmacı kimliklendirme için kullanılabilir.

Önceden duyduğumuz konuşmaların sonraki karşılaşmalarda kime ait olduklarını rahatlıkla hatırlayabiliriz. Telefonda konuşurken, telefon hattında gürültü olsa bile pek çok zaman karşıdaki kişiyi tanıyabiliriz. Özel olarak konuşan kişinin kimliğini bulmak için kullanılan diğer ipuçları hatalı veya çok belirsiz olduğu durumlarda ses ile konuşan kişiyi tanıma oldukça çok kullanılan bir yöntemdir.

Genel olarak konuşmacı tanıma, konuşmacı grubunun üyeleri (kimliklerinin doğru bilindiği) ve yanıtıcılar (kimliklerinin bilinmediği) olarak ikiye ayrılmasıyla 'Kapalı-küme' ve Açık-küme' olmak üzere iki alt bölüme ayrılabilir. Kapalı-küme

durumunda kimliđi saptanmış konuşmacının referans konuşmacılardan biri olduđu bilinir ve test verisi (hece, kelime veya cümle) üzerinde en iyi sonucu veren konuşmacı tanımlanır. Doğal olarak, geniş bir toplulukta bu iş daha zordur. Açık-küme durumunda kimliđi saptanan kişi bu topluluktan biri olmayabilir ve eđer bir konuşmacı test verisi üzerinde yeterince iyi sonuç verirse, o zaman konuşmacı tanınmış kabul edilir. Bu durumda yeterince iyi sonuç elde edilip elde edilmediđinin belirlenmesinde bir eşik deđer tanımlanması gerekmektedir. Açık-küme durumunda bir ek karar alternatifi istenir ve bu 'böyle bir kişi yok' kararıdır (Gish ve Schmidt 1994).

Dinleyiciler, konuşulan metinler birbirinden farklı olsa bile kişilerin seslerinden konuşmacıları tanıyabilir. Konuşmacı tanıma metine bağımlılık yönünden iki alt gruba ayrılır. Bunlar metine bağımlı ve metinden bağımsız konuşmacı tanımadır (Reynolds ve Rose 1995, Kinnunen 2003). Metine bağılı bir uygulamada tanıyıcı sistem, konuşulan metin hakkında bir ön bilgiye sahiptir. Bu alan ile ilgili örnekler kullanıcı özel veya ifade çıkarımı şeklindedir. Metine bağılı sistemler tanınacak kişinin tanınmayı istediđi ve bu nedenle gönüllü olduđu giriş (kapı) kontrol uygulamaları gibi uygulamalarda kullanılır. Ön bilgi ve metin sınırlandırılması sistemin tanıma başarımını önemli ölçüde arttırmaktadır (Reynolds 2002).

Metinden bağımsız bir uygulamada sistem, konuşulan metin hakkında bir ön bilgiye sahip deđildir. Metinden bağımsız tanıma daha zor fakat bir konuşmacının doğrulanması gibi uygulamalarda daha esnektir. Metinden bağımsız konuşmacı tanıma sistemleri, konuşmacının aynı metni konuşacađının garanti olmadığı uygulamalarda örneđin adli gözaltı uygulamalarında kullanılır. Konuşma ve konuşmacı tanıma sistemlerinde, konuşma doğruluđunun arttırılması ile metine bağımlı ve bağımsız uygulamalar arasındaki fark azalacađı düşünülebilir (Naik 1990). Bu tezde metinden bağımsız kapalı-küme konuşmacı tanıma problemine odaklanılmaktadır. Otomatik konuşmacı tanıma sistemi, Matlab programı kullanılarak hazırlanmıştır.

Bu tezin amacı, konuşmacı tanımada son on yılda en çok kullanılan Gauss karışım modelini, yine son yıllarda akademik çalışmalarda sıklıkla kullanılan TIMIT ve NTIMIT veritabanlarına uygulayarak tanıma başarımını etkileyen tüm parametreler için en iyi deđerlerini elde etmektir. Tezin diđer amacı mikrofondan ve telefon hattı üzerinden kaydedilen iki farklı veritabanı kullanarak, konuşmacı tanıma sisteminde, özellikle telefon hattı kullanıldıđında tanıma oranında iyileştirme sağlayabilmektir.

1.1 Tezin Katkısı

Bu çalışmada birinci olarak, TIMIT ve NTIMIT veritabanları kullanılarak öznitelik vektörü oluşturma aşamalarının her biri için parametre değişiminin konuşmacı tanımayaya etkisi incelenmiş ve tanımayı arttırıcı en iyi parametre değerleri bulunmuştur. Bu veritabanları ile yapılan diğer konuşmacı tanıma çalışmaları için, bilhassa telefon hattı üzerinden kayıt yapılmış olan NTIMIT veritabanı için, öznitelik vektörü elde edilirken diğer çalışmalara kaynak olabilecek en iyi parametre değerleri belirlenmiştir. Bu sayede bu modeli kullanan diğer araştırmacılara en ideal parametreleri bulmak için yol gösterecektir.

İkinci olarak, öznitelik vektörleri kümelenecek ağırlıklandırılmakta ve bu şekilde konuşmacı tanıma oranı arttırılmaktadır. Bölüm 3.4.3'de görüleceği üzere ağırlıklandırma işleminin süzgeç bankaları çıkışı yerine kepsrum katsayıları ile yapılması önerilmekte bu şekilde küme sayılarına bağlı olarak, TIMIT veritabanı için % 5, NTIMIT veritabanı için % 9'a varan başarımların sağlanmaktadır.

Üçüncü olarak, öznitelik vektörü elde edilirken, etkin frekans bölgeleri F-oranı analizi ile bulunarak, etkin frekans bölgelerine daha fazla süzgeç yerleştirilmiştir. Bu sayede ayırt ediciliğin fazla olduğu frekans aralıkları etkinleştirilmiştir. Bu öznitelik elde etme yöntemi bölüm 3.4.4'de tanımlanmaktadır. Bu yöntem, her iki veritabanı içinde tanıma oranı klasik öznitelik elde etme yöntemine göre % 10'a varan tanıma artışı sağlanmaktadır.

Son olarak, bürünsel özelliklerin, öznitelik vektörlerine eklenerek tanıma başarımlarını ölçülmektedir. Bu özelliklerden enerji ve formant frekansları tanıma oranını azaltmasına karşın, temel frekans, NTIMIT veritabanı için % 8.34 başarımların sağlanmaktadır. Konuşmadan sessiz kısımların atılması ile eşik değerine bağlı olarak % 4.46 tanıma oranında artış sağlanmaktadır.

1.2 Tez içeriği

Bu tezin bölümleri şu şekildedir: Bölüm 2'de ilk olarak, konuşmacı tanımda kullanılan algısal ipuçları tanımlanmakta ve bir otomatik konuşmacı tanıma sisteminin yapısı tanıtılmaktadır. İkinci olarak, otomatik konuşmacı tanımlarında kullanılan öznitelik vektörü üretme yöntemleri incelenerek bu tezde kullanılan öznitelikler kısaca

tanıtılmaktadır. Son olarak, konuşmacı tanıma sistemlerinde kullanılan temel konuşmacı modelleme teknikleri verilmektedir.

Bölüm 3.1’de Gauss karışım modeli tanıtılmaktadır. Gauss karışım yoğunlukları kullanılarak konuşmacı modellenmesi tanımlanmaktadır. Maksimum benzerlik sınıflandırıcısı kullanılarak tanıma kararının nasıl yapıldığı açıklanmakta, daha sonra maksimum benzerlik parametre kestiriminin denklemleri ve beklenti maksimumlaştırma eğitim algoritması tanımlanmaktadır. Son olarak bu tezde kullanılan veritabanları tanıtılmaktadır.

Bölüm 3.2’de metinden bağımsız konuşmacı tanıma için veritabanındaki kişilere ait cümlelerin, Gauss karışım modeli ile eğitimi esnasında oluşan bazı sorunlar belirtilmektedir. Bu sorunlara karşı çözümler tanımlanmaktadır. Büyük konuşmacı topluluğu ve telefon hattından geçirilen konuşmalar için GKM’deki karışım bileşen sayısının, kullanılan eğitim ve test verilerinin sürelerinin, konuşmacı tanıma sistemine etkisi incelenmektedir. Bu parametrelerin her birinin değiştirilmesi ile yapılan deneylere ait sonuçlar verilmekte ve modelin en iyi değerleri elde edilmektedir. Son olarak iki veritabanı için konuşmacı sayısının konuşmacı tanımaya etkisi incelenmektedir.

Bölüm 3.3’de Mel frekansı keppstrum katsayıları, öznitelik vektörü üretim yönteminin aşamaları teker teker tanımlanmakta, her aşamanın konuşma işaretine etkisi, elde edilen sonuçlar ile grafiksel olarak gösterilmektedir. Her aşama için en ideal parametrelerin bulunması için GKM ile konuşmacı tanıma deneyleri yapılmaktadır. Doğrusal, Mel, Bark, ERB frekans ölçekleri, TIMIT ve NTIMIT veritabanları için karşılaştırılmaktadır. Ayrıca insan kulağı yapısını en iyi modelleyen gamaton süzgeçlerin nasıl elde edildiği belirtilmekte ve bu süzgeçlerin öznitelik vektör üretiminde kullanılması ile elde edilen konuşmacı tanıma oranları verilmektedir.

Bölüm 3.4’de telefon hattından dolayı oluşan konuşma bozulmaları, zemin gürültüsü ve telefon ahizesinin doğrusal olmayan etkisinin konuşmacı tanıma oranını hangi oranda azalttığı incelenmektedir. Bu istenmeyen etkilerin giderilmesi için spektral değişim kompanzasyonu ve öznitelik vektörlerinin kümelenerek ağırlandırılması uygulanmaktadır. Son olarak öznitelik vektörleri elde edilmesinde Mel ölçekte dizilmiş süzgeçler yerine F-oranına bağlı olarak hazırlanan süzgeçler önerilmekte ve bu yöntemler ile yapılan deneyler ve elde edilen sonuçlar verilmektedir.

Bölüm 3.5’de ise bürünsel özellikler olarak verilen, enerji, f_0 ve formant frekanslarının konuşmacı tanıma üzerine etkisi incelenmektedir. Formant frekansları ve enerji ayırma algoritmalarının birlikte kullanılması ile elde edilen formant GM-FM öznitelik vektörlerinin telefon hattı üzerine etkisi incelenmektedir. Ayrıca formant GM-FM parametrelerinin özilinti zarfının polinom benzetimi yapılarak elde edilen polinom katsayıları, öznitelik olarak kullanılmaktadır. Bu yöntemlerin öznitelik vektörü olarak kullanılması ile elde edilen konuşmacı tanıma oranları verilmektedir.

Bölüm 4’de bu tezde elde edilen araştırma sonuçları özetlenmekte ve elde edilen sonuçlar daha önce yapılan çalışmalar ile karşılaştırmalı olarak verilmektedir. Ayrıca bu çalışma temelli geleceğe dönük öneriler yer almaktadır.

Ekler kısmında tezde kullanılan terimlere ait sözlük ve bölüm 3.1’de belirtilen GKM modeline ait, model parametrelerinin çıkartılması verilmektedir.

2. KAYNAK ARAŞTIRMASI

Bu bölüm, otomatik konuşmacı tanıma sistemi hakkında temel bilgi vermeyi amaçlamaktadır. İlk olarak konuşmacının kimliğinin belirlenmesinde kullanılan algısal ipuçları verilmekte ve bu ipuçlarının konuşma işareti ile ilişkisi belirtilmektedir. Daha sonra genel bir konuşmacı tanıma sistemi tanımlanmaktadır. Buna bağlı olarak konuşmacı tanıma sürecinde kullanılan öznitelik üretim yöntemleri belirtilmekte daha sonra bu yöntemler karşılaştırılmaktadır. Son olarak konuşmacı tanıma sistemlerinde kullanılan, sınıflandırma ve konuşmacı modelleme yöntemleri verilmekte ve bu yöntemlerin güçlü ve zayıf olduğu yönler belirtilmektedir.

2.1 Konuşmacı Tanımda Kullanılan Algısal İpuçları

Konuşma işareti, kelime veya konuşulan mesaj hakkında bilgi taşımakla birlikte ayrıca konuşanın kimliği hakkında bilgi taşır. Konuşma işareti konuşmacının psikolojik ve duygusal durumu, sağlığı ile sesin kaydedildiği ortam hakkında da bilgi içerir. Böylece, farklı konuşmacıların konuşma sinyalleri arasında çok fazla değişiklik vardır ve daha da önemlisi aynı konuşmacının değişik zamanlarda kaydedilmiş konuşma sinyalleri arasında farklılıklar bulunmasıdır.

Bir konuşmacının kimliğinin belirlenmesinde, insan kulağının algı mekanizmasının anlaşılması önemli yer tutmaktadır. İnsanların sadece sesleri kullanarak birbirini tanımalarının makinelerle nasıl uygulanacağı sorusu gündeme gelmektedir. İnsanlar konuşanın kimliğini belirlemek için sözle ilgisi olmayan pek çok ipucu kullanmaktadır. Bu ipuçları pek iyi anlaşılacakla birlikte kabaca anlam ile ilişkili olanlar “yüksek seviye”, konuşmanın akustik yanı ile ilişkili olanları “düşük seviye” ipuçları olarak gruplandırılmaktadır. Yüksek seviye ipuçları, kelime kullanımı, söyleyişteki kişisel özellik ve konuşma karakteristiği ile ilişkili olmayan konuşmacıya özel karakteristik özellikler içerir. Bu ipuçları kişinin konuşma söyleyiş biçimi dolayısıyla değişik yaşam biçimlerine bağlı olarak farklılıklar gösterir. Bu tip ipuçları öğrenilmiş davranış olarak ortaya çıkar (Reynolds 1992). Düşük seviye ipuçları kişinin sesiyle direkt ilişkili olup yumuşak, sert, kaba, açık, yavaş veya hızlı gibi nitelikler içerir. Düşük seviye ipuçları konuşmacının anatomik yapısı ile doğrudan bağlantılıdır. Konuşmacılar arasındaki anatomik farklılıklar, konuşmacıların ses sistemlerinde bulunan bileşenlerinin boyutları ve şekillerinin farklı olmasından kaynaklanır. Mesela

kısa ses yolu, yüksek formant frekansı oluştururken, ses tellerinin boyutlarındaki değişimler ortalama ses yüksekliğindeki farklılıklar ile ilişkilidir. Bundan dolayı, doğuştan gelen bu özellikler bir konuşmacı için oldukça sabit olmakla beraber bazı sağlık durumlarından etkilenebilirler (burun boşluğunda değişikliğe neden olan nezle gibi). Şekil 2.1’de konuşmanın taşıdığı bilgi seviyeleri ve ipuçları görülmektedir (Peskin ve ark. 2003).



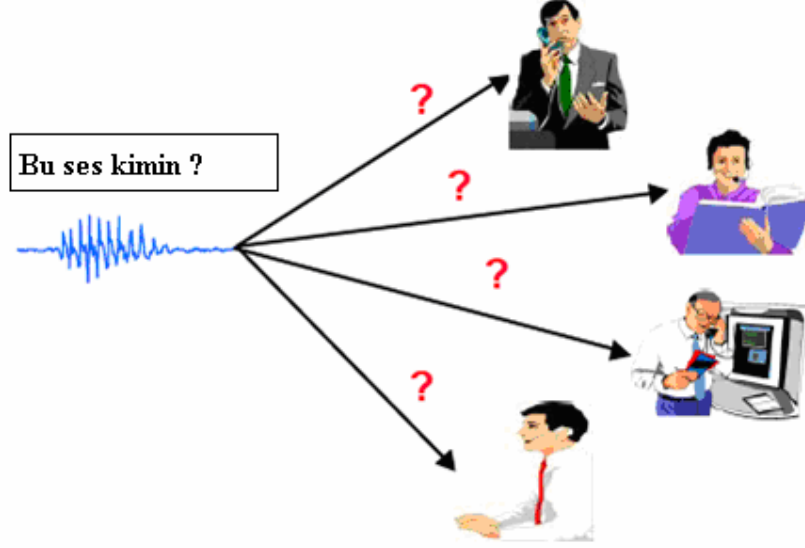
Şekil 2.1 Konuşmanın taşıdığı bilgi seviyeleri ve ipuçları

Bu ipuçlarının tamamı konuşmacının kimliğini belirlemeye yarayacak algısal bilgiler taşır. Ancak düşük seviye ipuçları konuşmacı tanıma sistemlerinde daha fazla uygulanmaktadır. Bunun iki sebebi vardır. Birincisi, yüksek seviye ipuçlarının konuşma işaretinden çıkartılması oldukça zordur. Bu durumda belirli kelimeler için güvenli konuşma tanıyıcı veya kelime çıkartıcı gerekir. Oysaki düşük seviye ipuçları, konuşma işaretinden akustik ölçümler ile çıkartılabilir. İkinci olarak düşük seviye ipuçları belirli kelimelere bağımlı değildir ve metinden bağımsız sistemler için daha kullanışlı olmaktadır (Reynolds ve ark. 2004).

2.2 Konuşmacı Tanıma Süreci

Konuşmacıyı tanıma sistemin görevi, o olduğunu iddia eden kişiyi bir grup içinden konuşanın kimliğini belirlemedir. Konuşanı tanıma esnasında konuştuğu metinden kişinin kimliği hakkında bilgi sahibi olunabilir ve “Kim konuşuyor ?” sorusuna otomatik olarak cevap verilir. Eğer gerekli ise “Ne söyledi ?” sorusuna da

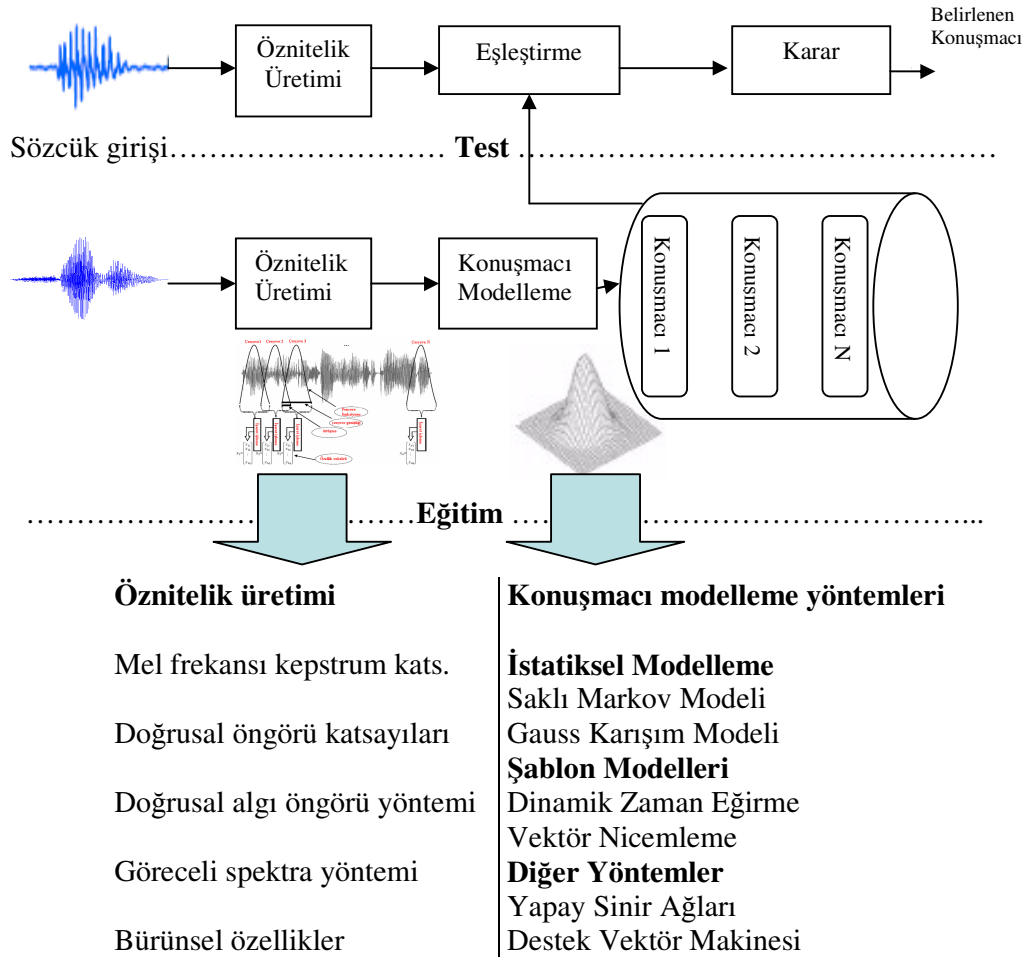
farklı bilgiler kullanarak cevap verilebilir. Gerçekte konuşmacıyı tanıma ile konuşulanı tanıma arasında yakın bir ilişki olup, aralarında büyük bir paralellik vardır (Ertaş 2000). Şekil 2.2’de bir konuşmacı tanıma sisteminin hedefi belirtilmektedir.



Şekil 2.2 Bir konuşmacı tanıma sisteminde hedef

Bütün konuşmacı tanıma sistemleri iki birbirinden bağımsız aşamaya hizmet etmek zorundadır. Bunlardan ilki eğitim aşaması iken ikincisi ise test aşamasıdır. Eğitim aşamasında tüm kullanıcılar, bir referans modeli oluşturmak için ses örnekleri verir, ikinci aşamada ise giriş sinyali referans modelleri ile karşılaştırılarak saptama yapılır. Bir konuşmacı tanıma sisteminin genel yapısı şekil 2.3’de verilmektedir.

Ses girişini alma işlemi farklı teknolojiler ve uygulamalar gerektirir. Konuşma girdi cihazı genellikle bir mikrofon veya bir telefondur. Konuşma çoğunlukla yüksek bir frekansta örneklenir (örneğin bir mikrofonda 16 kHz veya telefonda 8 kHz olarak). Bu, bize zaman üzerindeki bir dizi genlik değerini verir. Ses analogdur ve işlenebilmesi için öncelikle analog formdan sayısal forma dönüştürülmesi gerekir. Bunu yerine getirmek için geliştirilmiş olan farklı kodlama metotları vardır (Aydın 2005). Her bir konuşmacıya ait kodlanan ses girdileri belirli bir düzende bilgisayara alınıp saklanır, bu şekilde bir veritabanı oluşturulmuş olur. Saklanan bu ses girdileri şekil 2.3’de görüleceği üzere konuşmacı tanıma sistemine sözcük girişi olarak verilir. Bloklar halinde belirtilen bu sisteminin temel bileşenleri aşağıda incelenmektedir.



Şekil 2.3 Otomatik konuşmacı tanıma sistemi

2.3 Öznitelik Vektörleri

Konuşmacı tanımanın ilk aşamasında kullanılan tekniklerin amacı sınıflandırma için öznitelik vektörleri çıkarmaktır. Amaç çok fazla olan konuşma verilerinin, konuşmacıyı tanımlayabilecek vektörlere indirgenmesi ve bir sonraki aşama olan sınıflandırma için kullanışlı veriler üretmektir. Öznitelik vektörü üretimi için kullanılan yöntemler genel olarak iki grupta incelenir. Bunlar parametrik ve parametrik olmayan yaklaşımlardır.

Parametrik yaklaşım, konuşmanın üretiliş mekanizmasının tahmin edilmesine yönelik bir modeldir. Bir konuşma üretim sistemi öngörülür. Bu yöntemde giriş (kesin olarak bilinmez fakat tahmin edilir), ve çıkış (konuşmanın kendisi) arasında bir konuşma üretim fonksiyonu oluşturulur. Bu fonksiyonun parametreleri konuşmacı

tanıma sisteminde öznitelik vektörü olarak kullanılır. Parametrik teknikler böyle bir modelin varlığını kabul edip modeli tahmin etme temeline dayanır. Doğrusal tahmin bu model oluşturma yöntemlerinin bir alt kümesidir.

Parametrik olmayan yöntemler, konuşma işareti üzerinde pencereler halinde ilerleyerek işaret üzerinde bazı dönüşümlerin uygulanması temeline dayanır. Yöntemin başarısı, kullanılan pencerenin türünün ve uzunluğunun üzerinde yorum yapılabilecek nitelikte olmasına bağlıdır. Bu niteliğe sahip bir pencere türü ve uzunluğu için ayarlamalar yapmak bu tekniğin ilk aşamasını oluşturur. Pencere üzerinde daha sonra bir boyut dönüştürme işlemi yapılır. Örneğin fourier dönüşümü ile genlik-zaman boyutu, frekans-zaman boyutuna dönüştürülür. Daha sonra bu dönüşüm sonucu elde edilen veriler bazı iyileştirme yöntemleriyle sınıflandırma aşamasına hazır hale getirilir (Furui 1989).

2.3.1 İdeal öznitelikler

İdeal öznitelikler, konuşmacıyı tanımaya yardımcı olacak özelliklere sahip olmalıdır. Bu özellikler şunlardır.

- Kolay ölçülebilmeli
- Tabii olarak meydana gelmeli ve konuşmada sıkça oluşmalı
- Zamanla değişmemeli
- Konuşmacının sağlık değişimlerinden etkilenmemeli
- İletim şartlarından oluşan gürültüden etkilenmemeli
- Taklide karşı dayanıklı olmalıdır.

Pratikte, istenen bu özniteliklere ait özelliklerin eş zamanlı olarak elde edilmesi çok zordur (Reynolds 1992). Uygulamaya bağlı olarak bu öznitelik standartlarında kısmi değişimler oluşabilir.

İdealde istenen öznitelik özelliklerinden ilk ikisi göz önüne alındığında, eğer bir öznitelik, konuşmacı ayırımında yüksek oranda etkili olmasına rağmen az sıklıkta oluşuyor veya güvenli olarak çıkartılması zor ise bu öznitelik bir konuşmacı tanıma sisteminde az kullanılır veya hiç kullanılamaz. Sonraki üç madde özniteliklerin gürbüzlüğü ile ilgilidir. Pratikte, konuşma işaretinden elde edilen öznitelikler çıkartılırken pek çok değişikliğe uğrayacaktır. Bu değişiklikler anatomik sebeplerle

oluşabilir. Soğuk algınlığı ile veya zamanla bir kişinin sesinde değişimler olabilir. Bu değişimler, çoğunlukla mikrofon veya telefon ortamından ses kaydı esnasındaki akustik ortama (gürültülü veya sessiz) bağlı olmaktadır. Bir kişinin kaydedilen ses örneklerinden çıkartılan öznitelikleri ile sistem her zaman o kişiyi doğru tanıyabilmelidir. En güvenli konuşmacı tanıma başarımı elde etmek için konuşma işaretinden değişken şartlara karşı en tutarlı öznitelikler çıkartılmalıdır. İdeal öznitelik özelliklerindeki son madde güvenlik sistemleri için gereklidir. Eğer bir konuşmacı tanıma sistemi giriş kontrolünde kullanılıyorsa (örn. banka işlemleri, kişisel bilgi koruma) sistem yanıltıcı kişilere karşı korunmalıdır. Bununla birlikte özellikle konuşmacı doğrulama sistemleri için taklit problemi bir sorun teşkil etmektedir.

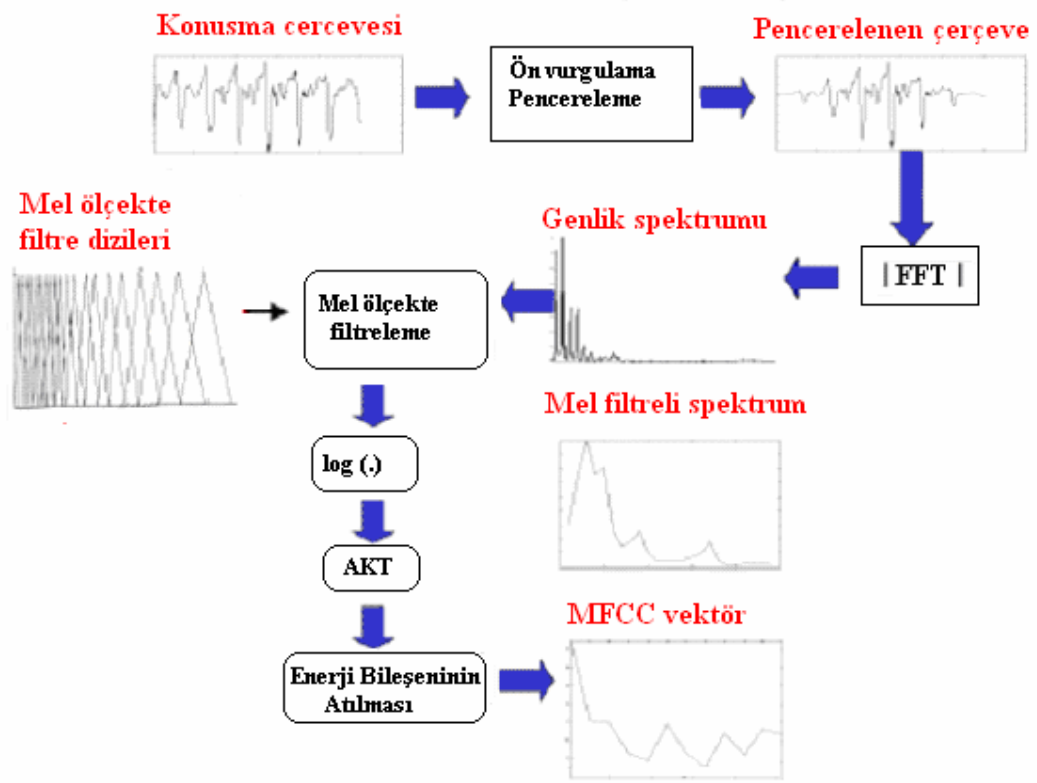
Konuşmacı tanıma sistemlerinde, öznitelik üretimi kısmında elde edilen öznitelik vektörleri özellik uzayı oluşturur. Uygun ve etkin özellikler seçilerek hem basit hem karmaşık sınıflama algoritmalarının uygulanabilmesine imkân verir (Rabiner ve Juang 1993). Öznitelik vektörü oluşturmada seçilen değişkenler, sınıflar arasında önemli farklılıklar gösteriyorsa, iyi başarıma sahip bir sınıflandırıcı ile konuşmacılar kolayca tanınabilir. Öznitelik vektörü elde edilmesinde aşağıda bazı temel yöntemler tanımlanmaktadır.

2.3.2 Mel frekansı keştrüm katsayıları (MFCC)

MFCC, insan kulağının kritik işitme bant genişliği ve frekansındaki değişimleri, düşük frekansta doğrusal süzgeçler ile yüksek frekanslarda ise logaritmik süzgeçler kullanılarak modellenmesi prensibine dayanır. Böylece önemli konuşma karakteristiklerinin yakalanması amaçlanır. Bu ölçekleme, mel frekansı ölçeği olarak adlandırılıp, 1000 Hz altı doğrusal frekans bölgesi ve 1000 Hz üstü ise logaritmik frekans bölgesi olarak tanımlanır (O'Shaughnessy 1987, Umesh ve ark. 1999, Kinnunen 2003). Şekil 2.4'de MFCC katsayılarının elde edilmesi işleminin blok diyagramı görülmektedir.

MFCC elde edilmesinde ilk olarak, sürekli konuşma işareti, N örnekten oluşan çerçevelere ayrılıp takip eden çerçeve M örnekten itibaren alınır ($M < N$). Her bir çerçevenin başından sonuna kadar işaret süreksizlikleri minimuma indirmek için her bir çerçeve pencereleme işlemine tabi tutulur. Hızlı fourier dönüşümü ile N örnekten oluşan zaman alanındaki her bir çerçeve frekans alanına çevrilir. Bu işaret, Mel frekans

ölçeğine göre dizilmiş süzgeç dizilerinden geçirilip logaritması alınır. Son olarak, logaritmik mel spektrumundan ayrık kosinüs dönüşümü kullanılarak zaman alanına geri dönülür. Sonuç olarak elde edilen katsayılara mel frekansı kepsrum katsayıları denir. Konuşma spektrumunun kepsral gösterimi, işaretin çerçeve analizi ile verilen yerel spektral özelliklerin iyi bir şekilde gösterimini sağlar. Çünkü Mel spektrum katsayıları (ve onun logaritması) gerçel sayılardır. Bu tezde öznitelik vektörü elde edilirken bu yöntem kullanılmaktadır. Bölüm 3.3’de MFCC çıkarımındaki her bir parametrenin analizi yapıp konuşmacı tanımayaya etkisi ayrıntılı olarak incelenmektedir.



Şekil 2.4 MFCC işleme blok diyagramı

2.3.3 Doğrusal öngörü katsayıları (DÖK)

Öznitelik vektörü üretme tekniklerinden en yaygın olanlarından biriside doğrusal tahmin yöntemidir (Rabiner ve Juang 1993). Bu yöntem konuşmacı parametrelerinin tahmininde kullanılan etkili yöntemlerden biridir. Doğrusal öngörülü öznitelik vektörü üretiminin dayandığı temel fikir, konuşma örneğinin geçmiş konuşma örneklerine dayanarak yaklaşık olarak elde edilebileceğidir. Şu andaki örnek konuşma ile doğrusal

olarak tahmin edilen konuşma arasındaki farkların karelerinin toplamı en aza indirilmeye çalışılarak, konuşmanın tahminini sağlayacak bir dizi birim katsayı bulunabilir (Ertaş ve Eskidere 2001). Bu katsayılara tahmin edici katsayılar denir ve tahmin edilen konuşmanın doğrusal olarak birleştirilmesi sırasında kullanılan ağırlıklandırma katsayıları olarak da tanımlanabilirler. DÖK yöntemi konuşmanın doğrusal, zamana bağlı değişen bir sistem olarak modellenmesine dayanır.

Konuşmalara ait öznitelik vektörü çözümleme bağlamında DÖK ses dalgasının formüle edilmesi olarak düşünülebilir. Bir sonraki konuşma örneğinin doğrusal olarak tahmini geçmiş örneklerin ağırlıklı toplamı denklem 2.1 ile ifade edilir (Atal 1974).

$$s_n = \sum_{i=1}^p a_i s_{n-i} \quad (2.1)$$

Doğrusal öngörü parametrelerinin tahmin edilmesi sürecinde, N değerden oluşan bir sesli ifade örneği verilmiş olsun. Amaç, en uygun sonucu üretecek olan a_i katsayılarını tahmin etmek için hesaplamalar yapmaktır. En uygun sonucu elde etmek için farkların karesini en aza indirme yöntemi kullanılır. Herhangi bir anda asıl konuşma ile tahmin edilen arasındaki hata denklem 2.2 ile hesaplanabilir.

$$e_n = s_n - \hat{s}_n = s_n - \sum_{i=1}^p a_i s_{n-i} \quad (2.2)$$

Bu durumda farkların kareleri toplamı denklem 2.3 ile hesaplanır (Lincoln 1999).

$$E = \sum_{n=0}^{N-1} e_n^2 = \sum_{n=0}^{N-1} \left(s_n - \sum_{k=1}^p a_k s_{n-k} \right)^2 \quad (2.3)$$

Burada E 'nin en küçük değerini aldığı an türevinin sıfır olduğu andır. Yani yukarıdaki formülün a_k için türevini alıp sıfıra eşitlersek gerçek sesli ifade ile tahmin edilen arasında en az hata olduğu durumu belirlemiş oluruz. Buna göre;

$$\frac{\partial E}{\partial a_j} = 0 = - \sum_{n=0}^{N-1} \left(2 \left(s_n - \sum_{k=1}^p a_k s_{n-k} \right) s_{n-j} \right) = -2 \sum_{n=0}^{N-1} s_n s_{n-j} + 2 \sum_{n=0}^{N-1} \sum_{k=1}^p a_k a_{k-j} s_{n-j}$$

eşitliğinde yeniden bir düzenleme yapılarak denklem 2.4 elde edilir.

$$\sum_{n=0}^{N-1} s_n s_{n-j} = \sum_{n=0}^{N-1} \sum_{k=0}^p a_k a_{k-j} s_{n-j} \quad (2.4)$$

Bu eşitlik, konuşma örneği $s_{-p} \dots s_{-1}$ noktaları için doğrusal öngörü katsayılarını (a_k) bulmayı sağlar. Bu eşitliğin çözümü için özilinti ve kovaryans yöntemleri kullanılmaktadır (Robinson ve ark. 1991). Bu yöntemlerin ayrıntıları burada verilmeyecektir. Konuşmacı tanımada genellikle özilinti yöntemi kullanılır. Bunun sebebi kullanılan etkin hesaplama yöntemi ve ürettiği daha durağan sonuçlardır. Bu yöntemlerle amaçlanan, doğrusal öngörü katsayılarının tahmini için yukarıda belirtilen doğrusal denklemin çözümünü yapmaktır. Bu denklemin çözümü sonucunda elde edilen doğrusal öngörü katsayıları (a_k) konuşmacı tanımada kullanılabilir.

Kepstrum değerleri, doğrusal öngörü katsayılarının doğrudan kullanılmasıyla elde edilir. Bunun için denklem 2.5'de tanımlanan özyineli yaklaşım kullanılır (Rabiner ve Juang 1993, Mengüşoğlu 1999).

$$c_k = a_k + \frac{1}{k} \sum_{i=1}^{k-1} i c_i a_{k-i} \quad (2.5)$$

Burada c_k , kepstrum değerlerinden k indisine sahip olanını temsil etmektedir. a_{k-i} ise ilgili doğrusal öngörü katsayısını göstermektedir.

2.3.4 Doğrusal algı öngörü yöntemi (PLP)

Doğrusal algı öngörü yöntemi, ayrık Fourier dönüşümü ve doğrusal öngörü tekniklerinin birleştirilmesi ile konuşma parametrelerinin hesaplanmasıdır. Bu yöntem insan kulağının duyma sistemini DÖK yönteminden daha iyi modellemeye yöneliktir.

DÖK tekniğinde konuşma modellenirken tüm frekanslardaki sesler eşdeğer tutulmaktadır. Bu durum insan kulağıyla uyumlu değildir. 800 Hz değerinden daha düşük frekanslarda duyma miktarı frekansla birlikte düşer. İnsan kulağı daha çok duyma frekans aralığının ortasındaki frekanslara duyarlıdır. Bu sorunu çözmek için birçok çalışma yapılmıştır. Bu çalışmalardan biri de, bulunan doğrusal öngörü katsayılarının mel skalasına uyarlanması olmuştur. Bir başka yaklaşım da DÖK tekniği uygulamadan önce konuşmanın güç spektrumunun alınmasıdır (Hermansky 1990). Doğrusal algı öngörü yöntemi de bu yaklaşımı kullanmaktadır.

2.3.5 Göreceli spektra yöntemi (RASTA)

Özellik vektörü oluşturmada kullanılan göreceli spektra yönteminde, konuşma içindeki çevresel etkilerin, yani gürültünün, modellenmesine dayalı bir konuşma modelleme yöntemi kullanılır. Yukarıda belirtilen doğrusal algı öngörü yöntemi üzerine gürültü modelleme tekniği eklenerek elde edilen bir yöntemdir.

Göreceli spektra yönteminin dayandığı temel, insan kulağının bir sözcüğü algılamasının daha önceki seslerden önemli derecede etkilendiğidir. Yani sözcüğün algılanması daha önce duyulan seslere bağlıdır. Daha değişik bir ifadeyle algılama şu andaki ses ile önceki ses arasındaki spektral farka bağlıdır. Bu durumda insan kulağı yavaş değişen seslere daha az duyarlıdır denebilir.

Yapılan konuşma çözümlemesinin yavaş değişen seslere daha az duyarlı yapılması insan kulağının bu özelliğinin de modellenmesini sağlar. Bunu yapmak için daha önce belirtilen doğrusal algı öngörü yönteminde kullanılan süzgeçten geçirme yönteminde değişiklikler yapılmıştır. Kullanılan süzgeçler spektral sıfır değeri keskinleştirilmiş, yani sıfır frekans düzeyine aniden inen süzgeçlerle değiştirilmiştir. Böylece frekanslardaki yavaş değişimlerin etkisi azaltılmıştır (Hermansky 1994).

2.3.6 Formant frekansları

Konuşma sinyalleri durağan olmayan ve zamanla yavaş değişim gösteren sinyallerdir. Eğer bir ses sinyalinin 5 ile 10 milisaniye gibi küçük zaman dilimlerine bakacak olursak birbirleri ile çok benzer karakteristiklerinin bulunduğunu görebiliriz. Fakat zamanın daha uzun periyotlarında bu karakteristik özelliklerin değiştiği gözlemlenebilir. Bu yüzden kısa zaman ölçekli spektral analizler konuşmacıyı tanımlamada kullanılan en yaygın yöntemdir.

Formant frekansları, konuşmacıları sesleri vasıtasıyla tanımak için oldukça elverişli konuşma parametrelerinden birisi olarak ele alınmıştır. Ancak, bunların çıkarılması ve ölçülmesinde, bilhassa konuşmacı bağımlı bilginin çoğunun bulunduğu yüksek formant bölgesinde bir takım güçlüklerle karşılaşılır. Formant tahmininin, zaman alıcı bir süreç olmasına ve yüksek dereceli formant frekanslarının çıkarılmasında problemlerle karşılaşılmasına rağmen, bunların parçalara ayrılması ve tekrarlanabilirlikleri sebebiyle, formant frekansları konuşma ve konuşmacı tanıma uygulamalarında kullanılabilme potansiyeline sahiptir (Broad 1972). Formant

frekansları ya yalnız başına ya da diğer özelliklerle birlikte kullanabilirler. Bölüm 3.5’de formant frekanslarının tek başına ve öznitelik vektörleri ile beraber kullanılarak konuşmacı tanımaya etkisi incelenmektedir.

2.3.7 Temel frekans

Ses telleri periyodik darbeler oluşturur ve bu darbelerin frekanslarına temel frekans adı verilir. Gırtlak darbelerinin frekansı ağızdan çıkan sesleri karakterize eden parametrelerin en önemlilerindedir ki bu, sesin temel frekansına ya da perde frekansına karşılık gelir. Her insanın, ses üretme mekanizmasına bağlı olarak bir perde frekansı aralığı vardır ki bu, erkekler için 50-250 Hz ve kadınlar için 120-500 Hz arasında değişir. Birçok araştırmacı perde özelliğini konuşmacı tanıma işleminde oldukça kullanışlı bulmuşlardır. Bunun sebebi ise, spektral bilgilerin aksine sesin temel frekansının kaydetme ve taşıma sisteminin frekanslarından bağımsız olmasıdır. Değişik konuşmacıların perde modelleri birbirlerinden farklı ise sadece perde özelliğine dayanan bir tanıma sistemi oluşturmak mümkün olur. Perde frekansı, formant frekansları ile karşılaştırıldığında, perde bilgisinin elde edilmesi daha kolaydır ancak gizlenebilirlikleri ve tekrarlanabilmeleri hususunda dezavantajları vardır. Stres, vurgu ve ruh haline bağlı olarak perde frekansı değişir. Perdenin belki de en büyük dezavantajı saklanabilir olmasıdır. Bu yüzden sadece perde özelliğini kullanan bir sistem taklide karşı savunmasızdır. Bir konuşmacının ortalama perdesi her ne kadar taklit edilebilirse de, bunu bir taklitçinin zamanın bir fonksiyonu olarak (yani belirli bir müddet) başarması beklenemez. Atal (1974), perde frekans değişimlerinin konuşmacıya bağlı olduklarını tespit etmekle beraber tek bir konuşmacının konuşmaları için oldukça sabit kaldıklarını gözlemlemiştir (Ertaş 2001). Her ne kadar perde yalnız başına konuşmacı tespitinde yeterli olmasa da başka parametreler ile birlikte kullanılabilir. Bölüm 3.5’de perde frekanslarının tek başına ve öznitelik vektörleri ile beraber kullanılarak konuşmacı tanımaya etkisi incelenmektedir.

2.3.8 Yoğunluk

Herhangi bir işaretin en basit özelliklerinden birisi, yoğunluğudur. Konuşma işaretinin yoğunluğu, zamanın bir fonksiyonu olarak tanımlanmalıdır. Konuşma işaretindeki yoğunluğun değişmesinin sebebi hem nefes borusu altındaki basınçta hem de sesle alakalı bileşenlerde meydana gelen zamana bağlı değişimdir ki bunlar ses

işaretinin konuşmacıya göre değişmesinin önemli sebeplerindendir. Kolaylıkla ölçülebildiği için birçok sistem yoğunluğu, diğer parametrelerle birlikte kullanılmışlardır. Bölüm 3.5’de yoğunluğun konuşmacı tanımaya etkisi incelenmektedir.

2.3.9 Öznitelik seçimi

Bölüm 2.3.1’deki ideal özniteliğe ait özellikler aslında genel olarak nasıl öznitelikler istendiğini göstermekte ve farklı öznitelikler ile karşılaştırma yapılmasını sağlamaktadır. Bu tezde yüksek doğrulukta konuşmacı tanıma elde edilebilmesi için öznitelik parametrelerine odaklanılmaktadır.

Burundan çıkan sesler ve ünlü seslerin spektral ölçümler sonucu, konuşmacı ayırımında daha etkili olduğu görülmüştür (Wolf 1972, Sambur 1975). Konuşma üretim mekanizması incelendiğinde, ünlü sesler, gırtlığa akciğerden hava gönderilmesi sonucu ses yolu ve burunda oluşan rezonanslar ile üretilmektedir. Ses tellerinin olduğu bölge gırtlak olarak adlandırılır ve sesin perdesini yansıtır. Perde çeşitli faktörlerden etkilenir. Bu faktörler anatomik yapıdan öte, çevresel şartlar ve konuşma başarımı gibi faktörler olup metinden bağımsız konuşmacı tanıma uygulamaları için güvenilir olmaz.

Bununla birlikte konuşma spektrumu, bir kişinin ses yolu ve burunun anatomik yapısını yansıtmaktadır. Bunların ışığında sabitleşmiş bir ses yoluna bağlı olarak üretilen ünlü sesler ve burun eğriliğindeki rezonanslara bağlı olarak üretilen burundan çıkan sesler konuşmacı ayırımında etkili olmaktadır. Konuşma spektrumunun hangisinin öznitelik olarak kullanılacağı konusunda bir fikir birliği bulunmamaktadır. Yaygın spektrum gösterimleri; doğrusal öngörü katsayıları ve onun çeşitli dönüşümleri (yansıma katsayıları, kepsral katsayıları v.b.) ve süzgeç bankası enerjileri ve onun kepsral gösterimi sayılabilir. DÖK gösteriminin bir dezavantajı konuşma karakteristiği gürültü ile önemli oranda bozulmaktadır (Tierney 1980). Süzgeç bankası enerjileri, farklı frekans bantlarındaki enerjilerin ölçümü olup herhangi bir model sınırlamasına bağlı değildir. Ayrıca süzgeçlerin bant genişlikleri ve merkez frekansları, kulağın kritik bant değerlerine uygun olarak ayarlanır. Bu şekilde süzgeç enerjileri ile konuşma işaretinden algılanan önemli karakteristikler daha iyi tutulur. Kritik bant veya Mel ölçek süzgeç bankası enerjileri ve onların kepsral dönüşümleri GKM konuşmacı tanıma sistemlerinde kullanılan özniteliktir. MFCC, konuşmacı tanıma için en çok istenen nitelikteki özelliklere sahiptir. Bu öznitelikler, sağlık şartlarındaki ve iletim ortamındaki

değişimlere karşı gürbüz olmamakla birlikte bölüm 3.4'de verilen çeşitli yöntemler ile bu etkiler en aza indirilmektedir.

2.4 Sınıflandırma Teknikleri

Sınıflandırma aşaması, her bir konuşmacıyı genellikle metine bağlılığa dayanarak şablona dayalı yada istatistiksel olarak modellemektedir. Sınıflandırıcı, öznitelik vektörü üretici tarafından hesaplanan özellikleri alarak başvurulan algoritmaya göre toplanan özellikler üzerinde ya şablon eşleştirmesi yada olasılık hesabı yapar. Genellikle, metine bağımlı sistemler şablon kullanırken, metinden bağımsız sistemler istatistiksel metotlar kullanırlar (Karpov 2003).

2.4.1 Şablon temelli yaklaşım

Olasılığa dayalı algoritmaların geliştirilmesinden önce, metine bağımlı konuşmacı tanıma sistemine klasik bir yaklaşım, ya spektral şablon eşleştirme yada spektrogram yaklaşımıdır. Bu yaklaşımda, her bir konuşmacı, genelde kısa zamanlı spektral özellik vektörleri olan bir dizi halindeki öznitelik vektörleri ile temsil edilir ve konuşmacının her bir kelimesi veya sözcüğü teker teker analiz edilir. İki konuşmacı aynı şeyleri konuştuklarında, ifade tarzları birbirine yakın olur ancak aynı olmaz, bu nedenle bu iki ifadenin spektrogramları birbirine benzerdir fakat aralarında mutlaka farklılıklar mevcuttur. Hatta bir konuşmacının, farklı zamanlarda kaydedilmiş ifadeleri arasında hem benzerlikler hem de farklar vardır. Bu farklar, sesin kaydedilmesi işleminden, iletilme şartlarından ve bizzat sesin kendisinden kaynaklanabilir. Fakat bu farklılıkların en önemlisi, aynı konuşmacı tarafından bilerek yada bilmeyerek yapılan değişikliklerdir. Bu tür farklılıklar, konuşmacı tanıma sistemini tamamen başarısız yapacak dereceye ulaşabilirler. Hatta aynı ortam veya şartlar altında, bir konuşmacı aynı ifadeyi tekrarlasa bile bunlar birbirinin tamamen aynısı olmaz (Ertaş 2001). Metinden bağımsız konuşmacı tanıma sistemlerinin önemli özelliklerinden biri, bir konuşmacıya ait farklı denemelerde söylenmiş olan aynı sözcüklerin zamanlama farklarını normalize edebilmesidir.

Şablon temelli yaklaşımda test sözcükleri, özellik ortalamaları arasındaki mesafeyi kullanarak eğitime şablonları ile karşılaştırılırlar. Bu teknikteki mevcut değişimler, öznitelik vektörleri ile mesafe matrislerinin seçiminden kaynaklanmaktadır. Minimum mesafe bulmak için birçok matris kullanılabilir ve bunlar arasında en yaygın olan ve hesaplanması en kolay olan öklit uzaklığıdır. Daha sonraları, Mahalanobis ve

ağırlıklaştırılmış mesafe metotlarının tanıma yeteneklerini arttırdıkları ortaya koyulmuştur. Bu yaklaşımda, dinamik zaman eğirme ve vektör nicemleme en çok kullanılan yöntemlerdir.

2.4.1.1 Dinamik zaman eğirme (DZE)

Aynı sözcüğü aynı kullanıcı tekrar seslendirdiğinde bile bir seslendiriliş daha önceki söyleyişlere benzemeyebilir. Sözcüğün uzunluğu doğrusal olmayan bir biçimde genişleme ve daralma gösterir. Zaman eğirme yöntemi sözcüğün ya da fonem sinyalinin, referans şablonu ile aynı zaman aralığında olabilmesi için zaman ekseninde daralma yada genişleme yapmayı amaçlar. DZE yönteminde zaman eksenini doğrusal olmayan bir biçimde genişletilip daraltılarak referans şablonu ile tanınacak olan konuşma kesiminin başlangıç ve bitiş zamanları çakıştırılmaya çalışılır. Amaç karşılaştırmanın aynı zaman aralıkları için yapılmasını sağlamaktır.

DZE işlemi, devingen programlama tekniği kullanılarak gerçekleştirilir. Dvingen programlamanın uygulanışı için iki örnek zaman dizisi düşünelim;

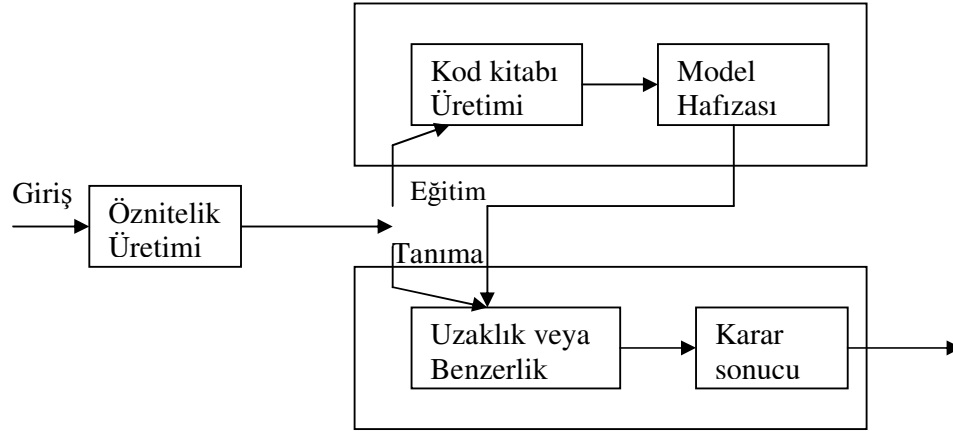
$$A = a_1, a_2, \dots, a_i \text{ ve } B = b_1, b_2, \dots, b_j$$

Bu iki farklı zaman dizisinin başlangıçlarının aynı olduğunu düşünürsek amaç a_i ve b_j zamanlarının çakıştırılmasıdır. Bunun için devingen programlama yöntemi kullanılarak bir fonksiyon tanımlanır. Bu fonksiyon yinelemeli bir yaklaşımla konuşmayı daraltarak ya da genişleterek referans şablonu ile aynı zaman aralığına getirir.

2.4.1.2 Vektör nicemleme (VN)

Vektör nicemleme algoritması, temelde en yakın komşu algoritmasını kullanarak aynı sınıfa dahil olan vektörlerin birbirine yakınlaştırılmasını ve farklı sınıfların birbirinden uzaklaştırılmasını hedefler. VN temelli konuşmacı tanıma sistemlerinde, bir konuşmacı kendisine ait ses örneklerinden oluşturulan öznitelik vektörleri ile modellenir. Eğitim aşamasında, öznitelik vektörlerinden oluşan N adet ayrı kümenin bir araya toplanması ile konuşmacı modelleri oluşturulur. Bu kümelere hücre adı verilir. Her bir hücrenin ortalama vektörünün alınır ve bir kod vektörü ile gösterilir. Sonuç olarak elde edilen kod vektörlerine kod kitabı denir ve referans vektör olarak saklanır. Tanıma aşamasında, bir giriş ses ifadesi kod kitabındaki her bir referans

konuşmacı ile vektör nicemleme yapılır. Kod kitabındaki VN bozulması toplanır (tüm sözcük üzerinde) ve karar verme için kullanılır (Deng 2003). VN temelli bir tanıma sistemi şekil 2.5’de görülmektedir.



Şekil 2.5 VN temelli bir konuşmacı tanıma sisteminin blok diyagramı

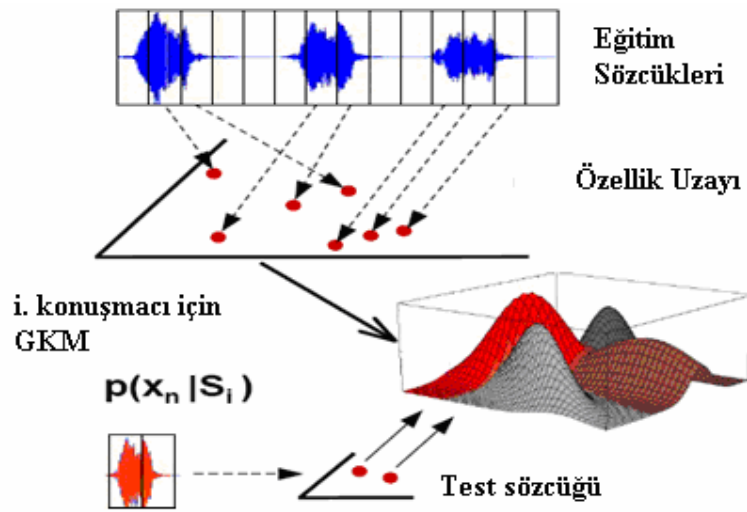
2.4.2 İstatistiksel Yaklaşım

Metinden bağımsız metotta, tanıma için kullanılan cümleler veya kelimeler tahmin edilemez. Bu yüzden, kelimeler veya sözcükler seviyesinde, konuşma eylemlerini modellemek mümkün olmamaktadır. Olasılığa dayalı konuşmacı modellemeden amaç, konuşmacının ortalama ifade özelliklerini kullanmak yerine olasılık dağılımını kullanarak modellemektir ve sınıflandırmayı ortalama özelliklere göre yapmak yerine olasılığa göre yapmaktır. Metinden bağımsız konuşmacı tanıma sistemlerinde genelde istatistiksel yaklaşım kullanılmaktadır (Reynolds ve ark. 2000, Wildermoth 2001, Karpov 2003). Gauss karışım modeli konuşmacı tanıma uygulamalarında en çok kullanılan istatistiksel yaklaşımdır.

2.4.2.1 Gauss karışım modeli (GKM)

Bu yaklaşım ilk olarak 1990 yılında Rose ve Reynolds tarafından kullanılmaya başlanıp son yıllarda konuşma tanıma (Fujimoto ve Ariki 2004.), konuşmacı tanıma (Reynolds ve Rose 1995), insan yüzü tanıma (Cardinaux ve ark. 2003) ve konuşma kodlama (Tancerel ve ark. 2000) gibi alanlarda kullanılan bir yöntem olmuştur. Gauss Karışım Modeli ile bir sınıfa ait örneklerin dağılımı Gauss temel işlevlerin doğrusal karışımı ile gösterilmektedir. Bir sınıfa ait örneklerin yoğunluk dağılımı Gauss işlevi

şeklinde olmayabilir. Bu durumda bu sınıfın örnek dağılımı modellemek için Gauss Karışım Modelleri kullanılır. Bu modelle herhangi bir dağılıma sahip yoğunluğun düzgünleştirilmiş yaklaşımı elde edilir. Modeldeki bileşen sayısı artırılarak istenilen hassasiyette örneklerin dağılımı modellenebilir. Gauss Karışım Modeli, konuşmacı tanımada çok yüksek doğrulukta sonuçlar vermektedir (Reynolds 1992, Reynolds ve Rose 1995). Bu nedenle, bu tezde konuşmacı modellemek amacıyla GKM kullanılmaktadır ve bu model ayrıntılı olarak bölüm 3.1’de incelenmektedir. Şekil 2.6’da GKM ile bir konuşmacı tanıma sistemi gösterilmektedir.



Şekil 2.6 GKM ile konuşmacı tanıma sistemi

Şekil 2.6 da görüleceği üzere bir konuşmaya ait sözcükler özellik uzayına indirgenip GKM ile modellenmektedir. Burada $p(x_n | S_i)$, x_n aday konuşmacıya ait öznitelik vektörlerinin, S_i konuşmacıya ait olma olasılığıdır. Test aşamasında ise aynı konuşmacıya ait farklı sözcükler özellik uzayına indirgenip GKM’deki her bir konuşmacıya ait modele uygulanıp en yüksek olasılıktaki model bulunur. Bu modelin ait olduğu konuşmacı, aday konuşmacı olarak atanır.

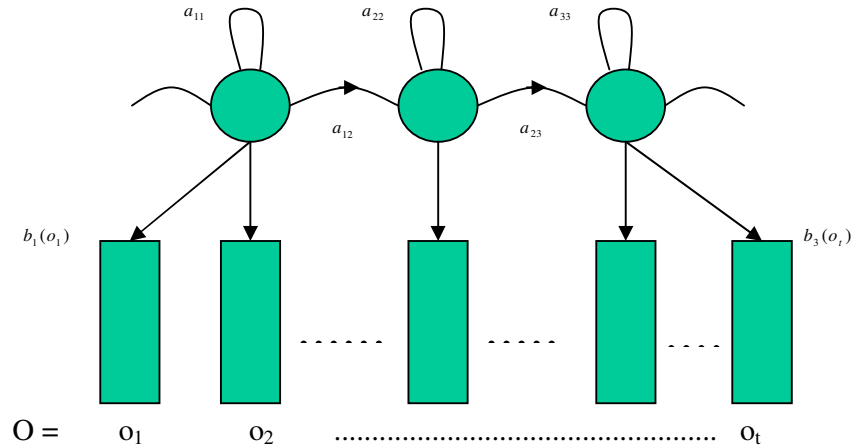
2.4.2.2 Saklı markov model (SMM)

Bu yaklaşım ilk olarak 1965-70 yıllarında kullanılmaya başlanıp ve 1985-90 yıllarında konuşmacı tanımada çok kullanılan bir yöntem olmuştur. SMM yönteminin sahip olduğu zengin matematiksel yapıdan dolayı çok çeşitli uygulamalarda

kullanılabilmesi için bir teorik temeli vardır (Rabiner 1989). Ayrıca SMM yöntemi uygun bir biçimde uygulandığında, başarılı sonuçlar elde edilmesini sağlamaktadır.

SMM'in yapısı Şekil 2.7 de görüldüğü gibi bir durumlar zincirinden meydana gelir. SMM zinciri üzerindeki her durum kelimenin bir parçasına karşılık gelir. Her durum bir diğerine geçişlerle bağlıdır. Geçişler, geçiş olasılıklarına (a_{ij}) bağlı olarak durum değiştirmeye imkan verir. Durumlara iliştilen emisyon olasılıkları (b_j) bir özellik vektörünün, referansın belirli bir zaman aralığıyla olan spektral benzerliğini gösterir. Sistem girdisine göre oluşturulan özellik vektörleri dizisine bağlı olarak, model üzerinde birinci durumdan başlayan farklı yollar izlenebilir. Bazı durumların tekrarı veya atlanması kullanıcının konuşma hızındaki değişimlere sistemin adaptasyonunu sağlar (Yapanel 1997).

Bir SMM modeli her anda durumu değişen birimleri olan bir sonlu durum makinesidir. Her t ayrık zaman anında, i durumundan j durumuna geçiş gerçekleşir ve gözlem vektörü o_t yoğunluk vektörü $b_j(o_t)$ ile dışarı verilir. Bundan başka i durumundan j durumuna geçiş aynı zamanda rasgeledir ve a_{ij} yoğunluğu ile olur. Uygulamada sadece gözlem dizisi bilinir fakat bu gözlem dizisini üreten durum dizisi bilinmez. Şekil 2.7'de, üç durumlu soldan sağa SMM atlamasız olarak verilmiştir (Rabiner ve Juang 1993, Aydın 2005).



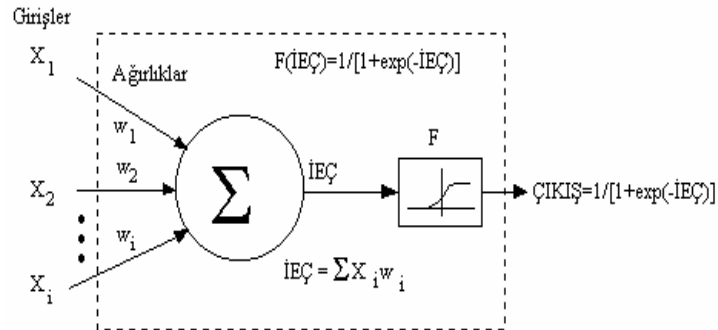
Şekil 2.7 Gözlem vektörlerinin her biri bir durum tarafından üretilen soldan sağa üçlü bir SMM

SMM, bir işaretin istatiksels olarak modellenmesidir. Konuşmacı tanımada kullanılması kısaca, ardışık kısa süreli sözcük kesimlerinin birlikte ele alınması ile ard arda gelebilecek bu kesimler için bir model oluşturmak ve bu modelden yararlanarak konuşmacı tanımayı sağlamak şeklinde özetlenebilir. Saklı markov modelinde, model parametreleri eğitim kümesi üzerinden Baum-Welch yöntemi, gradient tekniği veya bölütsel K-ortalama algoritmaları kullanılarak elde edilir. Bu yöntemlerin hepsinde özyineleme işlemleri kullanılır (Rabiner ve ark. 1985).

2.4.3 Yapay sinir ağları (YSA)

Yukarıda belirtilen tüm konuşmacı sınıflandırma teknikleri arasında, konuşmacı modeli, kişinin eğitim konuşmasından popülasyondaki diğer konuşmacılardan bağımsız olarak üretilmesi ortak özelliktir. Sınıflandırma, test sözcükleri arasında en olası veya en yakın konuşmacı modelinin bulunmasıdır. Son zamanlarda ayırım tekniği olarak yapay sinir ağı sınıflandırıcılar kullanılmaktadır (Lippmann 1987).

Beynin bütün davranışlarını tam olarak modelleyebilmek için fiziksel bileşenlerinin doğru olarak modellenmesi gerektiği düşüncesi ile çeşitli yapay hücre ve ağ modelleri geliştirilmiştir. Böylece YSA denen yeni ve günümüz bilgisayarlarının algoritmik hesaplama yönteminden farklı bir alan ortaya çıkmıştır. Genel anlamda YSA, beyin bir işlevi yerine getirme yöntemini modellemek için tasarlanan bir sistem olarak tanımlanabilir. YSA, yapay sinir hücrelerinin birbirleri ile çeşitli şekillerde bağlanmasından oluşur ve genellikle katmanlar şeklinde düzenlenir. Şekil 2.8'de yapay bir nöron görülmektedir.

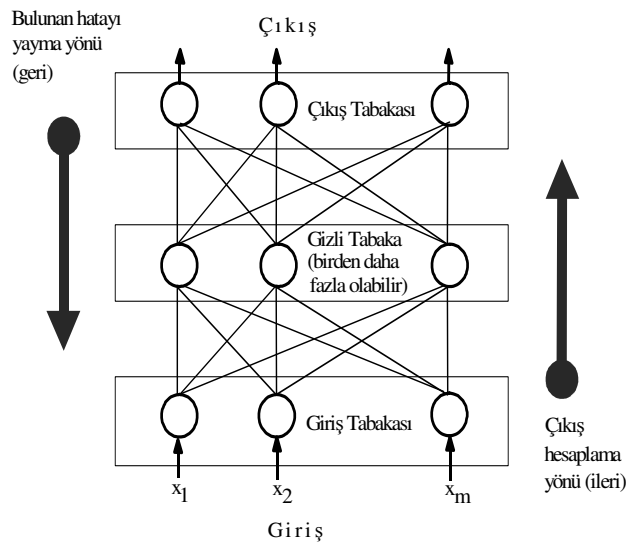


Şekil 2.8 Bir Yapay Nöron

Bir yapay nöron girişler, ağırlıklar, toplama fonksiyonu, transfer fonksiyonu ve çıkış olmak üzere altı kısımdan oluşmaktadır. Beynin bilgi işleme yöntemine uygun olarak YSA, bir öğrenme sürecinden sonra bilgiyi toplama, hücreler arasındaki bağlantı ağırlıkları ile bu bilgiyi saklama ve genelleme yeteneğine sahip paralel dağılmış bir işlemcidir. Öğrenme süreci, arzu edilen amaca ulaşmak için YSA ağırlıklarının yenilenmesini sağlayan öğrenme algoritmalarını içerir.

İleri beslemeli bir ağda nöronlar genellikle katmanlara ayrılmışlardır. İşaretler, giriş katmanından çıkış katmanına doğru tek yönlü bağlantılarla iletilir. Nöronlar bir katmandan diğer bir katmana bağlantı kurarlarken, aynı katman içerisinde bağlantıları bulunmaz. İleri beslemeli ağlara örnek olarak çok katmanlı algılayıcı (MLP) ağı verilebilir (Oglesby ve Mason 1990, Sağıroğlu 2001).

Bir MLP modeli, bir giriş, bir veya daha fazla ara ve bir de çıkış katmanından oluşur. Bir katmandaki bütün nöronlar bir üst katmandaki bütün nöronlara bağlıdır. Bilgi akışı ileri doğru olup geri besleme yoktur. Bunun için ileri beslemeli sinir ağı modeli olarak adlandırılır. Giriş katmanında herhangi bir bilgi işleme yapılmaz. Buradaki nöron sayısı tamamen uygulanan problemlerdeki giriş sayısına bağlıdır. Ara katman sayısı ve ara katmanlardaki nöron sayısı ise, deneme yanılma yolu ile bulunur. Çıkış katmanındaki nöron sayısı ise yine uygulanan probleme dayanılarak belirlenir. Şekil 2.9'da genel bir YSA modeli görülmektedir (Oglesby ve Mason 1990, Krauss 1996).



Şekil 2.9 Genel YSA Modeli

Bennani ve Gallinari (1991), aynı cinsiyetteki konuşmacıları ayırmak için YSA'ları küçük parçalara bölmektedir. Burada bir zaman gecikmeli sinir ağı (ZGSA) için alt ağlar ağaç yapısı oluşturulmaktadır. Bir ZGSA, sınıflandırma kararı oluşturmak için öznitelik vektörlerine ait pencereler kullanılarak YSA'ya geçici bilgi eklenerek bir MLP yapısı oluşturulur. Bu model ile 20 kişiden oluşan bir grupta metinden bağımsız konuşmacı tanıma için yüksek tanıma oranı elde edilmiştir.

YSA temelli konuşmacı tanıma sistemleri, VN temelli sistemler kadar başarımlı göstermektedir. Bununla birlikte YSA'ların konuşmacı tanıma uygulamalarında hala pek çok sorun (ağ, topolojileri, eğitim prosedürleri v.b.) vardır. Pek çok YSA temelli konuşmacı tanıma sisteminin dezavantajı, sisteme yeni bir konuşmacı eklendiğinde tüm konuşmacı modelinin veya ağın tamamının tekrar eğitilmesi gerekliliğidir.

2.4.4 Destek Vektör makinesi (DVM)

Destek vektör makinesi, konuşmacı tanıma amacıyla tek başına veya GMM ile birlikte son zamanlarda çok kullanılan bir genelleme yaklaşımıdır (Drucker ve ark. 1997, Cristopher 1999, Cristianini ve Taylor 2000). Bu yöntem veri örneklerinde hatayı minimumlaştırarak en iyi genellemeyi üreten sonucu elde etmeyi amaçlar. DVM iki sınıflı veriye ait bir nokta kümesini ayıran bir en iyi altdüzlem bulmaya çalışır. Bu kısımda, ilk olarak doğrusalca ayrılabilir veri basit durumu açıklanmış, sonra destek vektör kavramı ve ayrılabilen veri için genel durum anlatılmıştır.

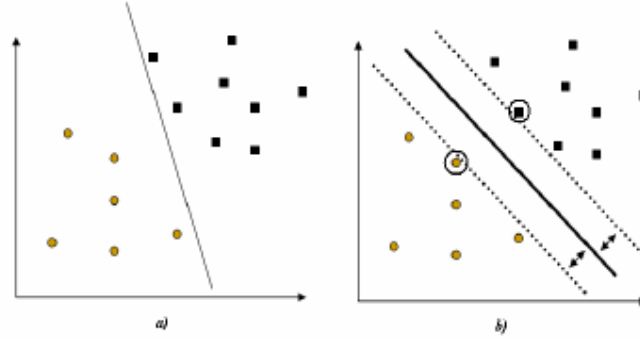
Doğrusal ayrılabilir veri durumunda her biri $y_i = \{-1, 1\}$ ile gösterilen sınıflardan birine ait olan, R_n 'in elemanı olan x_i 'ler, $i=1, \dots, N$, kümesi S verilmiştir. Amaç, veri kümesini verilen etiketlere göre bir altdüzlemle ayırıp, aynı sınıfa ait bütün veri noktalarını altdüzlemin aynı tarafında bırakmaktır.

Bir x_i 'ler veri kümesi, eğer $i=1, \dots, N$ için denklem 2.6'da belirtilen koşulu sağlayan bir w varsa doğrusalca ayrılabilir (Sezer ve ark. 2005).

$$y_i (w \cdot x_i + b) \geq 1 \quad (2.6)$$

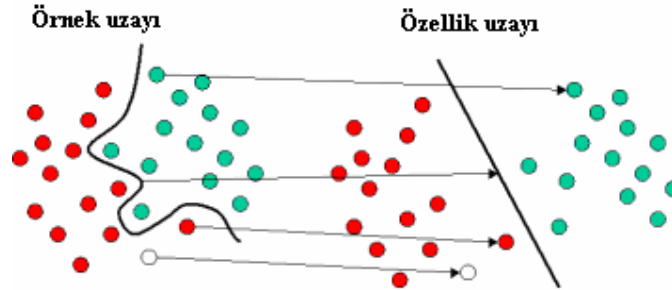
Burada $w \cdot x + b = 0$ olan bir altdüzlem tanımlanıp ayıran altdüzlem olarak adlandırılır ve denklem 2.6 daki çarpım, veri noktası ile etiketinin altdüzlemin aynı tarafında olmasını belirler. Şekil 2.10 (a)'da iki sınıfı ayıran altdüzlemlerden biri gösterilmiştir. Tabii ki, iki sınıfı ayırabilen sonsuz sayıda altdüzlem vardır. DVM bu çizgilerden her

iki sınıfa en uzak olanını bulur. Bu sayede hata azalımı maksimum olacaktır. Şekil 2.10 (b)'de iki sınıfı birbirinden ayırmak için DVM tarafından oluşturulan sınır çizgisi görülmektedir. Eğitim verileriyle sınır çizgisi bulunduğundan sonra test verileri sınırın hangi tarafında kaldıklarına göre sınıflandırılır.



Şekil 2.10 (a) İki sınıflı veriyi ayıran bir altdüzlem, (b) en iyi altdüzlem

Örnekler her zaman düzgün dağılmaz bu durumda örnekler çekirdek olarak bilinen matematiksel fonksiyonlar kullanılarak başka bir uzaya taşınır. Bu durum şekil 2.11'de görülmektedir. Bu şekilde örnekler daha iyi ayrılabilir.



Şekil 2.11 Düzgün dağılımlı olmayan örneklerin çekirdek fonksiyonları düzenlenmesi

3. MATERYAL ve YÖNTEM

Bu bölümde ilk olarak gürbüz konuşmacı tanıma için, Gauss karışım modeli ve bu modelin parametre kestirimi için maksimum benzerlik sınıflandırıcı tanıtılmaktadır. GKM konuşmacı modelinde bir konuşmacıdan elde edilen öznelik vektörleri tamamen etiketlenmeden denetimsiz olarak elde edilmektedir. GKM’de maksimum benzerlik konuşmacı sınıflandırıcı kullanılması durumunda, yüksek konuşmacı tanıma oranı elde edilmektedir (Reynolds ve ark 1995). Ayrıca tezde kullanılan veritabanları tanımlanmaktadır.

Bölüm 3.2’de konuşmacı tanıma için GKM’nin deneysel değerlendirilmesi yapılmıştır. GKM için konuşmacı modellerken eğitim aşamasında oluşan sorunlar incelenmiş ve deneysel değerlendirmesi yapılmıştır. Konuşmacı tanıma sisteminin karışım bileşen sayısı, test süresinin ve eğitilen veri miktarının konuşmacı tanıma üzerindeki etkisinin deneysel değerlendirilmesi yapılmıştır. Son olarak konuşmacı tanıma sisteminin, fazla sayıda konuşmacı için başarımı incelenmiştir.

Bölüm 3.3’de öznelik vektörü çıkartma ve parametre kestirimi verilmektedir. Mel ölçek kepstrum katsayıları oluşturulurken tüm adımların konuşmacı tanımaya etkisi deneysel olarak değerlendirilmiştir. Daha sonra Mel, bark, doğrusal ve ERB frekans ölçeklerinin konuşmacı tanımaya etkisi deneysel olarak değerlendirilmiştir. Son olarak insan işitsel sisteminin gamaton süzgeç dizileri ile benzetimi yapıp konuşmacı tanımaya uygulanmıştır.

Bölüm 3.4’de, telefon hattı üzerinden gürbüz konuşmacı tanıma için değişik öznelik vektörleri oluşturma yöntemlerinin deneysel değerlendirilmesi yapılmaktadır. Öznelik vektörlerine spektral değişim kompanzasyonu uygulanıp konuşmacı tanıma başarımı ölçülmüştür. Bu bölümde, öznelik vektörü üretiminde konuşmacılardan elde edilen öznelik vektörlerinin kümelenerek ağırlıklandırılması, öznelik vektörü oluşturulurken Mel ölçekte dizilmiş süzgeçler yerine F-oranına bağlı olarak ağırlıklandırılmış süzgeçler kullanılması önerilmiştir. Bu yöntemlerin konuşmacı tanımaya etkisi incelenmiştir.

Bölüm 3.5’de bürünsel (prosodic) özelliklerden temel frekans, formant frekansları ve enerji parametrelerinin konuşmacı tanıma için deneysel değerlendirilmesi yapılmıştır. Ayrıca konuşmadaki formant frekanslarından elde edilen genlik ve frekans modülasyonu (GM-FM) özneliklerinin, konuşmacı tanımaya etkisi incelenmektedir.

GM-FM formantları, ses yolundaki rezonansların frekans ve genliklerinin modülasyonu olarak açıklanmaktadır (Jankowski ve ark. 1995). Son olarak GM-FM parametrelerinin özilintisinin genlik zarfının, polinom benzetimi yapılıp, polinom katsayıları öznitelik vektörü olarak alınıp konuşmacı tanıma etkisi incelenmektedir.

3.1 Gauss Karışım Modeli

Metinden bağımsız konuşmacı tanıma için GKM yapısı incelenecektir. GKM içindeki Gauss bileşenlerin her biri ile spektral yapı olarak bilinen geniş fonetik sınıflar kolayca karakterize edilir. Bu fonetik sınıflar bazı konuşmacı bağımlı ses yolu yapılarını yansıtır, konuşmacı kimlik modellenmesinde kullanılır (Reynolds 1992). Ayrıca Gauss karışım yoğunluğu, bir konuşmacıdan alınan sözcüklerle gözlemlerin uzun süreli dağılımında düzgün bir yaklaşım sağlamaktadır (Bhattacharyya ve ark. 2001).

3.1.1 Model tanımı

Bir Gauss karışım yoğunluğu, M bileşenli yoğunlukların toplamının ağırlıklandırılması olup denklem 3.1'deki gibi ifade edilir.

$$p(\bar{x} / \lambda) = \sum_{i=1}^M w_i b_i(\bar{x}) \quad (3.1)$$

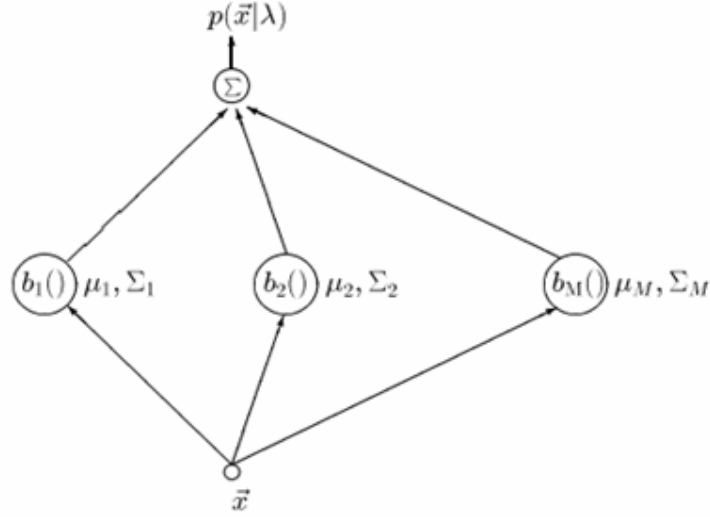
Burada \bar{x} , D boyutlu rastgele değişen vektör, $b_i(\bar{x})$, bileşen yoğunlukları ($i = 1, \dots, M$) ve w_i , karışım ağırlıklarıdır. Her bir bileşen için D boyutlu Gauss fonksiyonu denklem 3.2'de görülmektedir.

$$b_i(\bar{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\bar{x} - \bar{\mu}_i)' \Sigma_i^{-1} (\bar{x} - \bar{\mu}_i)\right\}, \quad (3.2)$$

Burada $\bar{\mu}_i$ ortalama vektör ve Σ_i ortak değişinti matrisidir. Karışım ağırlıkları $\sum_{i=1}^M w_i = 1$ şeklinde sınırlandırılır. Gauss karışım modeli, her bileşenin ortalama vektörü, ortak değişinti matrisi ve karışım ağırlık değerleri olarak denklem 3.3'deki gibi ifade edilmektedir.

$$\lambda = \{w_i, \bar{\mu}_i, \Sigma_i\} \quad i = 1, \dots, M \quad (3.3)$$

Konuşmacı tanıma için her bir konuşmacının GKM'si λ ile gösterilir. M bileşenli bir Gauss karışım yoğunluğu şekil 3.1'de görülmektedir (Reynolds 1992).



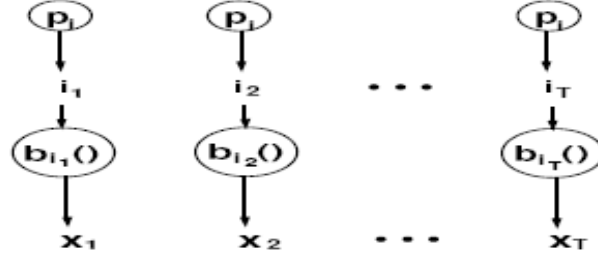
Şekil 3.1 M bileşenli Gauss karışım yoğunluğunun gösterimi

3.1.2 Akustik sınıf modelleme

Konuşma işaretindeki fonetik durumların modellenmesi ile bir kişinin sesi karakterize edilebilir. Bir kişinin sesindeki akustik sınıfların konuşmacı bağımlı modellenmesi ile eğitim ve test sözcükleri arasında metin farklarından dolayı oluşacak etkiler en aza indirilir ve karar sürecinde, ses yolunda konuşmacılar arasındaki fiziksel farklara odaklanılır.

Bir konuşmacının GKM'sindeki bir bileşen yoğunluğu, akustik sınıf veya bir akustik sınıftaki öznitelik vektörlerinin dağılımı olarak düşünülebilir. GKM'de bir konuşmacının sesi, ünlü sesler, burundan çıkan sesler veya sürtünmeli sesler gibi bazı fonetik durumlar, M akustik sınıf ile karakterize edilir. Bir akustik sınıfın olasılık modellenmesi önemlidir, çünkü bir akustik sınıf ile spektral özelliklerin bazı karakteristikleri ve kişilerin sözcükleri söyleyiş farklılıklarından dolayı aynı sınıftan öznitelik vektörlerinde değişim gözlenir. Bu değişimler; $\vec{\mu}_i$ ortalama vektör ile bir akustik sınıfın öznitelikleri, ortak değişinti matrisi Σ_i ile akustik sınıf içindeki özniteliklerin değişimi gösterilmektedir.

Bir konuşmacı model eğitimi için veriler etiketlenmediğinden dolayı akustik sınıflarda, saklı markov modelde olduğu gibi “saklı” süreç kullanıldığı düşünülebilir. Diğer bir deyişle konuşmacı model eğitiminde kullanılan öznitelik vektörler, akustik vektörlerin etiketlenmediği gözlemlerdir. Bir GKM ile bir konuşmacıdan gözlenen öznitelik vektör yoğunluğunun modellenmesi şekil 3.2’de görülmektedir (Reynolds 1992).



Şekil 3.2 Gizli akustik sınıflardan elde edilen gözlem vektörleri

Burada gözlem dizileri, $\{\bar{x}_1, \dots, \bar{x}_T\}$, gizli bir istatistiksel süreçten üretildiği düşünülebilir. Bir akustik sınıf i_T ’nin her bir zaman aralığında, i akustik sınıfın seçimi için bir ayrık olasılık dağılım ön bilgisi $\{p_i\}$ ye bağlı olarak M akustik sınıftan biri seçilir. Buna bağlı olarak \bar{x}_i gözlemi, Gauss dağılımı $b_{iT}(\bar{x})$ ye göre akustik sınıflardan üretilir. Öznitelik vektörlerinin gözlem dizisi, bir anlamda farklı istatistiksel kümelerin (akustik sınıflar) öznitelik vektörlerini içerir.

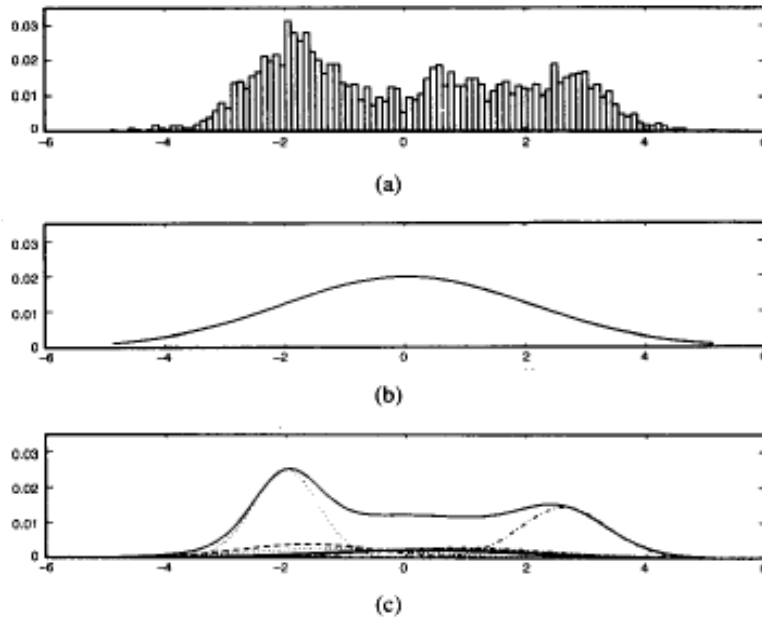
Bu yapı, gözlenen öznitelik vektörlerinin dağılımının biçimini belirler. Gözlem vektörlerinden üretilen akustik sınıflar, gizli ve bağımsız Gauss karışım yoğunluğu ile verilen gözlem vektör dağılımı denklem 3.4’deki gibidir.

$$p(\bar{x}|\lambda) = \sum_{i=1}^M p(\bar{x}, i|\lambda) = \sum_{i=1}^M w_i b_i(\bar{x}). \quad (3.4)$$

Böylece GKM sonuçları, bir kişinin konuşmasındaki akustik sınıfların olasılık modellenmesinin bir sonucudur.

GKM’nin diğer bir özelliği de dağınık şekilli yoğunluklara düzgün yaklaşım şekli oluşturmasıdır. Oysaki klasik tek modlu gauss konuşmacı modelinde bir konuşmacının öznitelik dağılımının modellenmesi, sadece bir pozisyon (ortalama vektör) ve bir eliptik şekilden (ortak değişinti matrisi) oluşur. GKM’de ise M adet

ortalama vektörü, ortak deęişinti matrisleri ve aęırlık parametreleri daha fazla modelleme kapasitesine sahiptir. Şekil 3.3’de GKM’nin bir boyutlu modelleme kapasitesini göstermektedir. Şekil (a) bir erkek konuşmacının, 25 sn uzunluęundaki bir konuşmasından alınan bir tek kepstral katsayısının histogramı göstermektedir. Şekil (b) de aynı konuşmanın, bir tek kepstral katsayısının en iyi tek modlu Gauss daęılımı görülmekte şekil (c) de ise aynı konuşmanın kepstral katsayısının 10 bileşenli GKM daęılımını göstermektedir. Bu şekilde model sadece tepeleri deęil aynı zamanda daęılımın tamamını izleyebilmektedir (Reynolds ve ark. 1995).



Şekil 3.3 GKM’nin modelleme kabiliyeti örneęi

3.1.3 Maksimum benzerlik sınıflandırıcı

Bir konuşmacı gurubuna ait gözlemlerden maksimum benzerlik sınıflandırıcısı kullanılarak elde edilen Gauss karışım yoğunlukları $\lambda_\tau, \tau = 1, \dots, S$ ile ifade edilir. Burada S konuşmacı sayısını göstermektedir. Bir öznitelik vektörü, \bar{x} , ile bir konuşmacının belirlenmesinde konuşmacı gurubu içindeki bir konuşmacıya ait konuşmacı modeli λ_s ile gösterilir (Papoulis 1984). s konuşmacısından üretilen \bar{x} vektörüne ait olasılık Bayes kuralı ile denklem 3.5’deki gibi ifade edilir.

$$\Pr(\lambda_s|\bar{x}) = \frac{p(\bar{x}|\lambda_s)}{p(\bar{x})} \Pr(\lambda_s), \quad (3.5)$$

Burada $p(\bar{x})$, \bar{x} öznitelik vektörü için pdf ve $\Pr(\lambda_s)$, s konuşmacısına ait ön bilgi olasılığıdır. Karar aşamasında maksimum olasılığa sahip konuşmacı belirlenen kişidir. s konuşmacısından gelen \bar{x} vektörünün sınıflandırma kuralı denklem 3.6'daki gibi ifade edilir.

$$\frac{p(\bar{x}|\lambda_s)}{p(\bar{x})} \Pr(\lambda_s) > \frac{p(\bar{x}|\lambda_\tau)}{p(\bar{x})} \Pr(\lambda_\tau), \quad \tau = 1, \dots, S \quad (r \neq s) \quad (3.6)$$

Bu kuralda tüm bilinmeyen konuşmacılar eş olasılıkta olduğu varsayılırsa yani $\Pr(\lambda_s) = 1/S$ ve $p(\bar{x})$ iptal edilirse denklem 3.7 elde edilir.

$$p(\bar{x}|\lambda_s) > p(\bar{x}|\lambda_\tau) \quad \tau = 1, \dots, S \quad (r \neq s) \quad (3.7)$$

Gözlem vektörü \bar{x} için her bir konuşmacının pdf'inin değerlendirilmesi ve maksimum değer seçimi ile yapılır. Bu kural Bayes karar kuralının özel bir halidir ve sınıflandırmada hata olasılığını en aza indirmek için en iyi karar kuralıdır (Tou ve ark. 1974).

Bir gözlem vektörü dizisi, $\{\bar{x}_t\}_{t=1}^T$, her bir vektör arasında istatistiksel bir ilişki bulunmadığı düşünülürse, karar kuralı denklem 3.8 deki gibi genelleştirilir.

$$\prod_{t=1}^T p(\bar{x}_t|\lambda_s) > \prod_{t=1}^T p(\bar{x}_t|\lambda_\tau), \quad \tau = 1, \dots, S \quad (r \neq s) \quad (3.8)$$

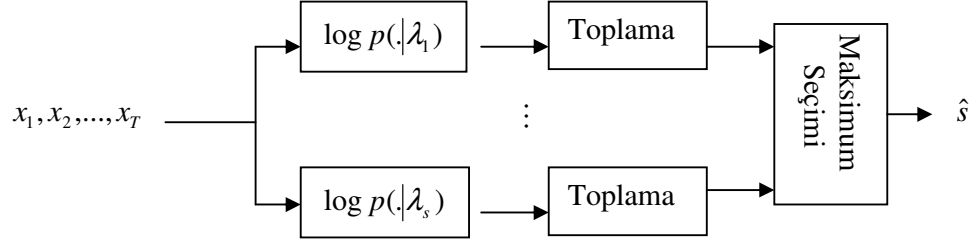
$\prod_{t=1}^T p(\bar{x}_t|\lambda_s)$ terimi, λ_s için benzerlik fonksiyonu olarak tanımlanır ve logaritmik benzerlik fonksiyonu olarak denklem 3.9'daki gibi ifade edilir.

$$L(\lambda_s) = \sum_{t=1}^T \log p(\bar{x}_t|\lambda_s) \quad (3.9)$$

Bu karar kuralı maksimum benzerlik karar kuralı olarak tanımlanmaktadır.

Şekil 3.4 bağımsız gözlem vektörleri için maksimum benzerlik karar kuralı, blok diyagram halinde gösterilmektedir. Konuşmadan sessiz kısımların kaldırılmasından

sonra öznitelik vektörlerinin üretilirse, test sözcüğündeki T öznitelik vektörleri dizisi $\{\bar{x}_t\}_{t=1}^T$, azalır. Her bir referans konuşmacı modeli benzerlik fonksiyonu denklem 3.9 kullanılarak öznitelik vektörleri üzerinden hesaplanır. Metinden bağımsız işlemlerde, öznitelik vektörlerinin derecesi konuşmacı belirleme hesabında çok önemli değildir. Belirlenen konuşmacı modeldeki en yüksek olasılığa sahip kişidir.



Şekil 3.4 Konuşmacı tanıma için kullanılan maksimum benzerlik sınıflandırıcı blok diyagramı

3.1.4 Maksimum benzerlik kestirimi

Bu kısımda, bir gözlem grubundan elde edilen maksimum benzerlik yöntemi tanımlanmaktadır. Bu maksimum benzerlik yöntemi beklentinin maksimumlaştırılması (BM) algoritmasının özel halidir. Ayrıca GKM parametre kestirimi için maksimum benzerlik kestirimi ve BM algoritmasının kullanılması tanımlanmaktadır. Karışım ağırlık, ortalama ve değişinti kestirim denklemleri çıkartılmaktadır.

Maksimum benzerlik parametre kestirimi, gözlenen örnek setlerinin parametreleri için kullanılan güçlü bir kestirim metodudur. Gözlenen örneklerden oluşan verilere en çok benzer olarak üretilen model parametreleri, λ , bulunur. Bir gözlem gurubundaki, modelin benzerlik fonksiyonunun maksimum olduğu parametrelerin bulunması ile bu sonuca ulaşılır. Bir gözlem gurubundaki veriler $X = \{\bar{x}_1, \dots, \bar{x}_T\}$ olarak ifade edilir. Model λ için benzerlik fonksiyonu, λ 'nın bir fonksiyonu gibi davranan X 'in, ek olasılık yoğunluğu olarak tanımlanır. Böylece $P(X|\lambda)$ olasılık yoğunluğu, λ bir fonksiyon olarak göz önüne alındığında, bir benzerlik fonksiyonu tanımlanmaktadır. Maksimum benzerlik parametre kestirimi için gerekli şart denklem 3.10'da belirtilmektedir.

$$\frac{\partial p(X|\lambda)}{\partial \lambda} = 0 \quad (3.10)$$

Maksimum benzerlik parametreleri kestiriminde asimptotik kararlılık gibi özelliklerin olması arzu edilir. Bunun anlamı, eğitim vektörlerinin örneklerinin sayısı çok fazla olması durumunda model olasılığı 1 olup gerçek model parametrelerine yakınsayacaktır. Denklem 3.10'u direkt olarak GKM parametrelerinden $\lambda = \{w_i, \mu_i, \Sigma_i\}$, $i = 1, \dots, M$, çözümü istenen yakınsamayı sağlamamaktadır (McLachlan 1988).

Maksimum benzerlik GKM parametreleri, BM algoritmasının özel bir hali olan döngüsel parametre kestirim yolu ile bulunur (Dempster 1977). BM algoritması, istatistiksel veri analizi (McLachlan 1988), konuşma tanıma (Rabiner 1989), gürültünün kaldırılması (Feder ve ark. 1988) gibi alanlarda kullanılır. BM algoritmasının geniş kullanılmasının nedeni her bir özyinelemeden sonra benzerlik fonksiyonu artışını garanti edip pek çok karışık kestirim problemleri için güçlü yapıya sahip olmasıdır. BM algoritmasının temelindeki iddia ilk model başlangıcı, yeni model $\bar{\lambda}$, $P(X|\bar{\lambda}) \geq p(X|\lambda)$ olarak kestirilir. Eski model yerine yeni model yerleştirilir bu işlem ve yakınsama süreci eşik değerine ulaşılan kadar devam edilir. GKM parametrelerinin maksimum benzerlik yöntemi ile kestirimi için EK 2 de belirtilen Baum yaklaşım fonksiyonu kullanılmaktadır. Bu şekilde denklem 3.11'de verilen sonsal olasılık parametresi elde edilmektedir.

$$p(i_t = i | \bar{x}_t, \lambda) = \frac{w_i b_i(\bar{x}_t)}{\sum_{k=1}^M w_k b_k(\bar{x}_t)} \quad (3.11)$$

Burada i . durumun sonsal olasılığı verilmektedir. Bu kestirim denklemi özyinelemeli parametre kestirim yönteminin temelini oluşturur. Karışım ağırlıkları, $\sum_{i=1}^M \bar{w}_i = 1$ sınırlama şartları altında denklem 3.12 deki gibi elde edilmektedir.

$$\bar{p}_i = \sum_{t=1}^T p(i_t = i / x_t, \lambda) \quad (3.12)$$

Burada $p(i_t = i | \bar{x}_t, \lambda)$ denklem 3.11'de verilmektedir. Bileşen yoğunluk ortalamaları denklem 3.13'deki gibidir.

$$\hat{\mu}_i = \frac{\sum_{t=1}^T p(i_t = i / \bar{x}_T, \lambda) \bar{x}_T}{\sum_{t=1}^T p(i_t = i / \bar{x}_T, \lambda)} \quad (3.13)$$

Tam ortak deęişinti matrisi denklem 3.14 deki gibidir.

$$\hat{\Sigma}_i = \frac{\sum_{t=1}^T p(i_t = i / \bar{x}_T, \lambda) \bar{x}_T x_T'}{\sum_{t=1}^T p(i_t = i / \bar{x}_T, \lambda)} - \bar{\mu}_i \bar{\mu}_i' \quad (3.14)$$

Köşegen ortak deęişinti matrisi için denklem 3.14 deki sadece köşegen elemanlar alınırsa denklem 3.15 deki gibi köşegen ortak deęişinti matrisi elde edilir.

$$\hat{\sigma}_i^2 = \frac{\sum_{t=1}^T p(i_t = i / \bar{x}_T, \lambda) x_T^2}{\sum_{t=1}^T p(i_t = i / \bar{x}_T, \lambda)} - \bar{\mu}_i^2 \quad (3.15)$$

3.1.4.1 Beklentinin maksimumlaştırılması

Yukarıda tanımlanan denklem (3.12), (3.13) ve (3.14) veya (3.15) kullanılarak, GKM parametreleri özyinelemeli olarak BM algoritmasından kestirilir. Şekil 3.5’de BM algoritmasının adımları görülmektedir. Algoritma aşağıdaki adımlardan oluşmaktadır.

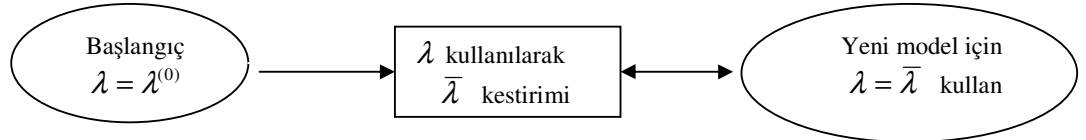
Başlangıç: $\lambda^{(0)}$ model parametreleri belirlenir.

B-Adımı: $\bar{\lambda}$ yeni model parametreleri, $\lambda^{(m)}$ kestirim denklemleri yardımıyla kestirimi yapılır.

M-adımı: Şu andaki model parametreleri yeni model parametreleri ile yer deęiştirilir:

$$\lambda^{(m+1)} \leftarrow \bar{\lambda}.$$

Özyineleme: Benzerlik fonksiyonundaki artış durana kadar B ve M adımlarına devam edilir.



Şekil 3.5 GKM konuşmacı modeli için BM algoritması adımları

Özyinelemeler arasındaki hatada değişim, denklem 3.16 kullanılarak bulunabilir. Hatadaki değişime bakılarak yerel bir en küçük noktaya ulaşıp ulaşılmadığı belirlenebilir.

$$\hat{E} - E = -\sum_n \ln \left\{ \frac{p(x / \hat{\lambda})}{p(x / \lambda)} \right\} \quad (3.16)$$

Kestirilen denklemden elde edilen yeni model parametreleri, benzerlik fonksiyonunda monoton artışı garanti eder, algoritma benzerlik fonksiyonunun sabitleştiği bir noktaya yakınsamayı sağlar (McLachlan 1988). Bu sabitleşme noktası bir eşik noktası veya benzerlik fonksiyonunun yerel bir maksimumu olabilir. BM algoritmasına uygulanmasına bağlı olarak yakınsama hızı, başlangıç, karışım bileşenlerinin sayısı ve gerekli eğitim miktarı gibi parametreler bölüm 3.2'de incelenmektedir.

3.1.5 Tezde kullanılan veritabanları

Konuşmacı tanıma amacıyla kullanılan ses örnekleri TIMIT ve NTIMIT veritabanlarına aittir. TIMIT veritabanı, konuşma ve konuşmacı tanıma sistemlerinin değerlendirilmesi ve geliştirilmesi ve akustik-fonetik çalışmalar için konuşma verileri sağlamak amacıyla hazırlanmıştır. TIMIT ve NTIMIT veritabanlarında konuşmalar tek oturumda kayıt edildiği ve iletişim ortamı farkları ve gürültü gibi gerçek ortam koşulları içermediği için konuşmacı tanıma için çok uygun bir veritabanı olmamasına rağmen literatürde konuşmacı tanıma uygulamalarında sıklıkla kullanılmıştır.

TIMIT veritabanı, Amerikan İngilizcesinin 8 ana lehçelerini konuşan, 438'i erkek, 192'si kadın olmak üzere toplam 630 konuşmacının her birinin 10 adet, fonetik olarak zengin cümlelerinin kaydedildiği geniş bant aralığını içerir. Bu cümlelerden ikisi her bir konuşmacı için aynı, cümlelerden beşi yaygın olarak kullanılan cümleler olup, diğer üç cümle ise fonetik olarak zor söylenen cümlelerden seçilmiştir.

Veritabanının %70-80'i eğitim için, geriye kalan %20-30'luk kısım test amacıyla hazırlanmıştır. Eğitim ve test amacıyla ayrı ayrı konuşmacılar kullanılmıştır. Test seti 168 kişi olup veri tabanının %27 sini oluşturmaktadır. Test setinde verilen 168 konuşmacının bölgelere dağılımı çizelge 3.1'de verilmektedir.

Çizelge 3.1 Test setinin tamamındaki konuşmacıların bölgelere göre dağılımı

Bölgeler	Erkek	Bayan	Toplam
1	7	4	11
2	18	8	26
3	23	3	26
4	16	16	32
5	17	11	28
6	8	3	11
7	15	8	23
8	8	3	11
Toplam	112	56	168

TIMIT veritabanındaki kişilerin ses örneklerinde, akustik gürültü, konuşmacı kayıtlarında oluşan oturumlar arası konuşmacı sesi değişimi ve mikrofondan dolayı seste bozulma gözlenmez. Bu nedenle temiz konuşma veritabanı olarak adlandırılır.

NTIMIT veritabanı, TIMIT veritabanına benzer bir bant genişliği sağlayacak şekilde Nytex firması tarafından üretilmiştir. NTIMIT veritabanı, TIMIT veritabanındaki cümlelerin karbondan yapılmış telefon ahizesi üzerinden bir yerel veya uzun mesafe merkez ofise iletilip ve aynı hat üzerinden tekrar kayıt için geri alınmış halidir. Çizelge 3.2’de her iki veritabanına ait temel özellikler verilmektedir (Campbell ve ark 1999).

Çizelge 3.2 TIMIT ve NTIMIT veritabanlarının karakteristikleri

Veritabanı	Konuşmacı sayısı	Cümle sayısı	Kanal	Mikrofon	Örnekleme hızı (kHz), formatı
TIMIT	630	10 (ortalama 3 sn cümle)	temiz	Sabit geniş bant	16, 1 kanal 16 bit doğrusal
NTIMIT	630	10 (ortalama 3 sn cümle)	yerel ve uzun mesafe telefon hattı	Karbon telefon ahizesi	16, 1 kanal 16 bit doğrusal

3.2 Konuşmacı Tanıma için GKM’nin Deneysel Değerlendirilmesi

Gauss karışım modelinin, metinden bağımsız olarak konuşmacı tanıma için deneysel değerlendirilmesi yapılacaktır. GKM konuşmacı tanıma sisteminde kullanılan cümleler, klasik geniş bant (0-8000 Hz), yüksek işaret gürültü oranına sahip (53 dB) kanal ve dar bant (300-3400 Hz) telefon kanalları şeklindedir. GKM’nin aşağıda belirtilen şartlarda konuşmacı tanıma başarımı belirlenecektir. Bunlar;

1. GKM ile konuşmacı modellerken eğitim safhasında oluşan sorunların belirlenip çözülmesi,
2. Konuşmacı tanıma sisteminin karışım bileşen sayısı, test süresi ve eğitilen veri miktarının konuşmacı tanıma üzerindeki etkisinin belirlenmesi,
3. Konuşmacı tanıma sisteminin fazla sayıda konuşmacı için konuşmacı tanıma başarımının belirlenmesi.

Yapılan deneylerin tamamında TIMIT ve NTIMIT veritabanına ait ses örnekleri kullanılmaktadır. TIMIT veritabanını ile yapılan konuşmacı tanıma deneylerinde, 168 konuşmacıdan oluşan test dizini ve veritabanının tamamı (630 konuşmacı) olmak üzere iki grup üzerinde çalışmalar yapılmaktadır. NTIMIT veritabanı ile 168 kişiden oluşan test dizini ile konuşmacı tanıma deneyleri yapılmaktadır. Eğitim ve test aşamalarında veritabanlarının cümle yapısı şu şekilde belirlenmektedir. Modelin eğitim safhasında 3 ayrı eğitim kümesi oluşturulmaktadır. Bunlar;

- Yaklaşık 24 saniye uzunluğunda 8 cümle (2 Sa, 3 Si, 3 Sx cümleleri),
- Yaklaşık 15 saniye uzunluğunda 5 cümle (2 Sa, 3 Si cümleleri),
- Yaklaşık 9 saniye uzunluğunda 3 cümle (2 Sa, 1 Si cümleleri).

Modelin test safhasında 3 ayrı test kümesi oluşturulmaktadır. Bunlar;

- 1 saniye uzunluğunda cümle parçası (1 Sx cümlesinin bir kısmı),
- 3 saniye uzunluğunda cümle parçası (yaklaşık 1 Sx cümlesi),
- 6 saniye uzunluğunda 2 cümle parçası (yaklaşık 2 Sx cümlesi).

Bölüm 3.2.1'de yapılacak olan deneylerde öznitelik vektörü elde edilmesinde Mel frekansı kepsrum katsayıları kullanılmaktadır. TIMIT veritabanı için 20 msn, NTIMIT veritabanı için 25 msn uzunluğunda çerçeve süreleri kullanılmaktadır. MFCC elde edilmesinde çerçevelerin örtüşme oranı 10 msn alınıp, çerçevelere Hamming pencereleme uygulanmaktadır. Pencereleyen sesin 512 örnek FFT si alınıp Mel ölçekte yerleştirilmiş üçgen süzgeç dizilerinden geçirilmiştir. Süzgeç dizileri, TIMIT veritabanı için 0-8000 Hz arasına, NTIMIT veritabanı için ise 300-3400 Hz arasına Mel ölçekte yerleştirilmiştir. Süzgeçten geçirilen işaretin logaritması alınıp ayrık kosinüs dönüşümü alınmıştır. Her bir çerçeveye karşılık olarak TIMIT veritabanı için 24, NTIMIT veritabanı için 20 boyutlu öznitelik vektörleri kullanılmaktadır.

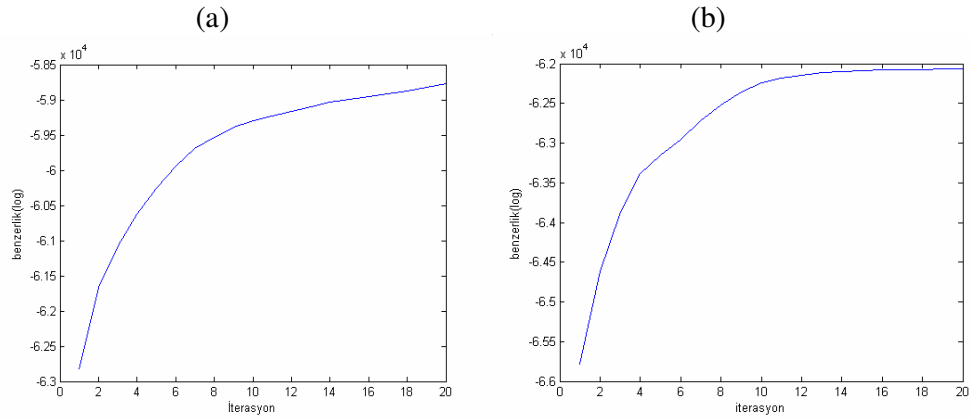
3.2.1 Model eğitimi aşamasında yapılan düzenlemeler

GKM konuşmacı modeli için eğitim yöntemi ve parametre kestirimi bölüm 3.1’de verilmektedir. GKM modelinde, BM algoritması ile konuşmacıların eğitiminde bazı düzenlemeler yapılması gerekmektedir ki bunlar, GKM konuşmacı model eğitiminde BM algoritmasının yakınsama miktarı, model başlangıç değerlerinin ayarı ve eğitim aşamasında model tekilliklerini önlemek için değışinti sınırlanmasıdır.

3.2.1.1 BM algoritmasının özyineleme sayısı

BM parametre kestirimi, benzerlik fonksiyonunun maksimum olduğu model parametre değerlerinin bulunmasıdır. BM algoritmasında her bir özyineleme benzerlik fonksiyonunun artışını sağlar. Özyineleme sayısı pratik anlamda benzerlik fonksiyonunun yeterli oranda yakınsayıp yakınsamadığını bulmak için gereklidir.

Gauss karışım modeli kullanılarak, bir konuşmacının eğitilmesinde BM algoritmasının 20 özyineleme için benzerlik fonksiyonunun değışimi şekil 3.6’da görülmektedir. Şekillerde Gauss karışım bileşen sayısı 32 ve 16 alınmaktadır. Her bir konuşmacı, 8 cümle’ye karşılık gelen yaklaşık 2400 öznitelik vektörü kullanılarak eğitilmektedir.



Şekil 3.6 GKM eğitim için BM algoritmasının benzerlik fonksiyonunun (a) karışım sayısı 32 (b) karışım sayısı 16 için değışimi

Şekil 3.6’dan görüleceği üzere, BM algoritmasına ait benzerlik fonksiyonu, alabileceği maksimum değerin % 90’ına ilk 5 özyineleme içerisinde ulaşmakta ve 15 özyineleme içerisinde belli bir değere yakınsamaktadır. Bu yakınsama, konuşmacı değışimlerinden, model başlangıcı için kullanılan yöntemlerden, karışım bileşen sayısı

ve eğitilen veri miktarından bağımsızdır (Reynolds 1992). Şekillerden de görüleceği üzere deneylerde BM özyineleme sayısı 15 alınması yeterlidir.

3.2.1.2 Model başlangıç değerleri

Gauss karışım modelinin eğitim safhasında başlangıç model değerlerine sahip olması gerekir. BM algoritması başlangıç değerlerinden (λ^0) bağımsız yerel maksimum benzerlik değerlerini garanti eder. Fakat GKM için benzerlik denklemi birkaç yerel maksimum değer ve farklı başlangıç değerine sahip olup, farklı yerel maksimum değere yönelebilir (Reynolds ve ark 1995). BM eğitim algoritması kullanılarak farklı başlangıç şartlarında konuşmacı tanıma başarımı etkilenme oranı bilinmemektedir. Bu sorunun çözümü için konuşmacı modeli farklı başlangıç şartlarında eğitilip konuşmacı tanıma oranı ölçülecektir.

Model başlangıç değerlerini belirlemek için üç farklı yöntem kullanılacaktır. Birinci yöntem olarak model başlangıç parametreleri tahmininde rastgele değişen değerler ile model ortalaması oluşturulur. Başlangıç ortak değişinti matrisi için birim matris kullanılır. İkinci yöntemde k-ortalama algoritması ile 32 bileşenli GKM modeline ait başlangıç değerleri oluşturulur. K-ortalama algoritması ile ilk olarak ortalamaları temsil eden rastgele merkez noktaları atanır. İkinci adımda öznitelik vektörleri, en yakın ortalamaların kümesine atanır ve ortalamalar tekrar hesaplanır. Ortalamalarda değişiklik olmayana kadar algoritma tekrarlanır (Sanderson 2002). Üçüncü yöntem olarak VN referans vektörleri (codebook) üretiminde kullanılan LBG algoritması uygulanmaktadır (Linde ve ark. 1980). Bu yöntemde ilk olarak öznitelik vektörlerinin ortalaması bulunmakta daha sonra ikili ayırma tekniği (binary splitting) kullanılarak ortalama değer 2'ye bölünmekte bu işlem istenen sayıda ortalama değer elde edilene kadar devam ettirilmektedir (Rabiner ve Juang 1993). Çizelge 3.3'de üç farklı model başlangıç metodu ile TIMIT ve NTIMIT veritabanları için konuşmacı tanıma oranları görülmektedir.

Çizelge 3.3'deki TIMIT veritabanı ile elde edilen sonuçlardan görüleceği üzere model başlangıç metotları tanıma oranını değiştirmemektedir. NTIMIT veritabanı ile elde edilen sonuçlardan, model başlangıç değerlerinin bulunmasında VN yönteminin daha iyi konuşmacı tanıma sonuçları verdiği görülmektedir. Çünkü rastgele değişim ve k-ortalama algoritması başlangıç değeri olarak rastgele değişen değerler kullanılırken,

VN yönteminde başlangıç değeri olarak öznitelik vektörlerinin ortalaması kullanılmaktadır. (Rabiner ve Juang 1993). Bu sayede model başlangıç değerleri daha iyi tanımlanabilmektedir.

Çizelge 3.3 Farklı model başlangıç metotları için konuşmacı tanıma oranları (%)

Model Başlangıcı	Veritabanları	
	TIMIT	NTIMIT
Rastgele değişim	99.4	66.07
k-ortalama	99.4	67.86
VN	99.4	68.45

Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168

3.2.1.3 Ortak değişinti matrisi seçimi

GKM, ortak değişinti matrisi seçimine bağlı olarak birkaç farklı yapıya sahiptir. Denklem 3.3'de görüldüğü üzere, GKM parametreleri, bileşen yoğunluğu başına bir tam ortak değişinti matrisi kullanılarak gösterilmektedir. Bununla birlikte bu tezde ve pek çok çalışmada (Reynolds 1992, Sanderson 2002) bileşen yoğunluğu başına köşegen ortak değişinti matrisi kullanılarak model basitleştirilir. Köşegen ortak değişinti matrisi kullanılmasının nedeni şu şekilde açıklanabilir.

- GKM de köşegen ortak değişinti matrisi kullanılarak, tam ortak değişinti matrisine göre daha az hesap yapılmaktadır. Denklem 3.2'de $D \times D$ boyutunda bir matrisin tersini almak yerine sadece aynı matrisin köşegen elemanlarının tersi alınmaktadır (Sanderson 2002).
- GKM de düşük karışım sayısına sahip bir tam ortak değişinti, daha yüksek karışım sayısına sahip bir köşegen ortak değişinti ile gösterilebilir (Reynolds ve ark. 2000).
- Köşegen ortak değişinti matrisi kullanılarak tam ortak değişinti matrisine nazaran bilinmeyen parametre sayısı azalır ve böylece daha az eğitim verisi kullanılır (Sanderson 2002).
- Ortak değişinti matrisinde köşegen dışı matris elemanları 0'a yakın değerler almaktadır (Kinnunen 2003).

Çizelge 3.4'de köşegen ve tam ortak değişinti matrisleri için konuşmacı tanıma oranları görülmektedir.

Çizelge 3.4 Köşegen ve tam değişinti matrisleri için konuşmacı tanıma oranları (%)

Ortak Değişinti Matrisi	Veritabanları	
	TIMIT	NTIMIT
Köşegen	99.4	68.45
Tam	97.02	37.5

Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168

Çizelge 3.4'den görüleceği üzere Gauss karışımlarının modellenmesinde tam ortak değişinti matrisi yerine köşegen ortak değişinti matrisi kullanılması tanıma oranını arttırmaktadır. Bu sonuçlara bağlı olarak her bir öznitelik vektörü, istatistiksel olarak bağımsız varsayılarak tam ortak değişinti matrisi yerine köşegen ortak değişinti matrisi kullanılmaktadır.

3.2.1.4 Değişinti sınırlanması

Gauss karışım modeli, MFCC kullanılarak eğitilirken değişinti vektörünün belirli bileşenlerinin oldukça küçük genlikte (sıfıra yakın) olduğu gözlenmektedir (Reynolds 1992). Bu durum özellikle Gauss karışım bileşen sayısı büyük olduğu (≥ 32) durumlar için geçerlidir. Gözlenen bu küçük değişintiler modelin benzerlik fonksiyonundaki tekilliklerin sonucudur. Bu durum maksimum benzerlik sınıflandırıcısı kullanıldığı durumlarda konuşmacı modelinde bozulmalar meydana getireceğinden konuşmacı tanıma başarımında düşmeler olur. Bu tekillikler eğitim için yeterli veri olmadığı durumlarda artar. Aynı durum, model çok fazla karışım bileşen sayısına sahip olduğu durumlarda da gözlenir. Ayrıca tekillikler, telefon veya gürültülü konuşma ortamlarında veri kırılması olduğu durumlarda da oluşabilir (Shannon 2003).

Modelin eğitimi esnasında oluşan model değişinti değerlerinin sıfıra yönelmesini önlemek için sabit değişinti sınırlaması uygulanır. Herhangi bir i . karışım bileşeninin değişinti vektörü, σ^2_i , olmak üzere değişinti sınırlaması denklem 3.17'deki gibi ifade edilir.

$$\overline{\sigma^2_i} = \begin{cases} \sigma^2_i > \sigma_{\min}^2 \Rightarrow \sigma^2_i \\ \sigma^2_i \leq \sigma_{\min}^2 \Rightarrow \sigma_{\min}^2 \end{cases} \quad (3.17)$$

burada σ_{\min}^2 , minimum değişinti değeri olup elde edilen $\overline{\sigma^2_i}$ değişinti değeri her BM özyinelemesi için bulunur.

Minimum değışinti değeri çok yüksek seçilmesi durumunda, bileşen değışintileri aynı σ^2_{\min} değeri ile sınırlandırılacak, buna bağı olarak tanıma başarımında düşme olacaktır (Reynolds 1992). Bu değerin çok küçük alınması durumunda da değışinti sınırlaması arzulanan işlevini göremeyecektir. Bu durumda en ideal minimum değışinti değeri denenerek bulunur. TIMIT ve NTIMIT veritabanları için çizelge 3.5’de çeşitli değışinti değerlerine bağı olarak konuşmacı tanıma oranları görülmektedir.

Çizelge 3.5 Farklı minimum değışinti değerleri için konuşmacı tanıma oranları (%)

Değışinti Sınırı	Veritabanları	
	TIMIT	NTIMIT
$\sigma^2_{\min}=0.001$	99.4	68.45
$\sigma^2_{\min}=0.01$	99.4	68.45
$\sigma^2_{\min}=0.1$	99.4	66.07

Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168

Elde edilen sonuçlar karşılaştırıldığında TIMIT veritabanında minimum değışinti sınırının değıştirilmesi tanıma oranını değıştirmemektedir. NTIMIT veritabanında σ^2_{\min} ’in 0.01 ve 0.001 alındığı durumlarda % 68.45 ile en yüksek tanıma oranı elde edilmektedir.

3.2.2 Karışım bileşen sayısı ve eğitilen veri miktarının konuşmacı tanıma etkisi

Bu kısımda yapılan deneylerde GKM karışım bileşen sayısı ve konuşmacı modeli eğitiminde kullanılan konuşma süresinin konuşmacı tanıma başarımına etkisi incelenmektedir. İlk olarak karışım bileşen sayısı değışiminin, daha sonrada 3 farklı eğitim süresinin, çeşitli karışım sayılarına bağı olarak konuşmacı tanıma başarımına etkisi incelenmektedir.

Bölüm 3.2.2’de yapılan deneylerde MFCC oluşturulurken şu işlemler uygulanmaktadır. Konuşmadan ortalama bileşenler atıldıktan sonra parçalara ayrılır. 10 msn de bir kaydırılan her bir 20 msn uzunluğunda çerçeveye Hamming pencereleme uygulanır. Elde edilen ses örneğinin 512 nokta ayrık fourier dönüşümü alındıktan sonra Mel ölçeğe yerleştirilmiş 40 adet üçgen süzgeç dizisinden geçirilir. Elde edilen işaretin logaritması alındıktan sonra ayrık kosinüs dönüşümü alınmaktadır. Sonuç olarak her bir

çerçeveye karşılık olarak, TIMIT veritabanı için 24 boyutlu, öznitelik vektörü elde edilmektedir.

3.2.2.1 İdeal karışım bileşen sayısının bulunması

Bir konuşmacıyı modellemek için, M bileşenli karışım sayısını belirlemek önemli bir problem olup konuşmacı tanıma başarımını doğrudan etkilemektedir. Karışım bileşen sayısını önceden belirlemenin teorik bir yolu yoktur (Reynolds ve Rose 1995). Az sayıda karışım bileşeni seçilirse bir konuşmacının dağılımının ayırıcı karakteristiği doğru olarak modellenemeyecektir. Çok fazla bileşen seçilmesi durumunda ise eğitim verisi ile ilişkili çok fazla model parametresi oluşup başarımı düşürecektir (Bhattacharyya ve ark. 2001). Ayrıca çok fazla bileşen, hem eğitim hem de test aşamasında işlem karmaşıklığına yol açacaktır.

TIMIT veri tabanına ait 168 kişiden oluşan konuşmacı kümesi kullanılarak GKM karışım bileşen sayısının, konuşmacı tanıma başarımına etkisi incelenmektedir. Konuşmacı modelleri 2, 4, 8, 16, 32 ve 64 bileşenli Gauss yoğunlukları ve köşegen ortak değişinti matrisi kullanılmaktadır. 24 saniye uzunluğunda (2 Sa, 3 Si, 3 Sx cümleleri) konuşma cümleleri kullanılarak Mel kepstrum katsayıları (2400 x 24 boyutlu) elde edilir. Eğitim için 15 BM özyinelemesi kullanılmaktadır. Değişinti sınırlaması olarak $\sigma^2_{\min}=0.01$ değeri alınmaktadır. Test aşamasında 1, 3 ve yaklaşık 6 saniye uzunluğundaki test cümle parçaları kullanılacaktır. Bu parametrelere bağlı olarak elde edilen konuşmacı tanıma oranları çizelge 3.6'da görülmektedir.

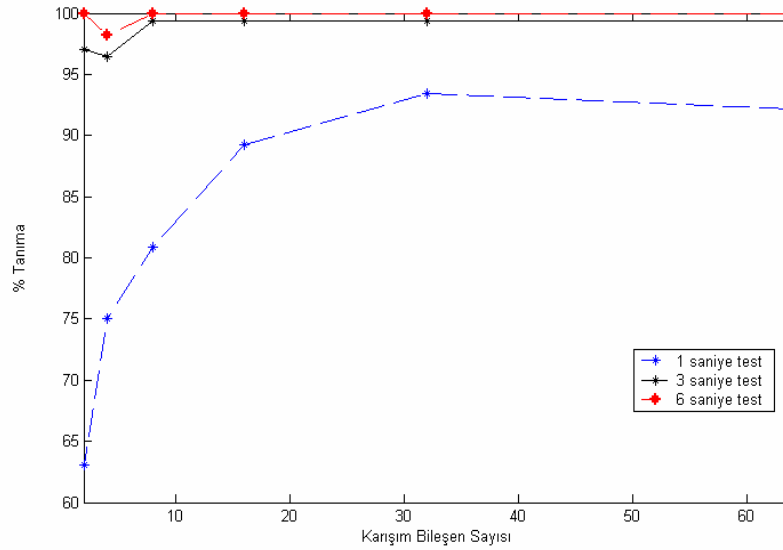
Çizelge 3.6 Karışım bileşen sayısına bağlı olarak konuşmacı tanıma oranları (%)

Karışım bileşen sayısı	Test süresi		
	1 saniye	3 saniye	6 saniye
$M=2$	63.1	97.0	100
$M=4$	75.0	96.4	98.2
$M=8$	80.9	99.4	100
$M=16$	89.2	99.4	100
$M=32$	93.4	99.4	100
$M=64$	92.2	99.4	100

Eğitim süresi 24 sn, kepstrum katsayı sayısı 24, TIMIT veritabanı

Çizelge 3.6'dan görüleceği üzere 1 saniye uzunluğunda test konuşmaları kullanıldığı durum incelendiğinde karışım bileşen sayısının 2 den 16'ya kadar artışına bağlı olarak konuşmacı tanıma oranında çok keskin bir artış gözlenmektedir. Konuşmacı tanıma oranında en iyi sonuç karışım bileşen sayısı 32 için gözlenmektedir. Bu değerden sonra karışım bileşen sayısı artmasına rağmen konuşmacı tanıma başarımı düşmektedir. Test cümlelerinin tamamı (~ 6 sn) kullanılarak test işlemi uygulandığında ($M=4$ hariç) % 100 doğrulukta konuşmacı tanıma oranı elde edilmiştir. Test süresi belli bir değer üzerine çıkartıldığında karışım bileşen sayısı değişimlerinden etkilenmemektedir.

Şekil 3.7'de üç farklı test süresi için karışım bileşen sayısına bağlı olarak konuşmacı tanıma oranları görülmektedir. Şekilden görüleceği üzere karışım bileşen sayısı ve test süresi arttıkça konuşmacı tanıma oranları artmaktadır. Test süresi 6 saniye için en yüksek tanıma oranı gözlenmektedir.



Şekil 3.7 TIMIT veritabanı için karışım bileşen sayısına bağlı olarak konuşmacı tanıma oranları (%)

NTIMIT veritabanı kullanılarak karışım bileşen sayısı değişiminin konuşmacı tanıma etkisi incelenmektedir. Öznitelik vektörleri elde edilmesinde konuşma 25 msn uzunluğunda çerçevelere ayrılıp 10 msn'de bir kaydırılır. İşaretin FFT'si alındıktan sonra 300-3400 Hz aralığına 70 Hz aralıklarla doğrusal ölçekte yerleştirilmiş süzgeç

dizilerinden geçirilir. Son olarak işaretin logaritması alındıktan sonra ayrık kosinüs dönüşümü uygulanır. Elde edilen kepstrum katsayıları 20 boyutludur.

Deneyde 1, 3 ve 6 saniye olmak üzere 3 farklı test süresi kullanılmaktadır. Konuşmacı modelleri 2, 4, 8, 16, 32 ve 64 bileşenli gauss yoğunlukları ve köşegen ortak değişinti matrisi kullanılmaktadır. 24 saniye uzunluğunda (2 Sa, 3 Si, 3Sx cümleleri) konuşmalar ile doğrusal olarak yerleştirilmiş kepstrum katsayıları (2400 x20 boyutlu) elde edilmektedir. Eğitim için 15 BM özyineleme kullanılıp değişinti sınırlaması olarak $\sigma^2_{\min}=0.01$ değeri kullanılmaktadır. Bu parametrelere bağlı olarak elde edilen konuşmacı tanıma oranları Çizelge 3.7’de görülmektedir.

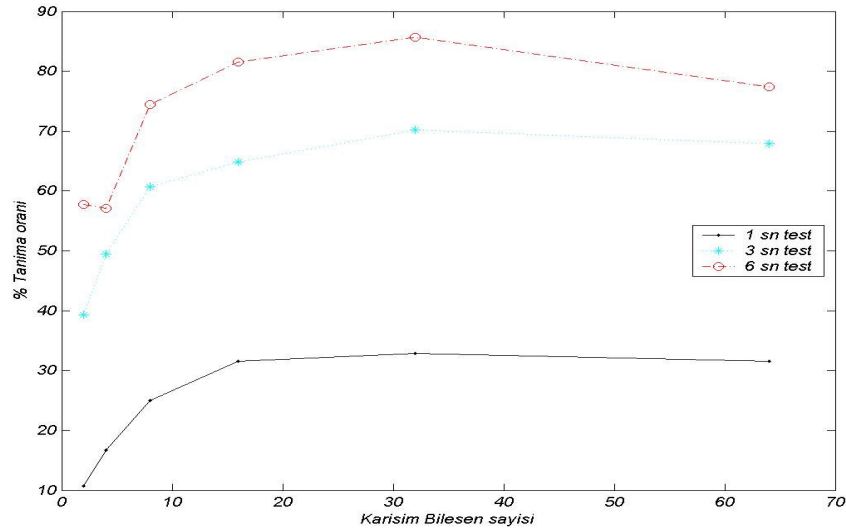
Her üç test süresi içinde karışım bileşen sayısı 32 olduğu durumda en yüksek tanıma oranı elde edilmektedir. Test süreleri karşılaştırıldığında test süresi 6 saniye alındığında NTIMIT veritabanı için tanıma oranı % 85.71’e çıkmaktadır.

Çizelge 3.7 Karışım bileşen sayısına bağlı olarak konuşmacı tanıma oranları (%)

Karışım bileşen sayısı	Test süresi		
	1 saniye	3 saniye	6 saniye
M=2	10.71	39.29	57.74
M=4	16.67	49.41	57.14
M=8	25	60.71	74.40
M=16	31.55	64.88	81.55
M=32	32.84	70.24	85.71
M=64	31.55	67.86	77.38

Eğitim süresi 24 sn, kepstrum katsayı sayısı 20, NTIMIT veritabanı

Şekil 3.8’de üç farklı test süresi için farklı model derecelerine bağlı olarak konuşmacı tanıma oranları görülmektedir. Şekil 3.8’den görüleceği üzere karışım bileşen sayısı ve test süresi artışına paralel olarak konuşmacı tanıma oranı artmaktadır.



Şekil 3.8 NTIMIT veritabanı için karışım bileşen sayısına bağlı olarak konuşmacı tanıma oranları (%)

3.2.2.2 Eğitim ve test süresi değişimi

TIMIT veritabanında farklı eğitim sürelerinin konuşmacı tanıma etkisini incelemek için konuşmacılar 9 saniye, 15 saniye ve 24 saniye eğitilmektedir. Her konuşmacıdan 24 boyutlu öznitelik vektörleri elde edilmektedir. Konuşmacılar 9 saniye, 15 saniye ve 24 saniye olmak üzere 3 farklı eğitim süresi için 2, 4, 8, 16, 32 ve 64 karışım bileşen sayıları kullanılarak, BM algoritması ile eğitilip 1 saniye, 3 saniye ve 6 saniye olmak üzere 3 farklı test süresi için ayrı ayrı test işlemlerine tabii tutulmaktadır. Çizelge 3.8’de GKM’nin 9 saniye eğitilmesi ile elde edilen konuşmacı tanıma oranları görülmektedir.

Çizelge 3.8 GKM’nin 9 saniye eğitilmesi ile elde edilen konuşmacı tanıma oranları (%)

Eğitim süresi	Karışım bileşen sayısı	Test süresi		
		1 saniye	3 saniye	6 saniye
9 saniye	$M=2$	55,4	90,5	97,0
	$M=4$	64,9	92,3	98,2
	$M=8$	72,0	94,0	99,4
	$M=16$	69,6	94,6	98,8
	$M=32$	70,8	94,6	97,0
	$M=64$	58,3	82,1	89,3

Kepstrum katsayı sayısı 24, konuşmacı sayısı 168, TIMIT veritabanı

Çizelge 3.8 incelendiğinde konuşmacıların eğitim süresi 9 saniye için, test süresi 1 ve 6 saniye olduğu durumlarda en yüksek tanıma oranı, karışım bileşen sayısı 8 olduğu durumda gerçekleşmektedir. Eğitim süreleri düştükçe konuşmacıları daha iyi modellemek için model derecesinin de düşürülmesi gerekmektedir. Aynı şartlarda konuşmacıların 15 saniye eğitilmesi ile elde edilen konuşmacı tanıma oranları çizelge 3.9’da görülmektedir.

Çizelge 3.9 GKM’in 15 saniye eğitilmesi ile elde edilen konuşmacı tanıma oranları (%)

Eğitim süresi	Karışım bileşen sayısı	Test süresi		
		1 saniye	3 saniye	6 saniye
15 saniye	M=2	60,7	91,7	99,4
	M=4	67,3	94,6	99,4
	M=8	77,9	97,6	100
	M=16	81,6	98,1	100
	M=32	83,3	98,9	100
	M=64	81,6	98,2	99,4

Kepstrum katsayı sayısı 24, konuşmacı sayısı 168, TIMIT veritabanı

Çizelge 3.9 incelendiğinde en yüksek konuşmacı tanıma oranının karışım bileşen sayısı 32 için elde edildiği gözlenmektedir. Çizelge 3.8 ve çizelge 3.9 karşılaştırıldığında eğitim süresinin 9 saniyeden 15 saniyeye çıkartılması her test süresi için tanıma oranını arttırmaktadır. Bu artış test süresi 1 saniye için daha keskin olmakta, test süresi artışına paralel olarak konuşmacı tanıma artış oranı da artmaktadır.

Son olarak konuşmacıların 24 saniye eğitilmesi ile elde edilen konuşmacı tanıma oranları çizelge 3.10’da görülmektedir.

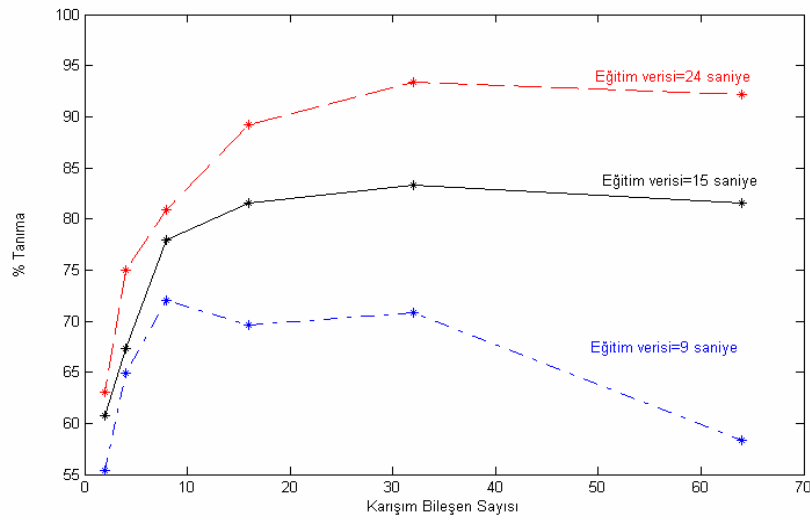
Çizelge 3.10 GKM’in 24 saniye eğitilmesi ile elde edilen konuşmacı tanıma oranları (%)

Eğitim süresi	Karışım bileşen sayısı	Test süresi		
		1 saniye	3 saniye	6 saniye
24 saniye	M=2	63.1	97.0	100
	M=4	75.0	96.4	98.2
	M=8	80.9	99.4	100
	M=16	89.2	99.4	100
	M=32	93.4	99.4	100
	M=64	92.2	99.4	100

Kepstrum katsayı sayısı 24, konuşmacı sayısı 168, TIMIT veritabanı

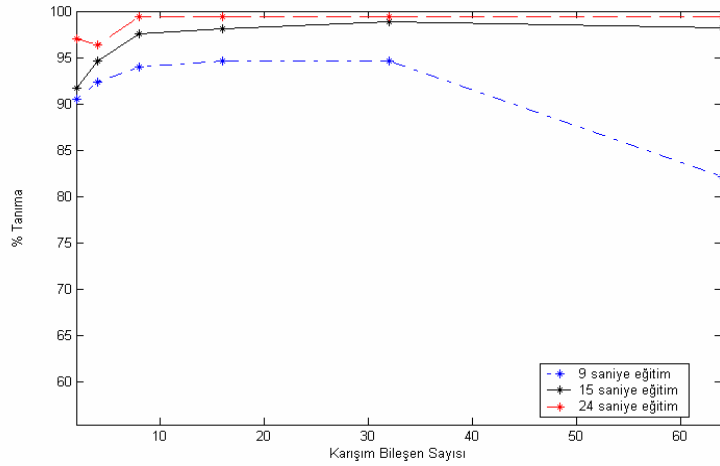
Eđitim süresi 24 saniye'ye ıkarıldıđında model dereceleri arasında zellikle 3 ve 6 saniyelik test süreleri için yakın sonuçlar elde edilmektedir. Ayrıca eđitim süresinin arttırılması ile düşük karışım sayılarında bile tanıma oranları artmaktadır. izelgelerden de görüleceđi üzere konuşmacıların Gauss karışım modeli oluşturulurken kullanılan eđitim sürelerinin arttırılması, konuşmacı tanıma oranını arttırmaktadır. Eđitim süresinin arttırılması ile konuşmacının ses karakteristiđi daha iyi modellenmektedir.

Şekil 3.9'da test süresi 1 saniye için 3 farklı eđitim süresinin karşılaştırılması görülmektedir.



Şekil 3.9 TIMIT için elde edilen üç farklı eđitim süresine bađlı olarak 1 saniye uzunluđunda test ifadesi için konuşmacı tanıma oranları (%)

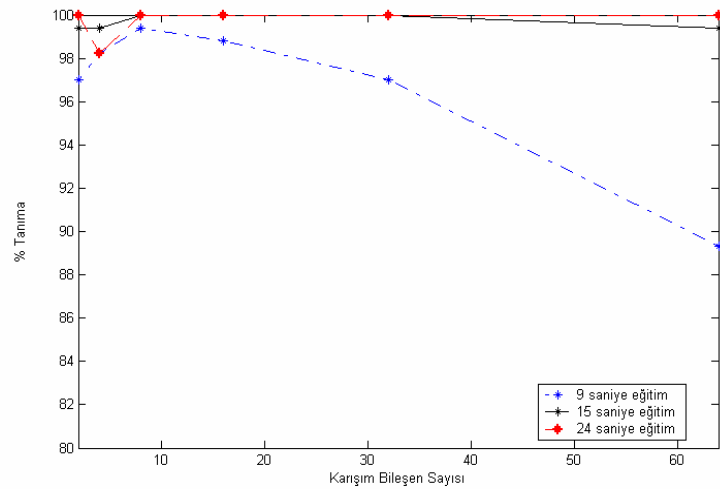
Şekil 3.9'da görüleceđi üzere, Gauss karışım sayısı 32 için 3 eđitim süresi karşılaştırıldıđında, eđitim süresi 9 saniye için konuşmacı tanıma oranı % 70.8, eđitim süresi 15 saniye için % 83.3, eđitim süresi 24 saniye için % 93.4 olmaktadır. Test süresi 1 saniye için, eđitim süresinin artışı ile konuşmacı tanıma oranında hızlı bir artış gözlenmektedir. Şekil 3.10'da test süresi 3 saniye kullanıldıđı durumda 3 farklı eđitim süresi için konuşmacı tanıma oranları görülmektedir.



Şekil 3.10 TIMIT için elde edilen üç farklı eğitim süresine bağlı olarak 3 saniye uzunluğunda test ifadesi için konuşmacı tanıma oranları (%)

Şekil 3.9 ve şekil 3.10 karşılaştırıldığında; eğitim süresi en düşük değeri 9 saniye alınıp test süresi 3 saniye için elde edilen değerler ile eğitim süresi 24 saniye alınıp test süresi 1 saniye için elde edilen değerlerden daha iyidir. Şekillerden test süresinin konuşmacı tanıma başarımına etkisi daha net görülmektedir.

Şekil 3.11’de ise test süresi 6 saniye için üç değişik eğitim süresi için konuşmacı tanıma oranları görülmektedir. Test süresi 6 saniye olduğu durumda eğitim süresi 15 saniye ve eğitim süresi 24 saniye olduğu durumlarda konuşmacı tanıma oranı % 100’e yaklaşmaktadır.



Şekil 3.11 TIMIT için elde edilen üç farklı eğitim süresine bağlı olarak 6 saniye uzunluğunda test ifadesi için konuşmacı tanıma oranları (%)

Şekil 3.9, şekil 3.10 ve şekil 3.11'den görüleceği üzere TIMIT veritabanı ile yapılan deneylerde üç farklı test süresi içinde eğitim sürelerinin artışına paralel olarak konuşmacı tanıma oranları artmaktadır. Sonuç olarak konuşmacı tanıma oranları kullanılarak test süresi, eğitim süresinden daha etkili olmaktadır.

NTIMIT veritabanı kullanılarak eğitim ve test süresi değişiminin konuşmacı tanıma oranına etkisi incelenecektir. Deneyde 9, 15 ve 24 saniye olmak üzere 3 farklı eğitim süresi kullanılmaktadır. Kepstrum katsayıları elde edilmesinde konuşmalar, 25 ms uzunluğunda çerçevelere ayrılıp 10 ms de bir bu çerçevelere hamming pencereleme uygulanır. Her bir çerçeve, doğrusal ölçekte dizilmiş süzgeç dizilerinden geçirilip ayrık kosinüs dönüşümü uygulanır. Her bir çerçeve 20 boyutlu kepstrum katsayıları ile ifade edilmektedir. Konuşmacı modelleri 2, 4, 8, 16, 32 ve 64 bileşenli Gauss yoğunlukları kullanılmaktadır. Eğitim için 15 BM özyineleme kullanılıp değişinti sınırlaması olarak $\sigma^2_{\min}=0.01$ değeri kullanılmaktadır. Deneyde 1, 3, ve 6 saniye olmak üzere 3 farklı test süresi kullanılmaktadır. Bu parametrelere bağlı olarak elde edilen konuşmacı tanıma oranları çizelge 3.11'de görülmektedir.

Çizelge 3.11 Değişik eğitim süreleri için elde edilen konuşmacı tanıma oranları

Eğitim Süresi	Model Derecesi	Test süresi		
		1 saniye	3 saniye	6 saniye
9 saniye	<i>M=8</i>	15.48	33.93	49.40
	<i>M=16</i>	17.26	36.31	45.83
	<i>M=32</i>	14.88	35.12	38.10
15 saniye	<i>M=8</i>	21.43	51.19	60.12
	<i>M=16</i>	23.81	53.57	67.26
	<i>M=32</i>	24.40	58.33	66.67
24 saniye	<i>M=8</i>	25	60.71	74.40
	<i>M=16</i>	31.55	64.88	81.55
	<i>M=32</i>	32.84	70.24	85.71

Kepstrum katsayı sayısı 20, konuşmacı sayısı 168, NTIMIT veritabanı

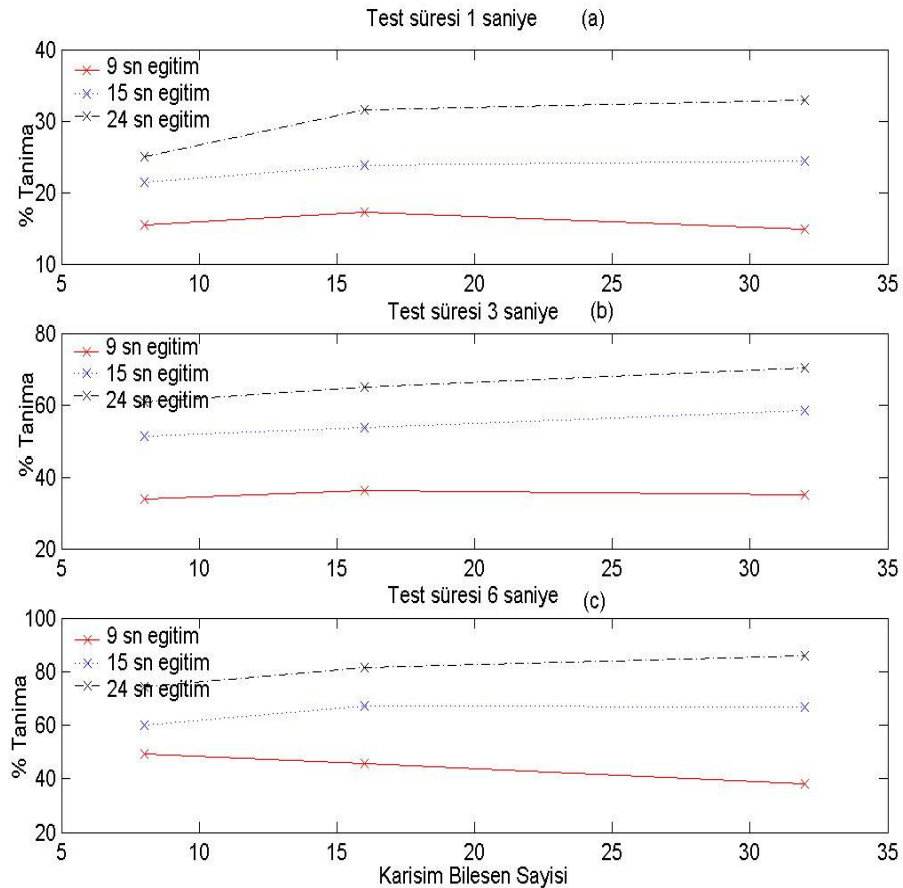
Yeteri kadar eğitim verisi bulunmayan veritabanları için karışım bileşen sayısının seçimi daha önemli olmaktadır. Bu durum şekil 3.12'de konuşmacı model değişimlerine bağlı olarak 3 farklı eğitim süresi için elde edilen tanıma oranları ile görülmektedir.

NTIMIT veritabanı için eğitim verisinin artışı ile tanıma başarımı artmaktadır. 1 saniye test süresi için konuşmacı tanıma oranı % 15 ile % 32 arasında değişirken test

süresi 3 saniye için ortalama 30 puan artış ile tanıma oranı % 70'e kadar çıkmaktadır. Test süresi 6 saniye alındığında ise tanıma oranı % 85'e kadar çıkmaktadır.

NTIMIT veritabanı gibi gürültü içeren veritabanları için test süresinin yüksek alınması tanıma başarımında önemli oranda artış (yaklaşık 15 puan) sağlarken, TIMIT veritabanı gibi gürültüsüz veritabanları için ise test süresinin 3 saniye yerine 6 saniye alınması tanıma oranını sadece 0.6 puan arttırmaktadır.

TIMIT ve NTIMIT veritabanları ile yapılan çalışmalar (Reynolds ve ark. 1995, Jankowski ve ark. 1995, Liu ve ark. 1996) incelendiğinde test süresi olarak 2 cümleye karşılık gelen 6 saniye uzunluğunda test süresi kullanılmayıp, genellikle yaklaşık 3 saniye uzunluğuna karşılık gelen 1 cümle kabul görmektedir. Bunun nedenini 6 sn lik veri örneği almanın geçek zamanlı sistemlerde her zaman mümkün olamayabileceği şeklinde açıklayabiliriz.



Şekil 3.12 Eğitim sürelerinin değişimine bağlı olarak (a) test süresi 1 saniye (b) test süresi 3 saniye (c) test süresi 6 saniye için konuşmacı tanıma oranları

3.2.2.3 Konuşmacı sayısı değişimi

Konuşmacı sayısına bağlı olarak GKM konuşmacı tanıma sisteminin başarımlarını incelemek için TIMIT veri tabanındaki konuşmacıların tamamı 630 kişi (438 erkek, 192 kadın) bu deneyde kullanılmıştır. Konuşmacı ses örneklerinden eğitim ve test için ayrı ayrı 24 boyutlu keştrüm katsayılı öznelik vektörleri çıkarılır. Eğitim süresi olarak 5 ve 8 olmak üzere iki farklı cümle kullanılmıştır. Test için 1 Sx cümlesi (~ 3 sn) kullanılır. Karışım bileşen sayısı 32 alınır. Eğitim, 15 BM özyineleme ile $\sigma^2_{\min}=0.01$ değışinti sınırlaması yapılır. 630 konuşmacının her biri GKM konuşmacı tanıma sistemi ile test edilerek elde edilen sonuç çizelge 3.12’de görölmektedir.

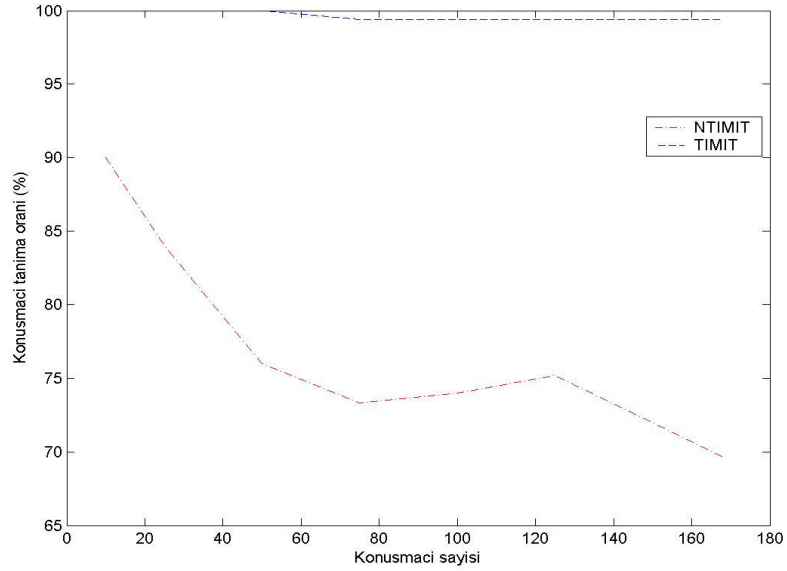
Çizelge 3.12 TIMIT veritabanının tamamı için konuşmacı tanıma oranı (%)

Eğitim Süresi	Test süresi	Konuşmacı tanıma oranı (%)
15 sn	3 sn	98.8
24 sn	3 sn	99.4

Karışım bileşen sayısı 32, konuşmacı sayısı 630, TIMIT veritabanı

Çizelge 3.12’den anlaşılacağı üzere çok fazla sayıda konuşmacı için konuşmacı tanıma başarımlarında düşme olmamaktadır. Diğer veri tabanlarında (örneğin KING) konuşmacı sayısı arttıkça tanıma oranı düşmektedir (Reynolds 1992). Bu durum daha çok TIMIT veritabanının yapısı ile ilgilidir. TIMIT veritabanında konuşmacılardan toplanan sesler çok temiz (gürültüsüz) olması, konuşmacıların farklı bölgelerden olması ve konuşmacılara ait konuşma-sessizlik arasının düzgün dağılması konuşmacı tanıma başarımlarının yüksek olmasında etkili olmaktadır.

Her iki veritabanı için 168 kişilik test kümesi ile konuşmacı sayısına bağlı olarak konuşmacı tanıma oranı ölçülecektir. Konuşmacı modeli 32 Gauss karışım bileşeni ile 2 Sa cümleleri, 3 Si cümleleri, 3 Sx cümleleri kullanılarak oluşturulmaktadır. TIMIT veritabanında, test için kalan 2 Sx cümlesinin ilk 3 saniyelik kısmı kullanılmaktadır. TIMIT veritabanı için her bir çerçeveye karşılık 24, NTIMIT veritabanı için ise 20 Mel keştrüm katsayısı kullanılmaktadır. NTIMIT veritabanında süzgeç dizileri, 300-3400 Hz arasına yerleştirilmektedir. Deneyde kullanılan konuşmacı sayıları 10, 25, 50, 75, 100, 125, 150, 168 şeklinde alınmaktadır. Bu durumda elde edilen konuşmacı tanıma oranları şekil 3.13’de görölmektedir.



Şekil 3.13 Konuşmacı sayısına bağlı olarak test kümesi için konuşmacı tanıma oranları (%)

TIMIT veritabanı konuşmacı sayısının artmasından çok az etkilenmektedir. TIMIT veritabanında ilk 50 kişide tanıma oranı % 100 iken bu oran konuşmacı sayısı artması ile % 99.4'e düşmektedir. NTIMIT veritabanında ise, 10 kişi için tanıma oranı % 90, 100 iken tanıma oranı % 74 olmakta, 168 kişi için tanıma oranı % 69.64'e kadar düşmektedir. NTIMIT veritabanında telefon hattı bozulmalarından dolayı tanıma oranı konuşmacı sayısının artmasıyla TIMIT veritabanına göre 30 puan düşmektedir.

Her iki veritabanında 10 cümlelerin 8'i eğitim için, 2'si test için kullanılmaktadır. Test için 2 Sx cümleleri kullanılmaktadır. Bu 2 cümle yaklaşık 6 saniye uzunluğundadır. Test için genellikle daha kısa cümleler kullanılmaktadır (Reynolds ve Rose 1995). NTIMIT veritabanında, test için 2 Sx cümlelerinden 4 çeşit test süresi kullanılacaktır. Bunlar; 2 Sx cümlesinin ilk 3 saniye, son üç saniyesinin test için alınması ve 2 Sx cümlesinin ilk ve son cümlelerinin test için kullanılmasıdır. Bu 4 durumda elde edilen konuşmacı tanıma oranları çizelge 3.13'de görülmektedir.

Öznitelik vektörleri elde edilirken konuşma 25 msn'lik çerçevelere ayrılıp 10 msn ilerlemeye tabii tutulmakta ve Hamming pencereleme uygulanmaktadır. süzgeçler 300-3400 Hz arasında Mel ölçeğe dizilmekte ve 28 adet üçgen süzgeç dizisi kullanılmaktadır. Sonuçta 20 boyutlu Mel ölçek kepstrum katsayıları elde edilmektedir.

Çizelge 3.13 Test süresi kullanılış biçimlerine göre tanıma oranı değişimi (%)

Konuşmacı sayısı	Test süresi alma yöntemi			
	İlk 3 sn	Son 3 sn	İlk cümle	Son cümle
10	80	90	80	90
25	72	84	72	84
50	62	64	58	76
75	68	69.33	64	73.33
100	70	71	66	74
125	72	72	66	75.2
150	72.66	70.67	68	72
168	69.64	68.45	64.88	69.64

Eğitim süresi 24 sn, kepstrum katsayı sayısı 20, karışım bileşen sayısı 32, NTIMIT veritabanı

Çizelge 3.13'den görüleceği üzere 4 değişik test süresi alma yöntemi içinde konuşmacı sayısı arttıkça konuşmacı tanıma oranı azalmaktadır. Veritabanında konuşmalar işlenirken ilk olarak kadın konuşmacılar (46 kişi) daha sonra erkek konuşmacılar sırayla işlenmektedir. İlk 50 kişinin 46'sını kadınlar oluşturmaktadır. Bu nedenle ilk 50 kişi ile yapılan deneylerde göreceli bir konuşmacı tanıma oranında düşüş gözlenmektedir. 168 kişi ile yapılan testte en yüksek tanıma oranı % 69.64 ile 2. Sx cümlesinin ilk 3 saniyesi ve 2. Sx cümlesi kullanıldığı durumlarda elde edilmektedir. Bundan sonraki çalışmalarda test için iki Sx cümlesinin ilk 3 saniyesi kullanılacaktır.

3.3 Öznitelik Vektörü Çıkartma ve Parametre Kestirimi

Konuşma spektrumunun öznitelik olarak kullanılma yöntemleri değişim göstermektedir. Yaygın spektrum gösterim yöntemleri; bölüm 2.3'de tanımlanan doğrusal öngörü katsayıları ve onun değişik dönüşümleri, süzgeç dizisi enerjileri ve onun kepsral gösterimleri sayılabilir. Doğrusal öngörü katsayılarının gösterimi, konuşmada gürültü olması durumunda konuşmanın spektral karakteristiğini modellemede yetersiz kalmaktadır (Reynolds ve Rose 1995). Kepstrum katsayıları elde edilirken farklı frekans bantların enerjileri doğrudan ölçülür ve herhangi bir model sınırlamasına bağlı değildir. Bununla birlikte süzgeçlerin bant genişlikleri ve merkez frekansları, kulağın seçici olduğu kritik bantlara uygun olarak ayarlanabilir. Bu sayede konuşma işaretinin önemli karakteristikleri daha iyi tutulur. Mel ölçek süzgeç dizisi enerjilerinin kepsral gösterimi konuşmacı tanıma için istenen öznitelik katkısını sağlar (Davis ve Mermelstein 1980, Reynolds 1992). Bu öznitelikler, bir dizi işaret işleme

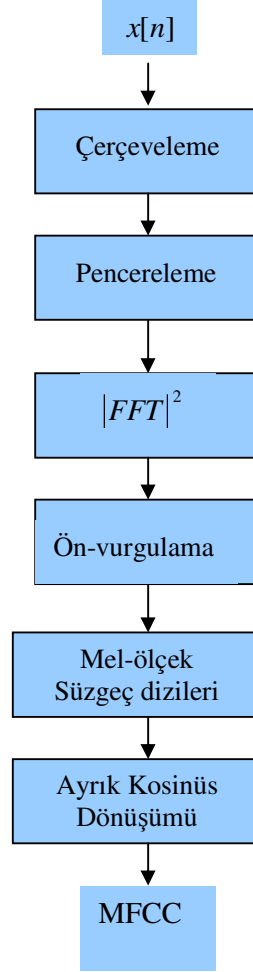
süreci kullanılarak çıkartılır. Konuşmacı tanıma sisteminin en önemli kısmı öznitelik vektörü elde etme işlemidir. Bu nedenle bu işlem adım adım incelenip her bir öznitelik vektörü parametresinin konuşmacı tanıma üzerine etkisi araştırılmaktadır. Tüm eğitim ve sınıflandırma adımları değişmeden sadece kullanılan öznitelik vektörleri değiştirilerek, öznitelik vektör setleri arasında kontrollü bir karşılaştırma yapılmaktadır.

3.3.1 Mel ölçek kepstrum katsayıları

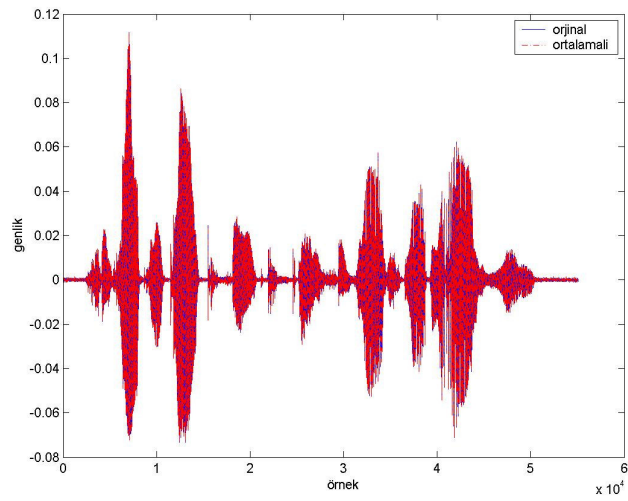
Öznitelik vektörü olarak kullanılan kepstrum katsayıları elde edilirken, genellikle konuşmacı tanıma uygulamalarında MFCC kullanılır (Matsui ve ark 1995). Bunun nedeni, MFCC insan kulağının frekans seçiciliğini taklit ederek iyi bir şekilde konuşmacıları ayırt edici değerler elde edilmesidir. Reynolds (1992), tarafından önerilen MFCC vektörü çıkartımı blok diyagramı şekil 3.14'de görülmektedir.

Bazı çalışmalarda (Davis ve Mermelstein 1980), Ön vurgulama çerçevelemeden önce uygulanıp, pencerelemeden sonra işaretin $|FFT|^2$ yerine $|FFT|$ 'si alınmakta ve farklı bir Mel ölçekte dizilmiş üçgen süzgeç dizileri kullanılmaktadır. MFCC elde edilirken kullanılan bu farklı yöntemlerin konuşmacı tanıma başarımına etkisi bu bölümde incelenmektedir.

Konuşmacı tanıma sisteminde ilk olarak veritabanındaki konuşmacılara ait cümleler hafızada kaydedildiği yerden okunur ve ortalama değerinden çıkarılır. Şekil 3.15'de TIMIT veritabanında bir konuşmacıya ait işaret ve ortalaması alınmış hali görülmektedir. Daha sonra öznitelik vektörlerinin çıkartımı için şekil 3.14'deki işlemler uygulanır.



Şekil 3.14 MFCC çıkarma işleminin blok diyagramı

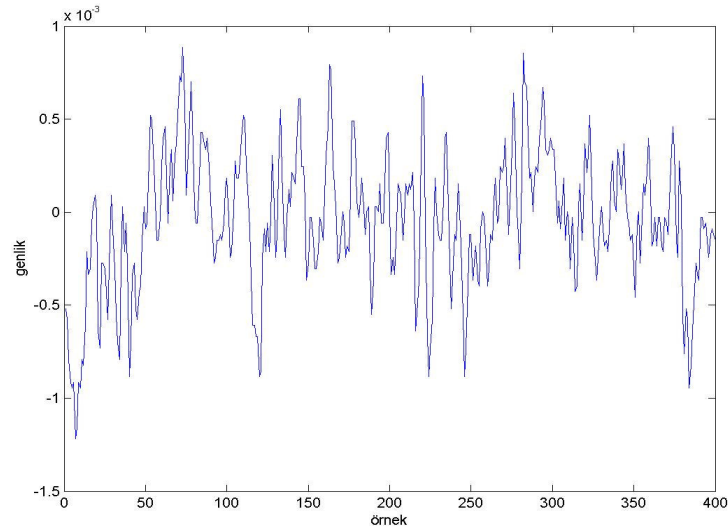


Şekil 3.15 Bir konuşma ve ortalaması alınmış hali

3.3.1.1 Çerçeveleme

Ses üretim organlarının sözcüklere bağlı olarak yer değiştirmesinden dolayı konuşma işareti de sürekli olarak değişir. Konuşma işareti, parametrelerin sabit kaldığı kabul edildiği çerçeve olarak adlandırılan küçük parçalara ayrılmalıdır. Çünkü tüm işaret boyunca FFT hesaplanırsa, farklı fonemlere ait spektral bilgilerin tutulmasında kayıplar oluşur. Tüm işaretin FFT'sini almak yerine çerçevenin FFT'si hesaplanır. Çerçeve uzunluğu 10-30 msn arasında değişir. Bu aralıkta konuşma oldukça sabit akustik karakteristik gösterir (Karpov 2003). Her bir çerçeveye örtüşme uygulanır. Çerçevelerin örtüşme oranı, çerçeve uzunluğunun % 30'u ile % 75' i arasında alınır (Kinnunen 2003). Örtüşme uygulanması ile çerçeve sonundaki işaretin önemlerini kaybetmemesi sağlanır.

Konuşma örneğinden ortalaması çıkartıldıktan sonra, konuşma değişimlerine karşı sabit kabul edilebilecek parçalar şu şekilde ifade edilir. Konuşma işareti N adet örnekten oluşan çerçevelere ve komşu M örnekten oluşan çerçevelere bölünür. ($M < N$) İlk çerçeve N örnekten oluşurken ikinci çerçeve ilk çerçeve den M örnek sonra başlar ve ilk çerçevenin $N-M$ çerçeve kadar üzerine biner (Rabiner ve Juang 1993). Şekil 3.16'da 25 msn (400 örnek) uzunluğunda konuşma parçası görülmektedir.



Şekil 3.16 Yirmi beş msn uzunluğunda konuşma parçası

Çerçeveleme süresinin değişiminin konuşmacı tanıma başarımına etkisi incelenecektir. Konuşmacı tanıma sistemine ait parametreler şu şekilde alınmaktadır.

TIMIT ve NTIMIT veritabanlarının test dizininden 168 kişinin her birine ait 10 cümleden 8'i (yaklaşık 24 saniye) eğitim için, kalan 2 cümlenin 3 saniye uzunluğundaki kısmı test için kullanılmaktadır. Gauss karışım sayısı 32 alınıp eğitim için BM algoritması kullanılır. Eğitim için 15 BM özyineleme kullanılıp değişinti sınırlaması olarak $\sigma^2_{\min}=0.01$ değeri kullanılmaktadır. Model başlangıcı olarak VN algoritması kullanılmaktadır. MFCC elde edilmesinde çerçevelerin örtüşme oranı 10 msn alınıp, çerçevelere Hamming pencereleme uygulanmaktadır. Pencereleyen sesin 512 örnek FFT'si alınıp, Slaney (1998) tarafından tanımlanan Mel ölçekte, üçgen süzgeç dizilerinden geçirilir. Süzgeçten geçirilen işaretin logaritması alınıp ayrık kosinüs dönüşümü alınır. Her bir çerçeveye karşılık olarak TIMIT veritabanı için 24, NTIMIT veritabanı için 20 boyutlu öznelik vektörleri kullanılmaktadır.

En ideal çerçeveleme süresi veritabanlarına ve kullanılan yöntemlere bağlı olarak değişmektedir. Çerçeveleme sürelerine bağlı olarak elde edilen konuşmacı tanıma oranları çizelge 3.14'deki gibidir.

Çizelge 3.14 Çerçeveleme sürelerinin konuşmacı tanımaya etkisi (%)

Veritabanları	Çerçeveleme süreleri (msn.)			
	30	25	20	15
TIMIT	99.4	99.4	99.4	99.4
NTIMIT	69.05	68.45	69.64	70.83

Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168

Çizelge 3.14'den görüleceği üzere TIMIT veritabanı için çerçeveleme süresi değişimi konuşmacı tanıma başarımını değiştirmez iken, NTIMIT veritabanı için 15 msn çerçeveleme süresi en yüksek tanıma oranını vermektedir.

Kolay uygulanabilir olmasından dolayı çerçeve uzunluğu genellikle sabit alınır. Oysaki sabit çerçeve uzunluğu, konuşma esnasında oluşan sesteki değişimleri tam olarak tutamaz. Perde periyodu değişimi (Huang ve ark. 2001), ardışıl çerçeve parametreleri arasında öklit uzaklığı hesabının ölçülmesi (Zhu ve Alwan 2000) gibi değişik metotlar ile uyarlamalı çerçeve uzunluğu kullanılabilir.

3.3.1.2 Pencereleme

Mel frekansı kepstum katsayılarını elde etmek için ikinci yapılan işlem pencerelemedir. Pencerelemenin amacı çerçeveleme işlemi sonucunda oluşan spektral etkilerin azaltılmasıdır. Pencereleme ile çerçevelerde süreksizliğin önüne geçilir (Rabiner ve Juang 1993). Bu sayede sesin orta bölgeleri güçlendirilirken kenar bölgeleri zayıflatılır. Yaygın olarak kullanılan Hamming, Hanning, Blackman, Gauss, dikdörtgen ve üçgen pencereleme fonksiyonlarının matematiksel ifadeleri aşağıdaki gibidir.

Hamming:

$$w[k+1] = 0.54 - 0.46 \cos\left(2\pi \frac{k}{n-1}\right) \quad k = 0, \dots, n-1 \quad (3.18)$$

Hanning:

$$w[k+1] = 0.5 \left(1 - \cos\left(2\pi \frac{k}{n-1}\right)\right), \quad k = 0, \dots, n-1 \quad (3.19)$$

Blackman:

$$w[k+1] = 0.42 - 0.5 \cos\left(2\pi \frac{k}{n-1}\right) + 0.08 \cos\left(4\pi \frac{k}{n-1}\right), \quad k = 0, \dots, n-1 \quad (3.20)$$

Gauss:

$$w[k+1] = e^{-\frac{1}{2} \left(\alpha \frac{k-N/2}{N/2}\right)^2} \quad 0 \leq k \leq N \quad \text{ve} \quad \alpha \geq 2 \quad (3.21)$$

Dikdörtgen:

$$w[k+1] = 1, \quad k = 0, \dots, n-1 \quad (3.22)$$

Üçgen:

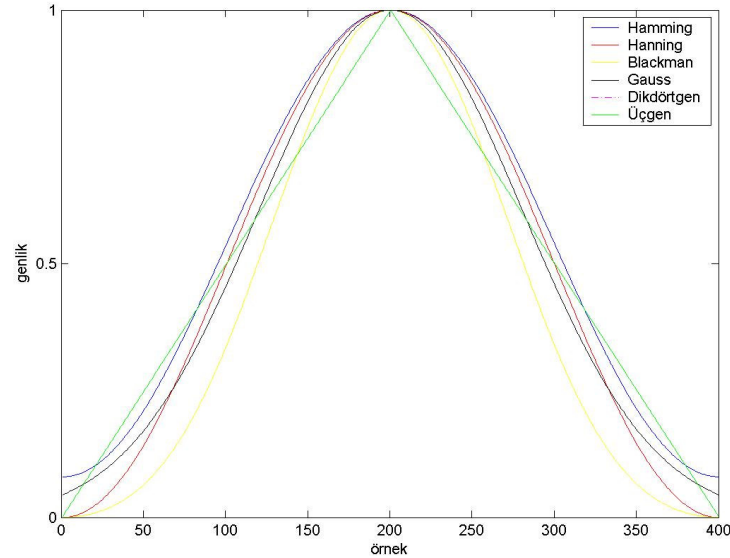
n tek için;

$$w[k] = \begin{cases} \frac{2k}{n+1}, & \dots, 1 \leq k \leq \frac{n+1}{2} \\ \frac{2(n-k+1)}{n+1}, & \dots, \frac{n+1}{2} \leq k \leq n \end{cases} \quad (3.23)$$

n çift için;

$$w[k] = \begin{cases} \frac{2k-1}{n}, & \dots, 1 \leq k \leq \frac{n}{2} \\ \frac{2(n-k+1)}{n}, & \dots, \frac{n}{2} + 1 \leq k \leq n \end{cases} \quad k = 0, \dots, n-1 \quad (3.24)$$

Şekil 3.17’de bu pencereleme fonksiyonlarının çerçeve süresi 400 örnek için eğrileri verilmektedir.



Şekil 3.17 Pencereleme fonksiyonları

Konuşmacı tanıma sisteminde başarıyı en yüksek pencereleme fonksiyonunu bulmak için Hamming, Hanning, Blackman, Gauss, dikdörtgen ve üçgen pencereleme fonksiyonları çerçevelere uygulanmaktadır. Konuşmacı tanıma sistemi parametreleri bir önceki deneyle aynı alınmıştır. Çerçeveleme süresi her iki veritabanı için de 20 msn alınmıştır. Pencereleme fonksiyonlarına bağlı olarak elde edilen konuşmacı tanıma oranları Çizelge 3.15’deki gibidir.

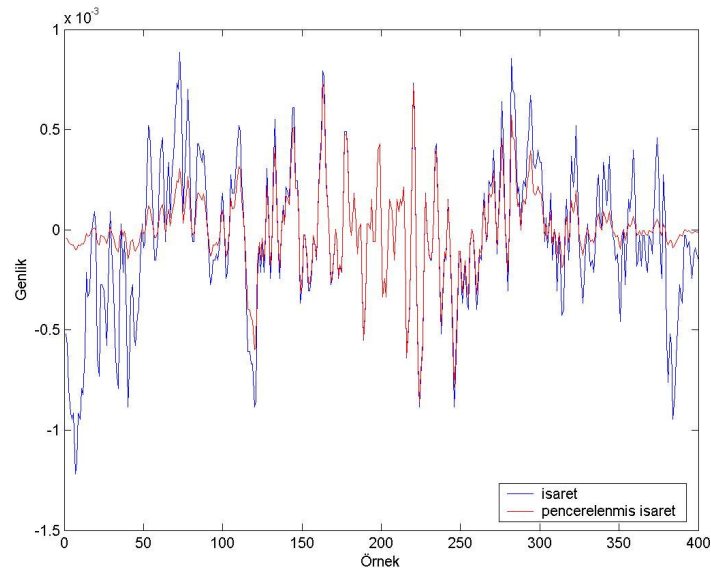
Çizelge 3.15 Pencereleme fonksiyonlarına bağlı olarak konuşmacı tanıma oranları (%)

Pencereleme fonk.	Veritabanları	
	TIMIT	NTIMIT
Hamming	99.4	69.64
Hanning	99.4	70.24
Blackman	99.4	66.67
Gauss	99.4	70.83
Dikdörtgen	99.4	64.88
Üçgen	99.4	70.83

Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168

Çizelge 3.15'den görüleceği üzere TIMIT veritabanı için pencereleme fonksiyonlarının değişimi ile tanıma başarımı değişmemektedir. NTIMIT veritabanı için ise Gauss ve üçgen pencereleme fonksiyonları kullanılarak en yüksek konuşmacı tanıma başarımı elde edilmiştir. En düşük tanıma başarımı pencereleme uygulanmama durumuna karşılık gelen dikdörtgen pencereleme ile elde edilmiştir.

Şekil 3.18'den görüleceği üzere Hamming pencerelenen bir çerçevelik konuşma parçası, sıfıra yakın bir değer ile başlayıp çerçeve süresinin yaklaşık 1/3'ünden itibaren çerçevelenen işaretin değerlerini takip etmekte ve sıfıra yakın bir değer ile sonlanmaktadır. Bu şekilde çerçevelerin sonunda oluşacak ani değişimlerin önüne geçilir (Karpov 2003) ve NTIMIT veritabanı için pencereleme uygulanmadığı duruma göre daha yüksek tanıma başarımı sağlanmaktadır.



Şekil 3.18 Konuşma parçası ve Hamming pencereden geçirilmiş hali

3.3.1.3 Hızlı Fourier Dönüşümü (FFT)

MFCC elde edilmesinde, pencereden geçirilen işaretin genlik spektrumu FFT ile hesaplanır. FFT ile N örnekten oluşan zaman alanındaki her bir çerçeve, frekans alanına çevrilir. FFT, ayrık fourier dönüşümünden üretilmiştir. N örnek $\{x_n\}$ olarak denklem 3.23'deki gibi tanımlanır.

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-2\pi jkn/N}, \quad n = 0,1,2,\dots,N-1 \quad (3.25)$$

Burada genellikle X_k 'ler kompleks sayılardır. Sonuç olarak elde edilen dizi $\{X_k\}$: sıfır frekansı $k=0$ a karşılık gelip, pozitif frekanslar ($0 < f < f_s/2$), $1 \leq k \leq (N/2) - 1$ değerlerine karşılık gelirken, negatif frekanslar ($-f_s/2 < f < 0$), $(N/2) + 1 \leq k \leq N - 1$ 'e karşılık gelir. Burada, f_s örnekleme frekansıdır (Claudio 1999).

Bir konuşma parçasının FFT'sinin k . harmonik bileşeni $X[k] = X_{re}[k] + jX_{im}[k]$ şeklinde bir kompleks sayı olarak ifade edilsin. Bu ifade kutupsal olarak denklem 3.26'daki gibi tanımlanır.

$$X[k] = |X[k]| e^{j\angle X[k]} \quad (3.26)$$

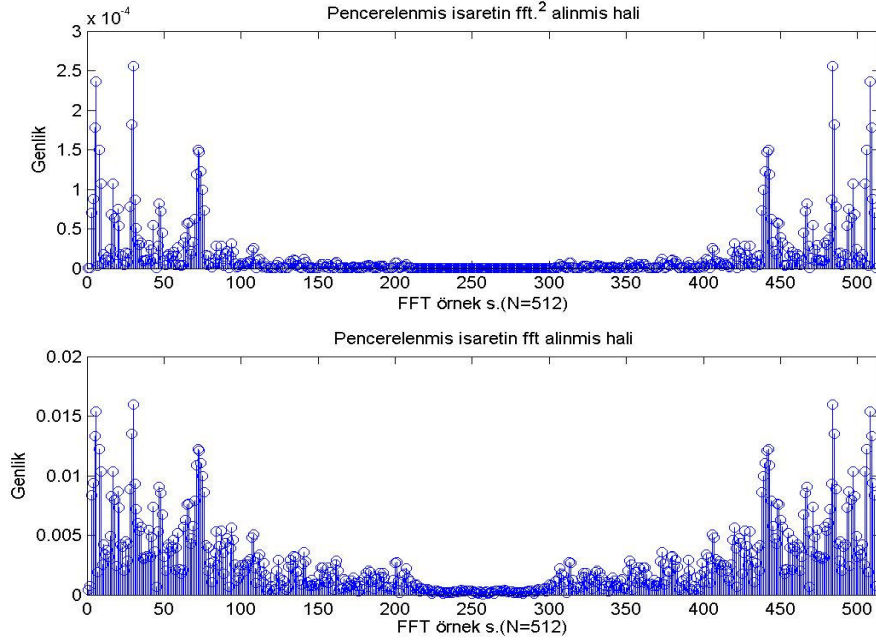
$$|X[k]| = \sqrt{X_{re}[k]^2 + X_{im}[k]^2} \quad (3.27)$$

$$\angle X[k] = \tan^{-1} \left(\frac{X_{im}[k]}{X_{re}[k]} \right) \quad (3.28)$$

Burada, $|X[k]|$ k .harmonik bileşene ait genlik, $\angle X[k]$ ise fazı olarak adlandırılır (Kinnunen 2003). Konuşma gibi gerçel işaretler için genlik spektrumu $N/2$ ile simetriktir. Konuşma analizinde faz spektrumu genellikle ihmal edilir. Çünkü konuşma ile ilgili önemli bilgi taşımamaktadır (Furui 1989).

Bir işaretin FFT'si hesaplanırken işaretin uzunluğu 2^M $M \in N_+$ şeklinde başka bir deyişle 2'nin kuvvetleri şeklinde olmalıdır. Örneğin işaret 400 örnekten oluşuyorsa işaretin uzunluğu 512 olana kadar işarete sıfır eklenir ve bu şekilde FFT'si hesaplanır. İşaretin başına veya sonuna sıfır eklenmesi FFT sonucunu değiştirmez.

Konuşmacı tanıma ile ilgili yapılan bazı çalışmalarda (Reynolds 1992, Reynolds ve Rose 1995, Sarma 1997, Besacier ve Bonastre 1998, Slaney 1998) FFT'nin güç spektrumu ($|FFT|^2$) alınmaktadır. Şekil 3.19'da NTIMIT veritabanında 25 msn'lik pencerelemiş konuşma parçasının $|FFT|^2$ ve $|FFT|$ alınmış hali bulunmaktadır. Şekilde işaretin $N/2$ 'ye göre simetrik olduğu görülmektedir.



Şekil 3.19 Pencerelemiş konuşma parçasının $|FFT|^2$ ve $|FFT|$ alınmış hali

Bir önceki deneyde belirtilen öznelik vektörü üretim, eğitim ve test şartlarında, konuşma işaretinin genlik ve güç spektrumu konuşmacı tanımaya etkisi incelenecektir. Pencereleme yöntemi olarak her iki veritabanı içinde çerçeveleme süresi 20 msn alınıp Hamming pencereleme kullanılmaktadır. FFT kuvvetlerinin konuşmacı tanımaya etkisi çizelge 3.16'da görülmektedir.

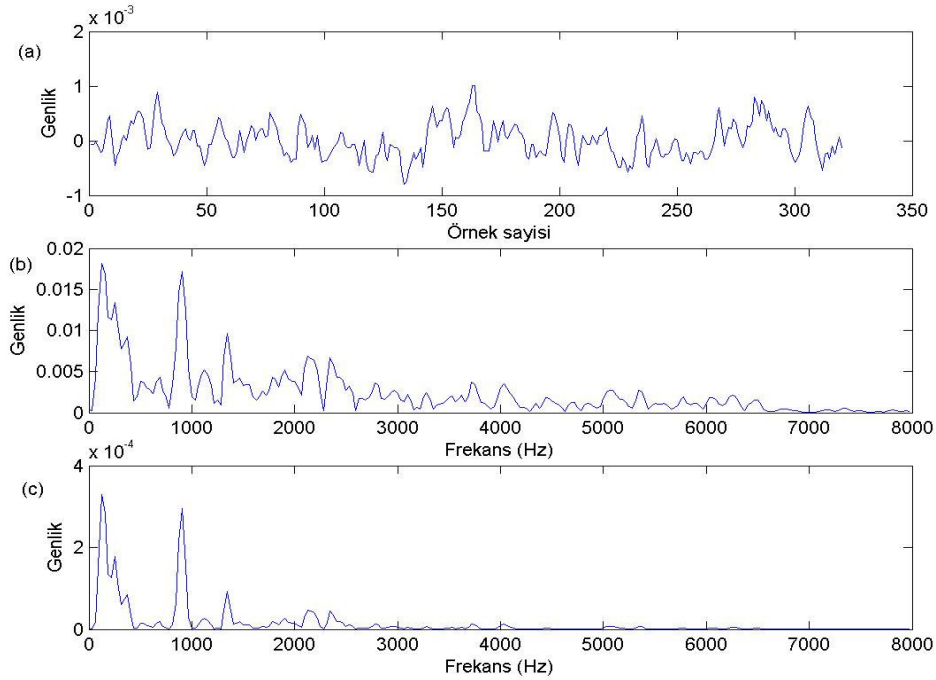
Çizelge 3.16 FFT kuvvetlerinin konuşmacı tanımaya etkisi (%)

Veritabanları	FFT kuvveti	
	$ FFT $	$ FFT ^2$
TIMIT	99.4	99.4
NTIMIT	66.07	69.64

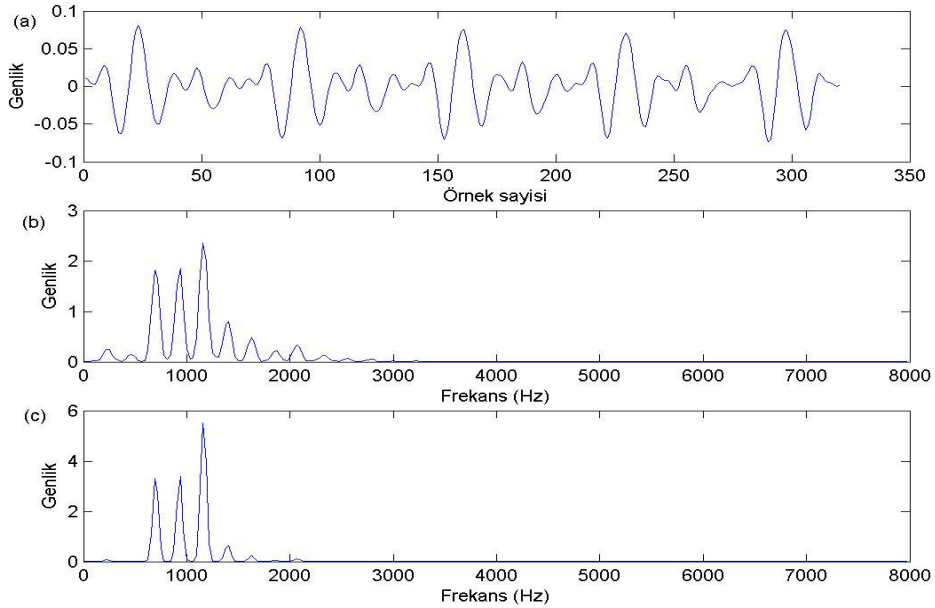
Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168

Çizelge 3.16'dan görüleceği üzere konuşma işaretinin FFT'sinin güç spektrumunu almak TIMIT veritabanı için tanıma başarımını değiştirmemektedir. NTIMIT veritabanı için konuşma işaretinin güç spektrumunu almak genlik spektrumu kullanılmasına göre daha yüksek konuşmacı tanıma başarımı elde edilmesini sağlamaktadır. Çünkü $|FFT|^2$ işlemi konuşma işaretini daha fazla pürüzsüzleştirilmektedir ve konuşmadaki düşük genlikli gürültü bileşenlerinin etkinliğini azaltmaktadır (Ganchev 2005). Bu şekilde düşük yoğunluktaki seslerin, özellikle ünsüz sürtünmeli seslerin zayıflatılması sağlanır. İngilizcedeki ünsüz sürtünmeli sesler sh, zh, jh, ch, s, z, f, th, v, dh olarak tanımlanmaktadır.

20 msn uzunluğunda ünsüz bir konuşma parçası için şekil 3.20'de, ünlü bir konuşma parçası için şekil 3.21'de $|FFT|$ ve $|FFT|^2$ alınmış hali görülmektedir. Şekillerden görüleceği üzere işaretteki düşük genlikli gürültü bileşenlerinin etkinliği azaltılmaktadır



Şekil 3.20 Yirmi msn uzunluğunda (a) ünsüz bir konuşma parçası (b) bu konuşma parçasının $|FFT|$ (c) $|FFT|^2$ alınmış hali



Şekil 3.21 Yirmi ms'n uzunluğunda (a) ünlü bir konuşma parçası (b) bu konuşma parçasının $|FFT|$ (c) $|FFT|^2$ alınmış hali

3.3.1.4 Ön vurgulama

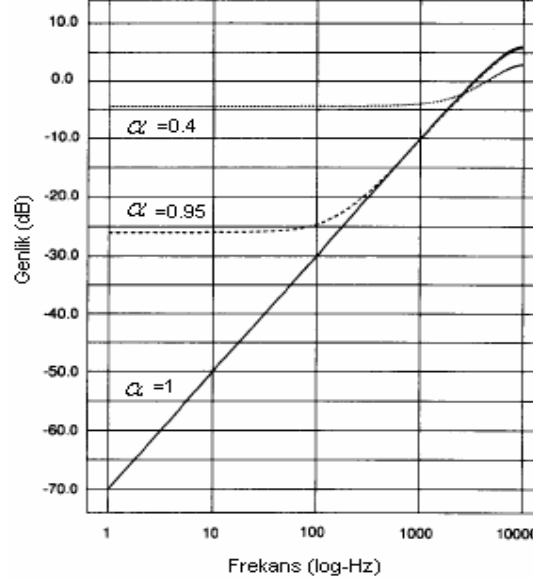
Ön vurgulama ile ses yolunun, yüksek frekansları, -6 dB/oktav zayıflatmasının telafi edilmesi amaçlanır. Ünlü sesler için gırtlak -12 dB/oktav yüksek frekansları zayıflatırken, dudaktan yayılma esnasında bu zayıflama 6 dB/oktav azaltılır. Sonuç olarak ses yolunda toplam -6 dB/oktav'lık zayıflama oluşur (Lincoln 1999). Ünlü sesler için bu zayıflamayı gidermek için genellikle birinci dereceden yüksek geçiren süzgeç kullanılır. Ünsüz sesler için spektrum düzgün olduğundan ön vurgulamaya ihtiyaç olmaz (Kinnunen 2003). Denklem 3.29'da verilen 1. dereceden süzgeç ile işaret 6 dB/oktav iyileştirilir.

$$H(z) = 1 - \alpha z^{-1} \quad (3.29)$$

Burada α , ön vurgulamanın derecesini yansıtır genellikle 0.9 ile 1 arasında alınır (Wildermoth 2001). Süzgeç $y(n)$, fark denklemi olarak denklem 3.30'daki gibi ifade edilir.

$$y(n) = x(n) - \alpha \cdot x(n-1) \quad n = 0, 1, 2, \dots, N-1 \quad (3.30)$$

Denklem 3.29’da verilen birinci dereceden süzgecin α değeri 0.4, 0.95 ve 1 için frekans cevabı şekil 3.22’de görülmektedir. Konuşma analizinde genellikle α değeri 0.95 alınmaktadır (Rabiner ve Juang 1993).

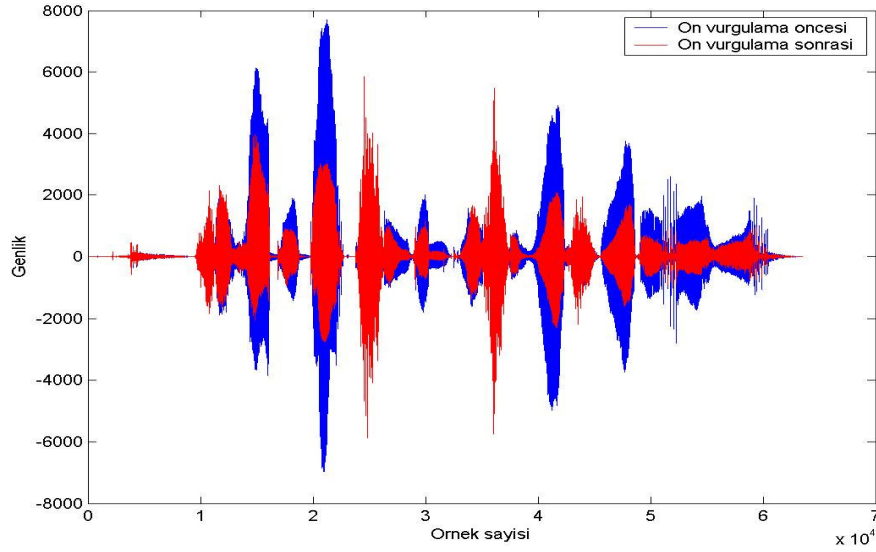


Şekil 3.22 Ön vurgulama süzgecinin değişik α değerleri için frekans cevabı

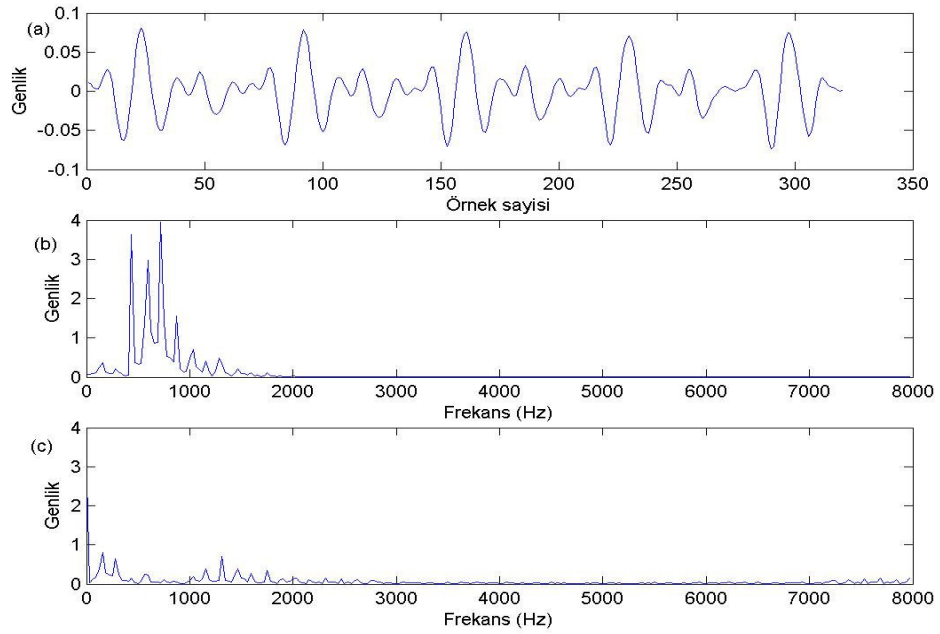
Genellikle konuşma işlemede ön vurgulama işaretin çerçevenmesinden önce, işareti spektral olarak düzleştirmek ve daha sonra oluşacak olan belli etkilere daha az duyarlı hale getirmek için kullanılmaktadır (Rabiner ve Juang 1993). Ön vurgulamanın spektral düzleştirme etkisi DÖK analizinde daha belirgin olarak görülmektedir (Kinnunen 2003).

Bazı konuşmacı tanıma uygulamalarında işaretin çerçevenmesi aşamasından önce ön vurgulama uygulamak yerine güç spektrumu alındıktan sonra ön vurgulama işlemi uygulanmaktadır (Reynolds 1992, Reynolds ve Rose 1995). Bu iki durumda yani birinci olarak işaret çerçevenmeden önce ön vurgulamanın etkisi incelenecek, ikinci olarak işaretin güç spektrumu alındıktan sonra vurgulama işlemi uygulanacaktır.

Konuşma işareti çerçevenmeden önce ön vurgulanması ($\alpha = 0.95$) durumunda işaretin genliğinin zamana bağlı değişimi şekil 3.23’de görülmektedir. Aynı konuşmanın bir çerçevesinin ön vurgulama uygulanmadan ve ön vurgulama uygulandıktan sonra değişimi şekil 3.24’de görülmektedir. Şekillerden görüleceği üzere düşük genlikli bileşenler zayıflatılmakta ve yüksek frekanslı bileşenlerin güçlendirildiği görülmektedir.



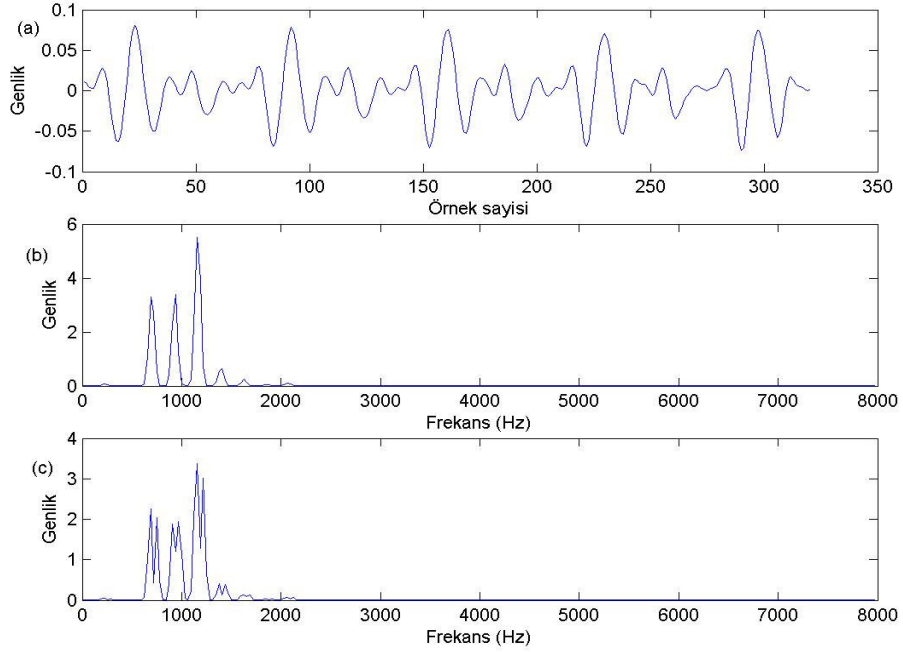
Şekil 3.23 Bir cümleye çerçevelemeden önce ön vurgulama uygulanması



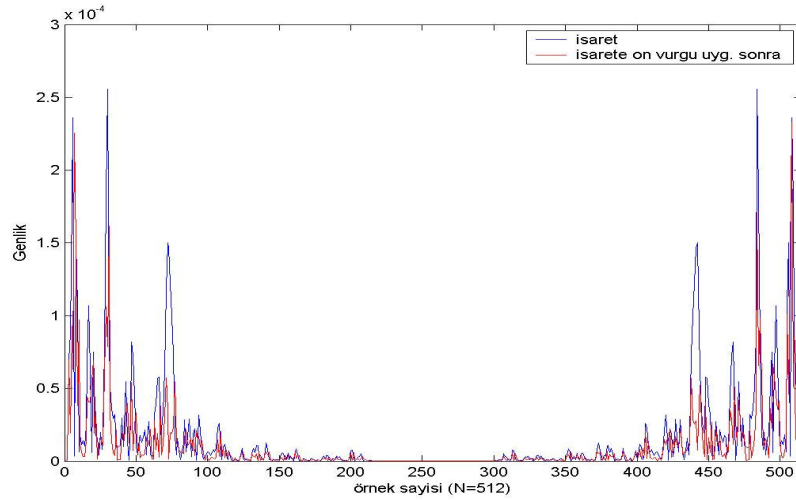
Şekil 3.24 (a)Yirmi ms'n uzunluğunda bir konuşma parçası (b) bu konuşma parçasının ön vurgulanmadan önce genlik spektrumu (c) ön vurgulama uygulandıktan sonra genlik spektrumu

Şekil 3.25'de ise güç spektrumu alınmış işarete ön vurgulanma uygulanması durumunda değişim görülmektedir. Şekil 3.25 (b) ile 3.25 (c) karşılaştırıldığında işaretin düşük frekanslı bileşenlerin genliğinin zayıflatıldığı görülmektedir. Şekil 3.26

da işaretin ön vurgulasız ve ön vurgulama uygulandıktan sonraki halleri üst üste çizdirilmiştir. Şekil'den görüleceği üzere işaretin düşük frekanslı bileşenlerinin genliği zayıflatılırken yüksek frekanslı bileşenlerinde fazla bir değişme olmamaktadır.



Şekil 3.25 (a)Yirmi msn uzunluğunda bir konuşma parçası (b) bu konuşma parçasının $|FFT|^2$ spektrumu (c) spektrumu alınmış işaretin ön vurgulanmış hali



Şekil 3.26 Konuşma parçasının güç spektrumun alındıktan sonra ön vurgulama yapılmadan önce ve sonraki halleri

TIMIT ve NTIMIT veritabanları kullanılarak ön vurgulamanın konuşmacı tanımaya etkisi incelenecektir. Ön vurgulama süzgeci olarak denklem 3.29 kullanılıp ve $\alpha = 0.95$ alınmaktadır. Birinci olarak işaret çerçevelemeden önce ön vurgulamanın etkisi incelenecektir. İkinci olarak işaretin güç spektrumu alındıktan sonra vurgulama işleminin konuşmacı tanımaya etkisi incelenecektir. Son olarak ön vurgulama uygulanmamasının konuşmacı tanımaya etkisi incelenecektir. Bir önceki deneyde verilen konuşmacı tanıma sistemi parametrelerine bağlı olarak elde edilen konuşmacı tanıma oranları çizelge 3.17’de görülmektedir.

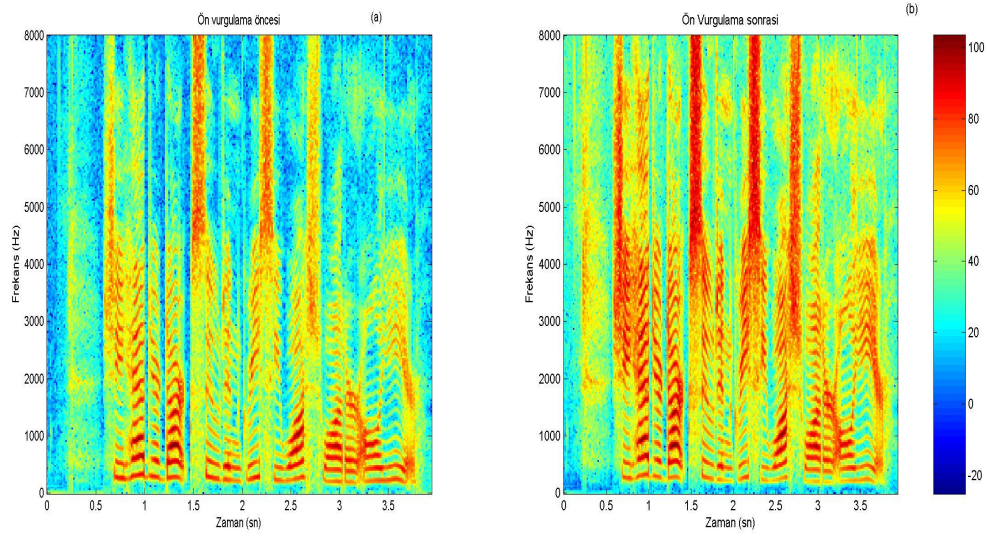
Çizelge 3.17 Ön vurgulamanın konuşmacı tanıma üzerine etkisi (%)

Veritabanları	Ön vurgulama uygulama şekilleri		
	Çerçevelemeden önce	Güç spektrumu alındıktan sonra	Ön vurgulama yok
TIMIT	100	100	99.4
NTIMIT	70.83	67.86	69.64

Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168

Çizelge 3.17’den görüleceği üzere TIMIT veritabanı için çerçevelemeden önce ve FFT alındıktan sonra ön vurgulama uygulandığında en yüksek konuşmacı tanıma başarımına ulaşılmıştır. NTIMIT veritabanı için ön vurgulama çerçevelemeden önce uygulandığı durumda en yüksek konuşmacı tanıma oranı elde edilmiştir. Deneysel çalışma çerçevelemeden önce ön vurgulamanın daha etkili olduğunu göstermiştir.

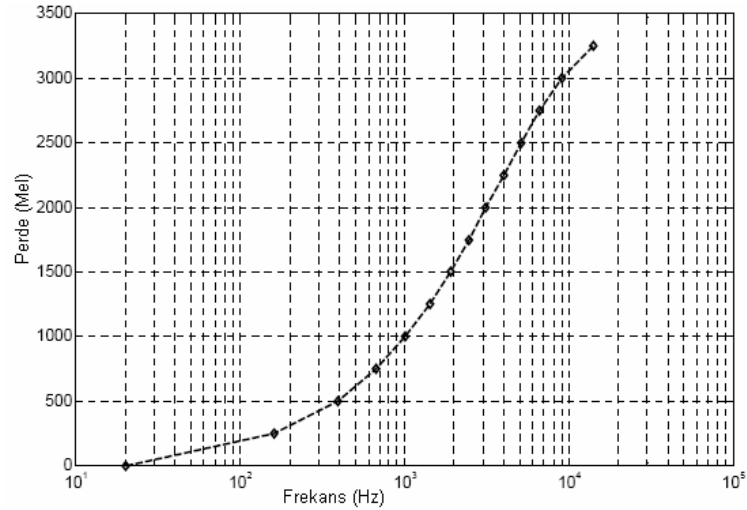
Bir cümle için yüksek frekanslı bileşenlerin güçlendirilmesi zaman-frekans eğrisinde (spektrogram) daha iyi belirlenebilmektedir. Şekil 3.27’de TIMIT veritabanına ait bir cümlenin ön vurgulamadan ($\alpha = 0.95$) önce ve sonra zamana bağlı olarak frekansındaki değişimler görülmektedir. Şekil 3.27 (b) den görüleceği üzere şekil 3.27 (a)’ya göre yüksek frekanslı bileşenler daha belirginleşmektedir.



Şekil 3.27 Bir cümlenin birinciden süzgeçten ($\alpha = 0.95$) (a) geçirilmeden (b) geçirildikten sonra zaman-frekans değişimi

3.3.1.5 Mel ölçekte dizilmiş süzgeç dizileri

Mel ölçek ilk olarak Steven ve ark. (1937) tarafından akustik frekans ile algılanan frekansın perdesi arasındaki ilişkiyi algısal bir ölçek olarak bulunmuştur. Normal frekans ile Mel ölçek arasındaki ilişkide referans noktası olarak 1000 Hz alınıp, 1000 Mel'e eşit olarak tanımlanmaktadır. Şekil 3.28'de Steven ve Volkman (1940) tarafından orijinal Mel ölçeğin güncellenmiş hali görülmektedir.



Şekil 3.28 Mel ölçek

Koeing (1949), Mel frekans ölçeğini 1000 Hz altı doğrusal, 1000 Hz'in üzeri ise logaritmik olarak tanımlamıştır. Bu tanımlama Mel ölçeği hesabında kolaylık sağlamasına rağmen orijinal ölçek ile bu tanımlama arasında önemli sapmalar oluşmaktadır. Fant (1949) daha doğru bir yaklaşım önermiştir. Bu yaklaşım denklem 3.31'de tanımlanmaktadır.

$$\hat{f}_{mel} = k \cdot \log_n \left(1 + \frac{f_{lin}}{F_b} \right), \quad F_b = 1000 \quad (3.31)$$

Fant (1973), bu ifadeyi denklem 3.32'deki gibi güncellemiştir.

$$\hat{f}_{mel} = \frac{1000}{\log_n 2} \cdot \log_n \left(1 + \frac{f_{lin}}{1000} \right) \quad (3.32)$$

Denklem 3.32, Koeing'in (1949) kullandığı ölçeğe göre Mel ölçeğe daha yakın bulunmuştur. Bu yaklaşım sadece [0, 5] kHz aralığı için geçerlidir. Denklem 3.31 kullanılarak Mel ölçek olarak yaygın olarak kullanılan denklem 3.33 ve 3.34 elde edilmiştir (O'Shaughnessy 1987).

$$\hat{f}_{mel} = 2595 \cdot \log_{10} \left(1 + \frac{f_{lin}}{1000} \right) \quad (3.33)$$

$$\hat{f}_{mel} = 1127 \cdot \ln \left(1 + \frac{f_{lin}}{1000} \right) \quad (3.34)$$

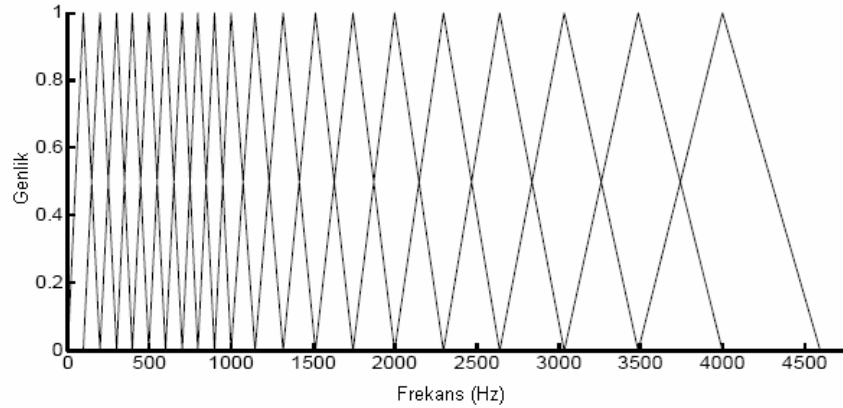
Denklem 3.33. ve 3.34, denklem 3.32'ye göre 1000 Hz altında Mel ölçeğe daha fazla yaklaşmakta, 1000 Hz üzerinde yakınsamada düşme olmaktadır.

Mel ölçek keprum katsayıları, ilk olarak Davis ve Mermelstein (1980) tarafından tanımlanmıştır. Davis ve Mermelstein (1980), işaretin genlik spektrumunu alıp üçgen şeklindeki süzgeç dizilerinden geçirmiştir. Süzgeç sayısı FS , seçilen işaret bant genişliği $[0, f_s/2]$ Hz ve f_s örnekleme frekansı olarak tanımlanmıştır. Üçgen süzgeç dizilerinden biri l olsun, $l \in [1, FS]$, bu süzgecin merkez frekansı f_{cl} olup alt ve üst bant geçiren frekansları ise; f_{cl-1} ve f_{cl+1} olarak ifade edilir. Buna bağlı olarak $f_{c0}=0$ ve $f_{cl} < f_s/2$ $\forall l$ olarak ifade edilir. Süzgeç dizileri, denklem 3.35'deki gibi ifade edilir.

$$F_l[k]=\begin{cases} \left(\frac{k}{N}\right)f_s - f_{cl-1} / (f_{cl} - f_{cl-1}) & L_l \leq k \leq C_l \\ f_{chl} - \left(\frac{k}{N}\right)f_s / (f_{chl} - f_{cl}) & C_l \leq k \leq U_l \end{cases} \quad (3.35)$$

Burada $C_l = \frac{f_{cl}}{f_s} N$, $U_l = \frac{f_{cl+1}}{f_s} N$ ve $L_l = \frac{f_{cl-1}}{f_s} N$ olup l 'inci süzgecin merkez, üst ve alt frekanslarıdır (Reynolds 1992). Kullanılan üçgen süzgeç dizilerinin merkez frekansları, Mel ölçeğinde eşit olarak yerleştirilir.

Şekil 3.29'da Davis ve Mermelstein (1980) tarafından tanımlanan üçgen süzgeç dizileri görülmektedir. Burada ilk 10 süzgecin merkez frekansları doğrusal olarak, sonraki 10 süzgeç ise logaritmik olarak yerleştirilmiştir. Tüm süzgeçler eşit genliğe sahiptir. Şekil 3.29'dan görüleceği üzere her bir süzgecin bitiş noktası, bitişindeki süzgecin merkez frekansına bağlıdır. Bu yüzden süzgeçlerin bant genişliği bağımsız bir değişken değildir. Süzgeçlerin bant genişliği ardışık süzgeçlerin merkez frekansları arası uzaklık olarak belirlenir.



Şekil 3.29 Mel ölçeğinde dizilmiş üçgen süzgeç dizileri (Davis ve Mermelstein (1980))

Davis ve Mermelstein (1980) makalesinde süzgeç şekli seçimi, süzgeçlerin örtüşme miktarı ve süzgeç sayısı hakkında açıklamada bulunulmamaktadır. Moore (2003) ve Skowronski ve ark. (2004) süzgeç dizilerinin üçgen seçilmesindeki amacın, insan algı sistemindeki kritik bandın kabaca üçgen süzgeç dizilerine benzetilmeye çalışıldığını savunmuştur. Süzgeçlerin merkez frekansı ve kritik bant genişliği arasındaki denklem 3.48 ile ifade edilen Eşdeğer dikdörtgensel bantgenişliği (ERB) ölçeği (Glasberg ve Moore, 1990), Davis ve Mermelstein (1980) makalesinden, daha sonra tanımlanmıştır. Kinnunen (2003), konuşmacı tanıma için Mel ölçeğinde dizilmiş

üçgen, dikdörtgen ve Hanning süzgeçleri karşılaştırmış ve süzgeç şekli değişiminin konuşmacı tanıma başarımına önemli bir etkisinin olmadığını göstermiştir.

Son yıllarda konuşmacı tanıma uygulamalarında Slaney'in (1998) MFCC elde etme yöntemi yaygın olarak kullanılmaktadır (Sarma 1997, Ganchev 2005). Slaney, 133-6854 Hz frekans aralığına 40 adet süzgeç yerleştirmiştir. İlk on üç süzgecin merkez frekansı 200-1000 Hz aralığında, 66.67 Hz aralıkla yerleştirilmiştir. Kalan yirmi yedi süzgecin merkez frekansları 1071-6400 Hz aralığında 1.0711703 logaritmik adımla yerleştirilmiştir.

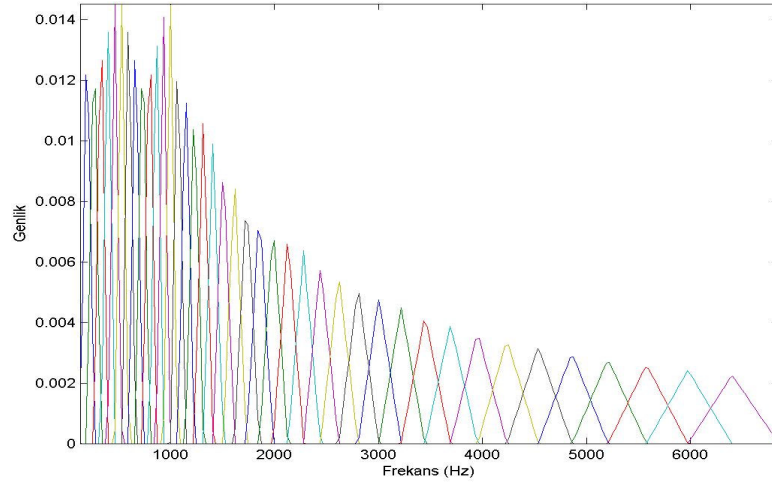
Slaney'in önerdiği süzgeç dizilerinin genliği, süzgecin bant genişliği ile ters orantılı olarak değişmektedir. Yani süzgecin bant genişliği küçük ise (1000 Hz altı doğrusal Mel ölçek bölgesi) süzgecin genliği büyük olmakta, süzgecin bant genişliği büyük olursa (1000 Hz üstü logaritmik Mel ölçek bölgesi) süzgecin genliği küçük olmaktadır. Süzgeç dizilerinin genliğinin bant genişliğine göre değişimi denklem 3.36'da tanımlanmaktadır.

$$Ag. = \frac{2}{f_{cl+1} - f_{cl-1}} \quad (3.36)$$

Burada f_{cl} 1. süzgecin merkez frekansıdır. Denklem 3.36 ile denklem 3.35'deki üçgen süzgeç dizileri çarpılmaktadır. Bu şekilde her bir süzgecin katsayılarının toplamı bire eşitlenerek normalize edilmektedir. Yani l . süzgeç denklem 3.37'deki gibi ifade edilmektedir.

$$\sum_{k=L_l}^{U_l} F_l[k] = 1 \quad l = 1, 2, \dots, FS \quad (3.37)$$

Burada FS toplam süzgeç sayısına karşılık gelmektedir. Şekil 3.30'da 40 adet Mel ölçekte yerleştirilmiş üçgen süzgeç dizisi görülmektedir. Şekilden de görüleceği üzere süzgeç dizilerinin genliği 1000 Hz'den itibaren bant genişliği artmasına bağlı olarak azalmaktadır.



Şekil 3.30 Mel ölçekte dizilmiş süzgeç dizileri (Slaney 1998)

Davis ve Mermelstein (1980) ve Slaney (1998) tarafından tanımlanan Mel ölçek süzgeç dizilerinin konuşma tanımadaki başarımları karşılaştırılacaktır. Bu Mel ölçeklerinin merkez frekansları ve bant genişlikleri çizelge 3.18'de görülmektedir. Deneyde mikrofon (TIMIT) ve telefon (NTMIT) ortamından verilerin toplandığı iki farklı veritabanı kullanılacaktır. Mikrofon ortamında (konuşma bant genişliği 0-8 KHz) yapılan deneylerde Davis ve Mermelstein'in (1980) tanımladığı Mel ölçeğinde 24 süzgeç, Slaney'in (1998) tanımladığı Mel ölçeğinde 40 süzgeç kullanılmaktadır. NTMIT veritabanı ile yapılan deneylerde Davis ve Mermelstein'in (1980) tanımladığı Mel ölçeğinde, 3-19 indisler arasında kalan süzgeçler, Slaney'in (1998) tanımladığı Mel ölçeğinde ise 3-31 indisleri arasında kalan süzgeçler kullanılmaktadır. Süzgeçler bu şekilde telefon ortamı bant genişliği olan 300-3400 Hz arasına sınırlandırılmaktadır.

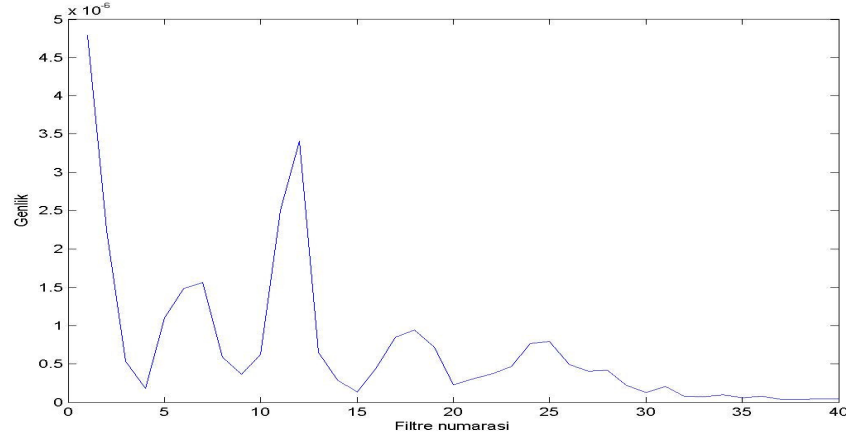
Davis ve Mermelstein (1980) konuşma çerçevesinin genlik spektrumunu ($|FFT|$) alırken Slaney (1998) konuşma çerçevesinin güç spektrumunu ($|FFT|^2$) almıştır. Güç spektrumunun konuşmacı tanıma başarımını arttırdığı çizelge 3.16'dan görülmektedir. Çizelge 3.18'de verilen Mel ölçekleri karşılaştırılırken güç spektrumu kullanılmaktadır.

Çizelge 3.18 İki farklı Mel ölçeğin merkez frekansları ve bant genişlikleri

Süzgeç indeksi	Mel Ölçek Davis ve Mermelstein (1980)		Mel Ölçek Slaney (1998)	
	Merkez fr. (Hz)	BW (HZ)	Merkez fr. (Hz)	BW (HZ)
1	100	100	200	66.67
2	200	100	266.7	66.67
3	300	100	333.3	66.67
4	400	100	400	66.67
5	500	100	466.7	66.67
6	600	100	533.3	66.67
7	700	100	600	66.67
8	800	100	666.7	66.67
9	900	100	733.3	66.67
10	1000	124	800	66.67
11	1149	160	866.7	66.67
12	1320	184	933.3	66.43
13	1516	211	999.8	71.15
14	1741	242	1070.9	76.22
15	2000	278	1147.1	81.64
16	2297	320	1228.8	87.45
17	2639	367	1316.2	93.68
18	3031	422	1409.9	100.34
19	3482	484	1510.2	107.48
20	4000	556	1617.7	115.13
21	4595	639	1732.9	123.32
22	5278	734	1856.2	132.11
23	6063	843	1988.3	141.51
24	6964	969	2129.8	151.57
25			2281.4	162.37
26			2443.7	173.92
27			2617.7	186.30
28			2804	199.55
29			3003.5	213.76
30			3217.3	228.98
31			3446.3	245.27
32			3691.5	262.73
33			3954.3	281.43
34			4235.7	301.46
35			4537.1	322.91
36			4860.1	345.89
37			5205.9	370.51
38			5576.5	396.88
39			5973.3	425.12
40			6398.5	455.38

Ön vurgulanan 1x512 boyutundaki konuşma işareti, şekil 3.31’de görülen Mel ölçek süzgeç dizilerine ait değerler (40x512 boyutunda) ile çarpılır. Mel ölçekte hazırlanan birinci süzgeç, 40x512 boyutundaki matrisin birinci satırı ile ifade edilir. Aynı şekilde kırkıncı süzgeç, matrisin 40. satırı ile ifade edilir. Çarpım sonucunda şekil

3.31’de görüldüğü gibi 1x40 boyutunda çıkış değeri elde edilir ve her süzgeç çıkışı bir değer ile temsil edilmektedir.



Şekil 3.31 İşaretin süzgeç dizisinden geçirildikten sonraki durumu

Çizelge 3.18’de verilen Mel ölçeklerin konuşmacı tanıma etkisi incelenecektir. Bunun için konuşmacı tanıma sistemine ait parametreler şu şekilde alınmaktadır. Konuşmacıların eğitimi için BM algoritması, model başlangıcı olarak da VN algoritması kullanılmaktadır. Eğitim için 15 BM özyineleme kullanılıp değişinti sınırlaması olarak $\sigma^2_{\min}=0.01$ değeri kullanılmaktadır. MFCC elde edilmesinde 20 msn uzunluğunda çerçevenin, örtüşme oranı 10 msn alınıp, çerçevelere Hamming pencereleme yapılmaktadır. Her iki veritabanı için de ön vurgulama işlemi uygulanmamaktadır. Pencerelenen sesin güç spektrumu alınıp çizelge 3.18’de verilen Mel ölçekte yerleştirilmiş üçgen süzgeç dizilerinden geçirilmiştir. Süzgeçten geçirilen işaretin logaritması alınıp ayrık kosinüs dönüşümü alınmıştır. Her bir çerçeveye karşılık olarak TIMIT veritabanı için 24, NTIMIT veritabanı için 20 boyutlu öznitelik vektörleri kullanılmaktadır. Bu parametrelere bağlı olarak elde edilen konuşmacı tanıma oranları çizelge 3.19’da görülmektedir.

Çizelge 3.19 Çizelge 3.18’de tanımlanan Mel ölçeklerin konuşmacı tanıma oranı (%)

Veritabanları	Mel Ölçek	
	Davis ve Mermelstein (1980)	Slaney (1998)
TIMIT	99.4	99.4
NTIMIT	67.26	69.64

Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168

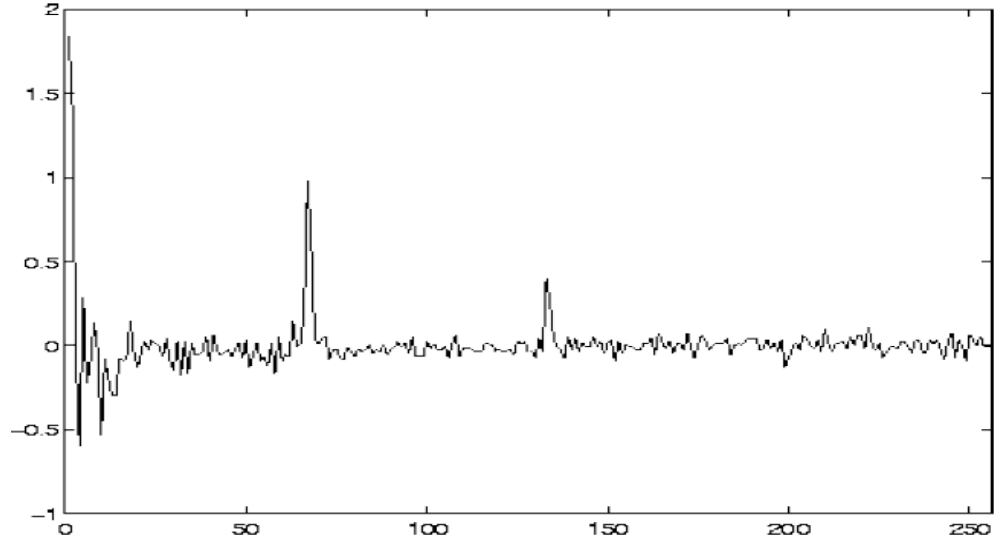
Çizelge 3.19'dan görüleceği üzere TIMIT veritabanında her iki ölçek için de tanıma oranı değişmemektedir. NTIMIT veritabanında Slaney (1998)'in önerdiği Mel ölçekte dizilmiş süzgeç dizileri, Davis ve Mermelstein (1980) tarafından tanımlanan Mel ölçege göre daha iyi başarımlar sağlamaktadır. Slaney (1998)'in önerdiği Mel ölçekteki süzgeç dizilerinin daha iyi olmasının temel nedeni, çizelge 3.18'den görüleceği üzere, süzgeçlerin bant genişliklerinin daha dar olması ve bu şekilde orta ve yüksek frekans bandının daha iyi modellenmesidir.

3.3.1.6 Logaritma alma

Konuşma işaretinin süzgeç dizisinden geçirildikten sonra logaritması alınır. Spektrum'un logaritması alınmasının nedeni şu şekilde açıklanabilir. Konuşma işareti, $|S(e^{j\omega})| = |X(e^{j\omega})||F(e^{j\omega})|$ olarak ifade edilir. Burada S , X ve F sırası ile konuşma işareti, kaynak ve süzgece karşılık gelmektedir. Kaynak, ses telleri tarafından üretilen ve değişime uğramamış ses işaretini temsil eder. Süzgeç ise ses yolu olarak ifade edilen sesin izlediği yola karşılık gelmektedir (Kinnunen 2003). Ses yolunun etkisini kaynaktan ayırmak için logaritma kullanılır. Logaritma alınarak, konuşma işaretinin bileşenlerinin çapımı, bileşenlerin toplamına $\log|S(e^{j\omega})| = \log|X(e^{j\omega})| + \log|F(e^{j\omega})|$ dönüştürülmüş olur. Logaritmik spektrum farklı frekanslara sahip bileşenlerin bileşimi olarak düşünülebilir. Daha sonra bu iki bileşene ters FFT uygulanarak hızlı ve yavaş değişen bileşenler hakkında bilgi sahibi olunabilir. Denklem 3.38 ile gösterilen işlem sonunda elde edilen katsayılar cepstrum katsayıları olarak adlandırılır.

$$ceps = FFT^{-1}(\log(|FFT(Hamm(512) \cdot x(n) |))) \quad (3.38)$$

Burada $x(n)$, çerçevelenmiş konuşma parçasına karşılık gelmektedir. Şekil 3.32'de bir konuşma parçasının denklem 3.38 uygulandıktan sonra elde edilen şekil görülmektedir.



Şekil 3.32 Konuşma parçasına denklem 3.38 uygulanması durumunda elde edilen kepstrum katsayıları

Kepstrum katsayılarına ait şekil 3.32'den görüleceği üzere orijin civarında çok fazla ayrıntı ve yüksek tepeler oluşmakta yani ses yolu (yavaş değişen bileşen olarak) bu kepstrum katsayılarına karşılık gelmektedir. Ses tellerinden geçirilmiş ses kaynağı (hızlı değişen bileşen olarak) yüksek sayılı kepstrum katsayıları karşılık gelmektedir. Bu bölgede en yüksek genliğe sahip (70. örnek civarı) perde periyodu hakkında bilgi vermektedir.

Konuşma spektrumu x , sifıra yakınsadığı durumlarda $\log(x)$ eksi sonsuza yönelir. Logaritma fonksiyonu x 'in küçük değerlerine karşı çok hassastır. Spektrumda düşük güce sahip yerler (SNR'in düşük olduğu) en hassas kısımlardır. Spektrumun küçük değerleri için $\log(x)$ yerine $\log(x+c)$ kullanılır (Hunter 1999). Burada c küçük bir sabittir. Bu fonksiyon $x \gg c$ için $\log(x)$ 'e benzemekte, bu durumda ($x \gg c$), $\log(x+c) \approx \log(c) + x/c$ olmaktadır. Bu durumda $\log(x+c)$, x küçük iken $\log(c)$ ile sınırlı ve x ile doğrusal olarak değişirken, x büyük iken $\log(x)$ 'e benzemektedir. Kinnunen (2003), konuşmacı tanıma deneylerinde $\log(x)$ değerine 1 sabitini eklemektedir.

Konuşma spektrumu x , sifıra yakınsadığında oluşan problemlerden dolayı konuşma güç spektrumunun logaritma fonksiyonu yerine güç $(.)^\gamma$ veya kök $(.)^{1/\gamma}$ fonksiyonu gösterimi önerilmiştir (Lim 1979). Ses şiddeti ve algılanan duyma düzeyi

arasındaki doğrusal olmayan bir ilişkinin varlığına dayanılarak bu ilişkinin modellenmesi yapılır. Algılanan sesin düzeyi sesin şiddetinin küp köküne eşittir. Spektrumun küp kökünü alma gürültü içeren konuşma tanıma deneylerinde (Alexandre ve Lockwood 1993, Chu ve ark. 2003) logaritma fonksiyonuna göre daha iyi sonuçlar elde edilirken, temiz konuşma için düşük başarımler elde edilmiştir. Sarıkaya ve ark (2001), kök değeri olarak 0.008 kullanarak MFCC'ye göre % 84 daha iyi fonemleri ayrıştırma başarımı elde etmiştir.

Kinnunen (2003), Helsinki ve TIMIT veritabanı ile VN konuşmacı tanıma yöntemini kullanarak yaptığı deneylerde spektrumun küp kökünü ve logaritma fonksiyonunu alarak karşılaştırmıştır. Değişik kod kitabı uzunluğu için küp kökünü kullanmanın tanıma başarımını arttırdığını göstermiştir.

Kök fonksiyonu $\sqrt[n]{p}$ şeklinde ifade edilsin. n değeri sonsuza yaklaştırıldığında alacağı değeri bulalım. $(1+x)^a$ denklem 3.39'daki gibi tanımlanmaktadır.

$$(1+x)^a = 1 + ax + \frac{a(a-1)x^2}{2!} + \frac{a(a-1)(a-2)x^3}{3!} \dots \quad (3.39)$$

a değerini 0'a yaklaştırıldığında denklem 3.40 elde edilmektedir.

$$\begin{aligned} a \rightarrow 0, (1+x)^a &\rightarrow 1 + ax - \frac{ax^2}{2!} + \frac{2ax^3}{3!} \dots \\ &= 1 + ax - \frac{ax^2}{2} + \frac{ax^3}{3} \dots = 1 + a \cdot \log(1+x) \end{aligned} \quad (3.40)$$

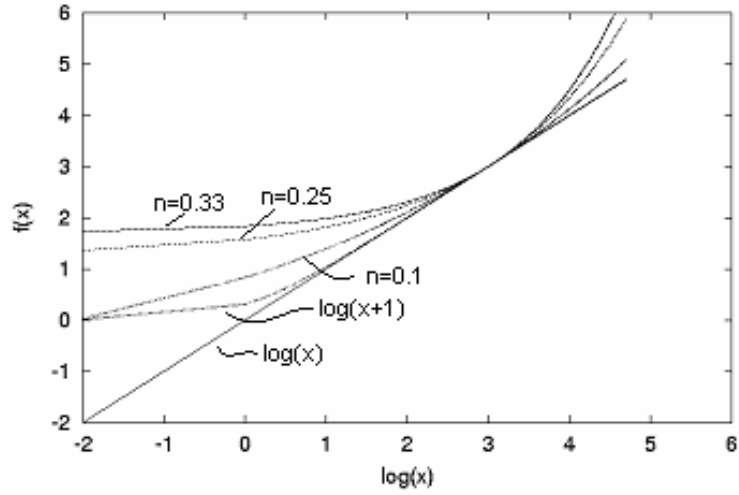
$(1+x)$, y ile ifade edilirse bu durumda denklem 3.41 elde edilir.

$$a \rightarrow 0 \quad \log(y) \rightarrow (y^a - 1)/a \quad (3.41)$$

Kök ifadesi için $a = 1/n$ olarak alınırsa denklem 3.42 elde edilir.

$$n \rightarrow \infty, \quad \log(y) \rightarrow n(\sqrt[n]{y} - 1) \quad (3.42)$$

Kökün derecesine karşılık gelen n sonsuza yaklaştırıldığında logaritmaya yakınsamaktadır. Şekil 3.33'de çeşitli güç ve logaritma fonksiyonları verilmektedir. Şekil 3.33'den kök fonksiyonunun değeri n , azaldıkça $\log(x+c)$ fonksiyonuna benzediği görülmektedir.

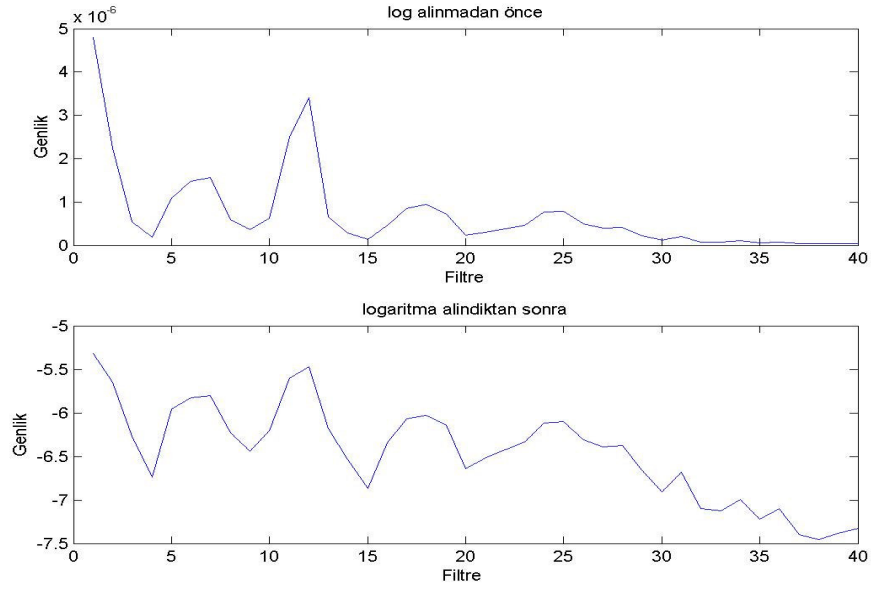


Şekil 3.33 Logaritmik ölçekte kök ve logaritma fonksiyonlarının değişimi

1. süzgeç için logaritmik enerji çıkışı denklem 3.43'de görüldüğü gibi $mfb(l)$ olarak ifade edilir.

$$mfb(l) = \log\left(\frac{1}{A_l} \sum_{k=L_l}^{U_l} F_l[k]X[k]\right) \quad (3.43)$$

Burada A_l süzgeçlerin bant genişliğine bağlı olarak kullanılan normalizasyon katsayısı olup $A_l = \sum_{k=L_l}^{U_l} F_l[k]$ olarak tanımlanır. Sonuç olarak elde edilen vektöre Mel-süzgeç dizisi vektörü denir. Logaritma alarak, dinamik sıkıştırma yapıp, öznelik vektörleri, dinamik değişimlere karşı daha az hassas olmaktadır (Claudio 1999). Şekil 3.34'de işaretin logaritması alındığında işarettaki değişimler görülmektedir.



Şekil 3.34 İşaretin süzgeç çıkışı ve logaritmali hali

Süzgeç çıkışlarının logaritmasının alınmasının konuşmacı tanıma üzerine etkisi incelenecektir. Süzgeç dizisinden geçirilen işaret x ile ifade edilsin. Süzgeç çıkışlarının logaritması alınması ve alınmama durumları ile konuşma tanımada başarımların artışı elde edilen süzgeç çıkışının $(1/3)$ ve (0.008) kuvvetlerinin alınmasının her iki veritabanı için sonuçları incelenecektir (Alexandre ve Lockwood 1993, Chu ve ark. 2003, Sarıkaya ve ark. 2001). Çizelge 3.20’de, bölüm 3.3.1.1’de tanımlanan konuşmacı tanıma sistemi parametrelerine bağlı olarak konuşmacı tanıma oranları görülmektedir.

Çizelge 3.20 Süzgeç çıkışlarının logaritması ve kuvvetleri alınmasının konuşmacı tanıma etkisi (%)

	Veritabanları	
	TIMIT	NTIMIT
$\log(x)$	99.4	69.64
$\log(x) + 1$	99.4	70.83
$x^{1/3}$	86.31	22.02
$x^{0.008}$	36.31	14.29
x	26.19	4.17

Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168

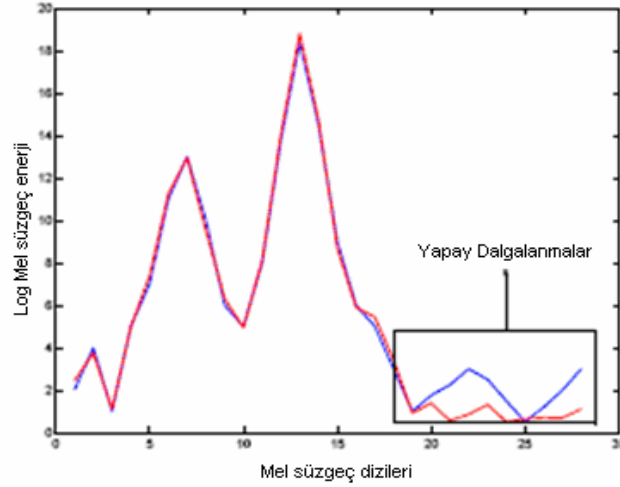
Çizelge 3.20’den görüleceği üzere süzgeç dizilerinin çıkışlarının logaritmasının alınması her iki veritabanı içinde tanıma oranını önemli oranda arttırmaktadır. Çünkü

işaret logaritma alınarak 10^{-6} lı değerlerden ∓ 20 aralığına kaymaktadır. Süzgeç çıkışının (1/3) ve (0.008) kuvvetlerinin alınması, her iki veritabanı için konuşmacı tanıma başarımını düşürmektedir. İşarete hiçbir işlem uygulamadan AKD alınarak MFCC elde edildiği durumda, iki veritabanı içinde en düşük tanıma başarımı elde edilmektedir.

Konuşma ve konuşmacı tanıma deneylerinde, MFCC ve PLP gibi öznelik çıkartma yöntemleri konuşmaların gürültüsüz temiz olarak adlandırılan ortamlarda kaydedildiğinde çok iyi başarımlar elde edilirken, gürültülü ortamlarda önemli oranda başarımları düşmektedir (Krishnakumar ve ark. 2003). Bu durumun nedenlerinden birisi spektrum alınarak elde edilen özneliklere gürültü bulaşmasıdır.

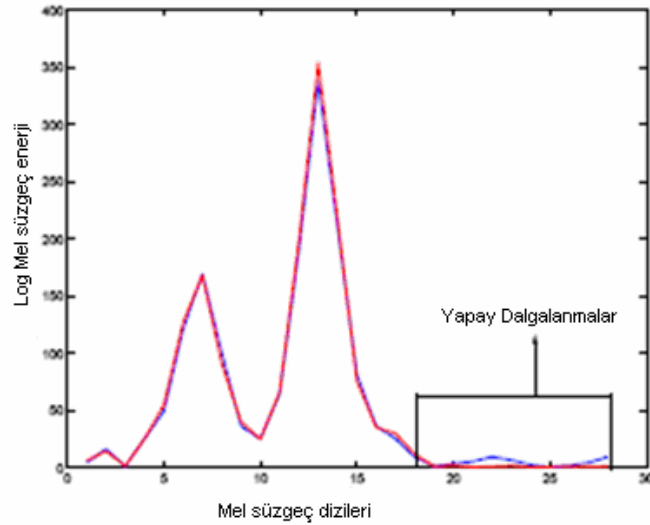
Temiz bir ünlü sözcük için perde harmoniklerinde (formantlar) yerel maksimum değerler oluşmaktadır. Sözcüğe gürültü eklenmesi durumunda sözcüğün yüksek genlikli kısımları diğer kısımlara göre daha gürbüz olmaktadır. Spektral çukurlar düşük SNR'a sahip olduklarından dolayı gürültü eklenmesinden önemli oranda etkilenmekte ve bu durumda yapay dalgalanmalar oluşmaktadır (Tyagi ve Wellekens 2005). Bu dalgalanmalar güç spektrumunu alındıktan sonra çok küçük oranlarda (10^{-6}) değişim göstermektedir. Şekil 3.34'de işaretin logaritması alınmadan önce genlik değerlerinin çok küçük olduğu görülmektedir. Konuşmanın logaritması alındıktan sonra yapay dalgalanmalar genlik olarak önemli değerler almaktadır. Bununla birlikte kestrum katsayıları elde edilmesinde işaretin logaritması alındıktan sonra ayrık kosinüs dönüşümü (AKD) bulunur. AKD ile süzgeç çıkışlarının düşük enerjiye sahip kısımlarının ağırlıklandırılması ile bu yapay dalgalanmalar önem kazanır. Şekil 3.35'de konuşma spektrumu Mel süzgeç dizilerinden geçirilip logaritması alındıktan sonra gürültülü ve temiz konuşma için elde edilen şekiller görülmektedir.

Şekil 3.35'den görüleceği üzere yüksek enerjili tepelere karşılık gelen formantlarda gürültü etkisi ile dalgalanma olmaz iken gürültü etkisi ile düşük enerjili çukur kısımlarda gürültü etkisi ile dalgalanmalar oluşmaktadır.



Şekil 3.35 Temiz (kırmızı) ve gürültülü (mavi) konuşmalar için logaritması alınmış Mel süzgeç dizilerinin enerjileri

Mel süzgeç dizilerinin logaritması alındıktan sonra ağırlıklandırma uygulanarak formantlar, düşük enerjili mel süzgeç dizisi örneklerden daha önemli hale getirilebilir. Tyagi ve Wellekens (2005), yapay dalgalanmaların etkisinin azaltılması için, Mel süzgeç dizilerinin logaritması alındıktan sonra P güç katsayısı kullanarak ağırlıklandırmayı önermiştir. Şekil 3.36'da konuşma spektrumu Mel süzgeç dizilerinden geçirilip logaritmasının karesi ($P = 2$) alındıktan sonra gürültülü ve temiz konuşma için elde edilen şekiller görülmektedir.



Şekil 3.36 Temiz (kırmızı) ve gürültülü (mavi) konuşmalar için logaritmasının karesi alınmış Mel süzgeç dizilerinin enerjileri

Şekil 3.36'dan görüleceği üzere yapay dalgalanmalara göre formantların genliği arttırılarak daha baskın ve önemli hale gelmektedir. Bu şekilde düşük enerji bölgesindeki yapay dalgalanmalardan oluşan düşük AKD katsayılarının hassasiyeti azaltılmaktadır.

Tyagi ve Wellekens (2005), MFCC elde edilmesinde konuşma spektrumunun logaritmasının karesi olarak öznitelik vektörü elde etmiştir. Gürültü içeren ortamlarda konuşma tanımada başarımların artışı sağlanmıştır. Aynı öznitelik vektörü elde etme yöntemi konuşmacı tanıma için uygulanacaktır. Bölüm 3.3.1.1'de tanımlanan parametreler kullanılarak elde edilen sonuçlar Çizelge 3.21'de görülmektedir. Veritabanlarına gürültü olarak 5 dB beyaz gürültü eklenmiştir.

Çizelge 3.21 Süzgeç çıkışlarının logaritması alınmasının tanımaya etkisi (%)

Veritabanları	$\log(x)$	$[\log(x)]^2$
TIMIT	99.4	97.62
TIMIT+ gürültü	88.1	90.48
NTIMIT	69.64	67.26
NTIMIT+gürültü	58.93	39.88

Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168

Çizelge 3.21'den görüleceği TIMIT veritabanına gürültü eklenmesi durumunda işaretin logaritmasının karesi alınması daha iyi sonuç verirken, diğer durumlarda işaretin logaritması alınması daha iyi konuşmacı tanıma başarımı sağlamaktadır.

3.3.1.7 Ayrık kosinüs dönüşümü (AKD)

MFCC elde edilirken en son olarak kepsrum katsayıları hesaplanır. Kepsrum gösterim ile kayıt ve iletim ortamından dolayı oluşan spektral şekil değişimleri kaldırılır. Ayrıca kepsrum katsayıları, yüksek derecede istatistiksel bağımsızlık gösterip genlik spektrum gösteriminden daha yüksek tanıma oranı verirler. Gerçek kepsrum, logaritmik genlik spektrumun ters fourier dönüşümü olarak tanımlanıp, gerçek işaretler için kosinüs dönüşümü kullanılarak hesaplanır.

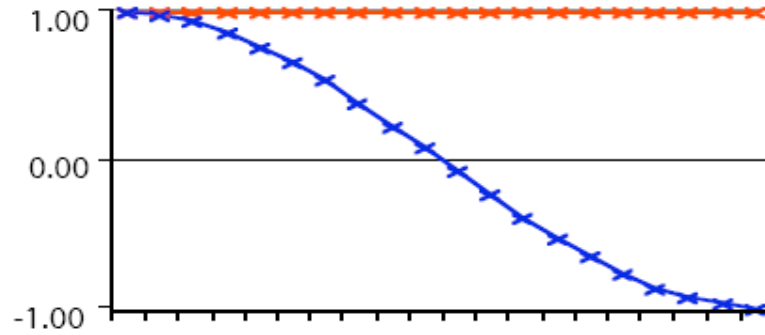
Mel ölçeğinde dizilmiş 14 adet süzgeç dizisi ele alalım. Bu süzgeç dizilerine AKD uygulandığında denklem 3.44 elde edilir.

$$c_i = \sum_{j=1,14} m_j \cdot \cos(p \cdot i \cdot (j - 0.5) / 14) \quad i = 1, \dots, N \quad (3.44)$$

Burada N istenen kepstrum katsayı sayısıdır. Kepstrum katsayıları elde edilirken logaritması alınmış Mel süzgeç dizileri (m_j) bir kosinüs eğrisi ile ağırlıklandırılıp toplanır. İlk kepstrum katsayısı c_0 , denklem 3.44'de $i = 0$ yazıldığında $\cos(0) = 1$ olacağından dolayı, süzgeç dizilerinin tümünün toplamına karşılık gelir. c_1 kepstrum katsayısı denklem 3.45'deki gibi ifade edilmektedir.

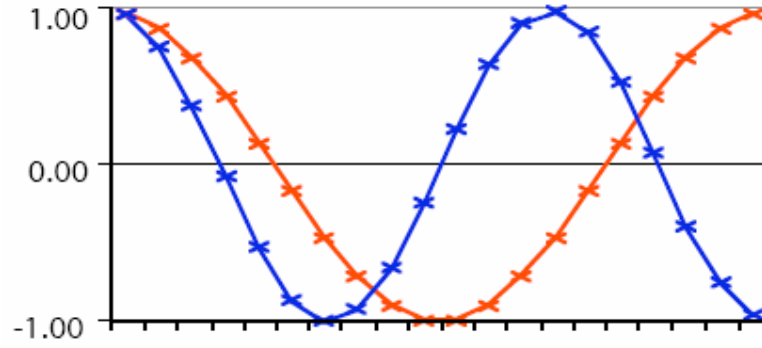
$$c_1 = 0.99 \cdot m_1 + 0.94 \cdot m_2 + 0.84 \cdot m_3 + 0.71 \cdot m_4 + 0.53 \cdot m_5 + 0.33 \cdot m_6 + 0.11 \cdot m_7 - 0.11 \cdot m_8 - 0.33 \cdot m_9 - 0.53 \cdot m_{10} - 0.71 \cdot m_{11} - 0.85 \cdot m_{12} - 0.94 \cdot m_{13} - 0.99 \cdot m_{14} \quad (3.45)$$

Elde edilen bu c_0 ve c_1 fonksiyonları şekil 3.37'de görülmektedir.



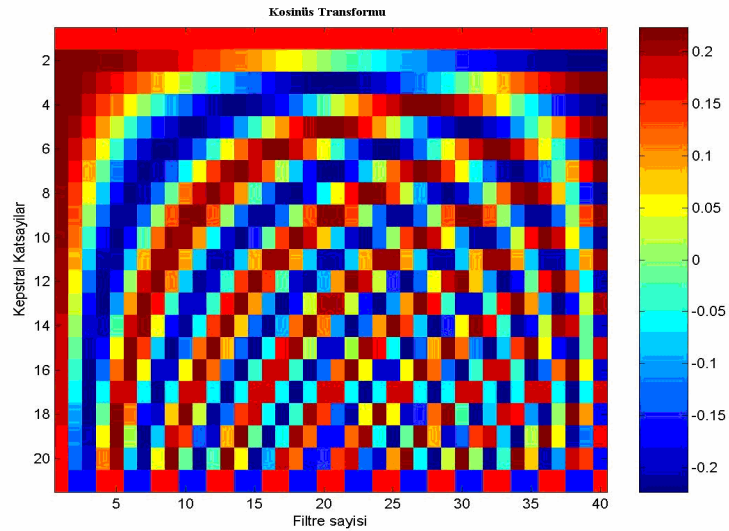
Şekil 3.37 c_0 (kırmızı) ve c_1 (mavi) fonksiyonları

Şekil 3.37'den görüleceği üzere, c_0 için spektrum yoğunluğunun ağırlıklandırılması tüm frekanslar için eşit olur. c_1 kepstrum katsayısı, ağırlıklandırma için bir kosinüs periyodunun yarısı kullanılır. c_2 için kosinüs fonksiyonunun bir periyodu, c_3 için kosinüs fonksiyonunun bir buçuk periyodu kullanılır. Bu işlem tüm kepstrum katsayıları için devam ettirilir. Bu şekilde her bir kepstrum katsayısının korelasyonu azaltılmış olur. Şekil 3.38'de c_2 ve c_3 fonksiyonları görülmektedir.



Şekil 3.38 c_2 (kırmızı) ve c_3 (mavi) fonksiyonları

Mel ölçek süzgeç sayısı 40, kepstrum katsayı sayısı 21 için ayırık kosinüs dönüşümü şekil 3.39'da görülmektedir. Şekilde verilen renk ölçeğine göre matrisin aldığı değerler görülmektedir.



Şekil 3.39 Ayırık kosinüs dönüşümü

Konuşmacı ses örnekleri Mel süzgeç dizisinden geçirildikten sonra AKD alınması ve AKD alınmayıp direkt olarak mel süzgeç çıkışlarının öznitelik vektörü olarak tanımlanması durumlarının konuşmacı tanıma etkileri incelenecektir. Deneyde ön vurgulama uygulanmayıp, Slaney (1998) tarafından Mel ölçekte dizilmiş süzgeç dizilerinin logaritması alınmaktadır. Bölüm 3.3.1.1'de verilen konuşmacı tanıma sistemi şartlarında elde edilen konuşmacı tanıma oranları, TIMIT ve NTIMIT veritabanları için çizelge 3.22'de verilmektedir. Çizelge 3.22'den görüleceği üzere her iki veritabanı için

öznitelik vektörü elde edilirken AKD kullanılması konuşmacı tanıma oranını yaklaşık 30 puan arttırmaktadır.

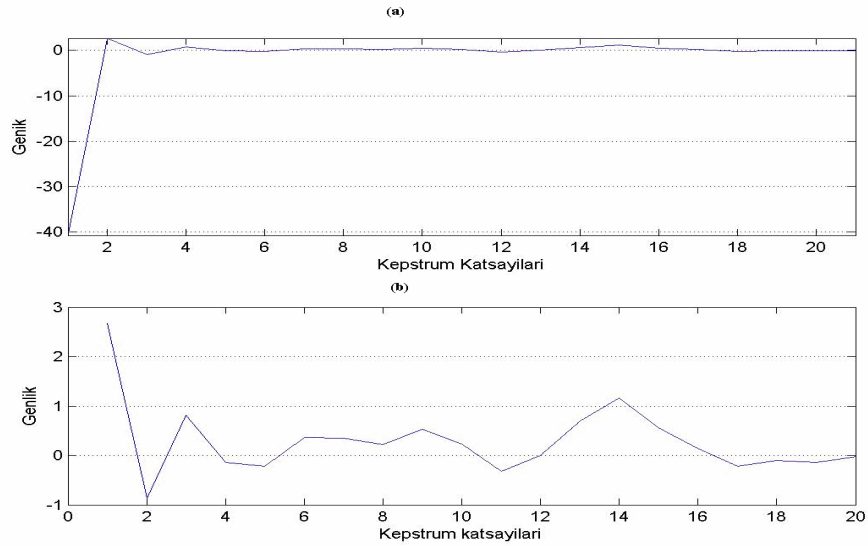
Çizelge 3.22 AKD'nin konuşmacı tanımaya etkisi (%)

Veritabanları	AKD var	AKD yok
TIMIT	99.4	68.15
NTIMIT	69.64	37.50

Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168

3.3.1.8 Sıfırıncı kepstrum katsayısı

Konuşma işaretinin logaritması alındıktan sonra elde edilen 1x40 boyutundaki matris, 40x21 (Süzgeçler x Kepstrum katsayıları) boyutunda oluşturulan ayırık kosinüs dönüşümü ile çarpılması sonucunda 1x 21 boyutunda 25 msn'lik konuşma çerçevesini karakterize eden kepstrum vektörü elde edilir. Elde edilen kepstrum vektörlerinden sıfırıncı kepstrum katsayısı, konuşmacı tanıma uygulamalarında genellikle alınmaz ve kaldırılır. Çünkü şekil 3.37'den görüleceği üzere c_0 , süzgeç dizilerinin tümünün toplamı olarak bulunmakta ve tüm konuşma frekans bandı için eşit ağırlıklandırma etkisine sahip olmaktadır. Bu nedenle bu öznitelik vektörünün ayırt edicilik özelliği fazla değildir. Şekil 3.40'da 25 msn'lik konuşma parçası için 0. kepstrum katsayısı (c_0) alınmış ve alınmamış hallerinin şekilleri görülmektedir.



Şekil 3.40 (a) c_0 çıkartılmadan elde edilen kepstrum katsayıları (b) c_0 çıkartıldıktan sonra elde edilen kepstrum katsayı eğrileri

Şekil 3.40'dan görüleceği üzere bu konuşma çerçevesi için, c_0 katsayısı diğer kepstrum katsayılarına nazaran -40 gibi yüksek bir değere sahip olmaktadır. Bölüm 3.3.1.1'de tanımlanan konuşmacı tanıma sistemi parametrelerine bağlı olarak c_0 'ın tanımaya etkisi çizelge 3.23'de görülmektedir.

Çizelge 3.23 Sıfıncı kepstrum katsayısının konuşmacı tanımaya etkisi (%)

Veritabanları	$c_0 - c_{20}$	$c_1 - c_{20}$
TIMIT	99.4	99.4
NTIMIT	66.07	69.64

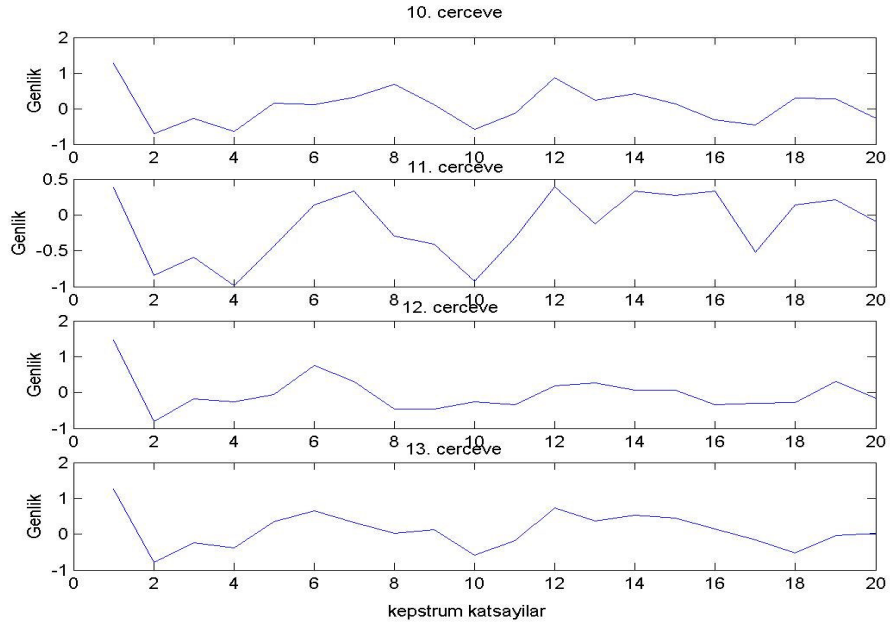
Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168

Çizelge 3.23'den de görüleceği üzere TIMIT veritabanı için öznitelik vektörlerinden c_0 'ın çıkartılması tanıma oranını değiştirmez iken, NTIMIT veritabanında tanıma başarımı artmaktadır. Sonuç olarak c_0 , çerçevenin ortalama logaritmik enerjisine karşılık gelir ve konuşmacıya ait çok az bilgi taşıdığından dolayı kullanılmaz.

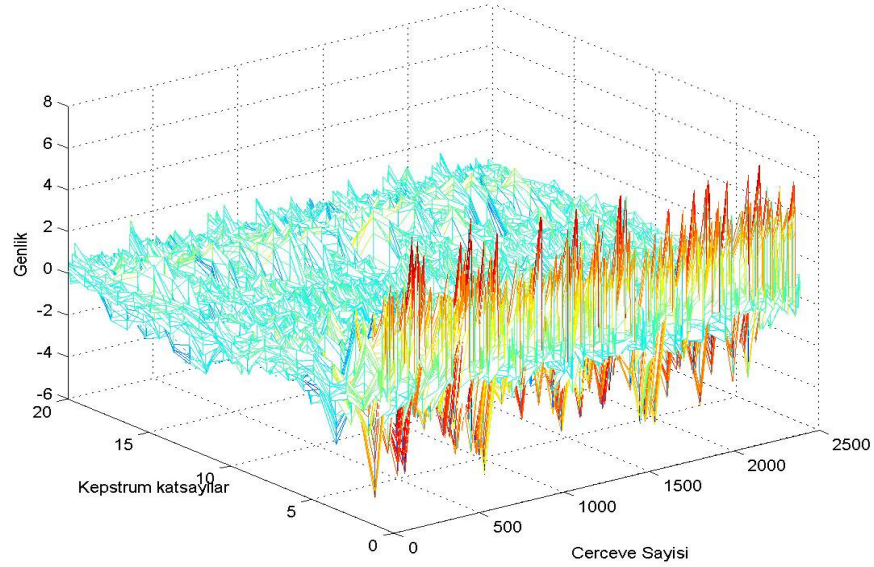
3.3.2 Kepstrum Katsayı Değişimlerinin Konuşmacı Tanımaya Etkisi

Kepstrum katsayıları kişinin ses yolu yapısı hakkında bilgi vermektedir. Bu nedenle kişileri sesinden ayırmada etkili olmakta ve konuşmacı tanımda yaygın olarak kullanılmaktadır (Reynolds ve Rose 1995). Konuşmacılara ait cümlelerden üretilen kepstrum katsayılarında, birbirini izleyen çerçeveler arasında çok hızlı bir değişim gözlenmemektedir. 20 adet kepstrum katsayısı için NTIMIT veritabanından bir cümlenin, 10, 11, 12 13. çerçeveler arasındaki kepstrum katsayılarının değişimi şekil 3.41'de görülmektedir.

Şekil 3.41'den görüleceği üzere yakın çerçeveler arası kepstrum katsayılarında çok hızlı bir değişim gözlenmemektedir. Şekil 3.42'de ise çerçeve sayısına bağlı olarak kepstrum katsayılarının değişimi görülmektedir.



Şekil 3.41 10-13. pencereler arası kepstrum katsayıları değişimi



Şekil 3.42 Çerçeve sayısına bağlı olarak kepstrum katsayıları değişimi

Şekil 3.42’de görülen kepstrum katsayıları, bir konuşmacıya ait 8 cümlesinden çıkartılan öznelik vektörleridir. İlk kepstrum katsayılarının genlik değeri diğerlerine nazaran daha büyüktür. Sonuç olarak kepstrum katsayılarındaki bu değişimler konuşmacının kimliğini ortaya çıkartacak bilgiler taşımaktadır.

Birbirini takip eden konuşma çerçeveleri arasındaki ilişki, ses yolu organlarının hareketiyle doğrudan ilişkilidir. Dinamik özellikler, konuşma işaretinin birbirini takip eden çerçeveleri arasındaki değişimin belirlenmesinde yardımcı olmaktadır (Lincoln 1999). Kepstrum katsayılarının birinci derece (Δk) ve ikinci derece ($\Delta\Delta k$) türevi alınarak dinamik katsayılar üretilmektedir. Birinci derece türev olarak elde edilen dinamik katsayılar denklem 3.46'daki gibi tanımlanmaktadır (Liu ve ark. 1996).

$$\Delta k(t) = \sum_{i=-M}^M k(t+i) \cdot i \quad (3.46)$$

Burada M çerçeve sayısına karşılık gelmektedir. İkinci derece dinamik katsayılar, denklem 3.46'daki k ile gösterilen katsayılarının yerine Δk katsayılarının uygulanması ile elde edilir.

Kepstrum katsayı değişimlerinin konuşmacı tanıma etkisi araştırılacaktır. Konuşmacı tanıma sistemine ait parametreler şu şekilde alınmaktadır. Gauss karışım sayısı 32 alınıp eğitim için BM algoritması kullanılır. Eğitim için 15 BM özyineleme kullanılıp değışinti sınırlaması olarak $\sigma^2_{\min}=0.01$ değeri kullanılmaktadır. Model başlangıcı olarak k-ortalama algoritması kullanılmaktadır. MFCC elde edilmesinde çerçevelerin örtüşme oranı 10 msn alınıp, çerçevelere Hamming pencereleme uygulanmaktadır. Pencerelenen sesin 512 örnek FFT'si alınıp, Slaney (1998), tarafından tanımlanan Mel ölçekte, üçgen süzgeç dizilerinden geçirilir. Süzgeçten geçirilen işaretin logaritması alınıp ayrık kosinüs dönüşümü alınmıştır. MFCC elde edilmesinde ön vurgulama işlemi uygulanmamaktadır. Her bir çerçeveye karşılık olarak TIMIT veritabanı için 24, NTIMIT veritabanı için 20 boyutlu öznitelik vektörleri kullanılmaktadır.

Her iki veritabanı için kepstrum katsayı sayısı ayrı ayrı incelenmiştir. TIMIT veritabanı için kepstrum katsayıları değışimlerinin konuşmacı tanıma etkisi çizelge 3.24'de görülmektedir. Deneyde kepstrum katsayıları sayısı 12, 19, 24, 30 şeklinde alınmaktadır. Ayrıca dinamik kepstrum katsayıları olarak ifade edilen Δk , 12 adet kepstrum katsayılarının birinci dereceden türevi olup $\Delta\Delta k$ ise 12 adet kepstrum katsayılarının ikinci dereceden türevini göstermektedir. Deneyde birinci derece dinamik katsayılar bulunurken çerçeve sayısı 9, ikinci derece dinamik katsayılarının elde edilmesinde çerçeve sayısı 5 alınmaktadır.

Çizelge 3.24 Kepstrum katsayıları ve test süresi değişimlerinin konuşmacı tanıma etkisi (%)

Kepstrum katsayı sayıları	Test süresi	
	1 saniye	3 saniye
k=12	86.9	98.21
k=19	92.9	99.4
k=24	93.4	99.4
k=30	86.9	99.4
k(12)+ Δk	77.4	98,2
k(12)+ $\Delta k + \Delta\Delta k$	75.6	97.6

Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168
Örnekleme frekansı 16 kHz, TIMIT veritabanı

Çizelge 3.24'den görüleceği üzere kepstrum katsayı sayısı 24 alındığında test süresi 1 saniye için % 93.4 ile en yüksek tanıma oranı elde edilmiştir. Test süresi 3 sn için kepstrum katsayı sayısı 19, 20, 24 alındığında % 99.4 ile en yüksek tanıma oranı elde edilmiştir. Elde edilen sonuçlara göre TIMIT veritabanı için en ideal kepstrum katsayıları sayısı 24 olduğu sonucu çıkarılabilir. Kepstrum katsayılarına, birinci ve ikinci dereceden türevlerinin eklenmesinin ise konuşmacı tanıma oranını azalttığı görülmüştür.

Delta katsayıları, telefon ahizesinden dolayı oluşan kayıpların azaltılmasında kullanılmakta (Sanderson 2002) ve DÖK katsayıları ile birlikte kullanıldığında metine bağımlı konuşmacı tanıma başarımını arttırmaktadır. (Soong ve Rosenberg 1988). Liu ve ark. (1996), DÖK ve MFCC vektörlerine delta katsayılarını ekleyip metinden bağımsız konuşmacı tanıma başarımında önemli oranda düşme gözlemiştir. Metinden bağımsız konuşmacı tanıma deneyleri için MFCC vektörlerine delta katsayılarının eklenmesi tanıma başarımını düşürmektedir (Liu ve ark. 1996, Kinnunen 2003). Çizelge 3.24, çizelge 3.25 ve çizelge 3.26'da dinamik katsayıların MFCC vektörlerine eklenmesi ile elde edilen sonuçlar, daha önce elde edilen sonuçları desteklemektedir.

TIMIT veritabanı ile telefon hatlarını modellemek için konuşmacıların ses örneklerinin örnekleme hızı, 16 Khz'den 8 Khz'e düşürülmektedir. Bir önceki deneyde verilen analiz şartlarında elde edilen sonuçlar çizelge 3.25'de görülmektedir.

Çizelge 3.25 Kepstrum katsayıları değişimlerinin konuşmacı tanıma etkisi (%)

Kepstrum katsayı sayıları	Test süresi	
	1 saniye	3 saniye
k=12	70.0	97.6
k=19	81.5	98.2
k=24	76.8	95.2
k=30	64.8	95.2
k(12)+ Δk	51.2	88.1
k(12)+ $\Delta k + \Delta\Delta k$	55.9	91.1

Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168
Örnekleme frekansı 8 kHz, TIMIT veritabanı

Çizelge 3.25’den görüleceği üzere örnekleme hızının 8 KHz’e düşürüldüğünde, Mel frekansı kepstrum sayılarının 19 alındığında test süresi 1 ve 3 sn için en yüksek konuşmacı tanıma oranı elde edilmiştir. Çizelge 3.24 ve çizelge 3.25 karşılaştırıldığında ise örnekleme frekansının düşürülmesinin konuşmacı tanıma oranını düşürdüğü görülmektedir. Kepstrum katsayısı 19 olduğu durumda test süresi 1 saniye için örnekleme hızının düşürülmesi tanıma oranını 11.4 puan düşürürken test süresi 3 saniye için tanıma oranı 1.2 puan azalmaktadır. NTIMIT veritabanı için kepstrum katsayı değişimlerinin konuşmacı tanıma etkisi incelenecektir. Deneyde kepstrum katsayıları sayısı 12, 15, 18, 20, 22, 24 şeklinde alındı. Ayrıca dinamik kepstrum katsayıları olarak Δk , ve $\Delta\Delta k$ kullanıldı. Test süresi 1 ve 3 sn için kepstrum katsayısı değişimlerine göre konuşmacı tanıma oranları çizelge 3.26’daki gibidir.

Çizelge 3.26 Kepstrum katsayıları değişimlerinin test süresine göre konuşmacı tanıma etkisi (%)

Kepstrum katsayı sayıları	Test süresi	
	1 saniye	3 saniye
k=12	28.57	54.17
k=15	29.76	60.12
k=18	30.36	63.69
k=20	27.98	69.05
k=22	28.57	61.31
k=24	26.19	61.90
k(12)+ Δk	19.05	50.60
k(12)+ $\Delta k + \Delta\Delta k$	20.24	51.79

Eğitim süresi 24 sn, test süresi 3 sn, karışım bileşen sayısı 32, konuşmacı sayısı 168
Örnekleme frekansı 16 kHz, NTIMIT veritabanı

Çizelge 3.26'dan görüleceği üzere test süresi 3 sn için kepstrum katsayı sayısı 20 alındığında en yüksek tanıma oranı % 69.05, test süresi 1 sn için kepstrum katsayı sayısı 18 alındığında en yüksek tanıma oranı % 30.37 elde edilmiştir. Ayrıca kepstrum katsayılarının birinci ve ikinci dereceden türevini alıp öznitelik vektörü olarak kullanmanın konuşmacı tanıma oranını negatif yönde etkilediği görülmektedir.

3.3.3 Süzgeç dizileri frekans ölçekleri

Öznitelik vektörleri, kepstrum katsayıları kullanılarak oluşturulurken ses örneği frekans alanında süzgeç dizilerinden geçirilmektedir. Bu süzgeç dizilerinin merkez frekanslarının yeri ve süzgeçlerin bant genişlikleri konuşmacı tanıma başarımını etkilemektedir (Claudio 1999). Süzgeçlerin hazırlandığı bu frekans ölçekleri insan işitme yapısı üzerine yapılan deneyler sonucu ortaya çıkmaktadır. İşitme yapısı ölçümlerine göre belirlenen frekans ölçekleri Mel, Bark ve ERB ölçekleridir. Süzgeçler, eşit aralıklarla yerleştirildiğinde doğrusal ölçek olarak ifade edilmektedir. İlk olarak bu ölçekler tanımlanacak daha sonra konuşmacı tanıma etkileri incelenecektir.

3.3.3.1 Mel ölçek

Konuşmacı tanıma uygulamalarında genellikle Mel ölçek süzgeç dizileri kullanılmaktadır. Slaney (1998), tarafından tanımlanan Mel frekans ölçeğinde süzgeç dizilerinin 1000 Hz altında doğrusal, 1000 Hz üzerinde ise logaritmik olarak düzenlenmesidir. Çizelge 3.18'de 133- 8 kHz konuşma frekans bandı için bu frekans ölçeğinin merkez frekansları ve bant genişlikleri görülmektedir.

3.3.3.2 Bark ölçek

Mel ölçek dışında bir başka süzgeç dizisi oluşturma yöntemi de Bark ölçek süzgeçler kullanmaktır. Ses frekansından belirli bir frekans aralığına bir eşleştirme yöntemi olan Bark ölçeği denklem 3.47'deki formülle açıklanabilir (Picone 1996).

$$\text{Bark}(f) = 13 \arctan\left(\frac{0.76f}{1000}\right) + 3.5 \arctan\left(\frac{f^2}{7500^2}\right) \quad (3.47)$$

Buradaki frekans ölçeğinin birimi kritik bant genişliği oranı ya da bark olarak adlandırılır.

3.3.3.3 ERB ölçek

Bir süzgeç için Eşdeğer dörtgensel bant genişliği (ERB), o süzgecin geçirdiği toplam beyaz gürültü gücüne eşit güçte gürültü geçiren ideal dörtgensel bir süzgecin bant genişliği olarak tanımlanmaktadır. Moore ve Glasberg (1983), deneysel ölçümlerle insan işitsel süzgeçlerinin ERB'si ile süzgeçlerin merkez frekansları arasındaki bağıntıyı denklem 3.44'deki gibi tanımlamaktadır (Julius ve ark. 1999).

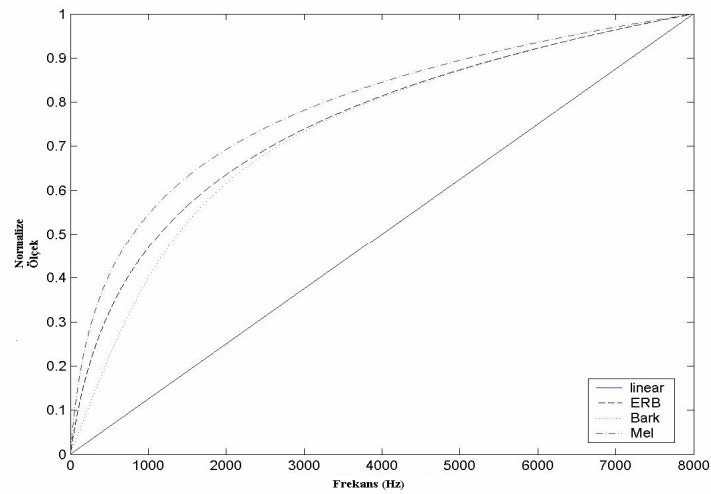
$$ERB(f) = 0.108f + 24.7 \quad (3.48)$$

Bu denklemde f in birimi Hz dir.

3.3.3.4 Doğrusal ölçek

Doğrusal ölçek, TIMIT veritabanı için 133-8000 Hz, NTIMIT veritabanı için 300-3400 Hz arasına eşit aralıklarla, merkez frekansları 66.67 Hz olan üçgen süzgeçlerin % 50 örtüşme uygulanarak düzgün aralıklarla yerleştirilmesi ile elde edilir. Slaney (1998), tanımladığı Mel ölçeği doğrusal aralıkta merkez frekansları arası uzaklık 66.67 Hz kullanmakta ve çizelge 3.19'dan görüleceği üzere iyi konuşmacı tanıma başarımı elde edilmektedir. Frekans spektrumunda 1000 Hz üstü logaritmik yerine doğrusal alınmakta, bu şekilde doğrusal ölçek elde edilmektedir.

Şekil 3.43'de 0-8000 Hz aralığında maksimum değerine normalize edilmiş Mel, doğrusal, Bark, ERB ölçekleri görülmektedir.



Şekil 3.43 Frekans ölçekleri karşılaştırması

Frekans ölçeklerinin konuşmacı tanıma etkisini incelemek için öznitelik vektörü şu şekilde oluşturulmaktadır. Konuşma, TIMIT veritabanında 20 ms uzunluğunda çerçevelere ayrılıp Hamming pencereleme uygulanmaktadır. Konuşma parçasının güç spektrumu alınarak frekans alanındaki ifadesi elde edilmektedir. İşarete ön vurgulama işlemi uygulanmamaktadır. Elde edilen bu işaret süzgeç dizilerinden geçirilir. Üçgen süzgeç dizileri Mel, Bark, ERB ve doğrusal frekans ölçeklerine göre yerleştirilir. Elde edilen işaretin logaritması alınıp AKD uygulanır. Her bir çerçeve 24 kepstrum katsayısı ile ifade edilmektedir. Süzgeçler TIMIT veritabanı için 0-8000 Hz arasına yerleştirilmiştir. Konuşmacıların eğitimi için 8 cümle test için 3 sn uzunluğunda cümle parçası uygulanmaktadır. Eğitim için BM algoritması uygulanıp, minimum değışinti sınırı 0.01 alınmaktadır. Model başlangıç değeri k-ortalama algoritması ile kestirilmektedir. Bu durumda aşağıdaki deneyler yapılmıştır.

1. İki değışik konuşmacı grubu için frekans ölçekleri değışimine göre doğru konuşmacı tanıma oranları incelenecektir. Konuşmacı grupları, 168 kişiden oluşan test dizini ve 630 kişiden oluşan TIMIT veritabanının tamamıdır. Çizelge 3.27’de bu iki konuşmacı grubu için yukarıda belirtilen frekans ölçeklerinde süzgeçlerin yerleştirilmesi ile elde edilen konuşmacı tanıma oranları görülmektedir.

Çizelge 3.27 Değışik süzgeç ölçekleri için konuşmacı tanıma oranları (%)

Konuşmacı sayısı	Doğrusal	Mel	Bark	ERB
168	100	99.4	98.81	100
630	100	99.4	99.68	99.68

Süzgeç aralığı 0-8000 Hz, kepstrum katsayı sayısı 24, örnekleme frekansı 16 kHz, karışım bileşen sayısı 32, TIMIT veritabanı

Çizelge 3.27’den görüleceği üzere konuşmacı sayısı 168 kişi için doğrusal ve ERB frekans ölçekleri kullanılarak % 100 lük konuşmacı tanıma oranı elde edilmektedir. Bu konuşmacılar test edilirken 3 sn uzunluğunda 2 ayrı cümle kullanılarak test işlemine tabii tutulmuştur. Veritabanının tamamı ile yapılan deneyde doğrusal frekans ölçeği ile test edilen konuşmacı grubu için % 100, Mel ölçek için %99.4 tanıma oranı elde edilmektedir. Bu deneyde ise test için sadece 3 sn uzunluğunda bir cümle kullanılmıştır.

2. TIMIT veritabanında Gauss karışım sayısı değişimine bağlı olarak frekans ölçeklerinin değişiminin tanıma üzerine etkisi incelenecektir. Konuşmacıların ses örneklerinin örnekleme hızı 16 kHz'den 8 kHz'e düşürüldüğünde çizelge 3.28'de görülen sonuçlar elde edilmektedir.

Çizelge 3.28 Karışım sayısına bağlı olarak değişik frekans ölçekleri için tanıma oranları

Karışım bileşen sayısı	Doğrusal	Mel	Bark	ERB
M=16	94.64	91.37	92.56	88.39
M=32	97.92	94.94	97.02	94.94
M=64	97.62	95.83	94.94	95.24

Süzgeç aralığı 0-8000 Hz, kepstrum katsayı sayısı 24, örnekleme frekansı 8 kHz, TIMIT veritabanı

Çizelge 3.28'den görüleceği üzere değişik Gauss karışım sayıları için en yüksek tanıma oranı doğrusal frekans ölçeğinde elde edilmektedir.

3. TIMIT veritabanı için süzgeç dizilerine bant sınırlama uygulanması durumunda tanıma oranı değişimi gözlenecektir. Süzgeç dizileri, 0-4000 Hz frekans aralığında hazırlanıp ses işaretine güç spektrumundan sonra ön vurgulama uygulanmasına bağlı olarak konuşmacı tanıma başarımı ölçülecektir. Örnekleme frekansı 16 kHz için bölüm 3.3.2'de verilen şartlarda elde edilen sonuçlar çizelge 3.29'da görülmektedir.

Çizelge 3.29 Süzgeç aralığı 0-4 kHz için değişik frekans ölçekleri için konuşmacı tanıma oranları (%)

	Doğrusal	Mel	Bark	ERB
Ön vurgulamasız	97.92	95.24	92.86	98.81
Ön vurgulamalı	96.43	96.73	95.54	96.73

Süzgeç aralığı 0-4000 Hz, kepstrum katsayı sayısı 20, örnekleme frekansı 16 kHz, TIMIT veritabanı

Çizelge 3.29'dan görüleceği üzere süzgeçler 0-4000 Hz aralığında yerleştirildiğinde Mel ölçeğinde ön vurgulamalı % 96.73, ERB ölçeği kullanılması durumunda ön vurgulamasız % 98.81 konuşmacı tanıma oranı elde edilmektedir. TIMIT veritabanında bant sınırlaması uygulanması durumunda ERB ölçek, Mel ölçeğe nazaran daha iyi tanıma sağlamaktadır. Reynolds ve ark. (1995), Mel frekans ölçeği kullanarak aynı şartlarda %95.2 konuşmacı tanıma oranı elde etmiştir.

4. Dört deęişik ölçekteki üçgen süzgeç dizileri bant sınırlamalı (0-4000 Hz) ve bant sınırlamasız (0-8000 Hz) frekans aralığında yerleştirilmektedir. Doğrusal, ERB, Mel, bark frekans ölçekleri için kepstrum katsayıları deęişimlerine göre konuşmacı tanıma oranları Çizelge 3.30'daki gibidir.

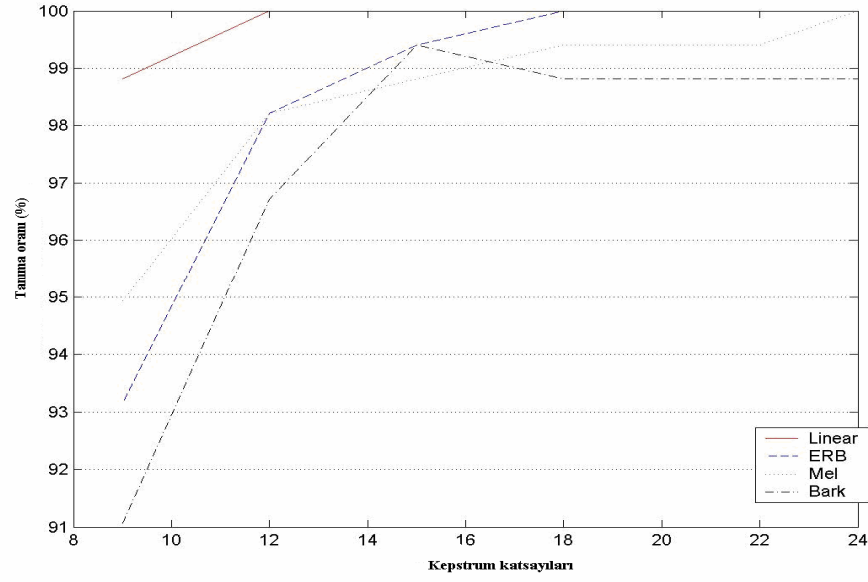
Çizelge 3.30 Deęişik frekans ölçekleri için konuşmacı tanıma oranları (%)

Kepstrum katsayıları	Doğrusal ölçek		Mel ölçek		Bark ölçek		ERB ölçek	
	0-8kHz	0-4kHz	0-8kHz	0-4kHz	0-8kHz	0-4kHz	0-8kHz	0-4kHz
k1-k9	98.21	92.86	94.94	90.48	91.07	90.48	93.15	96.72
k1-k12	100	94.64	98.21	92.56	96.72	94.94	98.21	98.81
k1-k15	100	95.24	98.81	93.45	99.4	93.15	99.4	97.02
k1-k18	100	97.92	99.4	97.32	98.81	96.43	100	97.32
k1-k20	100	97.92	99.4	96.73	98.81	95.54	100	98.81
k1-k22	100	92.86	99.4	95.54	98.81	94.64	100	95.24
k1-k24	100	91.96	99.4	96.13	98.81	88.10	100	95.83

Örnekleme frekansı 16 kHz, karışım bileşen sayısı 32, TIMIT veritabanı Mel, Bark ölçek ön vurgulamalı, Doğrusal ve ERB ölçek ön vurgulamasız, konuşmacı sayısı 168

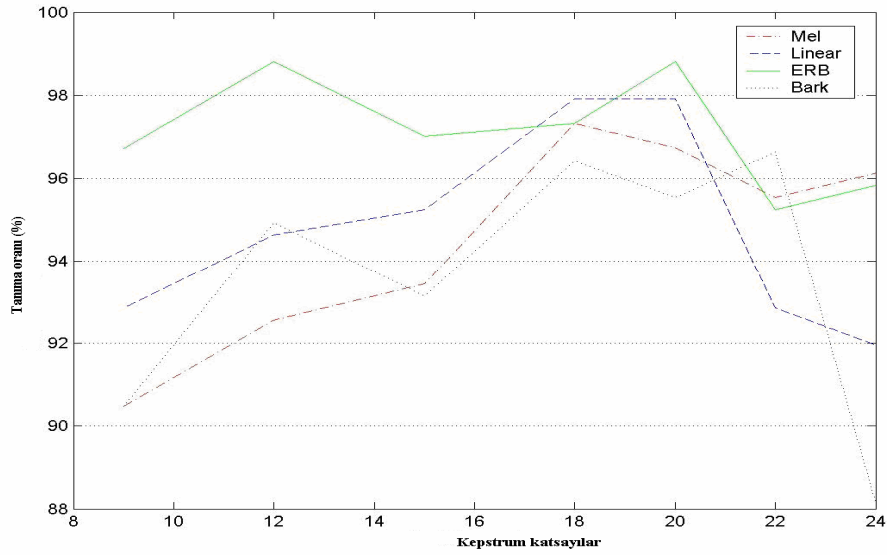
Çizelge 3.30'dan görüleceęi üzere süzgeç aralığı 0-8 kHz için en yüksek tanıma doğrusal ve ERB ölçeklerinde, süzgeç aralığı 0-4 kHz için en yüksek tanıma oranı ERB ölçeğinde gözlenmektedir. Frekans ölçeklerinin kepstrum katsayılarına baęlı olarak deęişimi Şekil 3.44'de daha ayrıntılı görülmektedir.

Süzgeçlerin yerleştirildięi bant aralığı 0-8000 Hz için, doğrusal ve ERB ölçekleri kepstrum katsayısı 18 ve üzeri olması durumunda % 100 lük konuşmacı tanıma elde edilmektedir.



Şekil 3.44 Değişik frekans ölçeklerinin kepstrum katsayıları değişimlerine bağlı olarak karşılaştırılması (0-8000 Hz)

Bant aralığı 0-4000 Hz için doğrusal, Mel, Bark, ERB frekans ölçeklerinde değişik kepstrum katsayıları için konuşmacı tanıma oranları şekil 3.45’de görülmektedir.



Şekil 3.45 Değişik frekans ölçeklerinin kepstrum katsayı değişimlerine bağlı olarak karşılaştırılması (0-4000 Hz)

Süzgeçlerin yerleştirildiği bant aralığı 0-4000 Hz için ERB ölçeğinde kepstrum katsayılarının 12 ve 20 olduğu durumlarda en yüksek (%98.81) konuşmacı tanıma oranı elde edilmiştir.

5. Doğrusal, Mel, Bark ve ERB frekans ölçeklerinin NTIMIT veritabanında karşılaştırılması yapılacaktır. Konuşma işareti 25 msn uzunluğunda çerçeveler ayrılıp 10 msn örtüşme uygulanmaktadır. İşaretin genlik spektrumu için 512 nokta ayrık fourier dönüşümü uygulanır. Üçgen süzgeç dizisi 300-3400 Hz frekans aralığında, 4 değişik frekans ölçeğine bağlı olarak yerleştirilmiştir. Süzgeçten geçirilen işaretin logaritması alınıp ayrık kosinüs dönüşümü alınmaktadır. Her bir çerçeve için 20 kepstrum katsayı alınıp, konuşma işaretine ön vurgulama uygulanmayıp, Gauss karışım bileşen sayısı 32 uygulanmaktadır. Her bir konuşmacı sekiz cümle (yaklaşık 24 saniye) kullanılarak eğitilmekte, 3 saniye uzunluğunda cümleler kullanılarak test edilmektedir. Çizelge 3.31’de NTIMIT veritabanı için değişik frekans ölçeklerinde konuşmacı tanıma oranları görülmektedir.

Çizelge 3.31 Değişik frekans ölçekleri için konuşmacı tanıma oranları (%)

Konuşmacı sayısı	Doğrusal	Mel	Bark	ERB
168	70.24	69.05	58.33	68.45

Kepstrum katsayı sayısı 20, karışım bileşen sayısı 32, NTIMIT veritabanı

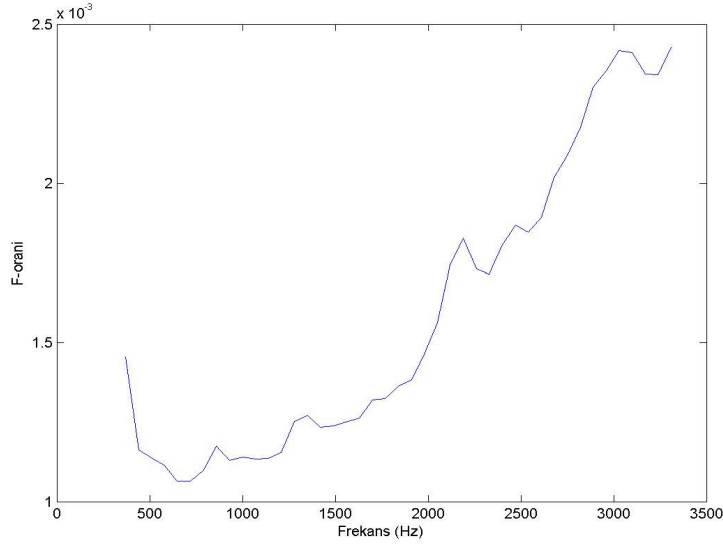
Çizelge 3.31’den görüleceği üzere doğrusal frekans ölçeği ile % 70.24 konuşmacı tanıma oranı elde edilmiştir. NTIMIT veritabanı için tanıma oranına göre süzgeç dizilerinin yerleştirildiği frekans ölçekleri; doğrusal, Mel, ERB ve bark olarak sıralanmaktadır.

Dört değişik frekans ölçeği ile yapılan beş deneyden de görüleceği üzere doğrusal frekans ölçeği diğer frekans ölçeklerine nazaran daha yüksek tanıma başarımı sağlamaktadır. Konuşmacı tanıma amacıyla yapılan bazı çalışmalar göstermiştir ki (Mokhtari 1998, Orman ve Arslan 2001, Kinnunen 2003), konuşma bandındaki orta ve yüksek frekanslar düşük frekanslara göre daha önemlidir. Oysaki MFCC’de düşük frekanslar vurgulanmaktadır.

F-oranı ölçütü kullanılarak frekans bandının etkili olduğu bölgeler bulunabilir. F-oranı konuşmacılar arası özniteliklerin ortalamasının konuşmacı içi değişimlerinin

ortalaması olarak tanımlanmaktadır. İyi bir öznitelik, konuşmacılar arasında büyük değişimlere, konuşmacı için ise düşük değişimlere sahip olmalıdır. Bu şekilde yüksek F-oranı olması arzu edilir. Bölüm 3.3.4’de F-oranı hesabı ayrıntılı anlatılmaktadır.

Şekil 3.46’da NTIMIT veritabanında konuşmacı 300-3380 Hz frekans bandına 70 Hz aralıkla doğrusal ölçekte 43 adet üçgen süzgeç yerleştirilmiştir. Süzgeç dizilerinden geçirilen konuşma işaretinin F- oranı hesaplanmaktadır.



Şekil 3.46 NTIMIT veritabanı test dizini (168 konuşmacı) için süzgeçlerin yerleştirildiği frekans bandı F-oranı

Şekil 3.46’dan görüleceği üzere 500-2000 Hz arası frekans bandı için F-oranı düşük olup, daha az ayırt ediciliğe sahiptir. 2000-3400 Hz arası diğer frekanslar ile karşılaştırıldığında daha fazla ayırt edicilik özelliği göstermektedir.

Şekil 3.46’da elde edilen sonuçlardan görüleceği üzere, NTIMIT veritabanı için süzgeçlerin yerleştirildiği orta ve yüksek frekans bölgeleri daha önemli olmaktadır. Mel ölçeğinde ise bu frekans bölgelerine daha az süzgeç yerleştirilmektedir. Doğrusal ölçekte bu frekans bandına diğer frekans ölçeklerine göre daha fazla süzgeç yerleştirildiği için tanıma başarımı daha iyi olmaktadır. Kinnunen (2003), Helsinki ve TIMIT veritabanları ile dört frekans ölçeğini VN algoritmasını kullanarak karşılaştırmıştır. Karşılaştırma sonuçlarına göre en yüksek konuşmacı başarımı doğrusal ölçek ile elde edilmiştir. Bu sonuçlar bu tezde GKM ile TIMIT ve NTIMIT veritabanı için elde edilen başarımların oranlarını desteklemektedir.

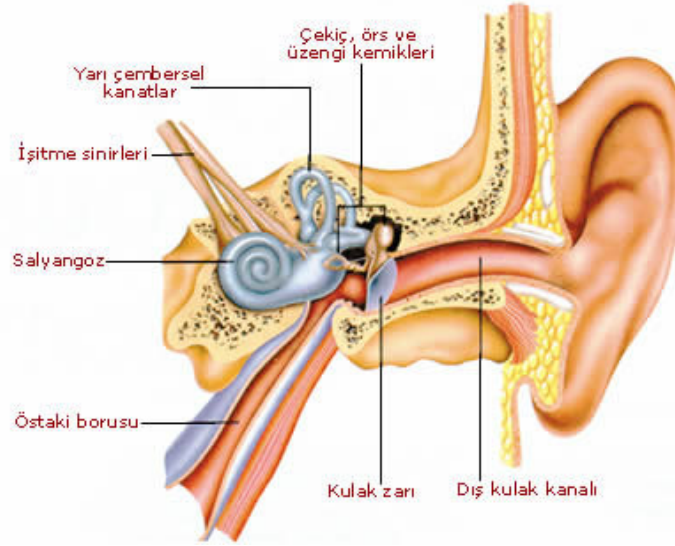
3.3.4 İnsan işitsel sistemi benzetiminin konuşmacı tanımaya uygulanması

İnsanın işitsel sistemi, duyabileceğimiz tüm karmaşık akustik işaretleri duyumsayacak, ayırabilecek ve tanımlayacak olağanüstü bir mekanizma örneği sergilemektedir. İnsan kulağının yaptığı analizin ses tanıma için en güçlü ön-işleme olduğuna inanan birçok bilim adamının varlığı, insanın ses üretme ve işitme mekanizmalarının önemle incelenmesine yol açmıştır. Ses işaretlerinin işlenmesinde işitsel sistemin matematiksel modelin kullanılmasının sağlayacağı yarar, oluşturulacak bir işitsel modelin gerçek durumu ne ölçüde yansıtacağı, bu kısmın gerçeğine ne derecede yakın temsil edildiğine son derece bağlı olacaktır (Ertaş 2002). Bu kısımda ilk olarak insan kulağının yapısı, basilar membranın bir gamaton süzgeç dizisi ile modellenmesi ve onun ses işaretine cevabı tanıtılmaktadır. Son olarak gamaton süzgeç dizilerinin konuşmacı tanıma başarımı ölçülmektedir.

3.3.4.1 İnsan kulağının yapısı ve işitme

İnsanın işitme sistemini bilgisayarda modelleyebilmek için, kulağın yapısı bilinmesi gerekmektedir. İnsan kulağında ses algılanması şu şekilde olur. Duyma işlemi, bilindiği gibi havada yayılan titreşimlerle başlar. Bu titreşimler kulak kepçesinde güçlendirilir. Kulak kepçesi bir tür megafon görevini yapar ve ses dalgaları, dış kulak yolunda yoğunlaştırılır. Bu şekilde ses dalgalarının şiddeti yaklaşık 17 desibel artar. Dış kulak yolundan bu şekilde geçen ses titreşimleri, kulak zarına gelir. Kulak zarı öylesine hassastır ki, molekül boyutundaki titreşimleri bile algılar. Zarın bir diğer özelliği ise, bir titreşim aldıktan sonra, hemen tekrar normal durumuna dönmesidir. Şekil 3.47'de kulak yapısı görülmektedir.

Titreşimler kulak zarından orta kulağa geçer. Orta kulak, aşırı derecede yüksek sesleri aşağı indirmek gibi bir tür "tampon" özelliği gösterir. Bu özellik, örs, çekiç ve üzengi kemiklerini kontrol eden, vücudun en küçük boyuttaki iki kası tarafından sağlanır. Bu kaslar, aşırı derecede yüksek seslerin iç kulağa geçirilmeden önce zayıflama sağlar. Bu sayede bizim için şok oluşturacak derecede yüksek sesleri daha alçak düzeylerde duyarız. Üzengi ise oval pencere yardımı ile salyangozun kemik kısmını aşarak zarsı labirente ulaşır

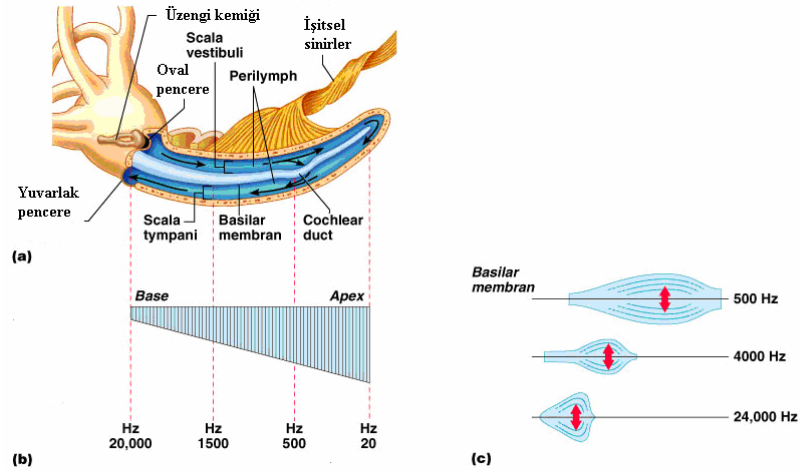


Şekil 3.47 Kulağın yapısı

Orta kulaktaki mekanik hareketlerin sese dönüştürülmeye başlaması iç kulak adı verilen bölgede olur. İç kulakta, içi sıvıyla kaplı olan spiral bir organ yer alır. Sahip olduğu şekil nedeniyle "salyangoz" olarak adlandırılır. Orta kulağın en son parçası olan üzengi kemiği, salyangozun başlangıcındaki bir zara bağlıdır. Orta kulaktaki mekanik titreşimler, bu bağlantıyla iç kulağın sıvısına yani basilar membrana aktarılmış olur. İç kulaktaki sıvıya ulaşan titreşimler, bu sıvının içinde dalgalanmalar oluşturur. Salyangozun iç duvarlarında ise, bu sıvının dalgalanmalarından etkilenen küçük tüycükler vardır. Bu tüycükler, sıvıdaki dalgalanmalara göre belli belirsiz şekilde hareketlenir. Eğer güçlü bir ses gelirse, daha fazla sayıdaki tüycük, daha güçlü bir biçimde eğilir. Dış dünyadaki her ayrı ses frekansı, bu tüycükler üzerinde ayrı etkileşimler oluşturmaktadır. Tüycükler, aslında salyangozun iç duvarını çevreleyen yaklaşık 20 bin ayrı hücrenin tepesinde yer alan birer mekanizmadır. Tüycükler bir titreşim algıladıklarında, birbirlerini iterek hareket ederler. İşte bu hareket, tüycüklerin altındaki hücrelerin kapılarını açar. Bu sayede hücrelere iyon girişi olur. Tüycükler ters yöne yattıklarında ise hücre kapıları bu kez kapanır. Bu sürekli hareket, hücrelerin kimyasal dengelerini de sürekli değiştirir ve elektrik uyarıları üretmelerini sağlar. Bu elektrik uyarıları, sinirler aracılığıyla beyine iletilir ve beyin de bunları yorumlayarak ses haline getirir.

3.3.4.2 Basilar Membran ve Gamaton Süzgeçler

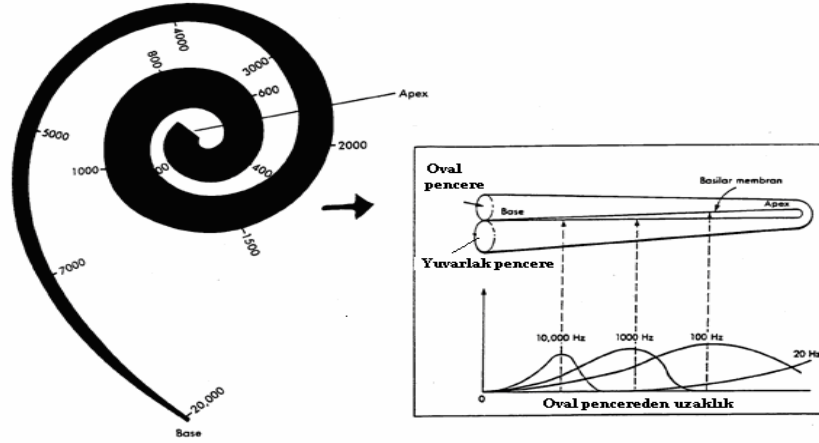
Basilar membran sesin sıklığını algılayan bir mekanik analizördür. Basilar membranın 33 mm' lik uzunluğu boyunca düzgün bir yapıda olduğunu varsayarsak bu durumda üzeninin salyangozun başlangıcındaki zarı içeriye doğru hareketlendirmesi basilar membranın aşağıya doğru hareketine yol açacaktır. Ancak basilar membranın mekanik öznitelikleri farklılık gösterdiği için Şekil 3.48'de görüleceği üzere üzenği kemiğinin taban parçasının hareketleri basilar membran içinde ilerleyen bir dalga serisi başlatır. Bu dalga önce bir doruk noktaya ulaşır daha sonra hızla düşer. Doruk nokta ile üzenği kemiği arasındaki uzaklık dalgayı başlatan titreşimlerin frekansı ile değişir. Örneğin düşük frekanslı sesler (100Hz) basilar membranın uç kısmında (apex), orta frekanslı sesler (1000 Hz) membranın ortalarında, yüksek frekanslı sesler ise (10000 Hz) membranın başlangıç kısmı (base) bölümünde dalga oluşumuna yol açarlar. İlerleyen dalganın bir diğer özelliği basilar membranın ilk bölümünde hızlı hareket etmesi, salyangoz içinde daha uzağa doğru ilerledikçe giderek daha yavaş hareket etmesidir. Bu olayın nedeni üzenği kemiğine yakın basilar liflerin esneme katsayısının yüksek olması ve bu katsayının ileri doğru gidildikçe giderek küçülmesidir.



Şekil 3.48 (a) Basilar membranın yapısı ve dalgaların hareket yönleri (b) basilar membranın duyarlı olduğu frekans bölgeleri (c) basilar membran boyunca ses dalgası hareketi

Basilar membran, karmaşık seslerin spektral analizini gerçekleştirmektedir. Şöyleki basilar membranın apeks bölgesi en iyi 20 Hz frekansındaki seslere yanıt verirken, bazal bölümü 20 kHz frekansındaki seslere duyarlıdır. Aynı şekilde, bir dizi

frekans bileşenleri olan karmaşık bir sinyal de, basilar membran boyunca farklı noktalarda her biri ayrı bir frekans bileşenine karşı gelen maksimum genlikli titreşimler oluşturacaktır, bu da basilar membranın bant geçiren süzgeç dizisi gibi davrandığı anlamına gelir. Şekil 3.49'da basilar membranın oval pencereden uzaklığına bağlı olarak bant geçiren süzgeç dizileri ile modellenmesi görülmektedir.

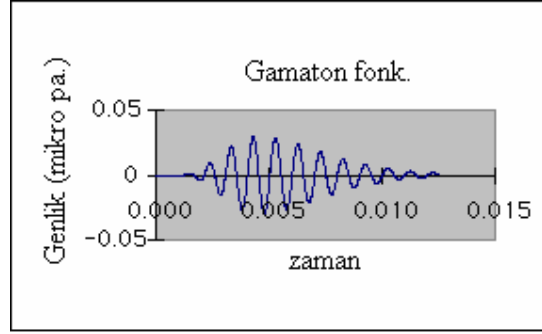


Şekil 3.49 Salyangoz yapı boyunca basilar membranın, duyarlı olduğu frekans bölgeleri ve bant geçiren süzgeç özelliği

Basilar membran üzerindeki bir işitsel sinir telinin dürtü cevabı denklem 3.49'da ifade edilen gamaton fonksiyonu ile temsil edilir.

$$g(t) = t^{n-1} \exp(-2\pi bt) \cos(2\pi f_0 t + \phi) u(t) \quad (3.49)$$

Bu ifade de $u(t)$ birim basamak fonksiyonu, f_0 sinir telinin rezonans frekansı, ϕ dürtü cevabının evresi, n fonksiyonun mertebesi ve b ise dürtü cevabının uzunluğunu belirleyen frekansa bağlı bir sabittir (Holdsworth 1988). Şekil 3.50'de gamaton fonksiyonunun dürtü cevabı görülmektedir. İşitsel sinir telleri kılcal sinir hücreleri vasıtası ile basilar membran üzerindeki, kendilerinin duyarlı olduğu frekanslara karşı gelen noktalara bağlıdır. İşitsel sinir tellerinin dürtü cevabı aslında basilar membranın ilgili noktalarının dürtü cevaplarıdır. Basilar membranın tamamı ise bir gamaton süzgeç dizisi ile temsil edilebilirler.



Şekil 3.50 Gamaton fonksiyonunun dürtü cevabı

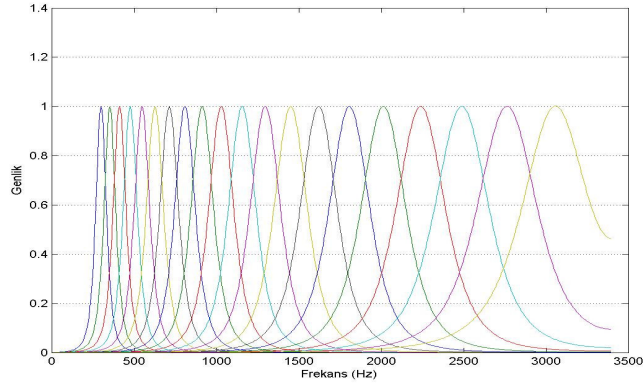
Gamaton süzgecin bant genişliğini belirleyen parametresi b , ERB cinsinden denklem 3.50'deki gibi ifade edilmektedir (Ertaş 2002).

$$b = a_n^{-1} ERB(f_0) \quad (3.50)$$

Burada f_0 süzgecin merkez frekansı ve a_n ise süzgeç mertebesine bağlı bir sabittir. 4. mertebeden ($n=4$) bir gamaton süzgecin ERB si ile b parametresi arasında $b = 1.019ERB$ şeklinde bir ilişki bulunmaktadır. 4. dereceden gamaton fonksiyonunun z-dönüşümü denklem 3.51'deki gibi temsil edilir (Slaney 1993).

$$z_{ir} = \frac{\alpha z e^{\phi} e^{(-2\pi b)} e^{(-2i\pi f_c)} \left(z^2 + \left(e^{(-2\pi b)^2} \right) \left(e^{(-2i\pi f_c)^2} \right) + 4z e^{(-2\pi b)} e^{(-2i\pi f_c)} \right)}{z - e^{(-2\pi b)} e^{(-2i\pi f_c)^4}} \quad (3.51)$$

Bu ifadeden her bir süzgeç için z dönüşümü pay ve payda katsayıları bulunur. Bu katsayılar; örnekleme frekansı, ERB bant genişliği ve ERB ölçekte süzgeç dizisi merkez frekanslarına dolayısıyla toplam süzgeç sayısı, en düşük ve en yüksek süzgeç frekansı değerlerine bağlı olarak elde edilir. Şekil 3.51'de yukarıda belirttiğimiz yöntemle elde edilen 300-3400 Hz bant genişliğinde 20 adet gamaton süzgeç dizisi görülmektedir.



Şekil 3.51 Yirmi adet gamaton süzgeç dizisi

3.3.4.3 Gamaton süzgeçlerin konuşmacı tanıma uygulaması

İnsan kulağının mekanik hareketini frekansa dönüştüren yapısı basilar membran, gamaton süzgeç dizileri ile ifade edilmektedir. Gamaton süzgeç dizileri ERB ölçeğinde tanımlanmaktadır. Gamaton süzgeçler ERB ölçeğine göre eşit aralıklarla yerleştirilmektedir. Süzgeçlerin yerleştirildiği frekans aralığı TIMIT için 100-8000 Hz, NTIMIT için ise 300-3400 Hz alınır.

Konuşmacıyı karakterize eden öznelik vektörleri elde edilirken konuşmacılara ait ses örnekleri parçalara ayrılıp pencerelendikten sonra pencere örtüşme süresi 10 ms alınır. FFT örnek sayısı (N) 512 alınıp, FFT çıkışının karesi alınmaktadır. Kepstrum katsayıları elde edilirken ön vurgulama uygulanmaz. Her bir çerçeve için toplam 20 adet kepstrum katsayı kullanılır ve 0. kepstrum katsayısı öznelik vektörü olarak alınmaz. Gauss karışım bileşen sayısı 32 alınmaktadır. TIMIT ve NTIMIT veri tabanlarının 168 konuşmacıdan oluşan test dizini ve TIMIT veritabanının tamamı (630 konuşmacı) olmak üzere iki grup üzerinde deneyler yapılmaktadır. Modelin eğitim safhasında yaklaşık 24 saniye uzunluğunda 8 cümle (2 Sa, 3 Si, 3 Sx cümleleri) kullanılarak eğitim kümesi oluşturulur. Modelin test safhasında 3 saniye uzunluğunda konuşma parçası (yaklaşık 1 Sx cümlesi) kullanılmıştır.

Yukarıda tanımlanan konuşmacı tanıma sistemi parametrelerine bağlı olarak elde edilen öznelik vektörleri her bir pencere için 20 katsayı ile ifade edilmektedir. Bu durumda 32 süzgeç için merkez frekansları aşağıdaki gibidir.

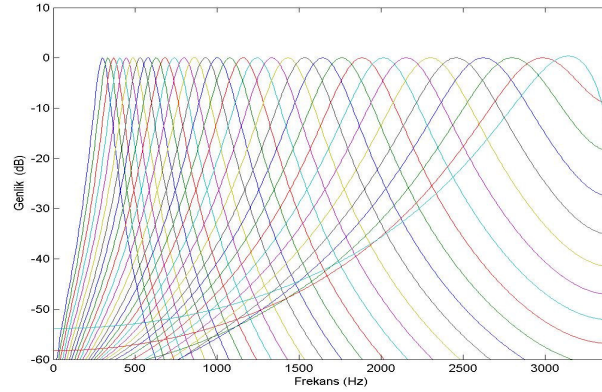
1.0e+003 *

0.3000 0.3328 0.3676 0.4047 0.4439 0.4857 0.5300 0.5771 0.6271 0.6802 0.7366
 0.7965 0.8601 0.9276 0.9994 1.0755 1.1565 1.2424 1.3337 1.4306 1.5335 1.6429
 1.7590 1.8823 2.0133 2.1524 2.3001 2.4570 2.6236 2.8005 2.9884 3.1880

ERB bant genişliği değerleri aşağıdaki gibidir.

57.0817 60.6228 64.3836 68.3776 72.6195 77.1244 81.9089 86.9901 92.3866 98.1179
 104.2047 110.6690 117.5344 124.8257 132.5694 140.7934 149.5275 158.8036 168.6550 179.1176
 190.2292 202.0302 214.5632 227.8738 242.0100 257.0232 272.9678 289.9014 307.8856 326.9854
 347.2701 368.8132

ERB bant genişliği değerleri incelendiğinde, frekans artışına bağlı olarak bant genişliği de artmaktadır. Belirlenen bu ERB ölçek ve bant genişliği değerlerine bağlı olarak denklem 4.18’de verilen gamaton süzgecin z-dönüşümü alınır. 300- 3400 Hz aralığında 32 adet gamaton süzgeç için frekans cevabının mutlak değeri aşağıdaki gibidir. 32 süzgeç için gamaton süzgeç dB cinsinden genlik spektrumu şekil 3.52’de görülmektedir.



Şekil 3.52 Gamaton süzgeç dizisi genlik spektrumu (dB)

Şekil 3.52’deki süzgeç dizileri, konuşmacı tanıma deneylerinde şu ana kadar uyguladığımız üçgen süzgeç dizileri yerine yerleştirilecektir. Deneyde kullanılan süzgeç sayısına bağlı olarak süzgeçlerin örtüşme oranları da değişmektedir. 32 süzgecin örtüşme oranı % 44.24, 64 süzgecin örtüşme oranı % 72, 128 süzgecin örtüşme oranı % 86 olmaktadır. Bu durumda gamaton süzgeç sayısına bağlı olarak elde edilen konuşmacı tanıma oranları çizelge 3.32’de görülmektedir.

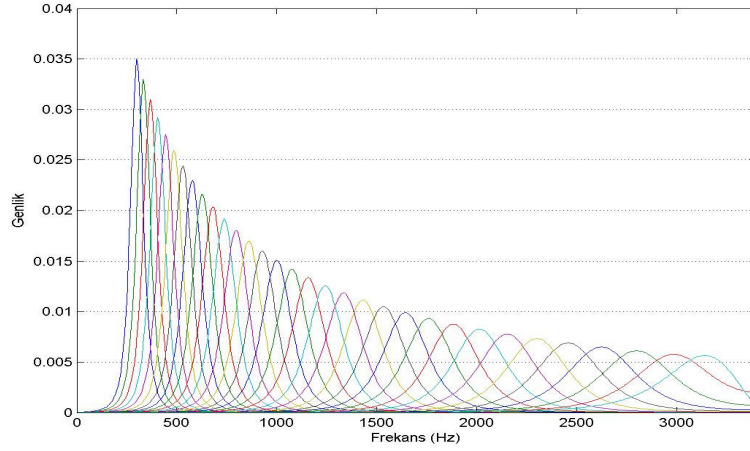
Çizelge 3.32 Gamaton süzgeç sayısına bağlı konuşmacı tanıma oranları (%)

Veritabanları	Gamaton süzgeç sayısı		
	32 süzgeç	64 süzgeç	128 süzgeç
NTIMIT	35.12	34.52	35.12

Kepstrum katsayı sayısı 20, karışım bileşen sayısı 32, konuşmacı sayısı 168

Gamaton süzgeçler, üçgen süzgeç dizileriyle karşılaştırıldığında tanıma oranında yaklaşık 30 puanlık düşüş gözlenmektedir.

Şekil 3.52’de görüleceği üzere süzgeç bant genişliğinin artmasına rağmen her bir süzgecin genliği 1 düzeyindedir. Her bir süzgecin genliği, bant genişliğinin değişimine göre düzenlenirse, şekil 3.53’deki bant geçiren gamaton süzgeçler dizisi elde edilir.



Şekil 3.53 Otuz iki adet gamaton süzgecin genliği bant genişliğine göre düzenlenmiş genlik spektrumu

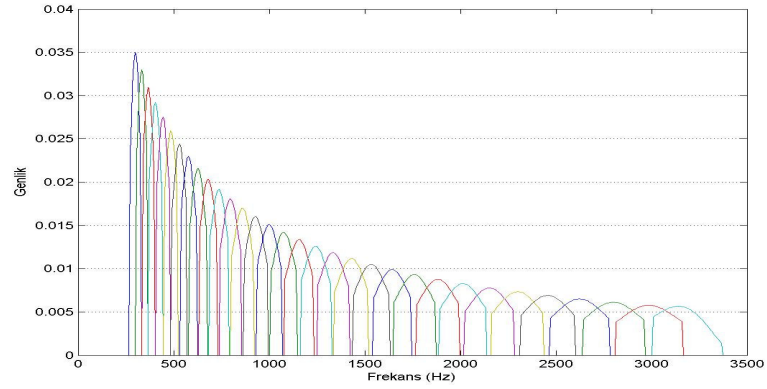
Süzgeçlerin genliği ERB bant genişliğine bağlı olarak değiştirilmesi TIMIT ve NTIMIT veritabanında incelenecektir. Şekil 3.53’deki gamaton süzgeçler dizisi kullanılarak elde edilen konuşmacı tanıma oranları çizelge 3.33’deki gibidir.

Çizelge 3.33 Genliği bant genişliğine göre düzenlenmiş gamaton süzgeçler için konuşmacı tanıma oranları (%)

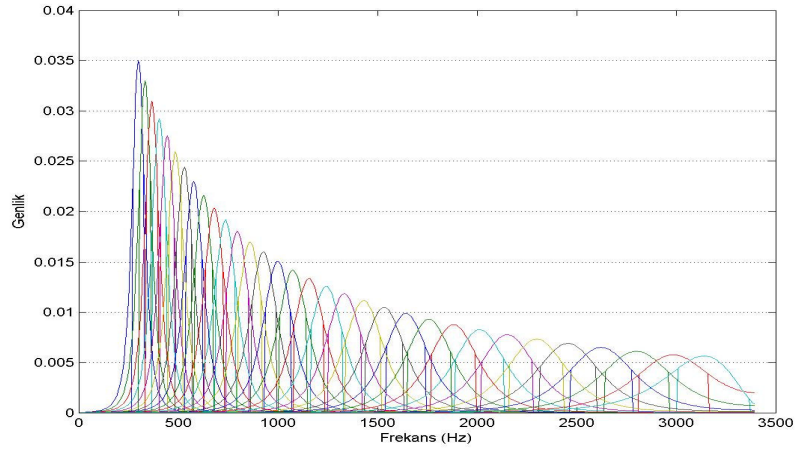
Veritabanları	Gamaton süzgeç sayısı		
	32 süzgeç	64 süzgeç	128 süzgeç
TIMIT	93.45	92.56	92.26
NTIMIT	31.55	35.12	35.12

Kepstrum katsayı sayısı 20, karışım bileşen sayısı 32, konuşmacı sayısı 168

Çizelge 3.32 ve çizelge 3.33 karşılaştırıldığında NTIMIT veritabanında genlik düzenlemesi konuşmacı tanıma oranını önemli oranda değiştirmemektedir. Sadece ERB bant genişliği içerisindeki süzgeç değerlerinin alınıp, gamaton süzgecin diğer kısımlar atılarak süzgecin seçiciliğinin artırılması amaçlanmaktadır. Gamaton süzgeçler için yukarıdaki düzenleme yapılırsa 32 adet süzgeç için şekil 3.54’de görülen süzgeç dizisi elde edilir. Şekil 3.53 ve şekil 3.54’deki süzgeç dizilerini üst üste çizersek ERB bant genişliği içerisindeki sınırlama daha net olarak şekil 3.55’de görülmektedir.



Şekil 3.54 Sadece ERB bant genişliği içerisindeki süzgeç değerlerine genlik düzenlemesi uygulanırsa elde edilen süzgeç dizisi



Şekil 3.55 Gamaton süzgeçlerin sınırlandırılmış ve sınırlandırılmamış halleri

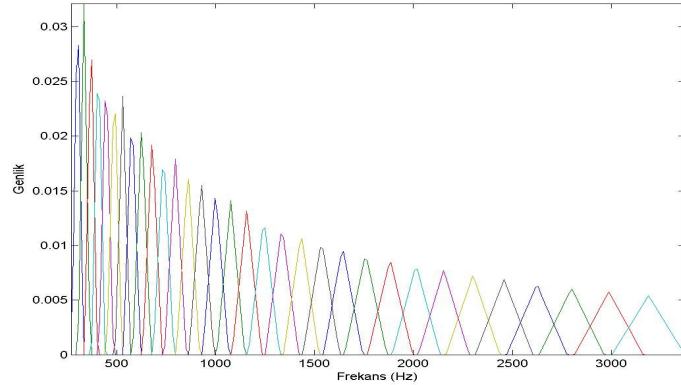
Şekil 3.55’deki gamaton süzgeçler dizisi kullanılarak elde edilen konuşmacı tanıma oranları çizelge 3.34’deki gibidir.

Çizelge 3.34 Sadece ERB bant genişliği içerisindeki süzgeç değerleri alınır elde edilen konuşmacı tanıma oranları (%)

Veri tabanları	Gamaton süzgeç sayısı		
	32 süzgeç	64 süzgeç	128 süzgeç
TIMIT	98.81	98.81	98.50
NTIMIT	42.26	48.21	39.88

Kepstrum katsayı sayısı 20, karışım bileşen sayısı 32, konuşmacı sayısı 168

NTIMIT veritabanında gamaton süzgeçlerde ERB bant genişliğine bağlı olarak yapılan kırpma, süzgeç sayısı 64 için tanıma oranını 13 puan arttırmaktadır. Aynı deney ERB ölçek ve bant genişliğinde, aynı sayıda üçgen süzgeç dizisi kullanılarak tekrar edilecektir. Kullanılan süzgeç dizisi şekil 3.56'da görülmektedir.



Şekil 3.56 ERB ölçek ve bant genişliğinde 32 adet üçgen süzgeç dizisi yerleştirilmesi

Şekil 3.56'daki ERB ölçekteki üçgen süzgeçler dizisi kullanılarak elde edilen konuşmacı tanıma oranları çizelge 3.35'deki gibidir

Çizelge 3.35 Üçgen süzgeç dizileri ile konuşmacı tanıma oranları

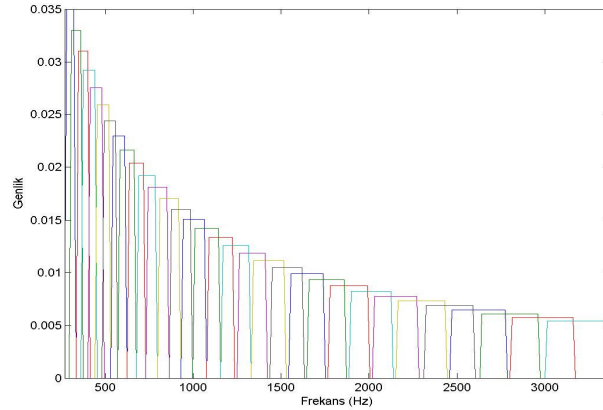
Veritabanları	Üçgen süzgeç sayısı		
	32 süzgeç	64 süzgeç	128 süzgeç
TIMIT	99.4	100	100
NTIMIT	64.29	65.48	63.69

Kepstrum katsayı sayısı TIMIT için 24 ve NTIMIT için 20, karışım bileşen sayısı 32

NTIMIT veritabanı için süzgeç sayısı 64 için gamaton ve üçgen süzgeçler karşılaştırıldığında, üçgen süzgeç dizisi kullanılması ile tanıma oranı 17 puan

artmaktadır. TIMIT veritabanında da tanıma oranı, üçgen süzgeçlerin kullanılması ile gamaton süzgeçlere nazaran belirgin bir şekilde artmaktadır. Sonuç olarak her iki veritabanında da üçgen süzgeç dizisi kullanılarak yapılan deneylerde, gamaton süzgeç dizilerine nazaran daha iyi sonuçlar elde edilmektedir.

Aynı deney ERB ölçek ve bant genişliğinde, aynı sayıda dikdörtgen süzgeç dizisi kullanılarak tekrar edilecektir. Kullanılan süzgeç dizisi şekil 3.57’de görülmektedir.



Şekil 3.57 ERB ölçek ve bant genişliğinde dikdörtgen süzgeç dizileri yerleştirilmesi

Şekil 3.57’deki ERB ölçekte yerleştirilmiş dikdörtgen süzgeçler dizisi kullanılarak elde edilen konuşmacı tanıma oranları çizelge 3.36’daki gibidir.

Çizelge 3.36 Dikdörtgen süzgeç dizileri ile konuşmacı tanıma oranları

Veritabanları	Dikdörtgen süzgeç sayısı		
	32 süzgeç	64 süzgeç	128 süzgeç
TIMIT	99.4	99.4	100
NTIMIT	62.50	64.29	66.67

Kepstrum katsayı sayısı TIMIT için 24 ve NTIMIT için 20, karışım bileşen sayısı 32

NTIMIT veritabanı için süzgeç sayısı 32 ve 64 için dikdörtgen süzgeçler üçgen süzgeçler ile karşılaştırıldığında, dikdörtgen süzgeç dizisi kullanılması ile tanıma oranı azalmakta, ancak süzgeç sayısı 128 için tanıma oranı yaklaşık 3 puan artmaktadır. TIMIT veritabanında dikdörtgen süzgeçlerin kullanılması ile üçgen süzgeçlere göre belirgin bir şekilde değişme gözlenmemektedir.

3.4 Telefon İletiminin Konuşmacı Tanıma Üzerine Etkilerinin Azaltılması

Telefon ortamının konuşmacı tanıma oranını düşürdüğü bilinmektedir (Reynolds 1996). Telefon ortamı konuşmalar üzerinde bant sınırlama, filtreleme ve gürültü ekleme etkileri yapmaktadır. Telefon ahizesi ve hattının konuşmacı tanımaya etkisini incelemek için NTIMIT veritabanı ve TIMIT veritabanının telefon ortamına benzetimi kullanılmaktadır. NTIMIT veritabanı konuşmaları üzerinde doğrusal olmayan bozulmalar ve var olmayan formantlar oluşmaktadır. Yapılan ölçümler, bu doğrusal olmayan bozulmaların karbondan yapıma telefon ahizesinden kaynaklandığını göstermiştir (Reynolds ve ark. 1995, Quatieri ve ark. 2000).

Bu bölümde, telefon hattı üzerinden gürbüz konuşmacı tanıma için değişik öznitelik vektörleri oluşturma yöntemlerinin deneysel değerlendirilmesi yapılmaktadır. Değerlendirilen öznitelik vektör setleri aşağıdaki yöntemler ile oluşturulmaktadır. Bunlar;

- 1.) Spektral değişim kompanzasyonu
- 2.) Konuşmacıların kümelenerek ağırlıklandırılması
- 3.) Öznitelik vektörü oluşturulurken konuşma frekans bandı parçalara ayrılıp, F-oranına bağlı olarak parçalara süzgeç yerleştirilmektedir.

Deneylerde, TIMIT veritabanını telefon ortamına benzetimi için, frekans ölçeği 0-8000 Hz'den 100-4000 Hz arasına sınırlandırılmakta ve örnekleme frekansı 16 kHz'den 8 kHz'e düşürülmektedir. İkinci olarak konuşmaların telefon ortamından kaydedildiği NTIMIT veritabanı kullanılmaktadır.

Tüm eğitim ve sınıflandırma adımları değişmeden sadece kullanılan öznitelik vektörleri değiştirilerek kontrollü bir karşılaştırma yapılmaktadır. Konuşmalar 32 elemanlı GKM ile modellenmektedir. NTIMIT ve TIMIT veritabanlarının 100 ve 168 konuşmacıdan oluşan test dizini üzerinde çalışmalar yapılmaktadır. TIMIT veritabanı için 15 saniye uzunluğunda 5 cümle (2 Sa, 3 Si cümleleri) kullanılarak eğitim kümesi oluşturulur. NTIMIT veritabanı için modelin eğitim safhasında yaklaşık 24 saniye uzunluğunda 8 cümle (2 Sa, 3 Si, 3 Sx cümleleri), Modelin test safhasında 3 sn uzunluğunda cümle parçası (2 Sx cümlesinin 3 sn uzunluğundaki kısmı) test kümesi olarak alınmaktadır.

3.4.1 Spektral deęişim kompanzasyonu

Telefon hattından iletilen konuşma işareti, hattın bant sınırlamasından dolayı doğrusal süzgeç etkisi oluşmaktadır. Bu durum konuşma işaretinin spektral şeklinde deęişimler oluşturmakta, hatalı konuşmacı atanmasına neden olup tanıma başarımını azaltmaktadır. Spektral deęişim kompanzasyon yöntemleri ile telefon konuşmaları için daha gürbüz öznitelikler oluşturulabilir.

Konuşma işareti, telefon hattını ifade eden doğrusal süzgeçten geçirildiğinde, konuşmanın genlik spektrumu, süzgecin genlik cevabı ile çarpılarak elde edilir. Eğer süzgecin genlik spektrumu düzgüne yakın ise kepsral öznitelik vektörleri \bar{x} ile süzgeç etkisini ifade eden \bar{h} vektörünün toplamı olarak denklem 3.52'deki gibi ifade edilir.

$$\bar{z} = \bar{x} + \bar{h} \quad (3.52)$$

burada \bar{z} , gözlenen kepsrum vektörüdür. Spektral deęişim kompanzasyonu ile hattın süzgeç etkisini gösteren \bar{h} öznitelik vektörü kaldırılması amaçlanmaktadır. Bu amaçla ortalama normalizasyonu, kepsrum fark katsayıları, frekans eğirme yaklaşımları incelenecektir.

3.4.1.1 Ortalama normalizasyonu

Ortalama normalizasyon yöntemi pek çok konuşmacı tanıma sistemlerinde uygulanmaktadır. Her bir öznitelik vektöründen elde edilen verilerin, global ortalama vektöründen çıkartılması ile telefon hattı süzgecinin etkisi kaldırılmaya çalışılır. Global ortalama vektör denklem 3.53'deki gibi ifade edilir (Reynolds ve ark. 1995).

$$\bar{m} = \frac{1}{T} \sum_{t=1}^T \bar{z}_t \quad (3.53)$$

Telefon hattı etkisi kaldırılmış kepsrum vektörü, denklem 3.54'deki gibi ifade edilir.

$$\hat{z}_t = \bar{z}_t - \bar{m} \quad (3.54)$$

Konuşmacıların eğitim ve test edilmelerinden önce kepsrum vektörlerinin her birinden global ortalama vektörü çıkartılır. Tüm öznitelik vektörleri aynı global ortalama ve konuşmacı ayırt ediciliğine sahip olacağından dolayı farklı telefon hattı

süzgeç etkilerine karşı gürbüz olur. Bu işlem sayesinde telefon hattının etkisinin yanında konuşmacıların öznitelik vektörlerinin global ortalama değerleri kaldırılmış olur. Bu durum konuşmanın ortalama spektrumunun tersinin süzgeçten geçirilmesi anlamına gelmektedir. Her ne kadar ortalama konuşma spektrumu, konuşmacıya özel bilgi içermese de, konuşmacının cümleleri arasındaki değişim hakkında bilgi verir.

3.4.1.2 Kepstrum fark katsayıları

Telefon hattının süzgeç etkilerini minimize etmek için, kanala göre değişmeyen öznitelikler kullanılır. Konuşmacı tanıma sistemlerinde uygulanan, kanala göre değişmeyen özniteliklerden biride kepstrum fark katsayılarıdır. Fark katsayıları hem dinamik bilgiyi tutmak hem de iletişim kanalındaki, zamanla değişmeyen spektral bilgiyi kaldırmak için kullanılır. Konuşmacının t . çerçevesinin fark katsayıları $\Delta \bar{z}_t$ ile gösterilip kepstral öznitelik vektörleri, w adet çerçeve parçası arasındaki fark olarak denklem 3.55'de görüldüğü gibi ifade edilir .

$$\Delta \bar{z}_t = \bar{z}_t - \bar{z}_{t-w} \quad (3.55)$$

Telefon hattı süzgeci zamanla değişmeyen veya yavaş değişen kabul edildiğinde temel terimi \bar{h} kaldırılır (Reynolds 1992). Bu durumda fark konuşma kepstrası denklem 3.56'daki gibi elde edilir.

$$\Delta \bar{z}_t = \bar{x}_t - \bar{x}_{t-w} \quad (3.56)$$

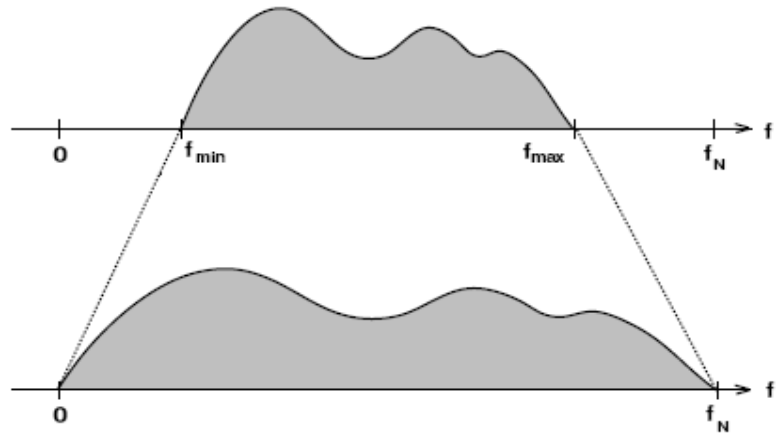
Fark katsayıları, zamanla değişen spektral değişimleri tuttuğundan dolayı, geçiş veya dinamik öznitelik olarak ifade edilir. Kepstral vektörler ise durağan öznitelik olarak adlandırılır. Öznitelik vektörü elde edilmesinde 20 adet Mel frekansı kepstrum katsayısına 20 fark katsayısı eklenir. Fark katsayıları 40 msn aralıkla (∓ 2 çerçeve) elde edilir. Elde edilen öznitelik vektörleri eğitim ve test aşamalarında kullanılır. Bu durumda konuşmacı tanıma başarımı incelenir.

3.4.1.3 Frekans eğirme

Telefon kanalının bant genişliğindeki, spektral farklılıklardan sakınmak için, FFT spektrum genliğine frekans eğirme uygulanır. Eğirme haritası frekans eksenindeki f , yeni frekans eksenindeki f' olarak denklem 3.57'deki gibi ifade edilir.

$$f' = \frac{f - f_{\min}}{f_{\max} - f_{\min}} f_N, \quad (3.57)$$

Burada f_N nyquist frekansıdır. $f_{\min}=300$ Hz, $f_{\max}=3400$ Hz alınıp $f_N=8000$ Hz alınmıştır. Şekil 3.58’de doğrusal eğirme örneği görülmektedir. Frekans eğirme ile hem $[f_{\min}, f_{\max}]$ frekans aralığı dışındaki spektral bileşenler atılmakta hem de spektrum bant genişliği büyümektedir (Reynolds 1992). Frekans eğirme sonucu elde edilen keprstrum katsayıları eğitim ve test aşamalarında kullanılır ve bu durumda konuşmacı tanıma başarımı incelenir.



Şekil 3.58 Frekans eğirme örneği

Spektral değişim kompanzasyonu yöntemlerine ait öznitelik vektörleri çıkarılırken konuşmalar 25 msn lik parçalara ayrılıp 10 msn örtüşme uygulanmıştır. Hamming pencereleyen konuşma parçaları 300-3400 Hz arası Mel ölçekte 28 adet üçgen süzgeçten geçirilip 20 adet öznitelik vektörü elde edilmiştir. Eğitim ve test için NTIMIT’in 168 kişilik test dizini kullanılmıştır. Gauss karışım bileşen sayısı 32 alınıp konuşmacılar BM algoritması ile eğitilmiş, eğitim için 8 adet cümle (yaklaşık 24 sn) test için 3 sn lik cümle parçaları kullanılmıştır. Spektral değişim kompanzasyonu yöntemlerinin konuşmacı tanıma sistemine uygulanması sonucu elde edilen tanıma oranları çizelge 3.37’de verilmiştir.

Çizelge 3.37 Spektral deęişim kompanzasyonu yöntemlerinin tanımaya etkisi

	Tanım oranı (%)
Kompanizasyon yok	69.64
Ortalama normalizasyonu	57.74
Kepsrum fark katsayıları	62.50
Frekans eęirme	42.86

Eęitim süresi 24 sn, karışım bileşen sayısı 32, NTIMIT veritabanı

Spektral deęişim kompanzasyon yöntemlerinden kepsrum fark katsayıları, ortalama normalizasyonu ve frekans eęirme yöntemlerine nazaran daha iyi sonuçlar vermesine rağmen spektral deęişim kompanzasyonu uygulanmadığı durum en iyi tanıma sağlamaktadır. Sonuç olarak, spektral deęişim kompanzasyon yöntemleri NTIMIT veritabanı için tanıma başarımını azaltmaktadır.

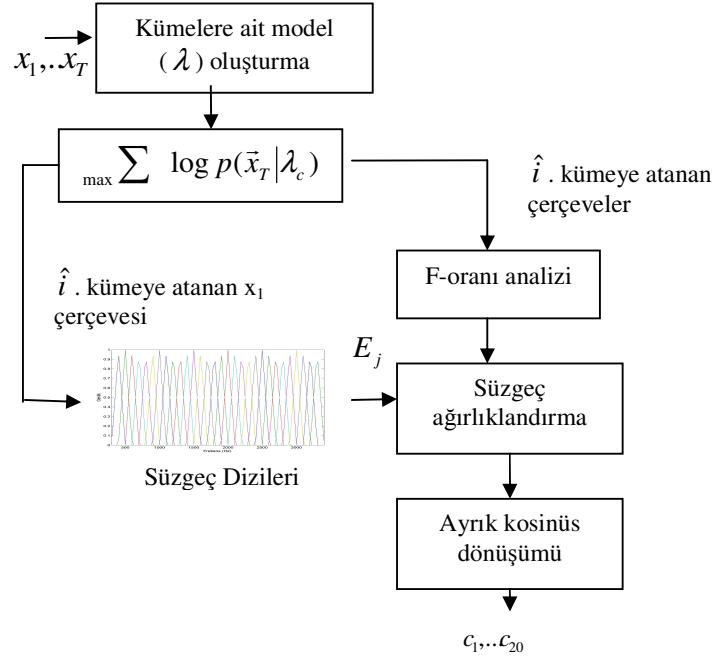
Genellikle spektral kompanzasyon yöntemleri, konuşmacıların birden fazla oturumda eęitildięi durumlarda ve kayıta farklı mikrofon veya telefon kanalı kullanıldığı veritabanları için konuşmacı tanıma oranını arttırmaktadır (Reynolds 1996, Quatieri ve ark. 2000). Oysaki NTIMIT veritabanı tek oturumda hazırlanıp, ses kaydında sadece tek bir çeşit telefon ahizesi kullanılmıştır.

3.4.2 Öznitelik vektörlerinin kümelenecek aęırlıklandırılması

Bazı yapılan çalışmalar göstermiştir ki, bazı sesler konuşmacılar arasında ayırt edicilik olarak farklı etkilere sahiptir (Morris ve ark. 2005, Rabiner 1993). Burundan çıkan sesler ve sesli harflerin oluşturduğu heceler dięer hece gruplarına göre daha ayırt edici bulunmuştur (Eatock 1994). Bu nedenle konuşmacılara ait öznitelik vektörleri çıkartılırken heceleri temsil eden her bir öznitelik vektörü gruplandırılmaktadır (Kinnunen 2002). Bu bölümde klasik mel ölçek süzgeç dizileri yerine, çerçeve temelli kümeleme sonucu aęırlıklandırılmış konuşmacı ayırt edici süzgeç dizileri kullanılmaktadır. Her bir süzgeç dizisi ile konuşmacıya ait her bir çerçevesinin karakteristik öznitelik deęerleri, ayırt edici alt bantlar ile vurgulanmaktadır.

Konuşmacı tanımada bazı frekans bölgelerinin farklı ayırt edicilik özelliklerine sahip olduęu bilinmektedir. Konuşmacıların seslerinin frekans bandı üzerinde etkin olduęu bölgeler F-oranı ölçütü ile bulunabilir. F-oranı, konuşmacılar arası öznitelik vektörlerinin ortalamasının deęişintisinin, konuşmacı içi öznitelik vektörlerinin deęişintilerinin ortalamasına oranı olarak ifade edilir (Paliwal 1992). F-oranı deęerinin

fazla olduğu yerlerde ağırlıklandırma ile süzgeçlerin etkinliği artırılır. Şekil 3.59'da öznitelik vektörlerinin kümelenecek ağırlıklandırılması blok diyagram halinde gösterilmektedir.



Şekil 3.59 Kümeleme ve ağırlıklandırma sonucu elde edilen öznitelik vektörleri

Öznitelik vektörlerinin kümelenecek ağırlıklandırılması için aşağıdaki adımlar uygulanır.

- Konuşmacıların kaç kümeye ayrılacağına belirlenmesi. 168 kişilik konuşmacı grubu için küme sayısı 2, 4, 8, 16, 32 alınmıştır.
- Küme sayısına bağlı olarak model oluşturulması. Her bir küme, k-ortalama algoritması ile başlangıç parametreleri belirlenip, BM algoritması ile bu model parametreleri kestirilir. Sonuçta her bir küme Gauss karışımları olarak modellenmektedir.
- Her bir konuşma çerçevesinin ait olduğu kümenin bulunması. Bunun için her bir çerçeve, oluşturulan küme modelleri ile karşılaştırılıp çerçeve maksimum olasılıklı kümeye ait olarak etiketlenmektedir. Örneğin bir konuşmacının eğitiminde, 8 cümleye karşılık olarak 2400 çerçeve üretilmektedir. Küme sayısı 4 için bu 2400 çerçeve sırayla hangi kümeye ait olduğu bulunur. 1. çerçeve 4. kümeye ait ise 4. kümeye ait olarak

etiketlenip daha sonra F-oranı hesabında kullanılmak üzere 4. kümeyle ait çerçevelerin bulunduğu bir grupta toplanır.

- Her bir küme için etiketlenen çerçeveler bir araya getirilmesi ve kümelerin F-oranı analizi yapılır.
- Çerçeveler, ait oldukları kümelere bağlı olarak süzgeç dizilerinden geçirilir ve süzgeç ağırlıklandırması uygulanır. Süzgeç ağırlıklandırılması çerçevelerin ait oldukları kümelere bağlı olarak farklılık göstermektedir. Bu ağırlıklandırma ile benzer çerçevelerin ayırt ediciliğinin fazla olduğu frekans bölgeleri etkinleştirilir.
- Bu işlemler sonunda elde edilen öznitelik vektörleri eğitim ve test için kullanılır.

Şekil 3.59'daki kümeleme ve ağırlıklandırma sonucu elde edilen öznitelik vektörlerine ait işlem basamakları aşağıda verilmektedir.

3.4.2.1 Spektral analiz

Konuşmacılara ait $\vec{x}_1, \dots, \vec{x}_7$ öznitelik vektörleri oluşturulur. Öznitelik vektörleri oluşturulurken şu parametreler kullanılmaktadır. Çerçeve uzunluğu 30 msn alınıp, 20 msn kaydırma uygulanır ve her bir çerçeveye hamming pencereleme uygulanır. 20 adet üçgen süzgeç dizisi ile 0. kepstrum katsayısı çıkartılıp 12 adet Mel-frekansı katsayısı elde edilir. Bu katsayılar sadece kümeleme için kullanılmaktadır.

3.4.2.2 Kümeleme

Uygulanacak küme sayısı belirlenir. Her bir kümenin ortalama değeri k-ortalama algoritması ile belirlenip, parametrelerin yakınsamasında BM algoritması kullanılmaktadır. Her bir küme, Gauss karışımları ile modellenip, ağırlık ortalama ve kovaryans model parametre değerleri (λ) oluşturulur (Antal 2004). TIMIT veritabanı için 15 saniye uzunluğunda 5 cümle (2 Sa, 3 Si cümleleri) kullanılarak eğitim kümesi oluşturulur. Veritabanlarından eğitim seti olarak 100 kişi kullanılıp TIMIT veritabanı aşağı örnekleme ile 8 kHz'e örneklenir.

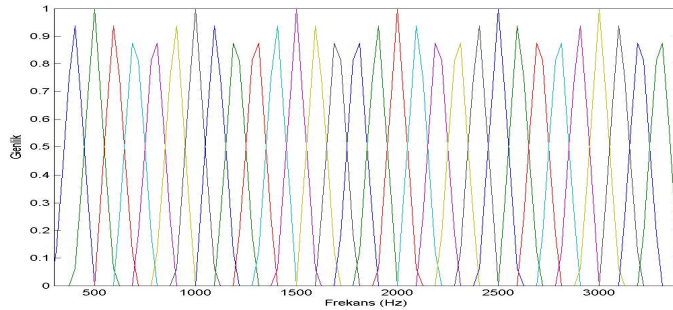
Her bir çerçevenin ait olduğu kümeyi bulmak için, Gauss karışımları kullanılarak oluşturulan kümelere ait parametreler kullanılır. \vec{x} , bir çerçeveye ait öznitelik vektörü olsun, bu öznitelik vektörünün ait olduğu küme denklem 3.43'deki gibi ifade edilir.

$$\hat{i} = \max_{\lambda} p(\vec{x}|\lambda) \quad \lambda \in GKM \quad (3.58)$$

Burada her bir kümeye ait model λ ile ifade edilip, \hat{i} öznitelik vektörünün atandığı kümeyi göstermektedir. \vec{x} öznitelik vektörü, modellenen maksimum olasılıklı kümeye atanır. Her bir öznitelik vektörü için aynı işlem tekrar edilir.

3.4.2.3 Süzgeç dizileri

Sınıflandırılan her bir çerçeve bir süzgeç dizisinden geçirilir. Süzgeçler eşit olarak dağıtılıp % 50 örtüşme ile yerleştirilir. Her bir çerçeveye Hamming pencereleme uygulanıp, 512 nokta FFT'si alınır. Süzgeç dizileri doğrusal olarak eşit aralıklarla dizilmektedir. Elde edilen işaretin genlik spektrumun, 10 tabanında logaritması alınıp dB genlik spektrumu elde edilir. Desibel genlik spektrum K adet süzgece uygulanıp her bir süzgecin enerjisi elde edilir. K adet süzgecin enerjisi E_j ($j = 1, \dots, K$) olup süzgeç dizisi vektörü $E = (E_1, E_2, \dots, E_K)^T$ olarak ifade edilir. Süzgeç dizisi Şekil 3.60'da görülmektedir.



Şekil 3.60 Eşit aralıklarla dizilmiş süzgeç dizileri

3.4.2.4 F-oranı analizi

Eğitilen konuşmacı sayısı K ile ifade edilirse, i . özelliğin F-oranı denklem 3.59'daki gibi ifade edilir (Paliwal 1992).

$$F_i = \frac{B_i}{W_i} \quad (3.59)$$

Burada B_i sınıflar arası deęişinti olup W_i ise i . özelliğın sınıf içi deęişintisi olarak tanımlanır. Sınıflar arası deęişinti ve sınıf içi deęişinti denklem 3.60 ve 3.61'de tanımlanmaktadır.

$$B_i = \frac{1}{K} \sum_{j=1}^K (\mu_{ij} - \mu_i)^2 \quad (3.60)$$

$$W_i = \frac{1}{K} \sum_{j=1}^K W_{ij} \quad (3.61)$$

Burada μ_{ij} ve W_{ij} , j . sınıfın i . özelliğın ortalama ve deęişintisidir. μ_i ise i . özelliğın ortalaması olarak denklem 3.62'deki gibi tanımlanır.

$$\mu_{ij} = \frac{1}{N} \sum_{n=1}^{N_j} x_{ijn} \quad (3.62)$$

$$W_{ij} = \frac{1}{N_j} \sum_{n=1}^{N_j} (x_{ijn} - \mu_{ij})^2 \quad (3.63)$$

$$\mu_i = \frac{1}{K} \sum_{j=1}^K \mu_{ij} \quad (3.64)$$

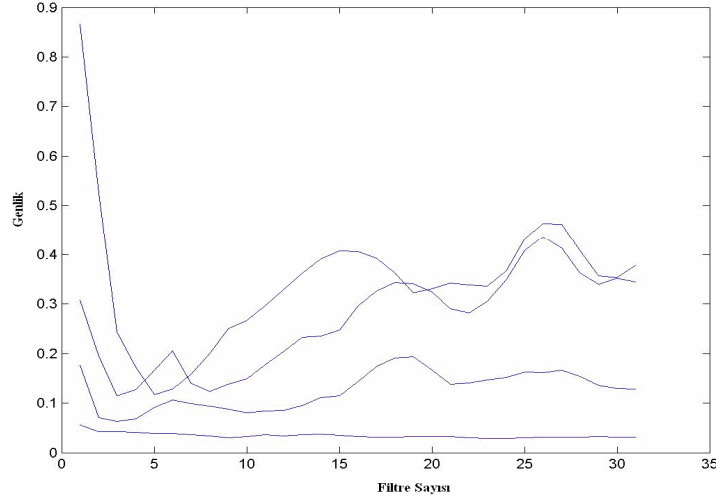
Burada x_{ijn} , j . sınıfın, n . öznitelik vektörünün i . elemanını göstermektedir. N_j , j . sınıfın toplam öznitelik vektörü sayısıdır.

Kümeleme yapıldıktan sonra her bir kümenin F-oranı denklem 3.59 kullanılarak elde edilir. i . öznitelik vektörünün j . alt bandının F-oranı deęeri, i . öznitelik vektörünün j . alt bandının, konuşmacılar arası ortalamasının deęişintisinin, aynı öznitelik vektörünün ve alt bandının konuşmacı içi deęişintisinin ortalamasının oranına eşittir. j . alt bant, bölüm 3.4.2.3'de tanımlanan E_j ye karşılık gelmektedir. Şekil 3.61'de 4 küme için F-oranı eğrileri görülmektedir. Şekilden de görüleceğı üzere her bir kümenin F-oranı deęerleri farklıdır.

Bu işlemlerden sonra her bir çerçeveye karşılık gelen alt bant için F-oranı deęeri hesaplanır. Bu deęerler öznitelik çıkartmada kullanılır. Her bir çerçeveye ait öznitelik vektörü bulunur. E_j vektörüne ait bileşenler, ilgili kümeye ait F-oranı deęerleri ile denklem 3.65'deki gibi ağırlıklandırılır.

$$\hat{E}_j = E_j \frac{F_{i,j}}{\sum_{m=1}^M F_{i,m}} \quad (3.65)$$

Ağırlıklandırılan süzgeç çıkışlarının ayrık kosinüs dönüşümü alınır. Elde edilen öznelik vektörleri C_1, \dots, C_{20} eğitim ve test için kullanılır.



Şekil 3.61 Dört küme için süzgeç çıkışlarına göre F-oranı değeri

Kümelenerek ağırlıklandırma ile öznelik vektörleri elde edilirken 3 farklı konuşmacı seti kullanılmıştır. Birinci konuşmacı seti, Bölüm 3.4.2.1’de belirtilen katsayılar oluşturularak GKM ile konuşmacıların kümelerine ait model oluşturulması amacıyla kullanılmıştır. İkinci konuşmacı seti, Bölüm 3.4.2.4’de belirtilen F-oranı değerleri hesabında kullanılmaktadır. Üçüncü konuşmacı seti, kümelenerek ağırlıklandırılmış öznelik vektörleri ile konuşmacıların eğitim ve testi için kullanılmaktadır. Bu sayede birbirinden bağımsız veriler kullanılarak, sistemin, konuşmacıları ezberlemesinin önüne geçilmektedir.

3.4.2.5 Öznelik vektörlerinin kümelenerek ağırlıklandırma deneysel sonuçları

Telefon ortamının bant sınırlama ve süzgeç etkilerinin TIMIT veritabanında benzetimi yapılabilir. Bu amaçla konuşmacı analizinde örnekleme frekansı 8 kHz’e düşürülür. Konuşma 20 msn ilerleme hızında, 30 msn (480 örnek) uzunluğunda parçalara ayrılıp Hamming pencereleme uygulanır. 100-4000 Hz arası 100 Hz Bant

genişliğinde 39 adet üçgen süzgeç kullanılmaktadır. Süzgeçlerde genlik sabit alınıp süzgeçler doğrusal ölçekte yerleştirilmektedir. Sonuçta 20 adet kepstrum katsayısı elde edilmektedir. Konuşmacıların eğitim ve testi için 100 kişi kullanılmaktadır. Eğitim için 15 saniye (5 cümle), test için 2 saniyelik konuşma parçaları alınmaktadır. F-oranı ağırlıklandırılması denklem 3.50 kullanılarak yapılmaktadır. Küme sayısına bağlı olarak elde edilen konuşmacı tanıma oranları çizelge 3.38’de görülmektedir.

Çizelge 3.38 Küme sayısına bağlı olarak tanıma oranları (%)

Küme sayısı	Tanıma oranı
2	86
4	84
8	79
16	73
32	68
MFCC	83

Örnekleme frekansı 8 kHz, 100-4000 Hz arası bant sınırlama
TIMIT veritabanı

En yüksek tanıma oranı küme sayısı 2 için elde edilip, kümeleme ile ağırlıklandırma uygulanmamış klasik MFCC katsayılarına nazaran 3 puan daha iyi tanıma oranı vermektedir.

Yukarıdaki deneyle aynı şartlarda eğitim ve test yapıp öznitelik vektörü elde edilmesinde farklı olarak örnekleme frekansı 16 kHz alınıp 0-8000 Hz aralığında 500 Hz Bant genişliğinde 31 adet süzgeç kullanılmaktadır. Küme sayısına bağlı olarak elde edilen konuşmacı tanıma oranları çizelge 3.39’da görülmektedir.

Kümeleme ve ağırlıklandırma uygulanmamış durum en yüksek tanıma oranını vermekle beraber diğer kümele sonuçlarıyla yakın sonuçlar elde edilmiştir.

Çizelge 3.39 Küme sayısına bağlı olarak tanıma oranları (%)

Küme sayısı	Tanıma oranı
2	96
4	97
8	98
16	94
32	89
MFCC	99

Örnekleme frekansı 16 kHz, bant aralığı 0-8000 Hz, TIMIT

Blok diyagramı şekil 3.59'da verilen öznitelik çıkartma yöntemi NTIMIT veritabanı için uygulanacaktır. Yaklaşık 24 sn (8 cümle) eğitilip 3 sn (~1 cümle) test edildi. 25 msn uzunluğundaki konuşma parçalarına 10 msn örtüşme uygulandı. 300-3400 Hz arasına 100 Hz aralıklarla süzgeçler yerleştirilip 20 adet kepstrum katsayısı kullanıldı. Ağırlıklandırma için denklem 3.65 kullanıldı. çizelge 3.40'da küme sayısına bağlı olarak konuşmacı tanıma oranları görülmektedir.

Çizelge 3.40 Kümeleme ile konuşmacı tanıma oranları (%)

Küme sayısı	Konuşmacı Tanıma oranı
2	63
4	65
8	61
16	61
32	57
MFCC	64

Örnekleme frekansı 16 kHz, bant aralığı 300-3400 Hz, NTIMIT

Denklem 3.65'de verilen F-oranı ağırlıklandırma ile elde edilen öznitelik vektörleri test edildiğinde birbirine yakın sonuçlar gözlenmekte, en yüksek tanıma oranı % 65 ile küme sayısı 4 için elde edilmektedir.

3.4.3 Kepstrum katsayıları ile F-oranı analizi

Öznitelik vektörü elde edilmesinde kümeleme ile ağırlıklandırma kullanımı Kinnunen (2002), tarafından yapılmıştır. Kinnunen, şekil 3.59'de verilen öznitelik vektörü elde edilmesinde, kümeleme için genelleştirilmiş fonem modeli, çerçevelere ait verilerin kümelere atanmasında minimum uzaklık, süzgeç ağırlıklandırma için denklem 3.65, konuşmacıların eğitim ve test aşamalarında vektör nicemleme algoritması kullanmıştır.

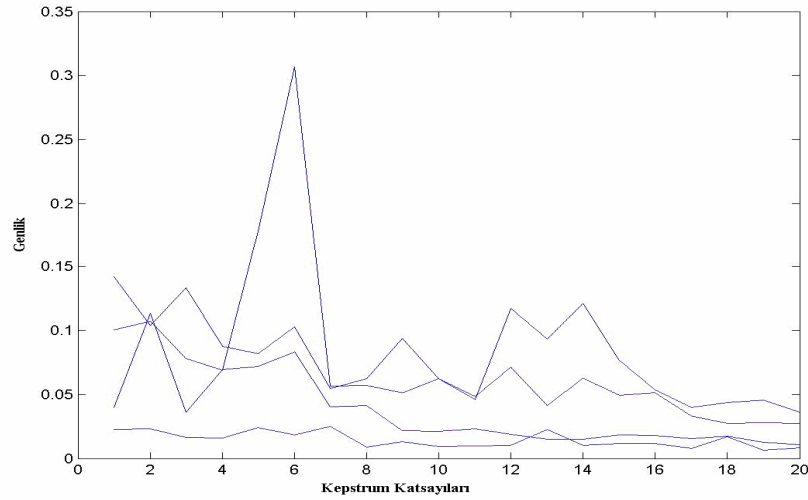
Bu tezde, şekil 3.59'da verilen öznitelik vektörü elde edilmesinde, kümeleme için k-ortalama ve beklentinin maksimumlaştırılması algoritması, çerçevelere ait verilerin kümelere atanmasında denklem 3.58'de önerilen maksimum olasılık, konuşmacıların eğitim ve test aşamalarında GKM kullanılmıştır.

Şekil 3.59'daki süzgeçlerin ağırlıklandırılması için Kinnunen'in (2002) önerdiği denklem 3.65'deki yöntem, iki veritabanı içinde çizelge 3.38, çizelge 3.39 ve çizelge 3.40'da görüleceği üzere konuşmacı tanımaya önemli oranda katkı sağlamamaktadır.

Bu tezde, süzgeç ağırlıkları hesabı için denklem 3.65'deki süzgeç enerjilerini ifade eden E_1, \dots, E_j kullanmak yerine aynı süzgeçlerin ayrık kosinüs dönüşümü uygulanarak elde edilen kepstrum katsayılarının c_1, \dots, c_{20} , F-oranı hesaplanması önerilmektedir. Her bir kepstrum katsayısı, ait olduğu kümeye karşılık gelen F-oranı değeri ile çarpılması sonucu elde edilen ifade denklem 3.66'da görülmektedir.

$$\hat{c}_j = c_j \frac{F_{i,j}}{\sum_{m=1}^M F_{i,m}} \quad (3.66)$$

Burada $F_{i,j}$, i . öznitelik vektörünün j . alt bandından oluşturulan kepstrum katsayılarının F-oranı değeri olup \hat{c}_j , j . alt banda ait ağırlıklandırılmış kepstrum katsayılarını göstermektedir. Bu denklem ile kişiyi ayırt edici bilgilerin daha fazla olduğu çerçevelere karşılık gelen, kepstrum katsayıları daha fazla ağırlıklandırılmış olur. Elde edilen kepstrum katsayıları öznitelik vektörü olarak, eğitim ve test aşamalarında kullanılır. NTIMIT veritabanında küme sayısı 4 için elde edilen F-oranı değerleri şekil 3.62'de görülmektedir.



Şekil 3.62 Kepstrum katsayılarına bağlı olarak F-oranı değerleri (küme sayısı 4)

Sonuçta F-oranı ve ağırlıklandırma işlemlerinde süzgeç dizilerinin enerjileri kullanılması yerine kepstrum katsayılarının kullanılması önerilmektedir.

3.4.3.1 Kepstrum katsayıları ile F-oranı analizinin deneysel sonuçları

TIMIT veritabanı için konuşmacı analizinde örnekleme frekansı 8 kHz'e düşürülür ve konuşma 20 msn ilerleme hızında 30 msn (480 örnek) uzunluğunda parçalara ayrılıp hamming pencereleme uygulanır. 100-4000 Hz arası 100 Hz Bant genişliğinde 39 adet üçgen süzgeç kullanılır. Süzgeçlerde genlik sabit alınıp süzgeçler doğrusal ölçekte yerleştirilmektedir.

Şekil 3.59'daki öznitelik vektörü çıkartma yöntemi ve önerilen denklem 3.51 kullanılarak kepstrum katsayıları elde edilmektedir. Kepstrum katsayısı 12 ve 20 olmak üzere iki farklı vektör dizisi alınıp eğitim ve test aşamalarından geçirilir. Eğitim için 15 saniye (5 cümle), test için 3 saniyelik (1cümle) konuşma parçaları alınır. Bu şartlar altında elde edilen tanıma sonuçları çizelge 3.41'de görülmektedir.

Çizelge 3.41 Küme sayıları değişimlerine bağlı olarak tanıma oranları (%)

Küme sayısı	Tanıma oranı (%)	
	Kepstrum Katsayı	
	12	20
2	93	72
4	85	85
8	86	83
16	90	88
32	88	86
12 adet MFCC	87	-
20 adet MFCC	-	83

Örnekleme frekansı 8 kHz, 100-4000 Hz arası bant sınırlama
Eğitim süresi 15 sn, TIMIT veritabanı

Çizelge 3.41'e göre; 12 kepstrum katsayısı kullanıldığında, küme sayısı 2 için tanıma oranı % 93 olup bu klasik MFCC'ye göre 6 puan, 20 kepstrum katsayısı kullanıldığında, küme sayısı 16 için tanıma oranı % 88 olup bu da klasik MFCC'ye göre 5 puan oranında artışa karşılık gelmektedir.

Öznitelik vektörü elde edilirken kümeleme ile ağırlıklandırmanın denklem 3.65 ile yapıldığı çizelge 3.38 ile denklem 3.66'nın kullanıldığı çizelge 3.41 karşılaştırılacaktır. 20 kepstrum katsayı sayısı için çizelge 3.41'de en yüksek konuşmacı tanıma oranı % 86 iken çizelge 3.41'de görüleceği üzere aynı durumda tanıma oranı % 88 olmaktadır. TIMIT veritabanı için önerdiğimiz denklem 3.66 ile F-oranı analizi yapılması denklem 3.65'e nazaran konuşmacı tanıma oranını 2 puan arttırmaktadır.

NTIMIT veritabanı için blok diyagramı şekil 3.59'da verilen öznelik çıkartma yöntemi kullanılmaktadır. 25 ms'lik konuşma parçalarına 10 ms'lik örtüşme uygulanır. 300-3400 Hz arasına 100 Hz aralıklarla süzgeçler yerleştirilip 20 adet kepstrem katsayısı kullanılır. Ağırlıklandırma için önerilen denklem 3.66 kullanılır. 8 cümle (yaklaşık 24 sn) eğitilip, 3 sn uzunluğunda (yaklaşık 1 cümle) test edilmektedir. Test dizininden 100 kişi ile deney yapılmaktadır. Çizelge 3.42'de küme sayısına bağlı olarak konuşmacı tanıma oranları görülmektedir.

Çizelge 3.42 Kümeleme ile konuşmacı tanıma oranları (%)

Küme sayısı	Konuşmacı Tanıma oranı (%)
2	70
4	65
8	73
16	65
32	66
20 adet MFCC	64

Örnekleme frekansı 16 kHz, bant aralığı 300-3400 Hz, NTIMIT veritabanı

Çizelge 3.42'ye göre; 20 kepstrem katsayısı kullanıldığında, küme sayısı 8 için tanıma oranı % 73 olup bu klasik MFCC'ye göre konuşmacı tanıma oranında 9 puan artışa karşılık gelmektedir.

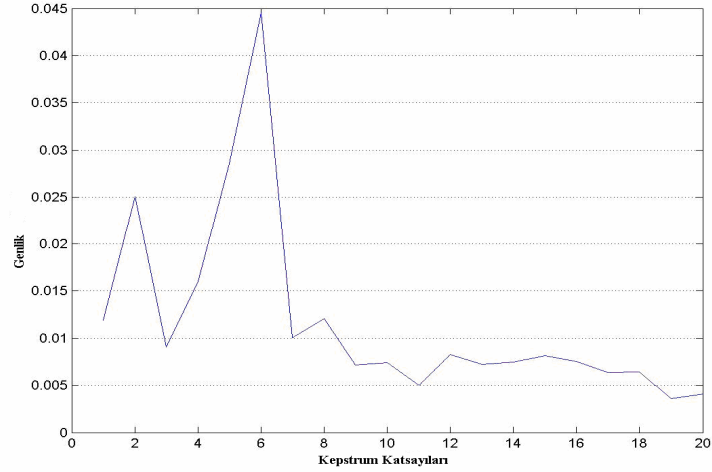
Öznelik vektörü elde edilirken kümeleme ile ağırlıklandırmanın denklem 3.50 ile yapıldığı çizelge 3.43 ile denklem 3.51'in kullanıldığı çizelge 3.45 karşılaştırılacaktır. Çizelge 3.43'de en yüksek konuşmacı tanıma oranı % 65 iken çizelge 3.45'de görüleceği üzere aynı durumda tanıma oranı % 73 olmaktadır. NTIMIT veritabanı için önerdiğimiz denklem 3.51 ile F-oranı analizi yapılması denklem 3.50'ye nazaran konuşmacı tanıma oranını 8 puan arttırmaktadır.

Sonuç olarak, her iki veritabanından elde edilen sonuçlar göz önüne alındığında, F-oranı hesabı ve kümelerin ağırlıklandırma işleminin önerilen kepstrem katsayıları ile yapılması tanıma başarımını önemli ölçüde arttırdığı görülmektedir.

3.4.4 Öznelik vektörleri oluşturulmasında F-oranına bağlı olarak süzgeç uygulanması

Bu tezde, öznelik vektörü elde edilmesinde F-oranına bağlı olarak süzgeç uygulanması önerilmektedir. Kullanılan süzgeçler ile konuşma frekans bandı, F-oranı

değerlerine bağlı olarak parçalara ayrılarak süzgeç dizileri oluşturulur. Veritabanındaki konuşmalara kümelere ayrılma işlemi uygulanmaksızın F-oranı değeri hesaplanır. Şekil 3.63'de NTIMIT veritabanında, F-oranı değerlerinin kepstrum katsayılarına bağlı olarak değişimi görülmektedir.



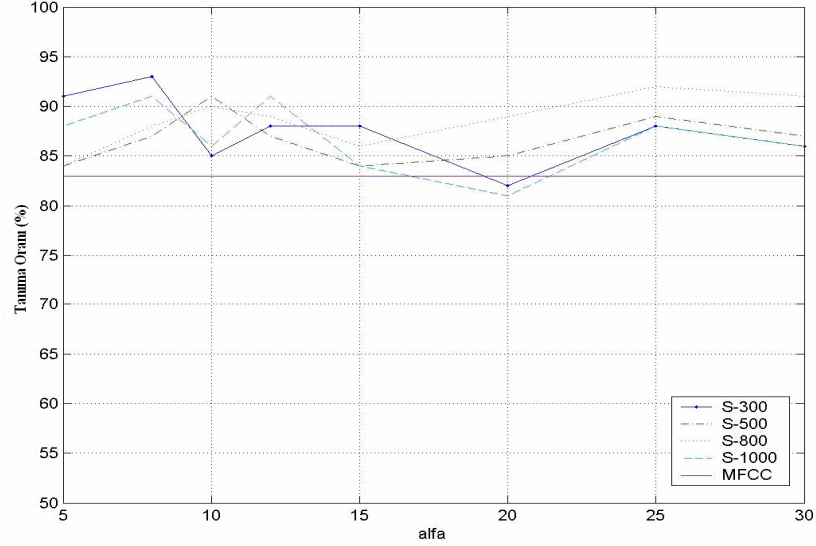
Şekil 3.63 F-oranı değerlerinin kepstrum katsayılarına bağlı olarak değişimi (NTIMIT)

Konuşmacı frekans bandı, S ile ifade edilen belirli parçalara ayrılıp o parçaya karşılık gelen F-oranı ($F_{k,i}$) değeri, α ağırlık etki faktörü ile çarpılması sonucu elde edilir. Frekans bölgesine uygulanacak süzgeç sayısı Fs_k , denklem 3.67'deki gibi önerilmektedir.

$$Fs_k = \alpha \cdot \sum_{i=S_1}^S F_{k,i} \quad (3.67)$$

Burada, $F_{k,i}$ F-oranı değeri, 0-1 arasında olduğundan, α katsayısı ile çarpılıp frekans bölgesine uygulanacak süzgeç sayısı bulunur. Her frekans bölgesi için tayin edilen süzgeç sayıları frekans bölgesine eşit aralıklarla sabit genlikte üçgen süzgeç dizisi olarak yerleştirilir. Bu şekilde konuşmacının sesinin ayırt ediciliği yüksek olan frekans bölgesine daha fazla süzgeç yerleştirilmiş olur. Sonuçta ayırt ediciliği F-oranı değerine göre hesaplanmış süzgeç dizileri ve bu süzgeç dizileri kullanılarak kepstrum katsayıları elde edilir. Kepstrum katsayıları kullanılarak konuşmacı tanıma başarımı ölçülür.

TIMIT veritabanında yapılan deneylerde GKM karışım sayısı 32 alınıp eğitim için 5 cümle (yaklaşık 15 saniye) kullanılıp test için ise 3 saniye uzunluğunda (yaklaşık 1 cümle) konuşma parçaları kullanılmaktadır. Her iki veritabanı eğitim ve test setleri 100 kişiden oluşturulmaktadır. Örnekleme hızı 8000 Hz e düşürülüp 0-4000 Hz süzgeç bant genişliğinde 20 boyutlu öznitelik vektörleri ile konuşmacılar modellenir ve test edilir. Elde edilen sonuçlar şekil 3.64’de görülmektedir.

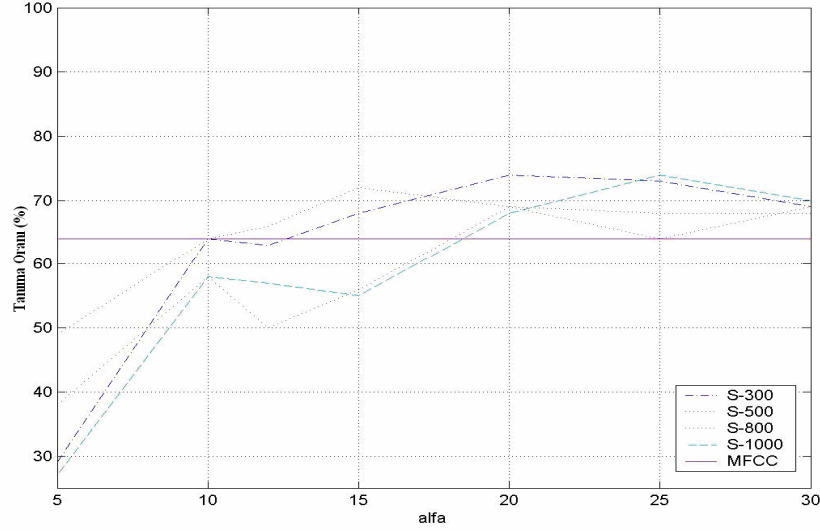


Şekil 3.64 Alfa ve parça genişliği değerlerine bağlı olarak konuşmacı tanıma oranları (TIMIT veritabanı, f_s 8 kHz, süzgeç bant genişliği 0-4000 Hz)

Denklem 3.67’de ifade edilen parça genişliği S ’e bağlı olarak uygun alfa (α) değerinde tanıma oranı % 93 olmaktadır ($S=300, \alpha=8$). Şekil 3.64’de de görüleceği üzere aynı şartlarda klasik MFCC kullanıldığında konuşmacı tanıma oranı % 83 elde edilmektedir. Önerilen F-oranına bağlı olarak süzgeç yerleştirilmesi yaklaşımı, NTIMIT veritabanına uygulanması sonucu şekil 3.65’de görülen konuşmacı tanıma oranları elde edilmiştir.

Aynı şartlarda klasik MFCC kullanıldığında elde edilen konuşmacı tanıma oranı % 64 olup α değeri 20’den itibaren tüm parça uzunlukları için bu değeri geçmektedir. En yüksek tanıma oranı parça uzunluğu 300 Hz, $\alpha=20$ değeri için % 74 olarak elde edilmiştir.

Frekans bandını parçalara ayırıp denklem 3.67 kullanılarak süzgeç yerleştirme her iki veritabanında başarımlarını artırarak sağlamaktadır. Bu şekilde konuşmacıyı ayırt edici öznelikler daha iyi modellenmektedir.



Şekil 3.65 Alfa ve parça genişliği değerlerine bağlı olarak konuşmacı tanıma oranları (NTIMIT veritabanı, f_s 16 kHz)

3.5 Bürünsel Özneliklerin (Prosodic Features) Konuşmacı Tanımaya Etkisi

Bürünsel öznelikler, kişinin ses yolunun fiziksel yapısı hakkında bilgi veren özneliklerdir. Bürünsel öznelikler; 3 ana grup altında toplanmaktadır (Reynolds ve ark. 2003). Bunlar;

- 1.) Kelime, hece ve cümle parçası süreleri
- 2.) Cümle içinde durma süreleri ve sıklıkları
- 3.) Temel (perde) frekans, formant frekansı ve enerji öznelikleri

Bu üç grup özneliğinden, konuşmacı tanıma üzerinde en fazla başarımların artışını 3. grup öznelikler sağlamaktadır (Peskin ve ark. 2003). Bu nedenle bu öznelikler üzerinde yoğunlaşılacaktır.

Gürültüsüz veritabanı olarak bilinen TIMIT veritabanı ile ideal şartlarda bürünsel parametrelerin öznelik parametreleri olarak kullanılmasının konuşmacı tanımaya etkisi incelenecektir. Ayrıca konuşmacı tanımadaki, telefon ahizesi ve hattından dolayı meydana gelen başarımların düşümlerinin bürünsel öznelikler ile

azaltılması amaçlanmaktadır. Bu amaçla konuşmaların telefon ortamında iletildiği NTIMIT veritabanı kullanılacaktır. Bürünsel özniteliklerden en yüksek tanıma oranı elde edilen yukarıdaki 3. grup içinde tanımlanan temel frekans, formant frekansları ve enerji parametreleri incelenecektir.

Ayrıca konuşma formantlarındaki genlik ve frekans modülasyonu (GM-FM) temelli özniteliklerin konuşmacı tanıma etkisi incelenecektir. Genlik ve frekans formantları, ses yolundaki rezonansların frekans ve genliklerinin modülasyonu olarak açıklanmaktadır (Jankowski ve ark. 1995). Son olarak GM-FM parametrelerinin özilintisinin genlik zarfının, polinom bantetimi yapıp polinom katsayıları öznitelik vektörü olarak alınmıştır. Elde edilen öznitelik vektörleri konuşmacı tanıma eğitim ve test aşamalarında uygulanacaktır.

3.5.1 Temel frekans (f_0)

Boğazda bulunan ses telleri, periyodik darbeler oluşturur ve bu darbelerin frekanslarına temel frekans adı verilir. Temel frekansın daha iyi anlaşılabilmesi için insan ses üretim mekanizmasının bilinmesi gerekmektedir. Ses üretiminde akciğerler hava üreten bir enerji kaynağı gibi davranır. Bir kişi konuşması ile birlikte hava akciğerden ses yolundaki boğaza doğru hareket eder. Konuşma sesi üretmek için ses telleri ve ses yolu belirli bir yapı alır. İnsan ses sisteminin en önemli parçası ses tellerini içeren boğazdır. Ses tellerinin aktivitesi üretilen sesin ünlü veya ünsüz olacağını belirler. Ünlü sesler için ses telleri hızlıca açılıp kapanarak havayı modüle eder.

Konuşma işaretinin temel frekansı, doğrudan boğaz kaynak işareti $s(t)$ ile ilişkilidir. Boğaz kaynak spektrumu frekans alanında uyarı dizisi $x(f)$ ve gırtlak kaynak karakteristiği $G(f)$ nin çarpımı olarak denklem 3.68'deki gibi gösterilebilir.

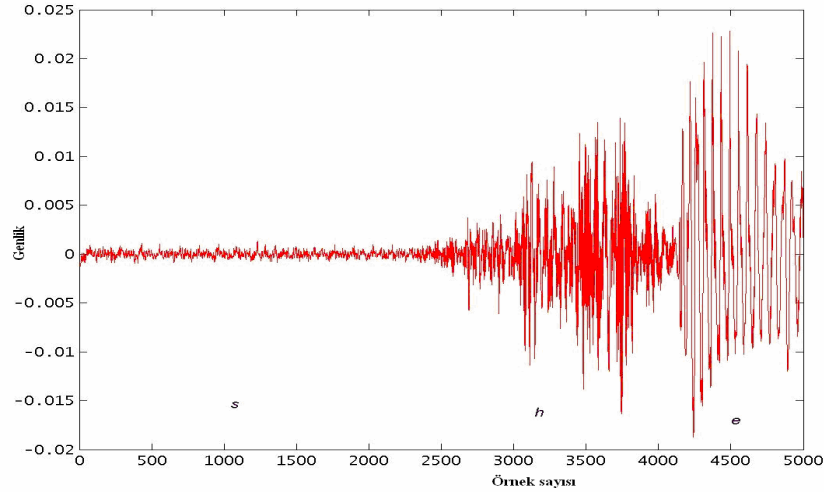
$$S(f) = G(f) \cdot X(f) = G(f) \cdot \sum_{k=-\infty}^{\infty} \delta(f - kf_0) \quad (3.68)$$

Burada f_0 , frekans alanından $X(f)$ ile gösterilen uyarılar arası aralık olup ses telleri titreşim hızıyla ilişkilidir. Şekil 3.66'da ses tellerinin darbe üretici olarak çalışması sonucu üretilen işaret görülmektedir.



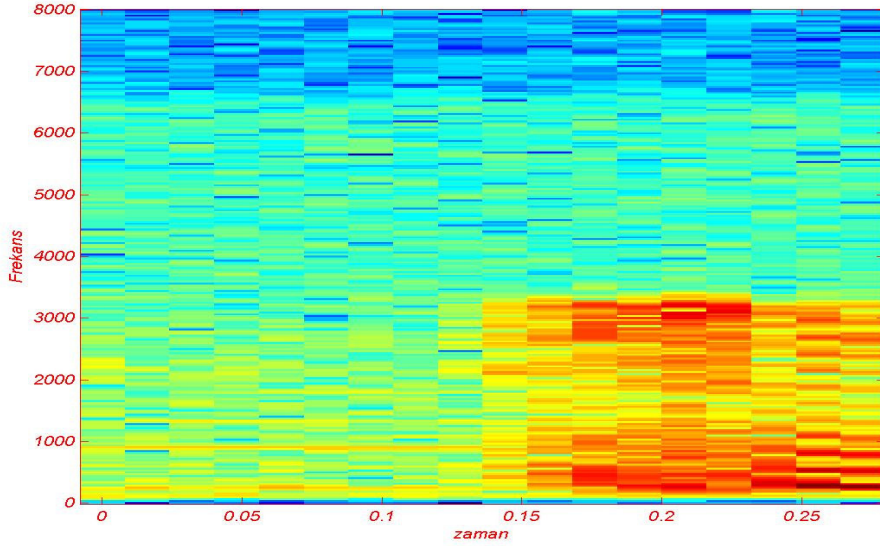
Şekil 3.66 Ses tellerinin darbe üretici gibi davranması

Ses tellerinin titreşim hızı tellerin gerginliğine ve kütlesine bağlıdır. Ünsüz sesler için ses telleri periyodik olmayan akış olacak şekilde pozisyon alır ve konuşma içerisinde periyodik olmayan bileşenler oluşur. Şekil 3.67’de bir işaretin (bir bayan konuşmacı tarafından söylenen “she” sözcüğü) zaman alanında gösterimi görülmektedir. Şekilden görüleceği üzere konuşmanın ünlü sese karşılık gelen bölgeleri periyodik iken ünsüz ses bölgeleri, gürültü benzeri bir yapıya sahiptir. İşaretin periyodik bölgeleri konuşmanın f_0 ’ının hesaplanmasına yardımcı olur.



Şekil 3.67 NTIMIT veritabanından alınmış “She” sözcüğü

Şekil 3.68’de ise aynı sözcüğün zaman-frekans-yoğunluk değişimi görülmektedir. Şekilde görülen yoğun bölgeler konuşmanın ünlü kısımlarını gösterirken az yoğun bölgeler ise ünsüz sesleri göstermektedir.



Şekil 3.68 “She” sözcüğünün zaman-frekans-yoğunluk değişimi

Ünlü sesler için f_0 , erkek konuşmacılarda 50-250 Hz arasında değişirken, bayan konuşmacılarda 120-400 Hz ve çocuk konuşmacılarda 150-450 Hz arasında değişir. Geniş değişim aralığı ve diğer faktörler f_0 'ın % 100 doğrulukta ayırt edilmesini zorlaştırmaktadır.

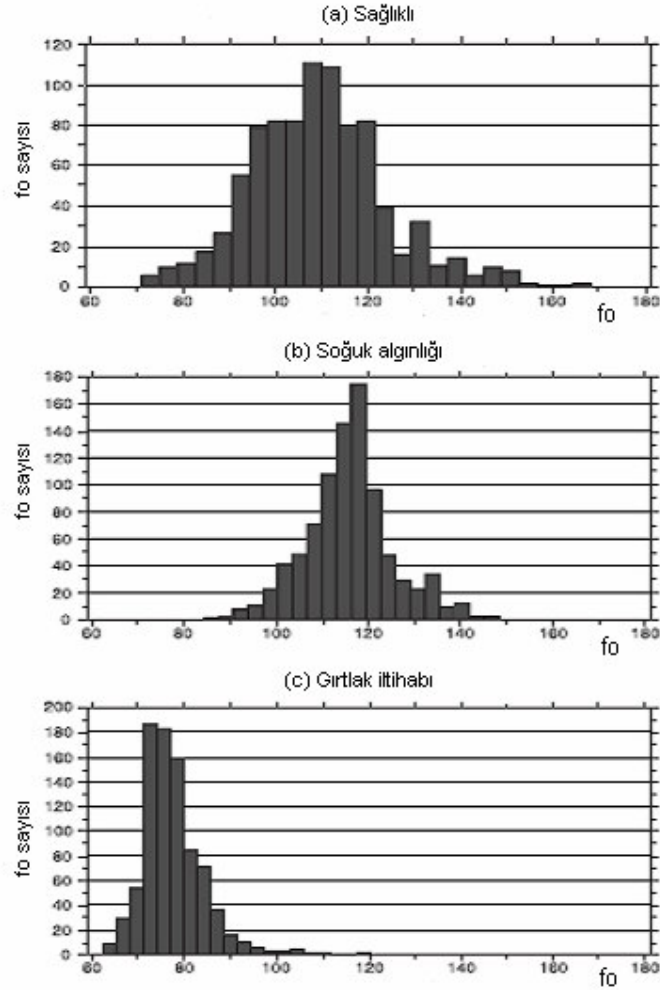
f_0 , ses üretim mekanizmasında temel frekans olarak adlandırılırken, algılanan temel frekans perde frekansı olarak adlandırılmaktadır. Atal (1974), perde frekansının konuşmacıya bağlı olduğunu söylemekle beraber tek bir konuşmacının söylediği sözcükler için fazla değişmediğini tespit etmiştir.

3.5.1.1 Perde (pitch) frekansı izlemenin zorlukları

Perde frekansı birincil olarak temel frekans ile ilişkili olup doğrudan ses tellerinin titreşimiyle belirlenir. Perde frekansı izlemede pek çok faktör etkili olduğu için zor bir problem olarak ifade edilir. En büyük problem, konuşma işaretinin gerçekte periyodik ve sabit olmayışıdır. Kısa zaman aralıklarında (50 msn), konuşma işaretinin f_0 'ı ve spektral karakteristiğinin genliği değişmektedir. Çok hızlı değişen konuşmada, konuşmanın kısa analiz aralıklarında sabit kabul edildiğinden f_0 'ın hesabı daha zor olmaktadır. Analiz aralığı en az 2-3 perde periyodu kullanılarak ortalama perde değeri belirlenmeye çalışılmaktadır. Perde frekansı belirleme probleminde bazı önemli

uygulamalar (telefon konuşması v.b.) için konuşmanın sesli ve sessiz kısımlarının ayrılması gerekir. Normal konuşmada bile bazı durumlarda ilk harmoniğin temel frekanstan büyük olması, pek çok perde frekansı izleyici algoritmada problemlere neden olmaktadır (Kasi 2002).

Perde frekansı kişinin içinde bulunduğu (stres, kızgınlık, üzüntü, sevinç v.b.) ruh haline bağlı olarak değişebilir. Perde frekansı soğuk algınlığı, gırtlak iltihaplanması gibi durumlardan etkilenir. Şekil 3.69'da değişik sağlık koşullarında bir kişinin f_0 değişimi görülmektedir (Rose 2002).



Şekil 3.69 Değişik sağlık koşullarında temel frekans değişimi

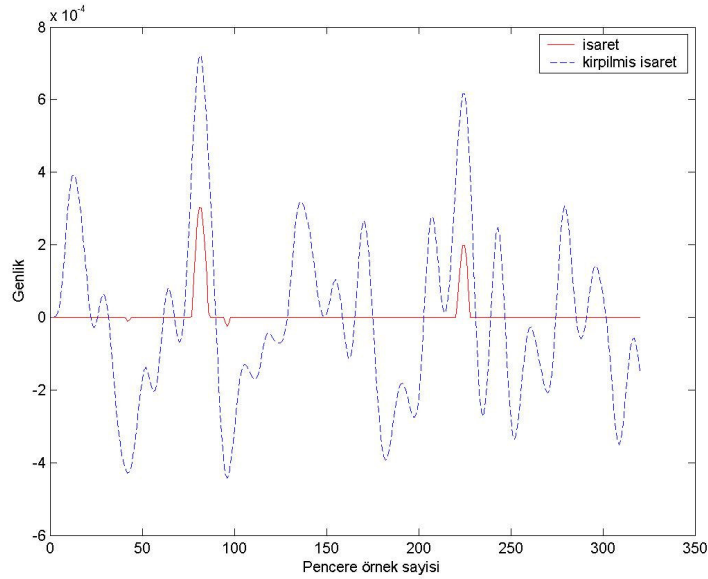
Şekil 3.69'dan görüleceği üzere sağlıklı bir kişinin temel frekansı 70-170 Hz arasında iken, aynı kişi soğuk algınlığına yakalanması durumunda temel frekans dağılımı 85-150

Hz, gırtlak iltihaplanması oluştuğunda ise 65-110 Hz arasında dağılmaktadır. Sağlık koşullarına bağlı olarak kişinin temel frekansı önemli oranda değişim göstermektedir.

3.5.1.2 Perde frekansı izleme aşamaları

Perde frekansı izleme algoritması üç temel adımdan oluşur. Bunlar ön işleme, f_0 kestirimi ve son işlemdir (Kasi 2002).

A. Ön işleme: Ön işlemede ilk adım, konuşma işareti öncelikle pencerelere ayrılıp, işaret 900 Hz kesim frekansında alçak geçiren süzgeçten geçirilerek yüksek frekanslı bileşenler atılır. İkinci olarak işarete merkez kırpması uygulanır. Merkez kırpması ile yüksek genlikli darbeler korunurken düşük genlikli darbeler işareten çıkartılır. Bu işlem harmonik yapıyı azaltırken periyodikliği korur. Şekil 3.70’de bir pencere ile temsil edilen işareten harmonik bileşenlerin çıkartıldığında elde edilen şekil görülmektedir.



Şekil 3.70 Merkez kırpması ile işaretin kırılması

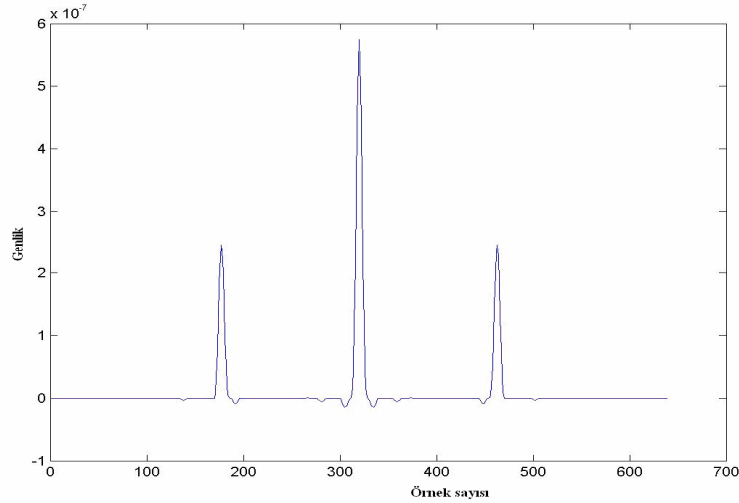
Şekil 3.70’den görüleceği üzere merkez kırpması ile temel frekans daha fazla belirginleşmektedir.

B. f_0 kestirimi: Perde frekansı izleme algoritmalarından özilinti ve kepstrum metotları en çok kullanılan iki algoritmadır. Özilinti yaklaşımı ile periyodik bir işaretin perde frekansı denklem 3.69’daki gibi kestirilir.

$$oto_kor(k) = \sum_{n=0}^{N-K} s(n) \cdot s(n+k) \quad (3.69)$$

burada $0 \leq k \leq K-1$ ve işaret $s(n)$ ile gösterilip $s(n+k)$ işaretin k kadar zaman gecikmeli hali olarak ifade edilmektedir. Eğer konuşma işareti periyodik ise özilinti fonksiyonu da periyodik olacaktır. Periyodik işaretler için özilinti fonksiyonu $0, \pm P, \pm 2P, \dots$ vb. işaret periyodu aralıklarla bir maksimum değere ulaşacaktır. Bu durum şekil 3.71'de görülmektedir.

Özilinti fonksiyonu, temel periyodik bileşenlerden farklı olarak başka tepeler içerebilir. Konuşma işaretleri için ses yolu cevabı sönümlü salınımlara sahip olmasından dolayı özilinti fonksiyonunda birçok tepe görünür. Konu ile ilgisi olmayan bu tepelerden dolayı tepe ayırt etmesi zorlaşır. Pencereleme işleminde büyük zaman değerleri alınması bu olumsuzluğu ortadan kaldırır.



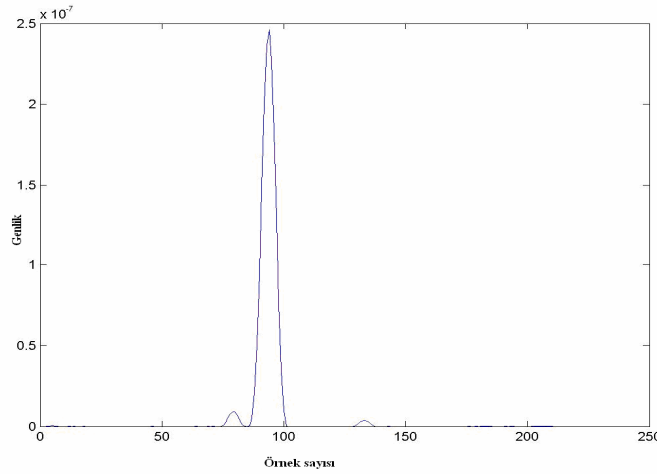
Şekil 3.71 Özilinti fonksiyonu ile elde edilen işaret

İkinci perde frekansı izleme tekniği kepstrum yöntemidir. Kepstrum yöntemi denklem 3.70'de görüldüğü gibi kısa zaman aralığında genlik spektrumun ters Fourier dönüşümü olarak tanımlanır.

$$c(k) = F^{-1}(\log|F(x(t))|^2) \quad (3.70)$$

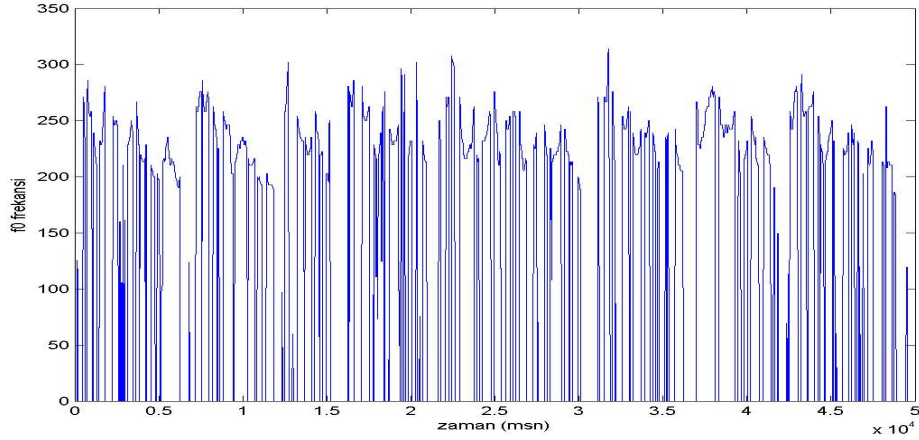
Burada $x(t)$, pencereleyen konuşma işareti olup $c(k)$, kepstrum katsayılarını ifade etmektedir. Denklem 3.70'de logaritma alınması, harmonik tepelerin daha düzleşmesi sağlanır (Rabiner ve Juang 1993, Kinnunen 2003).

C. Son işleme: Perde frekansı hesaplanırken büyük ve küçük hatalar oluşabilir. İlk harmonikten dolayı f_0 değeri, olması gerekenden iki kat daha fazla hesaplanabilir. Bir periyot yanlışlıkla iki periyot olarak anlaşıldığı zaman, perde frekansı değeri yarıya düşer. Diğer büyük hata ünlü sesin ünsüz, ünsüz sesin ünlü olarak karar verilmesidir. Perde frekansı değerlerinin medyan pürüzsüzleştirme işlemi ile ardışıl olarak çerçeve temelli ölçümler yapılır. Medyan pürüzsüzleştirme ile bazı istenmeyen düzeltmeler yapılırsa da doğrusal alçak geçiren süzgeçten daha etkilidir (Kasi 2002) Konuşmanın f_0 alt ve üst sınırlarını (60 ile 320 Hz arası) gösteren bölgelere karşılık gelen maksimum özilinti değeri alınması ile elde edilen grafik şekil 3.72'deki gibidir.



Şekil 3.72 f_0 alt ve üst sınırları içerisindeki tepe değerinin bulunması

Bu bölgedeki tepe değerin bulunduğu örneğin frekansı hesaplanır. Şekil 3.72'deki örnek için $f_0=114$ Hz olarak bulunur. Tüm bu işlemler her bir çerçeveye uygulanır ve her bir çerçeve için bir f_0 değeri hesaplanır. NTIMIT veritabanına ait bir konuşmacının toplam 8 cümlesi için elde edilen f_0 değerleri şekil 3.73'de görülmektedir. Şekilden görüleceği üzere konuşmada sesli-sessiz ayırımı yapılmaktadır. Konuşma olmayan sessiz kısımlara karşılık gelen çerçevelerin f_0 değeri, 0 olarak atanmıştır.



Şekil 3.73 Bir konuşmacının f_0 değerleri

3.5.1.3 Temel frekansın deneysel değerlendirilmesi

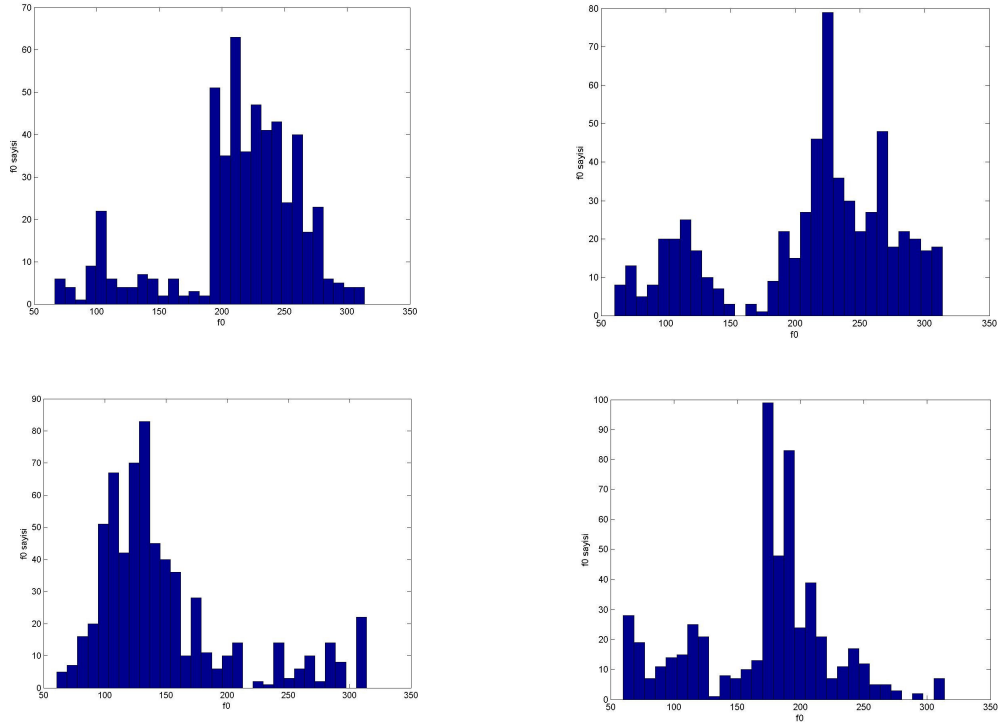
Temel frekans konuşmacıları birbirinden ayırmada etkilidir (Reynolds ve ark. 2003). Temel frekans, iletişim ortamı özellikleri ve gürültüden, spektral katsayılara nazaran daha az etkilenir (Arcienega ve Drygajlo 2001). Bu nedenle konuşmacı tanıma için öznel olarak kullanılabilir. Şekil 3.74'de NTIMIT veritabanında tüm konuşmacılar tarafından ortak olarak söylenen Sa1 cümlesinin 4 farklı konuşmacı için, temel frekansın histogram dağılımı görülmektedir. Şekil 3.74'den görüleceği üzere her bir konuşmacının söylediği cümleler aynı olmasına rağmen temel frekans dağılımında büyük farklılıklar oluşmaktadır.

f_0 'ın öznel olarak kullanılmasının konuşmacı tanıma etkisi incelenecektir.

Bu amaçla f_0 'ın içinde bulunduğu öznel vektörleri eğitim ve test için kullanılacaktır. Tüm eğitim ve sınıflandırma adımları değişmeden sadece kullanılan öznel vektörleri değiştirilerek, vektör setleri arasında kontrollü bir karşılaştırma yapılacaktır.

Konuşmacıların öznel vektörleri üretilirken süzgeçler 300-3380 Hz arasına doğrusal ölçekte 70 Hz aralıkla yerleştirilir. Her bir çerçeveye karşılık 20 adet kepstrum katsayısı elde edilir. Bu katsayılar 168 kişinin eğitim ve testi için kullanılır. Bir konuşmacıya ait 10 cümlelerin 8'i (2 Sa, 5 Si, 3 Sx cümleleri) eğitim aşamasında kalan 2 cümlelerin (2 Sx cümleleri) her biri ayrı ayrı test işlemine tabii tutulmaktadır. Temel frekans hesabında özilinti yöntemi kullanılmaktadır. Medyan pürüzsüzleştirme ile son

üç f_0 'ın ortalaması alınmaktadır. Belirtilen bu eğitim ve test şartlarında aşağıdaki deneyler yapılmaktadır.



Şekil 3.74 Dört farklı konuşmacının aynı cümleyi söylemesi ile elde edilen perde frekanslarının histogramları

1. İlk olarak konuşmadan sessiz kısımların atılmasının konuşmacı tanımaya etkisi araştırılacaktır. Konuşmada sesli sessiz ayırımında eşik değeri kullanılmaktadır. Konuşmadaki eşik değerinin altındaki sessiz çerçevelere karşılık gelen kısımlar atılır. Konuşmanın sessiz kısımlarına karşılık gelen çerçeveler, düşük seviye enerjiye sahip olan zemin gürültüsü içeren kısımlardır. Her bir konuşma çerçevesindeki konuşmanın enerjisi denklem 3.71 ile ifade edilmektedir.

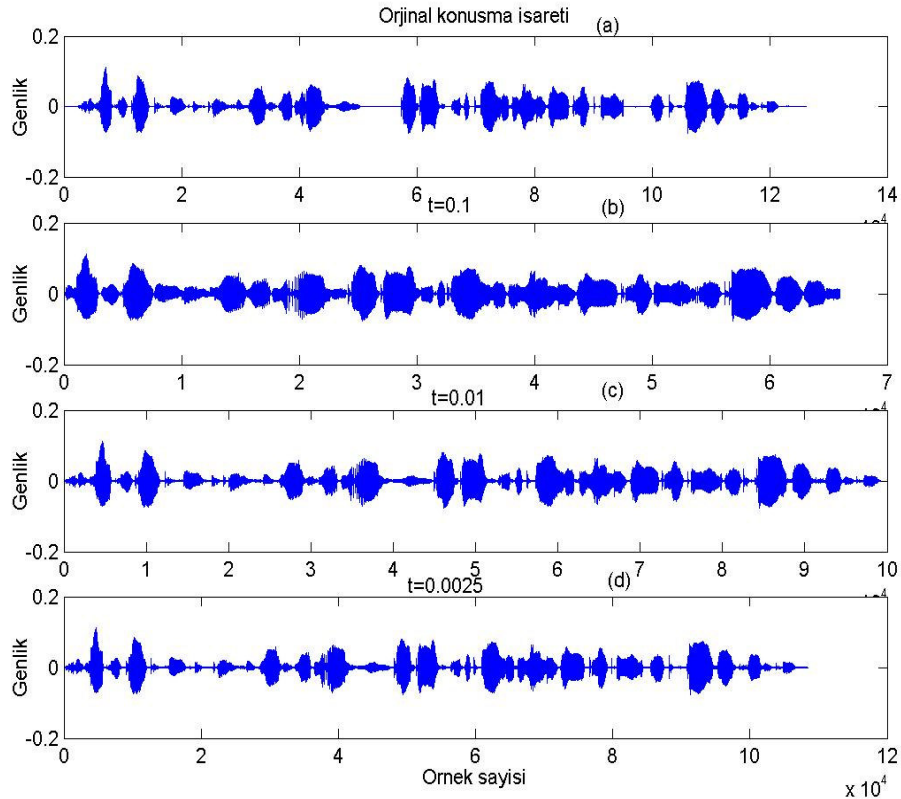
$$E_i = \sum_{k=1}^M x_i(k)^2 \quad i = 1, \dots, T \quad (3.71)$$

Burada M bir konuşma çerçevesindeki örnek sayısı ve T konuşma çerçevelerinin toplam sayısına karşılık gelmektedir. Sessiz çerçevelerin belirlenmesinde eşik değerinin doğru şekilde belirlenmesi gerekmektedir. Eşik değeri denklem 3.72'ye göre belirlenebilir (Aliaa ve ark. 2004).

$$E_{sik} = E_{min} + t \cdot (E_{max} - E_{min}) \quad (3.72)$$

Burada E_{min} ve E_{max} , T adet çerçevenin en düşük ve en büyük enerji değerleridir. t ise eşik parametresi olarak ifade edilen bir katsayıdır

NTIMIT veritabanında 8 cümleden oluşan bir konuşma için, çerçeve sayısı 2480, aynı işaretin denklem 3.72'deki eşik parametresi t 'ye bağlı olarak sessiz kısımların atılması durumunda; t değeri 0.1 için çerçeve sayısı 771, t değeri 0.01 için çerçeve sayısı 1780, t değeri 0.0025 için çerçeve sayısı 2067 olmaktadır. Konuşma 25 msn'lik çerçevelere ayrılıp, 10 msn'de bir konuşmalar yenilenmektedir. Şekil 3.75 (a)'da NTIMIT veritabanında bir konuşma işareti, şekil 3.75 (b)'de $t=0.1$ için işareten sessiz kısımların atılmış hali, şekil 3.75 (c)'de $t=0.01$ için işareten sessiz kısımların atılmış hali, şekil 3.75 (d)'de $t=0.0025$ için işareten sessiz kısımların atılmış hali görülmektedir.



Şekil 3.75 (a) NTIMIT veritabanında bir konuşma işareti (b) $t=0.1$ için işareten sessiz kısımların atılmış hali (c) $t=0.01$ için işareten sessiz kısımların atılmış hali (d) $t=0.0025$ için işareten sessiz kısımların atılmış hali

Konuşmadan sessiz kısımların atılmasının tanımaya etkisi incelenecektir. 168 kişiden oluşan NTIMIT veritabanı ile deney yapılmaktadır. Çizelge 3.43’de eşik parametresi t ye bağlı olarak konuşmacı tanıma oranları görülmektedir.

Çizelge 3.43 Eşik parametresi t ’ye bağlı olarak konuşmacı tanıma oranları (%)

t	Tanıma oranı
0.05	66.07
0.01	73.21
0.005	72.02
0.0025	73.51
0.001	69.05
MFCC	69.05

Çerçeve ve ilerleme süresi 25-10 msn, 20 adet MFCC vektörü,
karışım sayısı 32, NTIMIT veritabanı

Çizelge 3.43’den görüleceği üzere eşik parametresi t , 0.0025 alınması durumunda konuşmacı tanıma oranı % 73.51 olmakta ve bu klasik MFCC’ye nazaran 4.46 puan tanıma oranında artışa karşılık gelmektedir. Konuşmadan sessiz kısımların atılması ile eğitim ve test için kullanılan öznitelik vektörü sayısında $t=0.0025$ için yaklaşık 15 puan düşme olmasına rağmen tanıma oranı artmaktadır. Konuşmadan sessiz kısımların atılması ile konuşmadaki gürültü bileşenleri azaltılmaktadır. Bu sayede daha iyi konuşmacı tanınması sağlanmaktadır. Bundan sonraki deneylerde konuşmadan sessiz kısımların atılıp, t değeri 0.0025 alınacaktır.

2. Öznitelik olarak yalnız f_0 kullanıldığında % 4.16 gibi çok düşük bir konuşmacı tanıma oranı elde edilmektedir. f_0 , Mel frekansı kepstrum katsayıları ile birlikte kullanılmasının konuşmacı tanımaya etkisi incelenecektir. Denklem 3.72’deki eşik denklemine bağlı olarak konuşmadan sessiz kısımlar atılmaktadır. Öznitelik vektörlerine \log_2 tabanında f_0 eklenmesi ile elde edilen tanıma oranları çizelge 3.44’de görülmektedir.

Çizelge 3.44’den görüleceği üzere Mel frekansı kepstrum katsayılarına $\log_2(f_0)$ eklenmesi tanıma oranını 8.34 puan oranında arttırmaktadır.

Çizelge 3.44 Mel frekansı kepstrum katsayılarına f_0 eklenmesi ile elde edilen tanıma oranları (%)

Öznitelik vektörleri	Tanıma oranı
MFCC	73.51
MFCC+ $\log_2(f_0)$	81.85

Çerçeve ve ilerleme süresi 25-10 msn, 20 adet MFCC vektörü, karışım sayısı 32, NTIMIT veritabanı

3. Konuşma analizinde kullanılan çerçeve ve çerçeve ilerleme sürelerinin değişimine karşı MFCC ve f_0 'a bağlı olarak tanıma başarımı incelenecektir. 168 konuşmacının her biri 8 cümle ile eğitilip kalan 2 cümlenin her biri ile ayrı ayrı test edilmektedir. Konuşmadan sessiz kısımlar atılmasına bağlı olarak elde edilen sonuçlar çizelge 3.45'de görülmektedir.

Çizelge 3.45 Çerçeveleme sürelerine bağlı olarak temel frekansın tanımaya etkisi (%)

Öznitelik vektörü	Sessiz kısımlar	Çerçeve ve ilerleme süresi		
		30-20 msn	25-10 msn	20-10 msn
MFCC	yok	63.10	69.05	68.75
MFCC+ $\log_2(f_0)$	yok	67.86	77.68	75.89
MFCC	var	66.67	73.51	72.92
MFCC+ $\log_2(f_0)$	var	69.05	81.85	79.46

20 adet MFCC vektörü, karışım sayısı 32, konuşmacı sayısı 168, NTIMIT veritabanı

Konuşmadan sessiz kısımların atılmadığı durumda, çizelge 3.45'den görüleceği üzere en yüksek tanıma oranı % 77.68 ile 25-10 msn çerçeve ve ilerleme süresi için elde edilmektedir. Bu sonuç MFCC ye f_0 eklenmemiş duruma göre 8.63 puan daha iyidir. Konuşmadan sessiz kısımların atıldığı durumda ise, en yüksek tanıma 25-10 msn çerçeve ilerleme süreleri için % 81.85 ile MFCC ve f_0 'ın birlikte kullanıldığı durumda sağlanmaktadır.

4. Öznitelik vektörleri elde edilirken ön vurgulama uygulanmasının konuşmacı tanımaya etkisi incelenecektir. NTIMIT veritabanında konuşmalar 25 msn'lik çerçevelere ayrılıp 10 msn de bir çerçeveler yenilenir. Konuşmacıların öznitelik vektörleri üretilirken doğrusal ölçekte 300-3380 Hz arasına 70 Hz aralıklarla yerleştirilmektedir. 168 konuşmacının her biri 8 cümle ile eğitilip kalan 2 cümlenin her

biri ile ayrı ayrı test edilmektedir. Konuşmadan sessiz kısımlar atılmaktadır. Öznitelik vektörlerine ön vurgulama uygulanmasına bağlı olarak elde edilen sonuçlar çizelge 3.46’da görülmektedir.

Çizelge 3.46 Ön vurgulamaya bağlı olarak tanıma oranları (%)

Öznitelik vektörü	Ön vurgulama var	Ön vurgulama yok
MFCC	70.54	73.51
MFCC+ $\log_2(f_0)$	75.60	81.85

20 adet MFCC vektörü, çerçeve ve ilerleme süresi 25-10 msn, karışım bileşen sayısı 32, NTIMIT veritabanı

Çizelge 3.46’dan görüleceği üzere öznitelik vektörlerine ön vurgulama uygulanması, MFCC ve MFCC ye f_0 eklenmesi durumlarında konuşmacı tanıma oranını düşürmektedir.

5. MFCC’lerin, f_0 ile birlikte kullanıldığı durumda GKM karışım bileşen sayısının tanıma oranına etkisi incelenecektir. Öznitelik vektörü üretiminde konuşma örnekleri 10 msn’lik ilerleme aralıklarıyla 25 msn’lik parçalara ayrılır. Her bir çerçeve için 20 adet mel frekansı keppstrum katsayısı ve temel frekans kullanılarak öznitelik vektörü oluşturulmaktadır. Konuşmadan sessiz kısımlar atılmaktadır. Gauss karışım modelinde, MFCC ve f_0 ’ın birlikte kullanıldığı durum için, karışım sayısının tanımaya etkisi çizelge 3.47’de görülmektedir.

Çizelge 3.47 Karışım sayısının konuşmacı tanımaya etkisi

Karışım sayısı	Konuşmacı tanıma oranı (%)
8	68.45
16	77.68
32	81.85
64	75.30

20 adet MFCC vektörü ve f_0 , Ön vurgulama yok, NTIMIT veritabanı test süresi ayrı ayrı birer 1 cümle

Çizelge 3.47’den görüleceği üzere karışım sayısı 32 için en yüksek tanıma oranı elde edilmektedir. Konuşmacılara ait öznitelik vektörleri, 32 adet karışım ile daha iyi modellenmektedir.

6. Öznitelik vektörleri oluşturulurken kullanılan üçgen süzgeç dizisinin yerleştirildiği frekans ölçeği değişiminin konuşmacı tanıma etkisi incelenecektir.. Konuşmacıların öznitelik vektörleri üretilirken doğrusal, ERB, bark ve 3 değişik mel ölçekte süzgeçler 300-3380 Hz arasına yerleştirilir. Kulak tarafından algılanan frekansları ifade eden mel değerleri Steven ve Volkman (1940) tarafından tespit edilmiştir (Umesh ve ark. 1999). Bu mel değerleri O’Shaughnessy (1987), Fant (1960), Slaney (1998) tarafından tanımlanan mel ölçekleri ile yaygın olarak ifade edilmektedir. Bu ölçeklerin değişiminin konuşmacı tanıma etkisi incelenecektir.

Öznitelik vektörleri oluşturulurken NTIMIT veritabanında konuşmalar 25 msn’lik çerçevelere ayrılıp 10 msn de bir çerçeveler yenilenir. Her bir çerçeveye karşılık 20 adet kepstrum katsayısı elde edilir. Bu katsayılar 168 kişinin eğitim ve testi için kullanılır. Eğitim için 8 cümle, test için 2 cümlenin her biri ile ayrı ayrı kullanılır. Konuşmadan sessiz kısımlar atılmaktadır. Bu durumda elde edilen tanıma oranları çizelge 3.48’de görülmektedir.

Çizelge 3.48 Süzgeç dizileri frekans ölçeğine bağlı olarak tanıma oranları

Ölçek çeşidi	Konuşmacı tanıma oranı (%)	
	Kepstrum katsayı	Kepstrum katsayı+ $\log(f_0)$
Doğrusal	73.51	81.85
ERB	69.94	73.81
Bark	60.42	66.37
Mel¹ (O’Shaughnessy)	68.45	71.43
Mel² (Fant)	72.32	77.38
Mel³ (Slaney)	70.54	75.60

Ön vurgulama yok, sessiz kısımların atılması var, kepstrum katsayı sayısı 20
 $Mel^1 = a \log(1+f/b)$ $a=2595$, $b=700$ $Mel^2 = a \log(1+f/b)$ $a=1000/\log 2$, $b=1000$,
 $Mel^3 = 1000$ Hz altı 66.7 Hz bant genişliğinde doğrusal, 1000 Hz üstü logaritmik

Çizelge 3.48’den görüleceği üzere kepstrum katsayıları yalnız başına kullanıldığında, doğrusal ölçek için konuşmacı tanıma oranı % 73.51, MFCC vektörleri f_0 ile birlikte kullanıldığı durumda, doğrusal ölçek için en yüksek tanıma oranı % 81.85 olarak elde edilmektedir.

Sonuç olarak NTIMIT veritabanı için MFCC vektörlerine f_0 eklenmesi ile en yüksek tanıma oranı % 81.85 olup f_0 eklenmediği duruma göre 8.34 puan daha iyidir.

7. TIMIT veritabanında temel frekansın tanıma üzerine etkisi incelenecektir. TIMIT veritabanında sınırlandırmamış şartlarda (konuşmacı frekans bandı 0-8 kHz, örnekleme frekansı 16 kHz) MFCC özniteliklerine f_0 eklenmesi tanıma oranında (% 99.4) bir değişiklik oluşturmamaktadır. Bu nedenle bant sınırlaması ve örnekleme frekansının düşürülmesi gibi sınırlamalı şartlarda TIMIT veritabanı için, MFCC özniteliklerine f_0 eklenmesinin konuşmacı tanımaya etkisi incelenecektir.

Veritabanının frekans bandı 0-8000 Hz'den 300-4080 Hz bant aralığına sınırlandırılmaktadır. Bu aralığa, Mel ölçeğe üçgen süzgeçler yerleştirilir ve her bir çerçeve için 24 katsayı elde edilir. Bu katsayılar temel frekans eklenip öznitelik vektörü, eğitim ve test işleminde kullanılır. Eğitim için 5 cümle (15 saniye) kullanılmaktadır. Değişik çerçeve ve ilerleme süreleri için elde edilen tanıma oranları çizelge 3.49'da verilmektedir.

Çizelge 3.49 Bant sınırlamalı durumda tanıma oranları

Kepstrum katsayıları	Konuşmacı tanıma oranı (%)	
	20 -10 msn	40-20 msn
MFCC	98.81	98.81
MFCC +log(f_0)	97.62	98.81

Test süresi 3 sn, 24 adet MFCC vektörü, karışım sayısı 32, örnekleme frekansı 16 kHz eğitim süresi 15 sn, TIMIT veritabanı

TIMIT veritabanında bant sınırlamalı durumda çerçeve ve ilerleme süresi 20-10 msn için, MFCC ye f_0 eklenmesi konuşmacı tanıma oranını azaltmaktadır.

8. Yukarıdaki deneyle aynı şartlarda örnekleme hızı 16 kHz'den 8 kHz'e düşürülüp 300-4080 Hz bant genişliğinde tanıma oranı ölçülecektir. Çerçeveleme 40 msn, ilerleme hızı 20 msn alındığı durumda elde edilen sonuçlar çizelge 3.50'de görülmektedir.

Çizelge 3.50 Örnekleme hızının düşürülmesinin tanımaya etkisi

Kepstrum katsayıları	Tanıma oranı
MFCC	82.14
MFCC +log(f_0)	86.90

Test süresi 3 sn, 24 adet MFCC vektörü, karışım sayısı 32, örnekleme frekansı 8 kHz, Eğitim süresi 15 sn, TIMIT veritabanı

Çizelge 3.50 incelendiğinde örnekleme hızı düşürülmüş durumda MFCC katsayılarına f_0 eklemek tanıma oranını 4.76 puan arttırmaktadır.

9. TIMIT veritabanı için bir önceki deneyle aynı eğitim ve test şartlarında, örnekleme hızı 16 kHz alınıp, konuşmadan sessiz kısımların atılması durumunda tanıma başarımı ölçülecektir. Bu durumda elde edilen tanıma oranı çizelge 3.51’de görülmektedir.

Çizelge 3.51 Konuşmadan sessiz kısımların atılması

Kepstrum katsayıları	Tanıma oranı (%)
MFCC	92.86
MFCC +log(f_0)	88.69

24 adet MFCC vektörü, karışım sayısı 32, örnekleme frekansı 16 kHz, eğitim süresi 15 sn, TIMIT veritabanı

Konuşmanın sessiz kısımları atılıp, öznitelik vektörü olarak yalnız MFCC kullanılması durumunda tanıma oranı % 92.86 ve MFCC ye f_0 eklenmesi ile tanıma oranı % 88.69 olup konuşmacı tanıma oranı azalmaktadır.

10. Temel frekansın elde edilmesinde sıklıkla kullanılan özilinti ve kepstrum yöntemleri karşılaştırılacaktır. Bu amaçla NTIMIT veritabanı için öznitelik vektörleri üretilirken süzgeçler, 300-3380 Hz arasına, 70 Hz aralıkla yerleştirilir. Süzgeçler, doğrusal ölçekte % 50 örtüşme oranı ile yerleştirilmektedir. Her bir çerçeveye karşılık 20 adet kepstrum katsayısı elde edilir. Bu katsayılar 168 kişinin eğitim ve testi için kullanılır. Eğitim için 8 cümle, test için 2 cümlenin her biri ile ayrı ayrı kullanılır. Gauss karışım modelinde karışım sayısı 32 alınıp BM ile eğitilmektedir. TIMIT veritabanı için frekans bandı 0-8000 Hz’den 300-4080 Hz bant aralığına sınırlandırılmaktadır. Bu aralığa mel ölçekte üçgen süzgeçler yerleştirilir ve her bir çerçeve için 24 katsayı elde edilir. Bu katsayılar temel frekans eklenir ve elde edilen öznitelik vektörleri eğitim ve test işleminde kullanılır. Elde edilen sonuçlar çizelge 3.52’de görülmektedir.

Çizelge 3.52’deki tanıma oranları incelendiğinde, NTIMIT veritabanı için MFCC katsayılarına özilinti yöntemi ile perde frekansı eklenmesi sonucu tanıma oranı % 81.85 olmaktadır. Her iki veritabanı içinde özilinti yöntemi, kepstrum yöntemine nazaran daha iyi konuşmacı tanıma sağlamaktadır. TIMIT veritabanı için klasik MFCC’

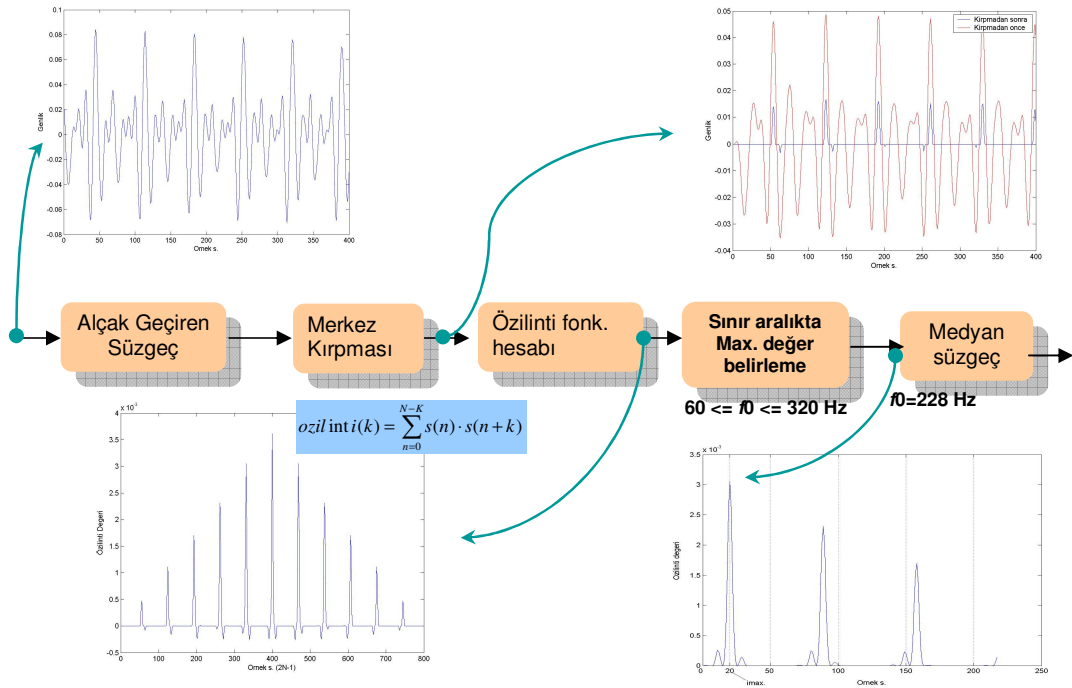
ye nazaran tanıma oranında 1.19 puan azalma olmasına karşın NTIMIT veritabanı için ise tanıma oranı 8.34 puan artmaktadır.

Çizelge 3.52 Özilinti ve kepstrum yöntemlerinin konuşmacı tanıma etkisi (%)

Öz nitelik Vektörü Hazırlanma Şartları				Konuşmacı Tanıma Oranı (%)		
Kepstrum katsayısı	Sessiz kısımlar atılması	Çerçeveleme ilerleme süresi (msn)	Veritabanı	MFCC+ f_0		Yalnız MFCC Kull.
				Özilinti	Kepstrum	
Mfcc20+log(f_0)	yok	25-10	NTIMIT	77.68	73.81	69.05
Mfcc20+log(f_0)	var	25-10	NTIMIT	81.85	80.06	73.51
Mfcc24+log(f_0)	yok	20-10	TIMIT	97.62	89.88	98.81
Mfcc24+log(f_0)	var	20-10	TIMIT	88.69	88.10	92.86

TIMIT veritabanı için parametreler: Eğitim 5 cümle, test 1 cümle, 24 adet MFCC
 NTIMIT veritabanı için parametreler: Eğitim 8 cümle, test 1 cümle, 20 adet MFCC

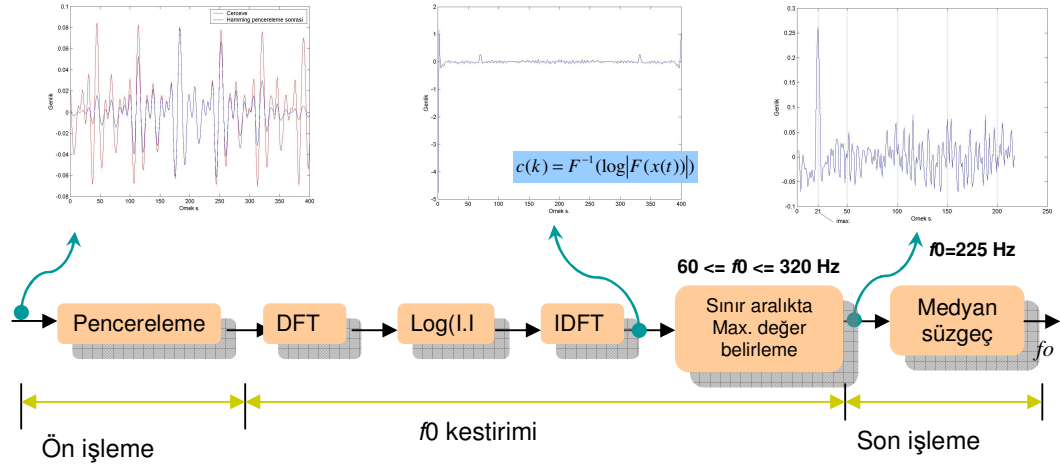
Özilinti yönteminin kepstrum yönteminden daha iyi başarımlar elde edilmesindeki temel neden kullanılan yöntem farklılığıdır. Şekil 3.76'da özilinti yöntemi, şekil 3.77'de ise kepstrum yöntemi görülmektedir.



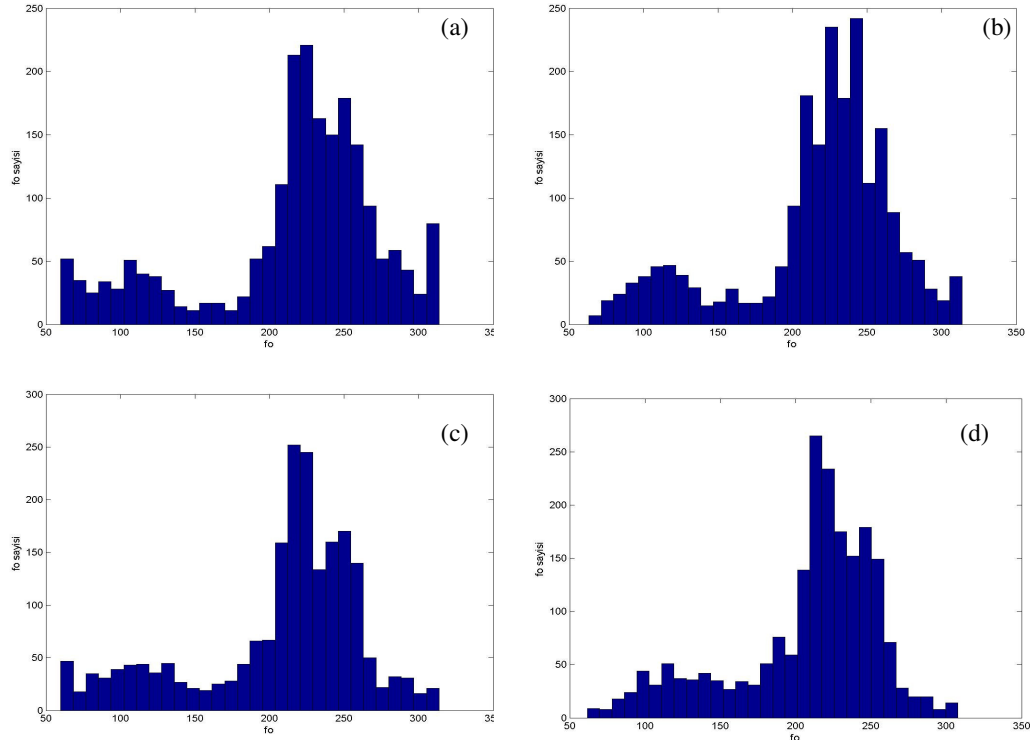
Şekil 3.76 Özilinti Yöntemi

Bölüm 3.5.1.2'de bu yöntemler perde frekansı izleme algoritması olarak üç temel adımda tanıtılmaktadır. Aynı çerçeve için özilinti yönteminde f_0 değeri 228 Hz

kestirilirken, kepstrum yönteminde ise 225 Hz kestirilmektedir. f_0 değerlerindeki bu farklılıklar şekil 3.78 (b)'de özilinti yöntemi için, şekil 3.78 (d)'de kepstrum yöntemi için histogram dağılımı olarak görülmektedir.



Şekil 3.77 Kepstrum Yöntemi



Şekil 3.78 Özilinti yöntemi ile elde edilen perde frekansın medyan süzgeç (a) öncesi (b) sonrası dağılımı (c) Kepstrum yöntemi ile elde edilen perde frekansın medyan süzgeç öncesi (d) sonrası dağılımı

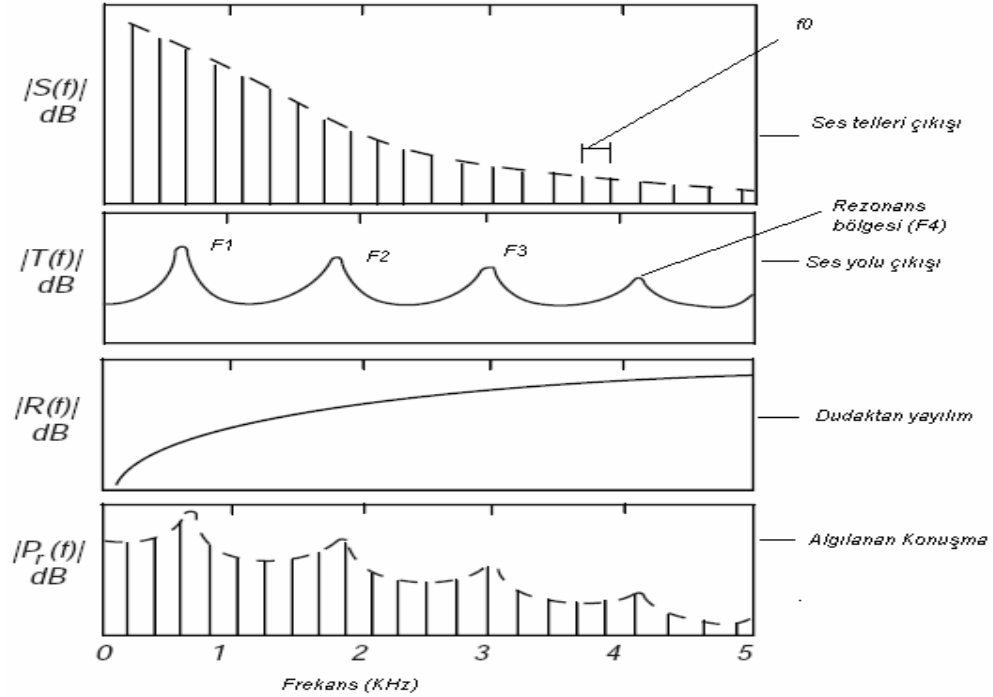
Konuşmacıların temel frekansındaki kaba hataların bir kısmını ortadan kaldırmak için, medyan süzgeçten geçirme uygulanıp, çerçeve tabanlı ardışık ölçümlerden elde edilen perde hatları yumuşatılır. Medyan süzgecin etkisi özilinti yöntemi için şekil 3.78 (a) ve (b)'de, kepstrum yöntemi için şekil 3.78 (c) ve (d)'de görülmektedir.

3.5.2 Formant frekansları

Konuşma işareti $P_r(t)$, gırtlaktan çıkan dalga $S(t)$, $T(f)$ ve $R(f)$ kaskat süzgeçlerinden geçirilerek denklem 3.73'deki gibi üretilir (Park 2002).

$$P_r(f) = S(f) \cdot T(f) \cdot R(f) \quad (3.73)$$

Burada $P_r(f)$ algılanan konuşma işaretinin spektrumunu ifade edip $S(f)$ boğaz kaynak spektrumu, $T(f)$ ses yolu transfer fonksiyonu, $R(f)$ yayılım karakteristiği olarak ifade edilir. Şekil 3.79'da bu ifadelerin spektral gösterimi görülmektedir. Formant frekansları şekilde de görüleceği üzere ses yolu transfer fonksiyonu $T(f)$ ile doğrudan ilişkilidir.



Şekil 3.79 Denklem 3.73'deki transfer fonksiyonlarının gösterimi

Akustik teoride ses yolu, deęişik genişlik ve boydaki tüplerin birleşimi olarak modellenir. Bu tüplerin alanı ve şekli konuşmacının ses üretme mekanizmasının fiziksel yapısına ve üretilen ses için ses yolunun alacağı pozisyona bağlıdır. Ünlü ses üretimi esnasında, ses yolunun akustik tüp modeli, ses yolu transfer fonksiyonunun köklerinde rezonans karakteristiğine sahiptir. Rezonans frekansları veya kök yerleri formant frekansları olarak adlandırılır. Her ne kadar formant frekansları sınırsız sayıda tanımlansa da konuşmacı tanıma çalışmalarında çoğunlukla 0-4 kHz aralığındaki ilk üç formant (F_1, F_2, F_3) değerleri kullanılmaktadır.

3.5.2.1 Formant frekansının etkisinin deneysel değerlendirilmesi

NTIMIT veritabanında formant frekanslarının konuşmacı tanımaya etkisi incelenecektir. Öznitelik vektörleri üretiminde konuşmalar 25 msn'lik çerçevelere ayrılıp 10 msn de bir çerçeveler yenilenir. Konuşmadan sessiz kısımlar atılmamaktadır. Konuşmacıların öznitelik vektörleri üretilirken süzgeç dizileri 300-3380 Hz arasına 70 Hz aralıkla % 50 örtüşmeli olarak yerleştirilir. Her bir pencereye karşılık 20 adet mel frekansı kepsrum katsayısı elde edilmektedir. Bu katsayılar 168 kişinin eğitim ve testi için kullanılır. Eğitim için 8 cümle (~24 sn), test için 1 cümle (~3 sn) kullanılmaktadır.

Konuşma formantlarının yerlerinin bulunmasında 19. dereceden doğrusal öngörü katsayıları (DÖK) kullanılır (Jankowski ve ark 1995). DÖK köklerinden genlik ve frekanslar kullanılarak formantlar seçilir. Formant frekansları olarak F_1, F_2, F_3 alınır. Formant frekanslarının tanımaya etkisi çizelge 3.53'de görülmektedir.

Çizelge 3.53 Formant frekansları için tanıma oranları (%)

Öznitelik vektörleri	Konuşmacı tanıma oranı(%)
$\log(F_1, F_2, F_3)$	17.26
$\log(F_1, F_2)$	5.95
MFCC	69.05
MFCC+ $\log(F_1, F_2, F_3)$	66.67

20 adet MFCC vektörü, karışım bileşen sayısı 32, NTIMIT veritabanı

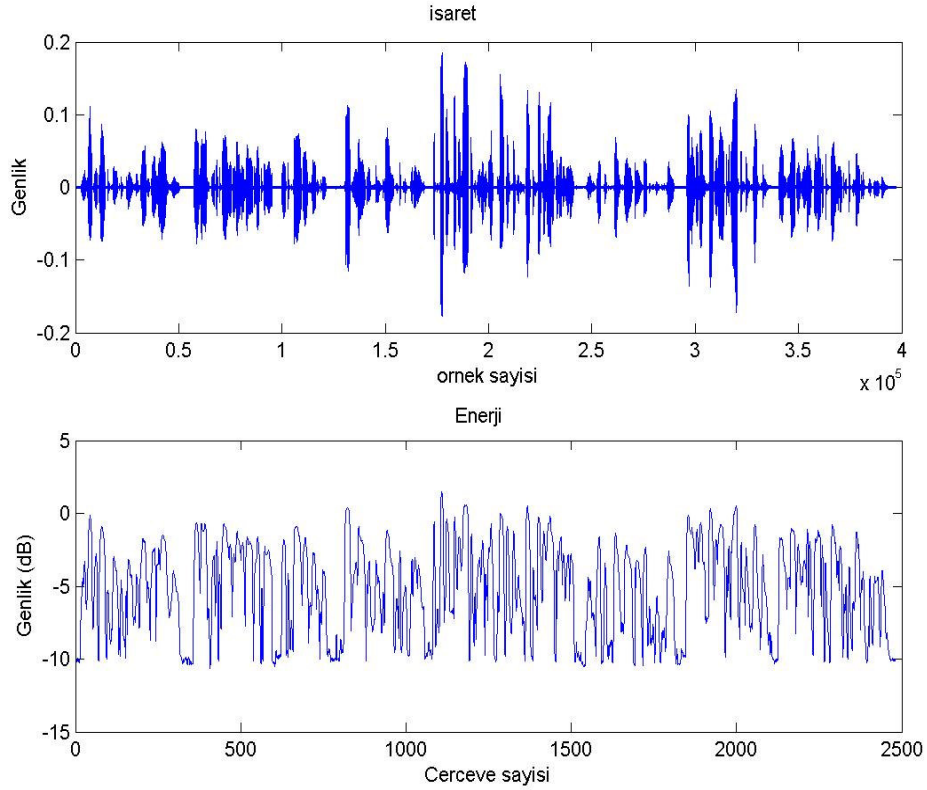
Çizelge 3.53'den görüleceği üzere öznitelik vektörü olarak yalnız formant frekansları kullanılması durumunda tanıma oranı oldukça düşük çıkmaktadır. MFCC ile birlikte formant frekansları kullanılması durumunda tanıma oranı 2.38 puan düşmektedir.

3.5.3 Enerji

Bir işaretin enerjisi, zaman alanında işaretin genliklerinin karelerinin toplamının ortalaması olarak ifade edilmektedir. Çerçvelenen konuşma $x(i)$ 'nin logaritmik enerjisi denklem 3.74'deki gibi elde edilir.

$$\log E = \ln \sum_{i=1}^N x(i)^2 \quad (3.74)$$

Burada N çerçeve uzunluğu olup, $x(i)$ giriş işaretine karşılık gelmektedir. Şekil 3.80'de bir konuşma örneği ve desibel olarak enerji değeri görülmektedir.

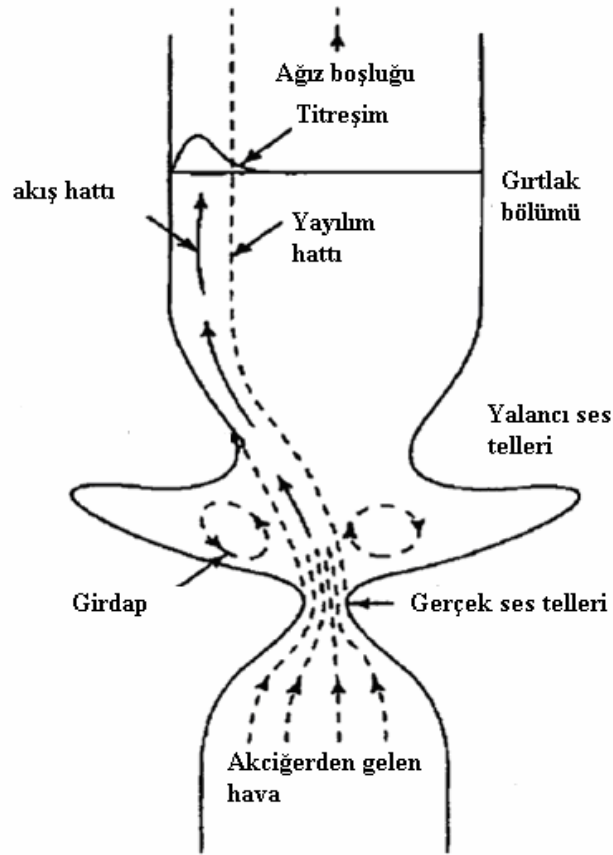


Şekil 3.80 Konuşma örneği ve enerjisi alınmış hali

Diğer bir enerji ölçme metodu da teager enerji operatörü kullanmaktır. Teager enerji operatörü ile ses yolundaki hava akışının doğrusal olmadığı kabul edildiği durumlarda anlık enerji değişimleri bulunur.

3.5.3.1 Teager enerji operatörü

Mel frekansı kepstrum katsayılarının da içinde bulunduğu yöntemlerin tamamı, doğrusal konuşma üretim modelini kullanmaktadır. Doğrusal konuşma üretim modelinde, havanın ses yolunda yayılımının bir düzlem boyunca olduğu varsayılır. Teager'in (1989), çalışmalarına göre bu akış ile birlikte oluşan girdaplar dolayısıyla hava, ses yolu boyunca dağılır. Şekil 3.81'de ses yolunda hava akışı ve girdap oluşumu görülmektedir.



Şekil 3.81 Ses yolunda girdap-hava akışı etkileşimi (Zhou ve ark. 2001)

Teager, ses üretimi esnasında meydana gelen girdap-ses akışı etkileşmesinden dolayı akışın doğrusal olmadığını önermiştir. Bu teori, akış mekanizması ve ses basıncı izlenerek desteklenmiştir (Hansen 1998). Kişinin içinde bulunduğu durum (stres, heyecan v.b.) fiziksel değişikliklere yol açtığı ve bu durumun ses yolunda girdap akışı etkileşmesine neden olduğu varsayılmaktadır (Plumpe ve ark. 1999). Şekil 3.81'de görülen doğrusal olmayan girdap-hava akışı etkileşiminin anlık enerji değişimi, Teager

tarafından, Teager enerji operatörü olarak denklem 3.75'deki gibi ifade edilmektedir (Zhou ve ark. 2001).

$$\Psi_c[x(t)] = \left(\frac{d}{dt} x(t) \right)^2 - x(t) \left(\frac{d^2}{dt^2} x(t) \right) \quad (3.75)$$

Burada $\Psi_c[\cdot]$ Teager enerji operatörü (TEO) olup ve $x(t)$ konuşma işaretinin zaman alanında bir bileşeni olarak ifade edilmektedir. Bu ifadenin ayrık zamandaki ifadesi denklem 3.76'daki gibi tanımlanmaktadır (Hamila ve ark. 1999).

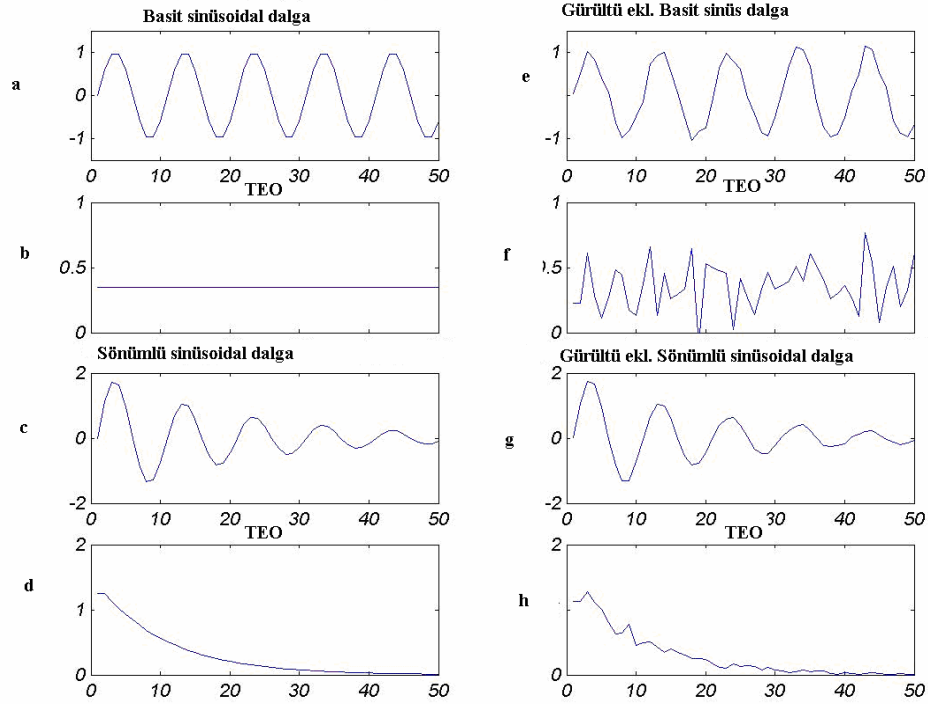
$$\Psi[x(n)] = x^2(n) - x(n+1) \cdot x(n-1) \quad (3.76)$$

Burada $x(n)$ örneklenmiş konuşma işaretini temsil etmektedir. Bir işarete Teager enerji operatörü uygulandığında, işarettaki süreksizlikler, sıçramalar gibi ani değişiklikler kuvvetlenirken, örnekler arasındaki yumuşak geçişler zayıflar (Duman ve ark. 2005).

Örneğin şekil 3.82 (a) da $\sin(n\frac{\pi}{5})$ sünisoidal dalgası görülmekte olup işarete TEO uygulandığında şekil 3.82 (b) de görülen sabit bir değer elde edilir. Sönümlü bir sinüs dalgası $2 \cdot e^{-0.005n} \sin(n\frac{\pi}{5})$ şekil 3.82 (c)'de görülmektedir. Bu işaretin genliği zamana bağlı olarak azalmaktadır. TEO uygulayarak zamana bağlı genlik değişimini Şekil 3.82 (d) de izlenebilir.

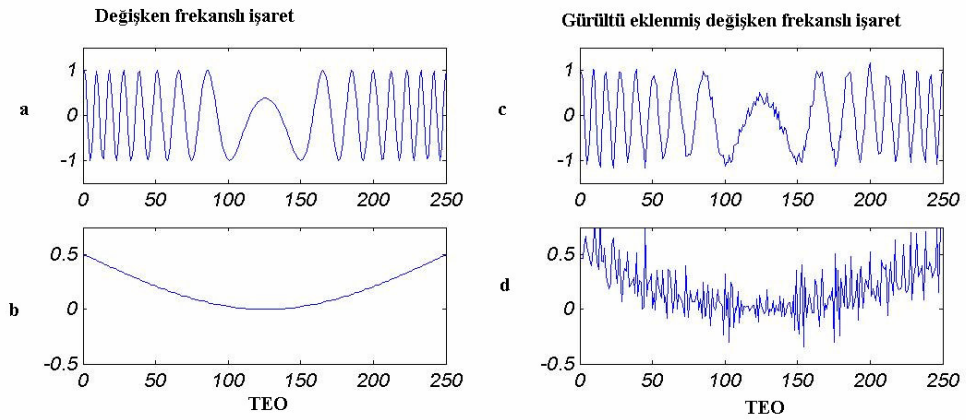
Sinüsoidal işarete SNR = 40 dB gürültü eklendiğinde şekil 3.82 (e) de görülen dalga şekli elde edilmektedir. Bu işarete TEO uygulanması durumunda elde edilen dalga şekli, şekil 3.82 (f)'de görülmektedir. Şekil 3.82 (g) ve (h) da gürültü eklenmiş sönümlü dalga ve TEO ile genlik değişimi izlenmiş işaret görülmektedir.

Şekil 3.83'de ise frekans değişiminin TEO ile izlenmesi görülmektedir. İzlenecek işaretin frekansı doğrusal olarak $\frac{\pi}{4}$ 'den 0'a doğru inmekte daha sonra tekrar $\frac{\pi}{4}$ 'e kadar artmaktadır. (MATLAB toolbox'ında bulunan chirp fonksiyonu ile işaret üretilmektedir)



Şekil 3.82 Bir sinüs işaretinin TEO ile genliğinin izlenmesi

Şekil 3.83 (a) değişken frekanslı bir sinüsoidal bir işaret (b)'de ise işarete TEO uygulandığında elde edilen dalga şekli görülmektedir. Frekans azaldıkça TEO uygulanması sonucu elde edilen şeklin genliği 0.5'den 0'a doğru azalmakta olup işaretin frekansı artması ile işaretin değeri de 0.5'e doğru artmaktadır. Şekil 3.83 (c) de 40 dB gürültü eklenmiş değişken frekanslı işaret ve bu işarete TEO uygulanmış hali şekil 3.83 (d) de görülmektedir.



Şekil 3.83 Bir sinüsoidal işaretin frekans izlenmesi

3.5.3.2 Enerji etkisinin deneysel deęerlendirilmesi

NTIMIT veritabanında konuřmacılara ait enerji deęerlerinin tanımaya etkisi incelenecektir. Öznitelik vektörleri üretiminde konuřmalar 25 msn'lik çerçevelere ayrılıp 10 msn de bir çerçeveler yenilenir. Konuřmacıların öznitelik vektörleri üretilirken süzgeç dizileri 300-3380 Hz arasına 70 Hz eşit aralıkla % 50 örtüşmeli olarak yerleştirilir. Her bir pencereye karşılık 20 adet Mel frekansı keprstrum katsayısı elde edilir. Bu katsayılar 168 kişinin eğitim ve testi için kullanılır. Eğitim için 8 cümle (~24 sn) test için 1 cümle (~3 sn) kullanılmaktadır. Her bir pencerenin logaritmik enerjisi ve denklem 3.61'de belirtilen teager enerjisi hesaplanır. Konuřmadan sessiz kısımlar atılmamaktadır. Elde edilen sonuçlar Çizelge 3.54'de görülmektedir.

Çizelge 3.54 Enerjinin konuřmacı tanımaya etkisi

Öznitelik vektörü	Tanım oranı (%)
MFCC	69.05
MFCC +log(E)	67.86
MFCC +Teager enerji	66.67
MFCC +log(f_0)	77.68
MFCC +log(f_0)+log(E)	75.60

20 adet MFCC vektörü, karışım bileşen sayısı 32, NTIMIT veritabanı

NTIMIT veritabanı için MFCC ile birlikte log enerji ve teager enerji konuřmacı tanıma başarımlarını azaltmaktadır.

3.5.4 Formant GM-FM Parametreleri

Konuřma rezonansları; ses yolundaki çukur ve tepeler ile belirli frekanslar vurgulanırken dięer frekanslar zayıflatıldığı osilatör sistemleri olarak ifade edilir. Doğrusal modelde, her bir konuřma rezonans işareti, 10-30 msn aralıklarda azalan genlikte sabit frekansta sönümlü kosinüs fonksiyonu olduğu varsayılır. Bir perde periyodu içinde konuřma rezonansının frekans ve genliğindeki anlık deęişimler genlik modülasyonu (GM) ve frekans modülasyonunu (FM) meydana getirir.

Ses üretimi esnasında sesin fıřkırtma şeklinde akışı nedeniyle hava düzensiz hareket eder ve ses yolu çeperleri arasında osilasyon yapar ve ayrılır bu durum hava basıncında modülasyona neden olur. Ayrıca bölüm 3.5.3.1'de belirtilen ses yolunda

oluşan girdaplarda bu durumda etkilidir. Maragos, Quatieri ve Kaiser (1993), GM-FM işaretini, denklem 3.77'deki gibi ifade etmektedir.

$$x(t) = a(t) \cdot \cos[\phi(t)] = a(t) \cdot \cos\left[\int_0^t w(\tau) d\tau + \phi(0)\right]; \quad w(t) = d\phi / dt \quad (3.77)$$

Her bir formant için toplam konuşma işareti, GM-FM işaretlerinin toplamı olarak ifade edilir. Burada zamanla değişen formant işareti için $a(t)$ anlık genlik işareti ve $w(t)$ anlık açısal frekansı göstermektedir. Kısa zamanlı ortalama formant frekansı, $w_c = (1/T) \int_0^T w(t) dt$ ifadesi ile bulunur. Burada T perde periyodu olup GM-FM işaretin taşıyıcı frekansı gibi görülür. Klasik doğrusal konuşma modelinde formant frekansları sabit kabul edilip kısa zaman periyotlarında (10-30 ms) w_c 'ye eşit kabul edilmektedir. Bununla birlikte GM-FM model, ortalama w_c , anlık formant frekans sapması $w(t)$ ve genlik yoğunluğu $|a(t)|$ hakkında ilave bilgi sağlamaktadır.

Orijinal konuşma işaretinden elde edilen tekil rezonans, formant merkez frekanslarının kestirimi için bant geçiren süzgece uygulanır. Teager doğrusal olmayan enerji izleme operatörü (denklem 3.77) kullanılarak denklem 3.78 geliştirilmiştir (Maragos ve ark 1993).

$$\frac{\sqrt{\dot{\psi}[x(t)]}}{\sqrt{\psi[x(t)]}} \approx w(t) \quad \frac{\psi[x(t)]}{\sqrt{\dot{\psi}[x(t)]}} \approx |a(t)| \quad (3.78)$$

Burada $\dot{x} = dx / dt$ olup, denklem 3.78 enerji ayırma algoritması olarak ifade edilir. Üretilen akustik rezonans işareti genlik ve frekans bileşenlerine ayrılır. Ayrık zamanda enerji ayırma algoritması (AEA-1), $x(n) - x(n-1) = y(n)$ olmak üzere denklem 3.79 ve 3.80'de belirtilmektedir.

$$\Omega_i(n) \approx \arccos\left(1 - \frac{\psi[y(n)] + \psi[y(n+1)]}{4\psi[x(n)]}\right) \quad (3.79)$$

$$|a(n)| \approx \sqrt{\frac{\psi[x(n)]}{1 - \left(1 - \frac{\psi[y(n)] + \psi[y(n+1)]}{4\psi[x(n)]}\right)^2}} \quad (3.80)$$

Frekans kestirimi (denklem 3.79), $0 < \Omega_i(n) < \pi$ aralığında yapılabilir. AEA-1 algoritması ile örnekleme frekansının yarısına kadar anlık frekans kestirimi yapılabilir. Maragos, Kaiser ve Quatieri (1993), tarafından geliştirilen diğer bir ayrık enerji algoritması (AEA-2) denklem 3.81 ve 3.82'deki gibi tanımlanmaktadır.

$$\Omega_i(n) \approx \frac{1}{2} \arccos \left[1 - \frac{\psi[x(n+1) - x(n-1)]}{2\psi[x(n)]} \right] \quad (3.81)$$

$$|a(n)| = \frac{2\psi[x(n)]}{\sqrt{\psi[x(n+1) - x(n-1)]}} \quad (3.82)$$

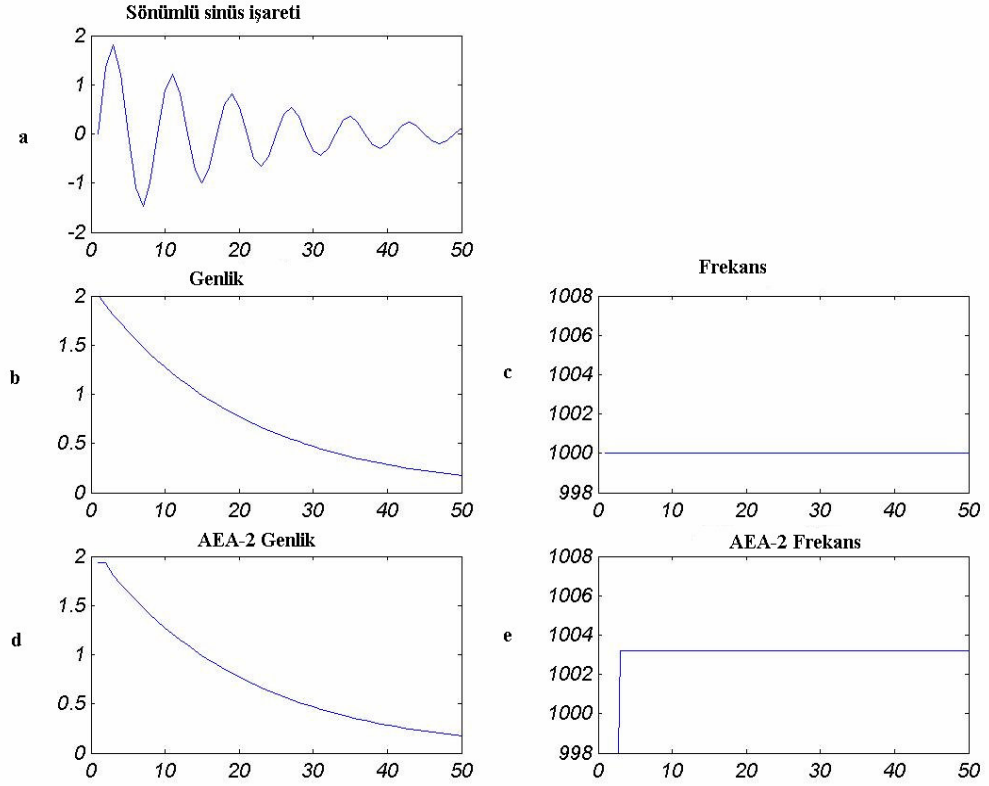
Bu algoritmanın frekans kestirimi (denklem 3.81), $0 < \Omega_i(n) < \frac{\pi}{2}$ aralığında yapılabilir.

AEA-2 kestirilecek anlık frekans örnekleme frekansının $\leq \frac{1}{4}$ olduğu durumlar için geçerlidir. GM-FM işaretler için AEA-2 pek çok durum için AEA-1'e yakın sonuçlar vermektedir (Quatieri 1997). Şekil 3.84'de sönümlü bir sinüs işaretinden $(2 \cdot e^{-0.005n} \sin(n \frac{\pi}{4}))$ AEA-2 uygulanarak genlik ve frekans kestirimi görülmektedir.

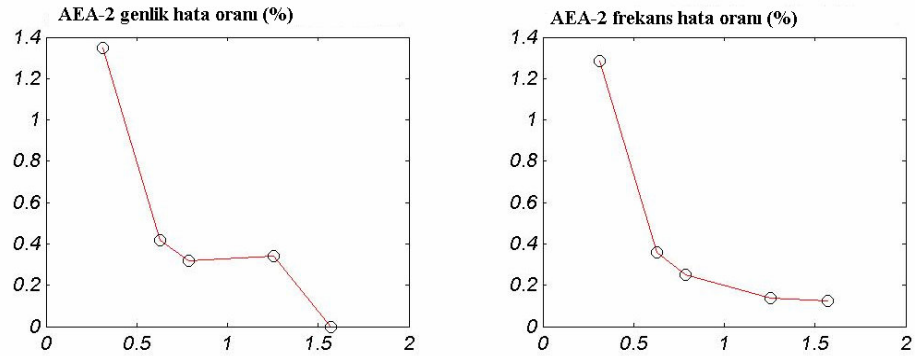
AEA-2 algoritması uygulanması ile genlik ve frekans kestiriminde oluşan hata oranları ise şekil 3.85'de görülmektedir.

Ayrık enerji ayırma algoritmaları ile gerçek konuşma rezonansları hakkında zengin bilgi elde edilebilmektedir. Sonuç olarak;

- i. GM-FM ayrıştırma için küçük hata değerleri vermektedir.
- ii. Düşük hesap kompleksliğine sahiptir.
- iii. Yüksek çözünürlüklü anlık zaman çözünürlüğüne sahiptir.
- iv. Akustik kaynağın gerçek fiziksel enerjisi izlenebilir.
- v. Patlamalı seslerin (p, t vb.) geçiş durumları ayırt edilebilir.



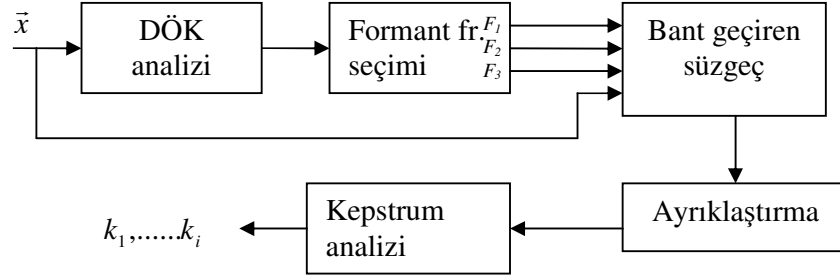
Şekil 3.84 (a) Sönümlü sinüs işareti (b) Sönümlü sinüs işaretinin genliği (c) Sönümlü sinüs işaretinin frekansı (d) AEA-2 algoritması ile kestirilen sönümlü sinüs işaretinin genliği (e) AEA-2 algoritması ile kestirilen sönümlü sinüs işaretinin frekansı



Şekil 3.85 Sönümlü sinüs işaretinden AEA-2 algoritması frekans ve genlik kestirimi sonucu oluşan hata oranları

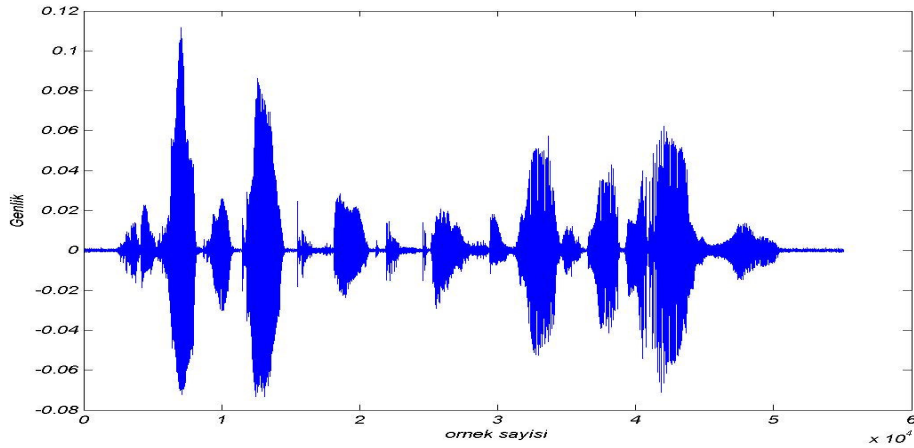
3.5.4.1 Formant GM-FM öznitelik vektörü oluşturma yöntemi

Formant GM-FM parametreleri kullanılarak öznitelik vektörü oluşturma yönteminin blok diyagramı şekil 3.86’da görülmektedir (Jankowski ve ark. 1995).



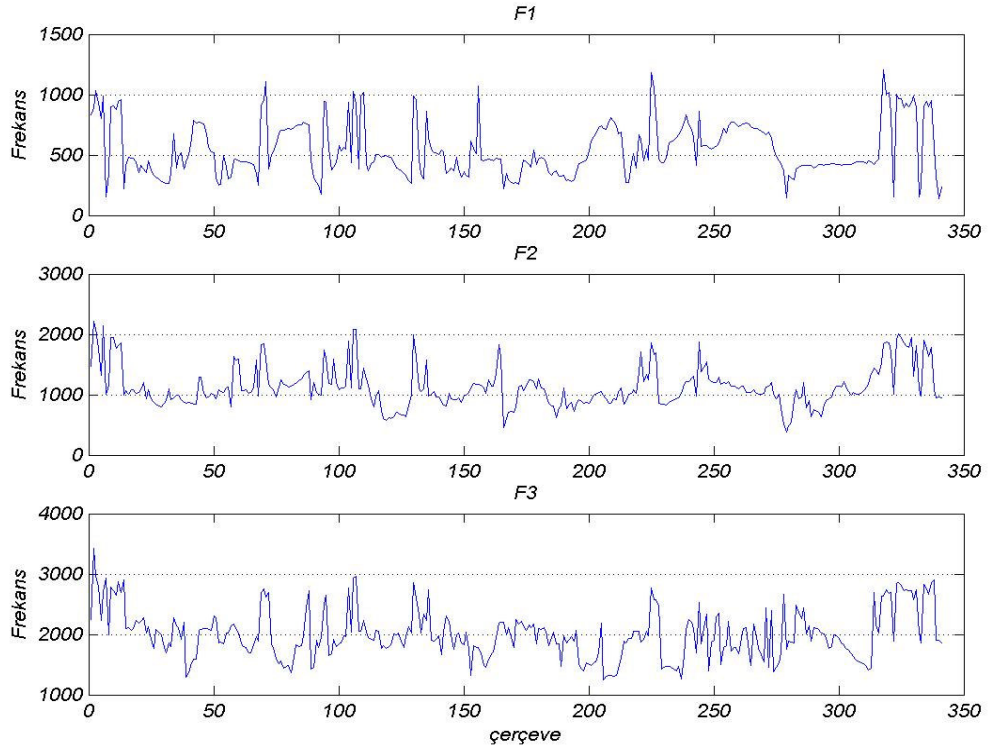
Şekil 3.86 Formant GM-FM parametrelerinin ölçümü için oluşturulan öznitelik vektörü oluşturma yönteminin blok diyagramı

Formant GM-FM öznitelik vektörü oluşturma yönteminde ilk olarak konuşma formantlarının olası yerleri DÖK analizi ile bulunur. DÖK köklerinin genlik ve frekansları kullanılarak ilk üç formant değerleri seçilir. Şekil 3.87’de NTIMIT veritabanından bir cümle görülmektedir.



Şekil 3.87 NTIMIT veri tabanından bir cümle

Şekil 3.87’deki cümleden üretilen ilk üç formant frekansı şekil 3.88’de görülmektedir. İlk formant frekansı F_1 , ortalama 500 Hz civarında, ikinci formant frekansı F_2 , 1000 Hz civarında, üçüncü formant frekansı F_3 , 2000 Hz civarında olduğu gözlenmektedir.



Şekil 3.88 Bir cümle için formant frekansları (bir çerçeve 25 msn)

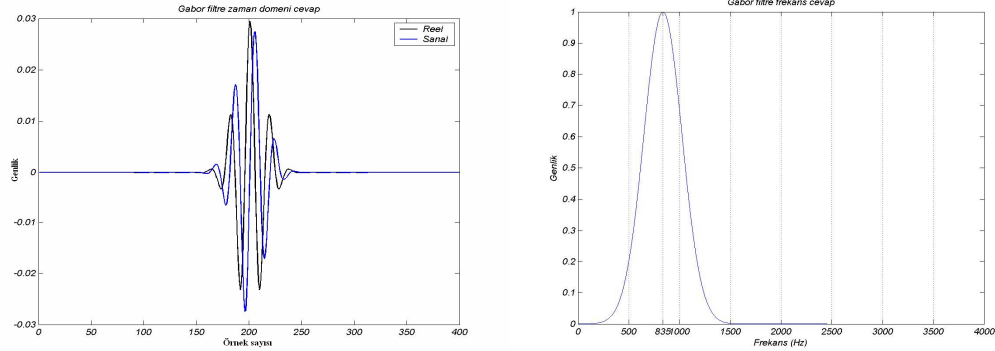
Her bir çerçeve için formant frekans değerleri bir bant geçiren süzgeçten geçirilir. Bu sayede konuşma rezonanslarındaki modülasyonlar ayırt edilebilir. Bant geçiren süzgeç olarak konuşma tanıma ve resim tanıma uygulamalarında sıklıkla kullanılan gabor süzgeci kullanılmaktadır (Quatieri 1997). Gabor süzgecin uyarı cevabı denklem 3.83'de ifade edilmektedir.

$$h(t) = \exp(-\alpha^2 t^2) \cos(w_c t) \quad (3.83)$$

$$H(w) = \frac{\sqrt{\pi}}{2\alpha} \left(\exp\left[-\frac{(w-w_c)^2}{4\alpha^2}\right] + \exp\left[-\frac{(w+w_c)^2}{4\alpha^2}\right] \right) \quad (3.84)$$

Gabor bant geçiren süzgecin frekans cevabı $H(w)$, gauss şeklinde olup kesim frekansı keskin olmayıp derece derecedir. Gabor süzgecin merkez frekansı, formant frekansının değeri olarak alınır. α değeri gabor süzgecin bant genişliği olup, bant

genişliği $\alpha/\sqrt{2\pi}$ değerine eşittir. $h(t)$, t yerine nT yerleştirilerek ayrıklaştırılır (T örnekleme periyodu). $h(n) = \exp(-b^2 n^2) \cos(\Omega_c n)$, $-N \leq n \leq N$, $b = \alpha T$, ve $\Omega_c = 2 \cdot \pi \cdot f_c \cdot T$ simetrik FIR süzgeç olarak ifade edilir. Şekil 3.89'da bir boyutlu gabor süzgecin zaman ve cevabı görülmektedir.

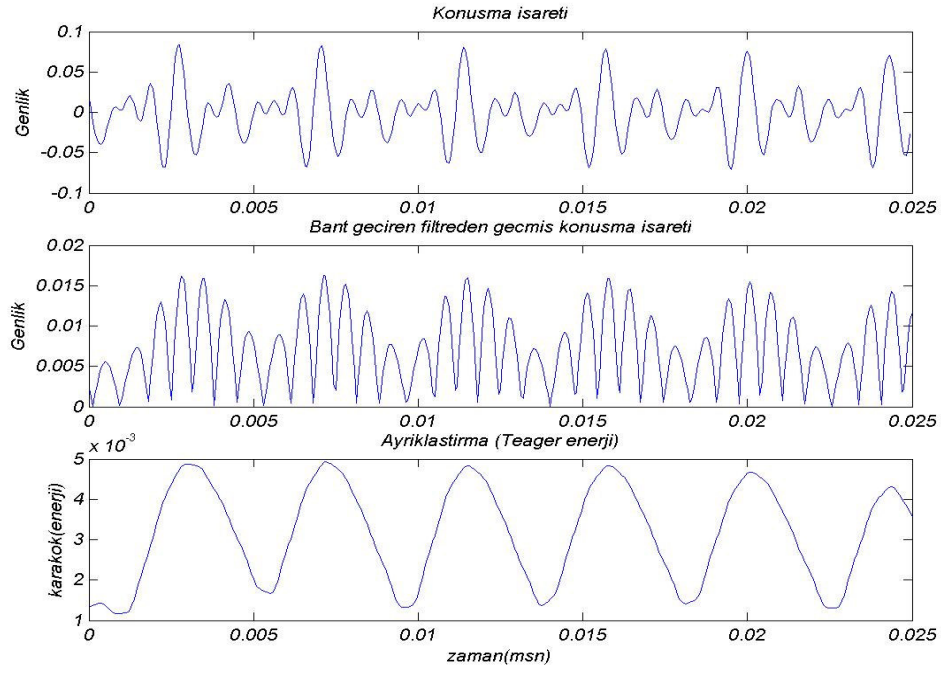


Şekil 3.89 Bir boyutlu gabor süzgeçlerinin zaman ve frekans cevabı
(merkez frekansı $f_c = 835$ Hz)

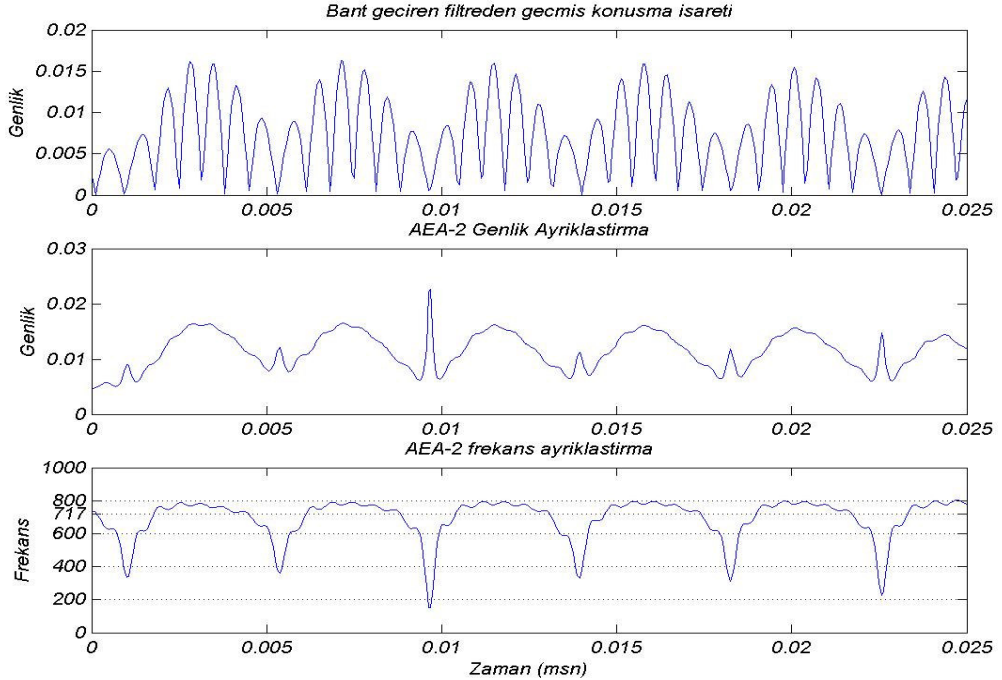
Her bir çerçeve için konuşma işareti ile merkez frekansı formant frekansı olan gabor bant geçiren süzgecin katlamalı integrali alınır. Gabor süzgeçten geçirilen dalga şekillerine ya Teager enerji operatörü (Enerji için) yada AEA-1, 2 enerji ayırma algoritmaları (genlik ve frekans için) uygulanarak ayrıklaştırılır. Bu sayede seçilen her bir formant frekansı için, enerji, genlik veya frekans ayrıklaştırmalardan biri kullanılarak üç dalga şekli oluşturulur.

Şekil 3.90 (a) da NTIMIT veri tabanından fadg0 klasöründeki Sa1 cümlesinin 80. çerçevesine ait konuşma işareti görülmektedir. Merkez frekansı 717 Hz olan gabor bant geçiren süzgeçten geçirilmiş işaret şekil 3.90 (b) de, teager enerji operatörü uygulanmış halinde şekil 3.90 (c) de görülmektedir.

Şekil 3.90 (a) da verilen konuşma işaretine ayrıklaştırma işlemi olarak AEA-2 algoritması uygulanması durumunda bir çerçeve için elde edilen değerler şekil 3.91'de verilmektedir. Şekil 3.91 (b) de AEA-2 algoritması ile elde edilen genlik zarfı görülmekte olup, şekil 3.91 (c) de ise aynı algoritma ile birkaç perde periyodu için elde edilen anlık frekans değerleri görülmektedir.



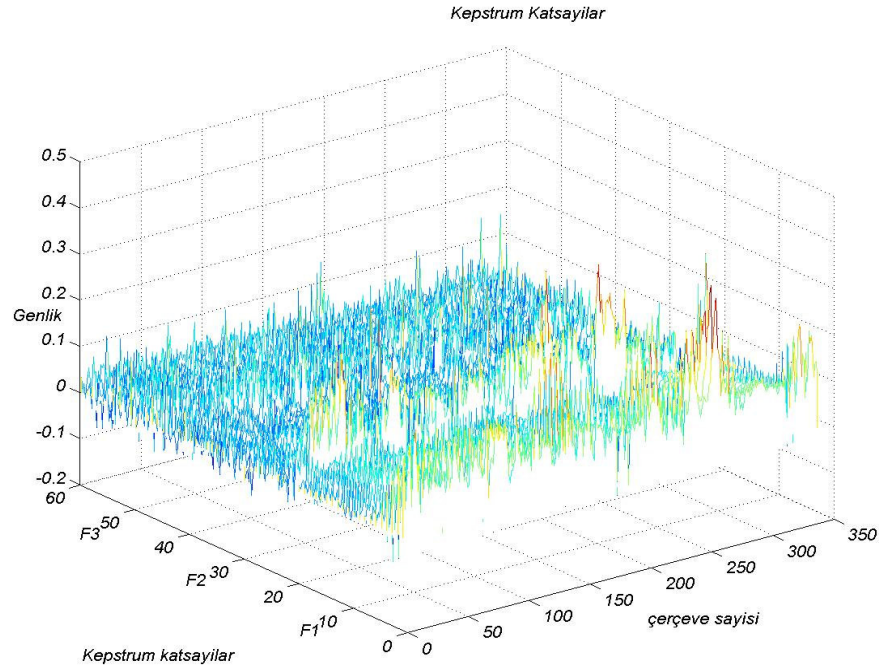
Şekil 3.90 (a) 25 msn uzunluğunda bir konuşma işareti parçası (b) gabor bant geçiren süzgeçten geçirilmiş konuşma işareti (c) Teager ayrıklaştırma ($\sqrt{\psi[x(n)]}$)



Şekil 3.91 (a) 25 msn uzunluğunda bir konuşma işareti parçasının gabor bant geçiren süzgeçten geçirilmiş hali (b) AEA-2 kullanılarak genlik zarfının kestirimi (c) AEA-2 kullanılarak anlık frekans kestirimi

Yukarıdaki şekillerden de görüleceği üzere AEA-2 genlik zarfı ile Teager enerji operatörü birbirine yakın şekiller sergilemektedir. AEA-2 ile kestirilen anlık frekans merkez frekansı etrafında salınım yapmaktadır.

Bu analizler, konuşma dalga şekli çerçevelere bölünerek her 10 msn'de bir yenilenir. Her bir 25 msn'lik çerçeve parçasının ortalama bileşenleri atılır. Konuşma parçası hamming pencereden geçirilir. $k[n]$ ile ifade edilen kepstrum katsayıları elde edilir. Konuşmacı tanıma sisteminde, öznitelik vektörü olarak orta frekansı gösteren kepstrum katsayıları kullanılır. Yüksek frekansı gösteren kepstrum katsayılar perde frekansı bilgisi taşıdığından dolayı kullanılmaz. Alçak frekansı gösteren kepstrum katsayıları, formantların enerji bilgisini taşıdığından dolayı silinir (Jankowski ve ark. 1995). Formant GM-FM işlemi sonucu 3 dalga şekli için elde edilen kepstrum katsayıları şekil 3.92'de görülmektedir. Her bir formant frekansı için k_9, \dots, k_{28} katsayıları alınmıştır. Çünkü formant frekanslarına karşılık gelen kepstrum katsayıları bu aralıkta bulunmaktadır. Şekil 3.32'de kepstrum katsayılarının karşılık geldiği frekans bölgeleri açıklanmıştır.



Şekil 3.92 Bir cümle için formant GM-FM işlemi sonucu elde edilen kepstrum katsayıları

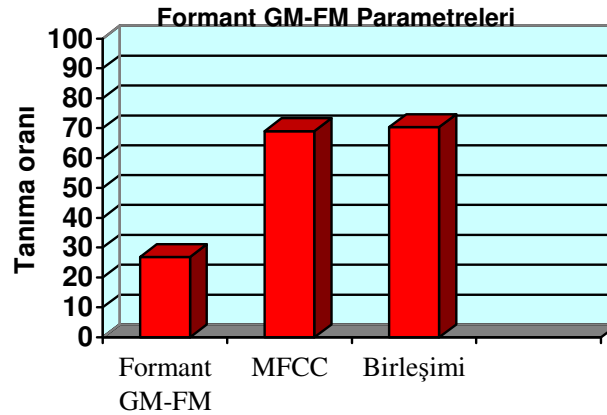
İlk 20 kepstrum katsayısı, F_1 'e bağlı olarak oluşturulan öznitelik vektörüdür. 20-40 arasındaki kepstrum katsayıları F_2 , 40-60 arası kepstrum katsayıları ise F_3 'e bağlı olarak oluşturulan öznitelik vektörleridir.

3.5.4.2 Formant GM-FM parametrelerinin deneysel değerlendirilmesi

Blok diyagramı şekil 3.86'da verilen, formant GM-FM parametreleri kullanılarak oluşturulan öznitelik vektörleri konuşmacı tanıma için kullanılacaktır. Her bir çerçeve için oluşturulan DÖK polinomun köklerinden F_1, F_2, F_3 formant frekansları seçilir. Konuşma dalga şekli, formant frekanslarına bağlı olarak gabor bant geçiren süzgeçten geçirilir ve ayırıştırma uygulanır. Ayırıştırma için teager enerji operatörü veya AEA 1-2 algoritmaları ile seçilen dalga şekillerine enerji, genlik veya frekans ayırıştırılmalarından biri kullanılarak 3 dalga şekli üretilir. Her bir dalga şekline kepstrum analizi uygulanarak öznitelik vektörleri üretilir.

Öznitelik vektörü üretiminde, her bir 25 ms'n'lik çerçevenin ayırıştırma operatörü olarak enerji, genlik veya frekans alınır. Konuşma parçası hamming pencereden geçirilir ve $k[n]$ ile ifade edilen kepstrum katsayıları elde edilir. Konuşmacı tanıma sisteminde, öznitelik vektörü olarak orta frekansı gösteren kepstrum katsayıları (k_9, \dots, k_{28}) kullanılır. Eğitim ve test için NTIMIT veri tabanından 168 kişi kullanılmaktadır. Bu konuşmacı kümesinin 56'sı kadın 112'si erkektir.

1. Ayırıştırma operatörü olarak teager enerji algoritması kullanılması durumunda elde edilen tanıma oranları şekil 3.93'de görülmektedir.



Şekil 3.93 Ayırıştırma teager enerji olması durumunda tanıma oranları

Formant GM-FM parametreleri yalnız başına kullanılması durumunda tanıma oranı %26.8, formant GM-FM parametrelerinin MFCC vektörlerine (% 69.05) eklenmesi ile konuşmacı tanıma oranı % 70.33 olmaktadır.

2. Ayırıklaştırma operatörü olarak teager enerji yerine AEA-1 veya 2 genlik kestirim algoritmaları kullanılması durumunda elde edilen konuşmacı tanıma oranları çizelge 3.55’de görülmektedir.

Çizelge 3.55 AEA-1 ve AEA-2 genlik kestirimi ile tanıma oranları (%)

Kepstrum katsayılarının kullanıldığı formant frekansları	AEA-1	AEA-2
F_1, F_2, F_3	16.07	19.88
F_1	15.71	17.98
F_2	7.74	8.33
F_3	5.95	5.14

Ayırıklaştırma operatörü olarak teager enerji yalnız başına kullanıldığında formant GM-FM tanıma oranı %26.8, AEA-1 için % 16.07, AEA-2 için % 19.88 olmaktadır.

Sonuç olarak formant GM-FM parametrelerinin MFCC vektörleri ile birlikte kullanılması durumunda, klasik MFCC’ye nazaran 1.28 puan tanıma artışı olmaktadır.

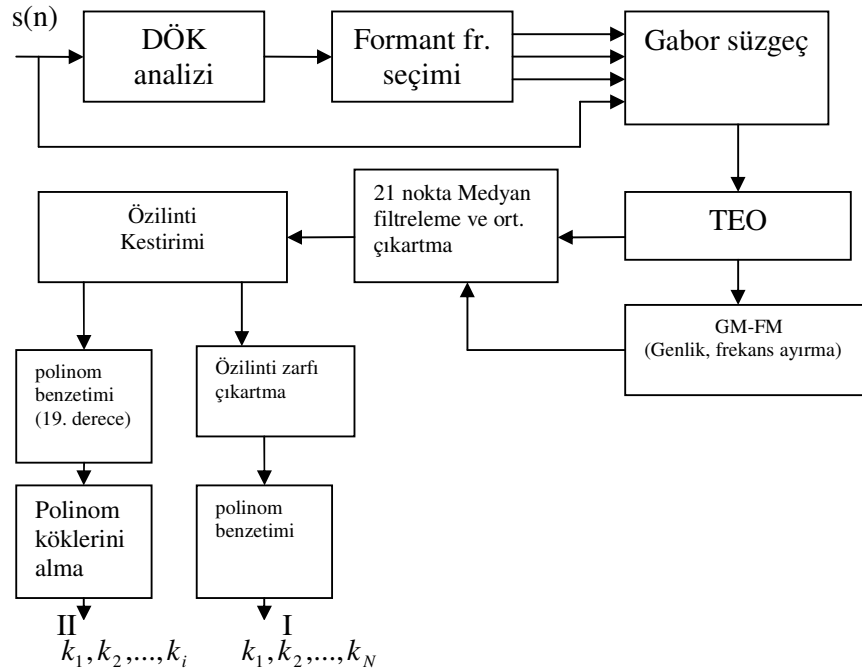
3.5.5 Doğrusal olmayan öznitelik parametreleri elde edilirken özilinti katsayılarının kullanılması ve polinom benzetimi

Konuşma, ses telleri ve ses yolu etkisi ile oluşumu esnasında doğrusal olmayan bir akış sergiler. Bu akışın modellenmesi zordur. Doğrusal olmayan konuşma özniteliklerinin analizinde formant GM-FM parametrelerine ek olarak özilinti katsayıları kullanılacaktır. Özilinti katsayılarının zarfı alınıp polinom benzetimi yapılacaktır. Bu sayede ses üretim mekanizması modellenecektir. Hansen ve ark. (1998), bu katsayıları kullanarak ses üretim yolundaki değişimlerin iyi modellenebildiğini bildirmekte ve ses üretim yolu hastalıklarındaki değişikliklerin izlenebileceğini önermiştir.

Son zamanlarda ses biliminde yapılan araştırmalarda ses üretim yolunda oluşan hastalıkların değerlendirmesinde, konuşma özniteliklerinin analizi ile ilgilenilmektedir. Konuşma özniteliklerinin analizi ile hastalıkta zamana bağlı değişimlerin takibi sağlanabilmektedir. Algılanan sesin niteliği, ses tellerinin açılıp kapanması ile doğrudan

ilişkilidir. Tabii olmayan söyleyiş biçimleri konuşma üretiminde oluşan problemlerden dolayı ortaya çıkar. Bunlar; fısıltı şeklinde konuşma, ses kısıklığı, ses tellerinin tamamen kapanmaması sonucu oluşan konuşma bozuklukları, ses tellerinin olduğundan fazla sıkıştırılması sonucu zayıf ses oluşumudur. Ses tellerinin tamamen kapanmaması farklı kas kasılmalarına neden olur. Ses tellerindeki asimetri iki farklı ses tonu oluşmasına neden olmakta buna bağlı olarak iki farklı temel frekans oluşmaktadır.

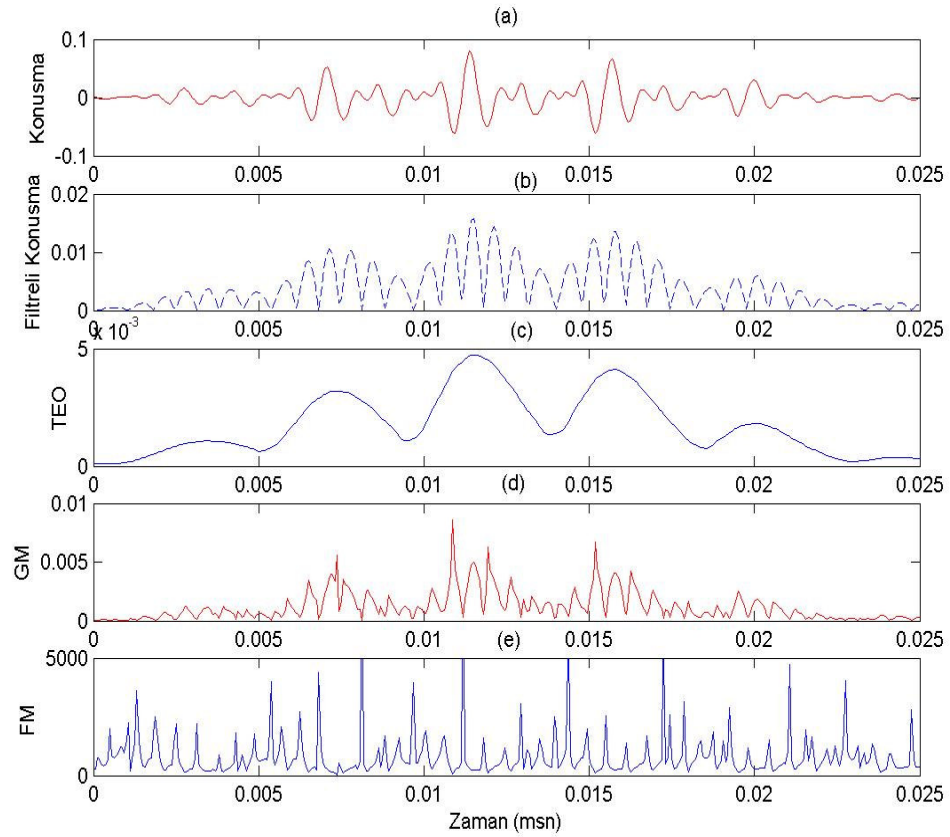
İnsan ses sisteminin anatomik ve fiziksel olarak kötü kullanılması sonucu seste bozulmalar oluşabilir. Bu bozulmaların genellikle görüntüleme ile belirlenmesi zordur. Çünkü gırtlak yapısı normal görünür. Sadece gırtlak kas gerilmelerinde ve ses kalitesinde azalma gözlenir (Hansen ve ark. 1998). Bu bozulmaların nedenini bulmak için ses terapisi deneyleri yapılır. Ses sisteminin yanlış kullanımdan dolayı oluşan bu durumlarda, doğru ses sistemi davranışları öğretilmesi gerekir. Bu durumda, şekil 3.94 de görüldüğü gibi konuşma özniteliklerinin analizi ile kişinin sesini karakterize eden katsayılar üretilmektedir.



Şekil 3.94 Doğrusal olmayan konuşma özniteliklerinin analizi blok diyagramı

Şekil 3.94'deki bloklar incelendiğinde ilk olarak konuşma işaretine formant frekansların yerini çıkartmak için 19. dereceden DÖK analizi uygulanır (Jankowski ve ark. 1995). Her bir çerçeve için F_1, F_2, F_3 formant frekanslarının yeri bulunur.

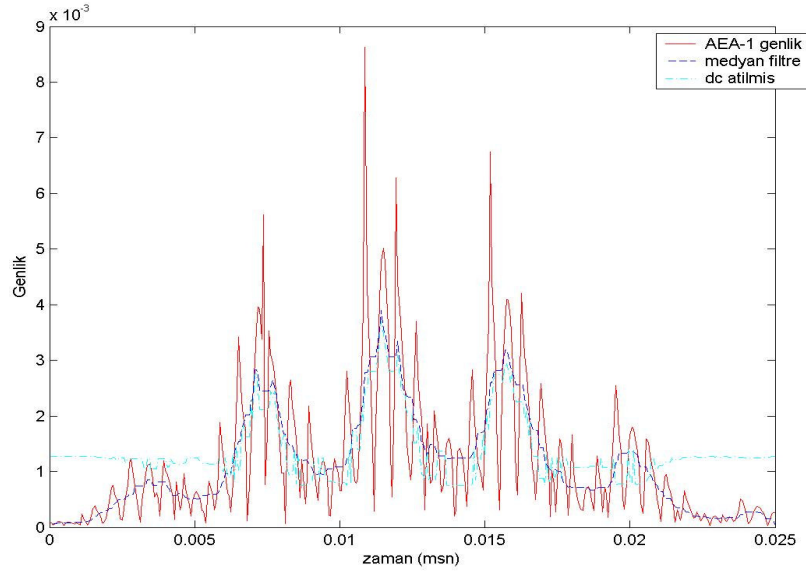
Formantları etkinleştirmek için işaret, merkez frekansı formant frekansları olan gabor bant geçiren süzgeçlerden geçirilmektedir. Daha sonra süzgeçlenen konuşmaya TEO uygulanıp GM ve FM modülasyonlu işaret elde edilir. Şekil 3.95 (a)'da NTIMIT veritabanından bir cümleye ait 25 msn'lik parça görülmektedir. Şekil 3.95 (b)'de ise konuşma parçasının gabor süzgeçten geçirilmiş hali görülmektedir. Gabor süzgecin merkez frekansı F_1 'in frekansına eşit olup 716.8 Hz dir. Şekil 3.95 (c)'de süzgeçlenen konuşmanın Teager enerjisi alınmış hali bulunmaktadır. Şekil 3.95 (d)'de teager enerjisi alınmış işaretin AEA-1 genlik kestirimi uygulanmış hali görülmektedir. Şekil 3.95 (e) de ise AEA-1 frekans kestirimi uygulanmış işaret görülmektedir.



Şekil 3.95 NTIMIT veritabanında bir konuşmacıya ait 25 msn lik çerçevede (a) orijinal konuşma (b) süzgeçlenmiş konuşma (c) TEO (d) AEA-1 ile genlik kestirimi (e) AEA-1 algoritması ile frekans kestirimi

Elde edilen genlik ve frekans bilgisi içeren işaretlere 21 nokta medyan süzgeçten geçirilerek pürüzsüzleştirilir ve işaret ortalama değerinden çıkartılarak işaretteki sabit

bileşenler kaldırılır. Şekil 3.96’da AEA-1 genlik kestirimi uygulanmış işaretin medyan süzgeci ve ortalama değerden çıkartılmış hali görülmektedir.

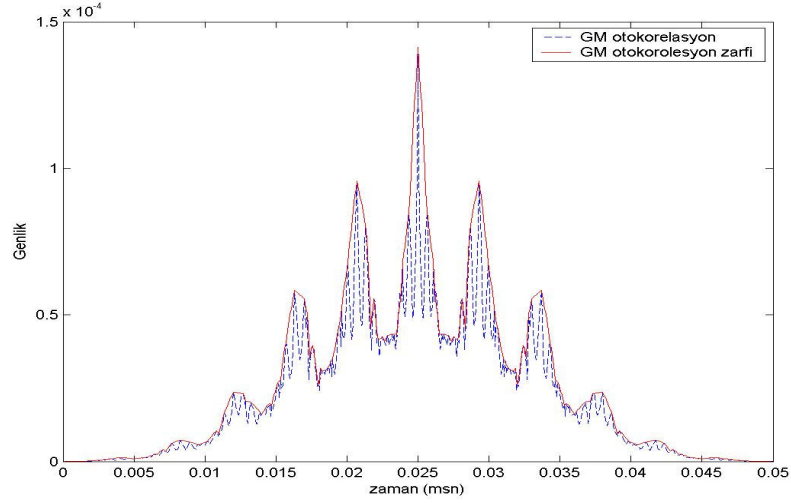


Şekil 3.96 AEA-1 genlik kestirimi uygulanmış 25 msn’lik işaretin 21 nokta medyan süzgeci ve ortalama bileşenler atılmış hali

Sonraki adımda genlik bileşenin özilinti değerleri hesaplanır. Bir konuşma işaretini temsil eden dizi $s(n)$ in uzunluğu N ile ifade edilir. Sonuçta $N-1$ uzunluğunda özilinti değerleri denklem 3.85’den elde edilir (Hansen ve ark. 1998).

$$R_s(k) = FT^{-1} \left[|S'(k)|^2 \right] \quad (3.85)$$

Burada $S'(k)$ değeri $s(n)$ dizisinin FFT alınmış halidir. $R_s(k)$, $s(n)$ dizisinin doğrusal özilintisine karşılık gelmektedir. Bundan sonra öznitelik katsayıları elde edilirken iki değişik yöntem uygulanmaktadır. Birinci yöntemde özilintisi elde edilen işaretin zarfı alınır. Şekil 3.97’de AEA-1 genlik kestirimi uygulanan işaretin özilintisi alınmış hali ve bu işaretin genlik zarfı görülmektedir. Şekilden görüleceği üzere genlik zarfı işaretin tepe noktalarını takip etmektedir.

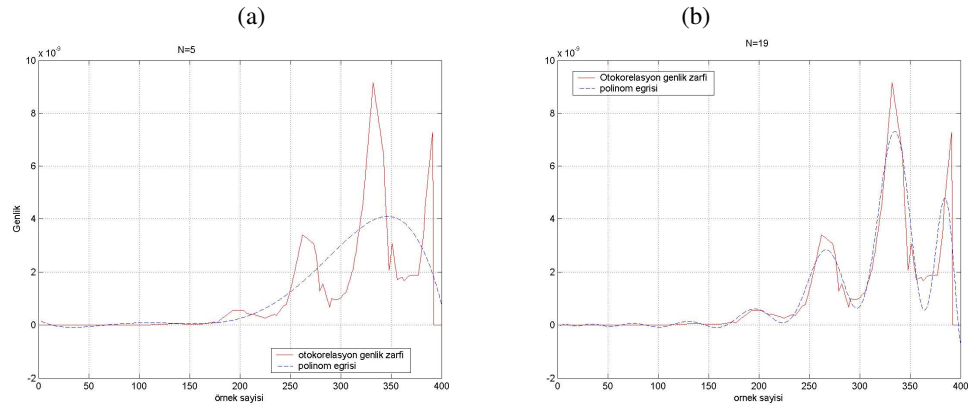


Şekil 3.97 AEA-1 genlik kestirimi uygulanan işaretin özilintisi ve özilintisinin genlik zarfı alınmış şekil

Son olarak elde edilen işaretin bir polinom ile benzetimi yapılır. Denklem 3.86’da genlik zarfı alınan katsayıların polinom ile ifade edilmesi görülmektedir.

$$P(x) = k_1x + k_2x^2 + \dots + k_Nx^N \quad (3.86)$$

Polinomun katsayıları öznitelik vektörü olarak atanır. İkinci yöntemde ise özilinti kestirimi yapılan işaretin polinom benzetimi yapılarak polinomun kökleri öznitelik katsayısı olarak alınır. Şekil 3.98’de özilinti genlik zarfı alınmış işaretin (a) polinom derecesi 5 için (b) polinom derecesi 19 için polinomlara ait eğriler görülmektedir.



Şekil 3.98 Özilinti genlik zarfı işaret ve bu işarete ait (a) $N = 5$ için (b) $N = 19$ için polinomlara ait eğriler

Şekil 3.98'den görüleceği üzere özilinti genlik zarfı alınan işaret polinom derecesi 5 için işaret tamamen modellenememekte polinom derecesi 19 için işaretteki tepe ve çukurların benzetimi yapılabilmektedir.

Şekil 3.94'de verilen doğrusal olmayan konuşma özneliklerinin analizi blok diyagramındaki I. ve II. Yöntemler ile elde edilen katsayılar öznelik vektörü olarak kullanılarak GKM konuşmacı tanıma sistemine verilmektedir. Öznelik vektörü olarak 19 katsayı alınmaktadır. NTIMIT veritabanından 168 kişi, 8 cümle ile eğitilip 1 cümle ile test edilmektedir. Polinom katsayıları yalnız başına kullanıldığında ve MFCC katsayıları eklenmiş durumlarda konuşmacı tanıma oranları verilmektedir. Elde edilen konuşmacı tanıma oranları çizelge 3.56'da görülmektedir.

Çizelge 3.56 I. ve II. yöntemler öznelik vektörleri olarak kullanılması durumunda konuşmacı tanıma oranları (%)

	I. Yöntem	II.yöntem
Polinom katsayıları	14.88	13.10
Polinom katsayıları +MFCC	15.48	35.71

20 adet MFCC, Karışım bileşen sayısı 32, NTIMIT veritabanı

Polinom katsayılarının tek başına ve MFCC katsayıları ile birlikte kullanılmaları durumunda konuşmacı tanıma oranı azalmaktadır. Ancak şekil 3.94'de I. yöntem olarak verilen polinom katsayıları, ses yolu hastalıklarında davranış düzeltilmesi amacıyla yapılan terapi öncesi ve sonrası farklılıkları ayırmada etkin olarak kullanılmaktadır (Hansen ve ark. 1998).

Formant GM-FM parametreleri ve doğrusal olmayan öznelik parametreleri elde edilirken özilinti katsayılarının kullanılması ve polinom benzetiminin konuşma tanımada iyi sonuçlar vermemesine sahte formantlar neden olmaktadır. TIMIT ve NTIMIT veritabanlarının formant yerlerini karşılaştırdığımızda çizelge 3.60 ortaya çıkmaktadır (Jankowski ve ark 1994). Sahte F_1 , NTIMIT'te formant izleyicide elde edilen F_1 , TIMIT formantlarından hiçbirine uymuyorsa oluşur.

Çizelge 3.57 incelendiğinde pek çok çerçevede sahte formantlar oluştuğu görülmektedir. Sahte formantlar erkeklerden çok kadınlarda oluşmaktadır. NTIMIT veritabanına ait konuşmalarda olan pek çok formant gözükmemektedir. Bu durum NTIMIT dalga formlarında oluşan harmonik bozulmalardan dolayı oluşmaktadır. Sahte formant frekansları, gerçek formantların toplamı veya farkı şeklinde oluşmaktadır. Bu

tip harmonik bozulmaların temelinde telefon ahizesinden dolayı oluşan doğrusal olmayan bozulmalardır (Reynolds ve ark. 1995).

Çizelge 3.57 TIMIT ve NTIMIT veritabanları için çerçeve başına formant karşılaştırmaları (Jankowski ve ark. 1994)

	Erkekler	Kadınlar
Doğru eşleştirme	70.0	20.3
Sahte F_1	0.0	0.3
Sahte F_2	9.8	57.2
Sahte F_3	11.5	15.5
diğer	8.7	6.7

4. ARAŞTIRMA SONUÇLARI VE TARTIŞMA

4.1 Araştırma Sonuçları

Bu tezde Matlab programı yardımıyla GKM konuşmacı tanıma sistemi oluşturulmuş ve değişik öznitelik vektörü oluşturma yöntemleri geliştirilerek Gauss karışım modeli ile konuşmacı tanıma etkisi incelenmiştir. Öznitelik vektörü elde edilirken orijinal olarak kullanılan kümeleme ile ağırlıklandırma ve konuşmacı frekans bandı parçalara ayrılıp, F-oranına bağlı olarak süzgeç yerleştirilmesi yöntemi ile konuşmacı tanıma oranında artış sağlanmıştır.

TIMIT ve NTIMIT veritabanlarının kullanıldığı bir konuşmacı tanıma sistemi, eğitim ve test aşamasında çeşitli parametre değişimlerine karşı incelenip, en yüksek başarıyı verecek ideal parametreler belirlenmiştir. TIMIT veritabanı, gürültü olmayan ortamda mikrofon ile veriler toplandığından dolayı temiz bir veritabanı olarak tanımlanmaktadır. NTIMIT veritabanı, konuşmalar telefon ortamından iletiildiğinden dolayı telefon ahizesi ve iletim hattının etkilerini içermektedir. TIMIT ve NTIMIT veritabanının bu farklılıklarından dolayı ideal parametreler, bu veritabanlarına bağlı olarak farklılık arz etmektedir.

NTIMIT veritabanında TIMIT veritabanında verilen parametrelerden farklı olarak şu parametreler kullanılmaktadır. Öznitelik vektörü elde edilirken konuşma 25 msn'lik çerçevelere ayrılıp 10 msn'de bir çerçeve ötelenir. İşarete ön vurgulama uygulanmaz. İşaret, 300-3400 Hz aralığında doğrusal veya ERB ölçekte dizilmiş üçgen süzgeç dizilerinden geçirilir. Doğrusal ölçekte dizilmiş süzgeçlere % 50 örtüşme uygulanır. En ideal kepsrum katsayı sayısı 20 olarak bulunmuştur. Konuşmadan sessiz kısımların atılması, NTIMIT veritabanında tanıma oranını arttırmaktadır.

Bu parametreler iki veritabanı için karşılaştırmalı olarak çizelge 4.1'de görülmektedir. Bu parametreler ile TIMIT veritabanında % 100, NTIMIT veritabanında ise % 73.51 konuşmacı tanıma oranı elde edilmiştir.

Çizelge 4.1 TIMIT ve NTIMIT veritabanı için ideal öznitelik parametreleri

Özellikler	TIMIT	NTIMIT
Çerçeveleme-ilerleme süresi	20-10 msn	25-10 msn
Pencereleme fonk.	Hamming	Hamming
FFT örnek sayısı	512	512
Ön vurgulama	var	yok
Süzgeç dizisi şekli	Üçgen	Üçgen
Süzgeç dizisi örtüşme oranı	% 50	% 50
Frekans ölçeği	Doğrusal, ERB, Mel	Doğrusal, Mel
Kepstrum katsayı sayısı	24	20
Sessiz kısımların atılması	yok	var

Eğitim 8 cümle, test 1 cümle, karışım bileşen sayısı 32, örnekleme frekansı 16 kHz

Bölüm 3.2.1’de görüleceği üzere, GKM eğitimi aşamasında bazı ayarlamalar yapılması gerekir. BM eğitim algoritması özyineleme sayısı 15 alınması modelin yeterli yakınsamasını sağlamaktadır. Model başlangıç değerleri VN algoritması ile bulunur. GKM’nin tekilliklerden korunması için minimum varyans sınırı $\sigma^2_{\min}=0.01$ alınır. Eğitim süresi 24 sn için, karışım bileşenleri sayısının 32 alınması ile konuşmacılar en iyi şekilde modellenmektedir. Eğitim süresinin düşmesi ile birlikte ideal karışım bileşen sayısı düşmektedir. Eğitim süresi 24 sn için (çizelge 3.10) ideal karışım bileşen sayısı 32, eğitim süresi 9 sn için ideal karışım bileşen sayısı 8 (çizelge 3.8) olmaktadır. İdeal eğitim parametreleri iki veritabanı için karşılaştırmalı olarak çizelge 4.2’de görülmektedir.

Çizelge 4.2 TIMIT ve NTIMIT veritabanı için ideal GKM eğitim parametreleri

Parametreler	İdeal değerler
BM algoritması özyineleme sayısı	15
Model başlangıç değerleri	VN
Varyans sınırı	$\sigma^2_{\min}=0.01$
Karışım bileşen sayısı	32

Eğitim süresi 24 sn, test süresi 3 sn, örnekleme frekansı 16 kHz

GKM kullanılarak yapılan deneylerde, model eğitimi aşamasında karışım sayısı ve eğitim süresi; test aşamasında ise test süresi değişimi, konuşmacı tanıma başarımı üzerinde belirleyici olmaktadır. Eğitim ve test süreleri artışına paralel olarak tanıma oranı artmaktadır. Bu durum şekil 3.9, şekil 3.10, şekil 3.11 ve şekil 3.12’de görülmektedir. Her iki veritabanı içinde en yüksek tanıma oranı eğitim süresinin 24 sn,

test süresinin 6 sn olduğu durumda olmaktadır (çizelge 3.10 ve çizelge 3.11). Ancak literatürde test süresi için bir cümle (~3 sn) kullanılmaktadır (Reynolds 1995). TIMIT veritabanının tamamı (630 kişi) için GKM konuşmacı tanıyan sistem test edildiğinde, kişi sayısının artmasına bağlı olarak, konuşmacı tanıma başarımında (% 98.81) önemli oranda düşme olmadığı görüldü. Ancak NTIMIT veritabanı 168 kişi ile test edildiğinde tanıma oranı % 69.64'e düşmektedir (şekil 3.13). Tanıma oranındaki bu düşüş, NTIMIT veritabanında konuşmaların telefon hattından kaydedilmesinden kaynaklanmaktadır.

Bu tezde, telefon ortamından dolayı oluşan bozulmaları azaltmak için denklem 3.66 kullanılarak özniteliklerin kümelenerek ağırlıklandırılması önerilmiştir. TIMIT veritabanının NTIMIT veritabanına benzetimini yapmak için örnekleme frekansı 8 kHz'e çekilmiş, 100-4000 Hz arası bant sınırlaması uygulanmıştır. Eğitim süresi 15 sn için, 12 kepstum katsayısı kullanıldığında, küme sayısı 2 için tanıma oranı % 93 olup bu klasik MFCC'ye göre 6 puan; 20 kepstum katsayısı kullanıldığında, küme sayısı 16 için tanıma oranı % 88 olup bu da klasik MFCC'ye göre 5 puan artış sağlanmıştır (çizelge 3.41). NTIMIT veritabanında, 20 kepstum katsayısı kullanıldığında, bant aralığı 300-3400 Hz alınıp, küme sayısı 8 için tanıma oranı % 73 olup bu klasik MFCC'ye göre konuşmacı tanıma oranında 9 puan artışa karşılık gelmektedir (çizelge 3.42). Ayrıca TIMIT ve NTIMIT veritabanı için spektral değişim kompanzasyonu yöntemlerinin konuşmacı tanıma oranını azalttığı gözlenmiştir.

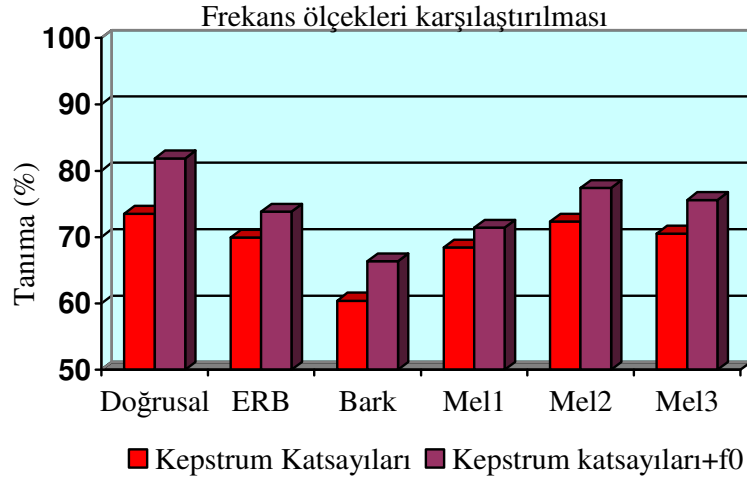
Öznitelik vektörü elde edilirken süzgeç dizilerinin denklem 3.67 ile tanımlanan konuşmacı frekans bandına F-oranına bağlı olarak süzgeç yerleştirilmesi önerilmiştir. Süzgeç dizilerinin yerleştirildiği frekans ölçeği belirli parçalara ayrılıp her bir parçaya yerleştirilecek süzgeç sayısı, F-oranına bağlı olarak tespit edilmiştir. Bu sayede ayırt ediciliği fazla olan frekans aralığına daha fazla süzgeç yerleştirilmiştir. TIMIT veritabanında yapay sınırlamalar yapılarak telefon ortamına benzetilmeye çalışılmıştır. Önerilen bu yöntem, bant sınırlamalı ortamda tanıma oranını arttırmaktadır (şekil 3.64). NTIMIT veritabanında bu yaklaşımda konuşmacı tanıma oranında klasik MFCC'ye nazaran 10 puan artış sağlanmıştır (şekil 3.65).

NTIMIT veritabanında uygun eşik değeri kullanılarak konuşmadan sessiz kısımların atılması ile konuşmacı tanıma oranı % 69.05'den % 73.51'e çıkmaktadır. TIMIT veritabanında ise konuşmadan sessiz kısımların atılması tanıma oranını

azaltılmaktadır. Çünkü TIMIT veritabanı gürültüsüz temiz bir veritabanı olduğu için konuşmadan atılan kısımlar konuşmacıyı ayırt edici bilgiler taşımaktadır.

MFCC parametrelerine, bürünsel özniteliklerin eklenmesi ile elde edilen öznitelik vektörleri kullanılarak yapılan deneylerde, temel frekans, tanımayı önemli oranda arttırırken, formant frekansları ve enerji tanıma başarımını düşürmektedir. NTIMIT veritabanında gürültü ve iletişim ortamının (telefon v.b.) doğrusal olmayan etkisinden (Reynolds ve ark. 1995) temel frekans daha az etkilenmekte (Arcienega ve Drygajlo 2001) bu nedenle konuşmacı tanıma başarımını arttırmaktadır.

NTIMIT veritabanında, konuşmadan sessiz kısımlar atılıp 32 karışım sayısı, 25 msn çerçeveleme, 10 msn ilerleme süresi ile 20 adet MFCC vektörü üretildiğinde, tanıma oranı % 73.51 olup bu katsayılara temel frekans eklenmesi ile tanıma oranı %81.85'e çıkmaktadır (çizelge 3.44). TIMIT veritabanında bant sınırlama durumunda ($f_s=16$ kHz), MFCC vektörlerine temel frekans eklenmesi konuşmacı tanıma oranını azaltılmaktadır. NTIMIT veritabanında kepstrum katsayılarına temel frekans eklenmesi durumunda frekans ölçeği değişiminin konuşmacı tanımaya etkisi şekil 4.1'de görülmektedir.



Şekil 4.1 Frekans ölçeklerinin tanımaya etkisi (NTIMIT veritabanı)

Şekil 4.1'den görüleceği üzere kepstrum katsayıları doğrusal ölçek ile yalnız başına %73.51 ve f_0 ile birlikte kullanıldığı durumlarda %81.85, en yüksek tanıma oranı elde edilmektedir. Öznitelik vektörleri üretilirken süzgeçlerin yerleştirildiği frekans ölçekleri konuşmacı tanıma oranına göre doğrusal, Mel2 (Fant (1960)), Mel3

(Slaney (1998)), ERB, Mell (O'Shaughnessy (1987)) ve Bark ölçeği şeklinde sıralanmaktadır (çizelge 3.48).

NTIMIT veritabanı için, formant frekanslarının (F_1, F_2, F_3) öznitelik vektörü olarak kullanılması durumunda tanıma oranı % 17.26, MFCC vektörlerine eklenmesi durumunda tanıma oranı % 66.67 olup bu tanıma oranında klasik MFCC ye göre 2.38 puan azalmaya karşılık gelmektedir (çizelge 3.53). Formant frekanslarından temel frekansa nazaran daha düşük tanıma oranı elde edilmesi, NTIMIT veritabanında formant frekanslarının yerinin bozulmalara uğramış olmasından kaynaklanmaktadır. Ayrıca düşük değerli formant frekansları (F_1, F_2) daha çok sözcüğe bağlı olarak değişmekte bu nedenle tanıma başarımını arttırmamaktadır (Rabiner ve Juang 1993). MFCC vektörlerine logaritmik enerji eklenmesi tanıma oranını 1.19 puan, teager enerji eklenmesi tanıma oranını 2.38 puan azaltmaktadır (çizelge 3.54).

Formant GM-FM parametrelerinde, teager enerji ile ayrıklaştırma yapılması, konuşmacı tanıma başarımını 1.28 puan arttırmaktadır (şekil 3.93). Ayrıca AEA-1 ve 2 kullanılarak bir çerçeve boyunca anlık genlik ve frekans değişimleri izlenebilmektedir. Ayrıklaştırma algoritmaları ile birlikte işaretin özilinti genlik zarfı alınması sonucu elde edilen polinom benzetimi katsayıları konuşmacı tanımda iyi sonuçlar vermemektedir. Ancak bu katsayıların değişimine bağlı olarak insan ses yolu davranış bozuklukları tespitinde kullanılmaktadır.

4.2 Tartışma

Bu bölümde, tezde elde edilen sonuçlar daha önce yapılan çalışmalar ile karşılaştırmalı olarak verilmektedir. Tüm çalışmalarda GKM kullanılmaktadır. Eğitim için BM algoritması kullanılmakta, test aşamasında ise maksimum olasılıklı konuşmacı tanınmaktadır.

TIMIT veritabanında 168 kişiden oluşan test dizini ile yapılan konuşmacı tanıma başarımları literatürde verilen başarımlar ile karşılaştırılacaktır. Bu tezde TIMIT veritabanı ile yapılan çalışmalarda kepstrum katsayılarının hazırlanmasında çizelge 4.1'de verilen parametreler kullanılmaktadır. Frekans ölçeği olarak doğrusal frekans ölçeği kullanılmıştır. Eğitim aşamasında çizelge 4.2'de verilen parametreler kullanılmaktadır. Her bir konuşmacı 8 cümle yaklaşık 24 sn ile eğitilmekte, 3 sn

uzunluğunda yaklaşık 1 cümle ile test edilmektedir. Elde edilen sonuçlar karşılaştırmalı olarak çizelge 4.3’de görülmektedir.

Çizelge 4.3 TIMIT veritabanında literatür karşılaştırması

	Konuşmacı tanıma oranı (%)		
	TIMIT	TIMIT+ Bant sınırlama	TIMIT + $f_s = 8kHz$
Bu tezde	100	98.81	94.94
Literatürdeki çalışmalar	99.7 ¹	95.2 ²	86.9 ³

¹Reynolds (1996), ²Reynolds ve ark. (1995), ³Grassi ve ark. (2002)

Çizelge 4.3’den görüleceği üzere bu tezde, literatürdeki çalışmalara göre daha iyi sonuçlar elde edilmiştir. TIMIT veritabanı ile yapılan deneyde frekans ölçeği olarak doğrusal veya ERB ölçeği kullanılması % 100 tanıma oranını vermektedir. Reynolds (1996), Mel frekans ölçeği kullanmaktadır.

TIMIT veritabanına 0-4 kHz bant sınırlama uygulanması durumunda ERB frekans ölçeği için % 98.81 tanıma oranı elde edilmiştir (çizelge 3.30). Örnekleme frekansının düşürülmesi durumunda da tanıma oranı diğer çalışmalardaki sonuçlara göre 8 puan daha iyidir. Sonuç olarak TIMIT veritabanında süzgeçlerin ERB ve doğrusal frekans ölçeğine göre yerleştirilmesi, Mel ölçeğine göre daha etkili olmaktadır.

NTIMIT veritabanı ile hem literatürde hemde bu tezde konuşmacı tanımada elde edilen sonuçlar karşılaştırılacaktır. Veritabanının 168 kişiden oluşan test dizini çizelge 4.1 ve çizelge 4.2’de belirtilen parametrelere bağlı olarak eğitim ve test işlemine tabii tutulmuştur. Her bir konuşmacı 8 cümle (~24 sn) ile eğitilmekte, 1 cümle ile (~3 sn) test edilmektedir. Elde edilen sonuçlar karşılaştırmalı olarak çizelge 4.4’de görülmektedir.

Çizelge 4.4 NTIMIT veritabanında literatür karşılaştırması

	Tanıma oranı
Bu tezde	73.51
Reynolds ve ark. (1995)	69
Reynolds (1996)	76.2

Çizelge 4.4 ‘de görüleceği üzere NTIMIT veritabanı için en yüksek tanıma oranı Reynolds’a (1996) aittir. Bu tezde elde edilen en yüksek sonuç % 73.51’dir.

Öznitelik vektörü elde edilirken kümeleme ile ağırlıklandırma kullanımı Kinnunen (2002), tarafından yapılmıştır. Kinnunen, (2002) şekil 3.47’de verilen öznitelik vektörü elde edilirken, kümeleme için genelleştirilmiş fonem modeli, çerçevelere ait verilerin kümelere atanmasında minimum uzaklık, süzgeç ağırlıklandırma için denklem 3.65, konuşmacıların eğitim ve test aşamalarında vektör nicemleme algoritması kullanmıştır. TIMIT veritabanı için elde ettiği sonuçlar çizelge 4.5’de görülmektedir.

Çizelge 4.5 Küme sayısına bağlı olarak tanıma oranları

Küme sayısı	Tanıma oranı (%)
4	69.37
8	74.85
16	67.07
32	58.49
64	55.73

Eğitim 15 sn, kepstrum katsayı sayısı 20, örnekleme frekansı 8 kHz
Konuşmacı sayısı 100, kod kitabı uzunluğu 64, TIMIT veritabanı

Bu tezde, şekil 3.59’de verilen öznitelik vektörü elde edilirken, kümeleme için beklentinin maksimumlaştırılması algoritması, çerçevelere ait verilerin kümelere atanmasında denklem 3.58’de önerilen maksimum olasılık, konuşmacıların eğitim ve test aşamalarında GKM kullanılmıştır.

Çizelge 4.6’da süzgeçlerin ağırlıklandırılması için denklem 3.65 ve önerdiğimiz denklem 3.66 kullanılarak elde edilen sonuçlar görülmektedir (çizelge 3.38 ve 3.41).

Çizelge 4.6 Küme sayısına bağlı olarak tanıma oranları (%)

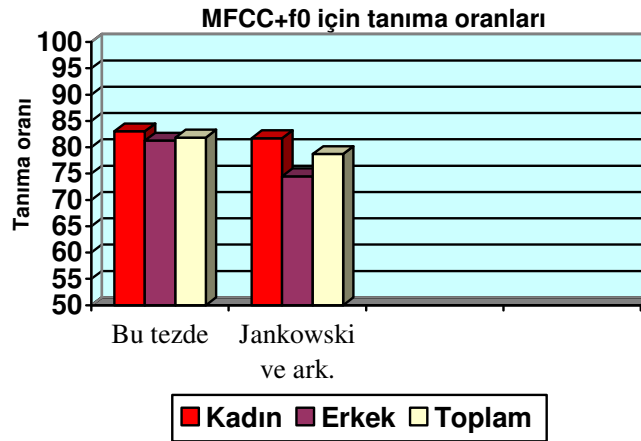
Küme sayısı	TIMIT ¹ veritabanı		NTIMIT ² veritabanı	
	Denklem 3.50	Denklem 3.51	Denklem 3.50	Denklem 3.51
2	86	72	63	70
4	84	85	65	65
8	79	83	61	73
16	73	88	61	65
32	68	86	57	66

¹Eğitim 15 sn, kepstrum katsayı sayısı 24, örnekleme frekansı 8 kHz, Konuşmacı sayısı 100, Karışım bileşen sayısı 32

²Eğitim 24 sn, kepstrum katsayı sayısı 20, örnekleme frekansı 16 kHz, Konuşmacı sayısı 100, Karışım bileşen sayısı 32

Öznitelik vektörlerinin elde edilirken kümeleme ile ağırlıklandırma uygulanmasında önerdiğimiz yöntem, Kinnunen'in (2002) yöntemine kıyasla daha iyi sonuçlar vermektedir. Kinnunen, (2002) kümeleme ile ağırlıklandırmayı sadece TIMIT veritabanında denemiştir. Ayrıca önerdiğimiz denklem 3.66, denklem 3.65'e nazaran konuşmacı tanıma oranını TIMIT veritabanında 2 puan, NTIMIT veritabanında ise 8 puan arttırmaktadır.

Konuşmacı tanıma için literatürdeki temel frekans kullanımını incelendiğinde en iyi sonuç (%78.7) Jankowski ve ark (1995) aittir. Bu çalışmada, dakikadaki perde frekansı periyodundaki değişimler, formantları 1500 Hz altı ve üstü şeklinde ayrılıp bu formantların bant geçiren süzgeçten geçirilip teager enerji operatörü uygulanması ile kestirilmektedir. Aynı eğitim ve test şartlarında, MFCC ye f_0 eklenmesi sonucu elde edilen konuşmacı tanıma oranları, karşılaştırmalı olarak şekil 4.2'de görülmektedir.

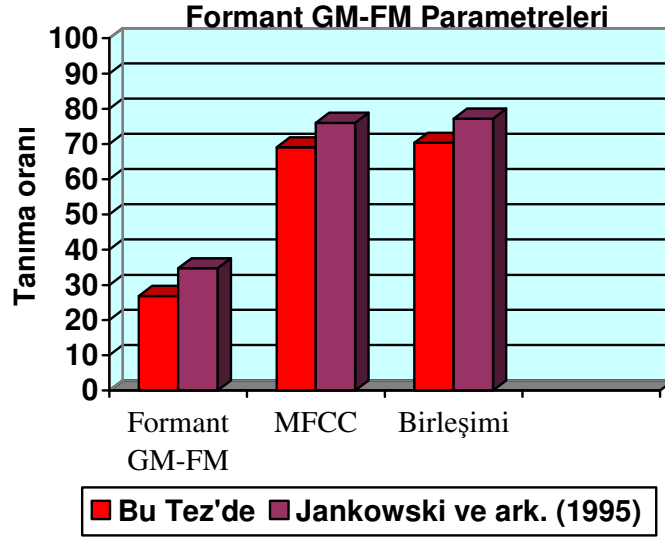


Şekil 4.2 MFCC ve f_0 'ın birlikte kullanıldığında konuşmacı tanıma oranları

MFCC vektörüne temel frekans eklenmesi ile tanıma oranı Jankowski ve ark. (1995) kullandığı yöntemde % 78.7 olurken, bu tezde tanıma oranı % 81.85'e çıkmaktadır (çizelge 3.45). Bu sonuç kullanılan temel frekansın elde edilmesinde kullanılan yöntem farklılıklarından kaynaklanmaktadır.

Formant GM-FM parametrelerinin öznitelik vektörü olarak kullanılması ile elde edilen konuşmacı tanıma oranları şekil 4.3'de görülmektedir. Jankowski ve ark. (1995), blok diyagramı şekil 3.86'da verilen, formant GM-FM parametrelerini kullanılarak her bir formant frekansı için 20 adet öznitelik vektörü oluşturmuşlardır. 168 kişiden oluşan NTIMIT veritabanı test dizini kullanarak mel kepstrum katsayılarına formant GM-FM parametreleri eklenip eğitim ve test işlemine uygulanmıştır. Aynı şartlarda formant GM-

FM parametreleri karşılaştırıldığında toplamda Jankowski ve ark. (1995) tanıma oranı % 77.2 bu tezde ise % 70.33 elde edilmiştir.



Şekil 4.3 Formant GM-FM parametrelerinin tanıma oranlarının karşılaştırılması

Jankowski ve ark. (1995), sadece MFCC katsayılarını öznitelik vektörü olarak kullandığında tanıma oranını % 76.2 olarak bulmuş, bu katsayılara formant GM-FM parametreleri eklendiğinde ise konuşmacı tanıma oranı 1.2 puan artış göstermiştir. Bu tezde, MFCC vektörlerinin yalnız kullanılması durumunda tanıma oranı % 69.05, bu katsayılara formant GM-FM parametreleri eklendiğinde ise tanıma oranı % 70.33 olmaktadır. Bu farklılık, Jankowski ve ark. öznitelik vektörlerini elde ederken konuşmadaki sesli harfleri ayırmasından kaynaklanmaktadır.

4.3 Öneriler

İnsan kulağının en iyi modellenmesinin ERB ölçeğinde gamaton süzgeçler ile olduğu bilinmektedir (Slaney 1993). Kepstrum katsayı elde edilmesinde gamaton süzgeçler iyi sonuç vermemektedir (çizelge 3.33). Öznitelik vektörlerinden tanımayı arttırıcı olanları bir araya getirilip tanıma üzerine etkisi olmayan öznitelikler çeşitli budama algoritmaları ile azaltılabilir.

Düşük seviye ipuçları, çeşitli öznitelikler olarak günümüzde konuşmacı tanıma sistemlerinde kullanılmaktadır. Yüksek seviye ipuçları; kelime kullanım şekli, hece ve cümle parçası süreleri ve cümle içinde durma süreleri ve sıklıkları şeklindeki

ölçümlerdir. Bu ölçümlerin konuşmacı tanıma katkısı incelenebilir. Yüksek seviye ipuçları, sadece tanıma sisteminin başarımını arttırmakla kalmayıp aynı zamanda ses iletim ortamı etkilerine karşı daha gürbüz olmasını sağlayabilir.

Konuşmacı tanıma gerçek şartlarda uygulanabilmelidir. Konuşmalara ait ses örnekleri, farklı mikrofon telefon veya cep telefonu ahizesi, farklı ses iletim ortamları ve farklı akustik çevrelerden toplanmalıdır. Gerçek dünyada oluşabilecek çeşitli gürültülere karşı sistemin dayanıklı hale getirilmesi gerekir. Bu şartlarda yeni veya var olan kompanzasyon teknikleri ve bürünsel yöntemler kullanılarak gerçek dünyada konuşmacı tanıma başarımı artırılabilir.

Destek vektör makinesi (DVM) son yıllarda popüler olan bir metot olup konuşmacı tanıma iyi başarımlar göstermektedir (Campbell 2003). DVM sınıflandırıcı olarak çalışma mantığı bölüm 2.4.4 de tanımlanmıştır. Bu yöntem, ses örneklerinde hatayı minimumlaştırarak en iyi genellemeyi üreten sonucu elde etmeyi amaçlar. Düzgün dağılmayan veriler için örnekler çekirdek olarak bilinen matematiksel fonksiyonlar kullanılarak başka bir uzaya taşınır. DVM, kullandığımız konuşmacı tanıma sisteminde Bayes karar kuralı yerine kullanılabilir. GKM, destek vektör makinesi ile birlikte karma olarak kullanılabilir.

Ses verisinde kişiye ait bilgilerin çoğunlukla taşındığı kısımların sesli harfler olduğu bilinmektedir (Sambur 1975). Uygun bir algoritma ile bu ses verileri seçilerek konuşmacı tanıma kullanılabilir.

KAYNAKLAR

Aliaa, A. Y., A. S. Ebada, W. H. El Behaidy. 2004. Development of Automatic Speaker Identification System. 21st National Radio Science Conference.

Alexandre, P., P. Lockwood, 1993. Root Cepstral Analysis: A Unified View Application to Speech Processing in Car Noise Environments. *Speech Communication*, Vol.12, p. 277-288.

Antal, M. 2004. A Comparison of Parametric Clustering Techniques used in Speaker Identification. *IJSIT Lecture Note of International Conference on Intelligent Knowledge Systems*, Vol. 1, No. 1. p. 19-25.

Arcienega, M., A. Drygajlo. 2001. Pitch-dependent GMMs for Text-Independent Speaker Recognition Systems. *Eurospeech'01, Scandinavia*, p. 2821-2824.

Atal, B. 1974. Effectiveness of Linear Prediction Characteristics of the Speech wave for Automatic Speaker Identification and Verification. *Journal of the Acoustical Society of America*, vol. 55, p. 1304-1312.

Aydın, Ö. 2005. Yapay Sinir Ağlarını Kullanarak Bir Ses Tanıma Sistemi Geliştirilmesi. Master Thesis Trakya University Graduate School of Natural and Applied Sciences Department of Computer Engineering, s. 20-24

Bhattacharyya, S. T. Srikanthan, P. Krishnamurthy, 2001. Ideal GMM. parameters & Posterior Log Likelihood for Speaker Verification, *Proceedings of the IEEE Signal Processing Society Workshop, USA*. ISBN: 0-7803-7196-8, p. 471-480.

Bennani, Y. Gallinari, P. 1991. On the use of TDNN-extracted Features Information in Talker Identification. in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, p. 385-388.

Besacier, L. J.F. Bonastre, 1998. Frame Pruning for Automatic Speaker Verification, *Proc. EUSIPCO'98, Greece, September 8-11, Vol.1*, p. 367-370.

Broad, D. J. 1972. Formants in Automatic Speech Recognition. *Int. J. Man-Machine Studies* (4) p. 411-412.

Campbell, J. P. and A. D. Reynolds. 1999. Corpora for the Evaluation of Speaker Recognition Systems. *IEEE Trans. Speech Audio Processing*, p. 829-832.

Cardinaux, F., C. Sanderson, S. Marcel, 2003. Comparison of MLP and GMM Classifiers for Face Verification on XM2VTS, *IDIAP-RR* p. 3-10.

Christopher J. C. Burges. 1998. A Tutorial on Support Vector Machines for Pattern Recognition". *Data Mining and Knowledge Discovery* 2:121 – 167.

- Cristianini N., J. Shawe-Taylor. 2000. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press.
- Chu, K. K., S. H. Leung and C. S. Yip, 2003. Perceptually non-uniform spectral compression for noisy speech recognition, Proc. ICASSP 2003, p. 404-407.
- Claudio, B. and L. P. Ricotti. 1999. Speech Recognition Theory and C++ Implementation. John WILEY&Sons, Ltd, p. 125-137.
- Deng, J. and Q. Hu. 2003. Open Set Text-Independent Speaker Recognition Based on Set-Score Pattern Classification. ICASSP, p. 73-77.
- Davis, S. B. and P. Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-28, p. 389-397.
- Duman, F. O. Eroğul., Z. Telatar., S. Yetkin. 2005. Uyku İçciklerinin Kısa ve Uzun Dönemli Karma Analizi. SIU, Kayseri.
- Drucker, H. Chris J.C. Burges, L. Kaufman, A. Smola, V. Vapnik, 1997. Support Vector Regression Machines. Advances in Neural Information Processing Systems 9, NIPS 1996, 155-161, MIT Press.
- Eatoock, J. P. and J. S. Mason. 1994. A Quantitative Assessment of the Relative Speaker Discriminating Properties of Phonemes. Proc. ICASSP'94, p. 133-136.
- Ertaş, F. 2000. Fundamantels of Speaker Recognition. Journal of Engineering Sciences, No. 2-3, Pamukkale, p. 185-193.
- Ertaş, F. 2001. Feature Selection and Classification Techniques for Recognition. Journal of Engineering Sciences, No. 1, Pamukkale, p. 47-54.
- Ertaş, F. ve Ö. Eskidere. 2001. Yazılım Tabanlı Sözcük Sentezleyici. DEÜ Mühendislik Fakültesi Fen ve Mühendislik Dergisi, Cilt:3, Sayı:1.
- Ertaş, F. 2002. Ses İşaretlerine Karşı Basilar Membran Hareketinin Yazılım Benzetimi. S.D.Ü. Fen Bilimleri Dergisi 6:1, s. 86-93.
- Fant, G. 1949. Analys av de svenska konsonantljuden. L.M. Ericsson protokol H/P 1064. p. 139.
- Fant, G. 1960. Acoustic Theory of Speech Production. Mouton & Co., The Hauge.
- Feder, M. Weinstein and A. Oppenheim. 1988. A New Class Of Sequential and Adaptive Algorithms with Application to Noise Cancellation. in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, p. 557-560.
- Furui, S. 1989. Digital Speech Processing, Synthesis, and Recognition. M. Dekker Inc.

- Fujimoto M., Y. Ariki, 2004. Robust Speech Recognition in Additive and Channel Noise Environments Using GMM and EM Algorithm.' Proc. ICASSP'04, Vol.I, p. 941-944.
- Ganchev, T. 2005. Speaker Recognition , PhD thesis, Dept. of Electrical and Computer Engineering, University of Patras, Greece. p. 61-82.
- Gish, H., M. Schmidt. 1994. Text Independent Speaker Identification. IEEE Signal Proc. Mag. Vol.11, No.4, p. 18-32.
- Glasberg, B. R., Moore, B. C. J. 1990. Derivation of auditory filter shapes from notched noise data, Hearing Research, Vol. 47. no. 1-2, p. 103-138.
- Grassi S., M. Ansorge, F. Pellandini, and P.-A. Farine 2002. Distributed speaker recognition using the ETSI AURORA standard, Proc. 3rd COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication, p.120-125.
- Hamila, R., J. Astola., F. A. Cheikh., M. Gabbouj. and M. Renfors. 1999. Teager Energy and the Ambiguity Function. IEEE Transactions on Signal Processing, Vol. 47, no. 1. p. 260-261.
- Hansen, J.H.L., L. Gavidia-Ceballos and J.F. Kaiser. 1998. A Nonlinear Based Speech Feature Analysis Method with Application to Vocal Fold Pathology Assessment. IEEE Transactions on Biomedical Engineering, vol. 45, no. 3, p. 300-313.
- Hermansky, H. 1990. Perceptual Linear Predictive (PLP) Analysis of Speech. Journal of Acoustic Society of America, vol. 87, no. 4, p. 1738-1752.
- Hermansky, H. and N. Morgan. 1994. RASTA Processing of Speech. IEEE Transactions on Speech and Audio Processing, vol. 2, no. 4, p. 578-589.
- Holdsworth, J., I. Nimmo-Smith., R. Patterson and Rice, P. 1988. Implementing a Gammatone filter bank", Draft manuscript.
- Huang, X., Acero, A., Hon, H.-W., 2001. Spoken Language Processing: a Guide to Theory, Algorithm, and System Development. Prentice-Hall, New Jersey.
- Hunt, M. J., 1999. Spectral Signal Processing for ASR. IEEE ASRU Workshop, Colorado, Keystone, U.S.A.
- Jankowski, C. R., T. F. Quatieri., D. A. Reynolds. 1994. Formant AM-FM for Speaker Identification. IEEE Transactions on Speech and Audio Processing, p. 608-611.
- Jankowski, C. R., T. F. Quatieri., D. A. Reynolds. 1995. Measuring Fine Structure in Speech: Application to Speaker Identification. IEEE Transactions on Speech and Audio Processing, p. 325-328.

- Julius, O., S. Jonathan., S. Abel. 1999. Bark and ERB Bilinear Transforms. IEEE Trans. Speech Audio Processing, Vol. 7, No. 6.
- Karpov, E. 2003. Real-Time Speaker Identification, Master thesis, University of Joensuu, Department of Computer Science p. 17-26.
- Kasi, K. 2002. Yet Another Algorithm for Pitch Tracking, Master thesis, Old Dominion University, p. 9-13.
- Kinnunen, Tomi. 2002. Designing A Speaker-Discriminative Adaptive Filter Bank for Speaker Recognition. in Proc. Int. Conf. on Spoken Language Processing ICSLP, Colorado, USA, p. 2325–2328.
- Kinnunen, T. 2003. Spectral Features for Automatic Text-independent Speaker Recognition, Ph.D. thesis, University of Joensuu, Department of Computer Science p. 49-115.
- Koeing, W. 1949. A new frequency scale for Acoustic Measurements. Bell telephone Laboratory Record, Vol. 27. p. 299-301.
- Krauss P., I. Shure., J.N. Little. 1996. MATLAB Signal Processing Toolbox User's Guide, The Mathworks Inc.
- Krishnakumar, S., Kumar, K. P., and Balakrishnan, N. 2003. Pitch maxima for robust speaker recognition. In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, vol. 2, p. 201-204.
- Lim, J. S. 1979. Spectral Root Homomorphic Deconvolution system, IEEE Trans. on ASSP, Vol. ASSP-27, No. 3.
- Linde, Y., A. Buzo., R. M. Gray. 1980. An Algorithm for Vector Quantization, IEEE Trans. Communications, Vol. 28, No. 1, p. 84-95.
- Liu, Li., J. He., G. Palm. 1996. Signal Modeling for Speaker Recognition. IEEE Trans. Speech Audio Processing, p. 665-668.
- Lincoln, M. 1999. Characterization of Speakers for Improved Automatic Speech Recognition. Thesis Doctor of Philosophy in the School of Information Systems, University of East Anglia, Norwich. p. 18-23.
- Lippmann, R. P. 1987. An introduction to computing with neural nets. IEEE ASSP Magazine, vol. 4. p. 4-22.
- Maragos, P., J. F. Kaiser. and T. F. Quatieri. 1993. Energy separation in signal modulations with application to speech analysis. IEEE Trans. Signal Processing, vol. 41, p. 3024–3051.

- Matsui, T. and S. Furui. 1995. Speaker Recognition Technology. NNT Review, Vol. 7, No. 2, p. 40-48.
- McLachlan, G. 1988. Mixture Models. New York, NY: Marcel Dekker, 1 st. ed.
- Mengüşoğlu, E. 1999. Bir Türkçe Sesli İfade Tanıma Sisteminin Kural Tabanlı Tasarımı ve Gerçekleştirimi. Hacettepe Üniversitesi Yüksek Lisans Tezi, s. 46-47.
- Mokhtari, P. 1998. An Acoustic-Phonetic and Articulatory Study of Speech-Speaker Dichotomy. PhD thesis, School of Computer Science, University of New South Wales, Australia.
- Moore, B. C. J. and B. R. Glasberg. 1983. Suggested Formula for Calculating Auditory Filter Bandwidths and Excitation Patterns. J. Acoust. Soc. Am., 74, p. 750-753.
- Moore, B. C. J. 2003. An Introduction to the psychology of hearing. Academic Press, London, 5th Ed.
- Morris, A., W. Dalei and J. Koreman. 2005. GMM Based Clustering and Speaker Separability in the Timit Speech Database. IEICE Trans. Fundamentals Commun. Electron. Inf. & Syst., Vol. E85-A/B/C/D, No. 1.
- Naik, J., L. Netsch. and G. Doddington. 1989. Speaker Verification Over Long Distance Telephone Lines. in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, p. 524-527.
- Naik, J. 1990. Speaker verification: A tutorial. IEEE Comm. Mag, Vol. 28, No. 1, p. 42-48.
- Oglesby, J. Ve J. S. Mason 1990. Optimisation of neural models for speaker identification. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, p. 261-264.
- O'shaughnessy, D. 1987. Speech Communication Human and Machine. Addison Wesley, New York.
- Orman, D., L. Arslan, 2001. Frequency analysis of speaker Identification. In Proc. Speaker Odyssey: the Speaker Recognition Workshop Greece, p. 219-222.
- Park, A. 2002. ASR Dependent Techniques for Speaker Recognition. Master of Engineering in Electrical Engineering and Computer Science at the Massachusetts Institute Of Technology, USA, p. 65-66.
- Papoulis, A. 1984. Probability Random Variables and Stochastic Processes. NY: McGraw-Hill, 2 ed. New York.
- Paliwal, K. K. 1992. Dimensionality Reduction of the Enhanced Feature Set for the HMM-Based Speech Recognizer. Digital Signal processing 2, p. 157-173.

Peskin, Barbara et al. 2003. Using Prosodic And Conversational Features for High-Performane Speaker Recognition. Report from JHU WS'02", IEEE Trans. Speech Audio Processing, p. 792-796.

Peskin, B., A. Adami., Q. Jin., D. Klusácek., J. S. Abramson., R. Mihaescu., J. J. Godfrey, D. A. Jones and B. Xiang. 2003. TheSuper SID Project: Exploiting High-level Information for High-accuracy Speaker Recognition. International Conference on Acoustics, Speech, and Signal Processing IEEE, Hong Kong, p. 784-787.

Picone, J. 1996. Fundamentals of Speech Recognition: a Short Course. Institute for Signal and Information Processing, p. 68-69.

Plumpe, M. D., T. F. Quatieri. and D. A. Reynolds. 1999. Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification. IEEE Tansactions on Speech and Audio Processing, vol. 7, no. 5.

Rabiner, L. R., B.H. Juang and S. E. Levinson. 1985. Some Properties of CDHMM Representations. AT&T technical Journal.

Rabiner, L. R. 1989. A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedins of IEEE, vol. 77, no. 2, p. 257-286.

Rabiner, L. R. and B. H. Juang. 1993. Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs.

Reynolds, D.A. 1992. A Gaussian Mixture Modeling Approach to Text Independent Speaker Identification. Ph.D. thesis, Georgia Inst. of Technology.

Reynolds, D. A., M. A. Zissman., T. F. Quatieri., G. C. O'Leary. and B.A. Carlson. 1995. The effects of Telephone Transmission Degrations on Speaker Recognition Performance. ICASSP, Detroit, MI. p. 329-332.

Reynolds, D. A., R. C. Rose. 1995. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Trans. Speech Audio Processing, 3. p. 72-83.

Reynolds, D.A. 1996. Speaker Identification and Verification using Gaussian Mixture Speaker Models. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, p. 27-30.

Reynolds, D. A. T. F., Quatieri, R. B. Dunn, 2000. Speaker verification using Adapted Gaussian Mixture Models, Digital Signal Processing, vol. 10, p. 19-41.

Reynolds, D.A. 2002. An Overview of Automatic Speaker Recognition Technology ICASSP, p. 4072-4076.

Reynolds D.A., et al. 2003. The SuperSID Project: Exploiting High-Level Information for High-Accuracy Speaker Recognition. in Proc. ICASS, p. 784-787.

- Reynolds, D.A., J. Campbell., B. Campbell., B. Dunn., T. Gleason., D. Jones., T. Quatieri., C. Quillen., D. Sturim., P. T. Carrasquillo. 2004. Beyond Cepstra: Exploiting High-Level Information in Speaker Recognition. Super SID Project Final Report, p. 223-229.
- Robinson, T. and F. Fallside. 1991. A Recurrent Error Propagation Network Speech Recognition System. *Computer Speech and Language*, vol 5, no. 3, p. 259-274.
- Rose, P. 2001. *Forensic Speaker Identification*, Taylor & Francis Forensic Science Series, ISBN 0-415-27182-7, p. 225-280.
- Rudasi, L. and S. A. Zahorian. 1991. Text Independent Talker Identification with Neural Networks. in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, p. 389-392.
- Sambur, M. 1975. Selection of Acoustical features for Speaker Identification. *IEEE Transactions on Acoustics, Speech and Signal Processing.*, Vol. ASSP-23, p. 176-182.
- Sanderson, C. 2002. *Automatic Person Verification using Speech and Face Information*, Ph.D thesis, Griffith University, p. 17-33.
- Sağıroğlu, Ş. 2001. *Yapay Sinir Ağı ve Uygulamaları Ders Notları*, Erciyes Üniversitesi, Kayseri.
- Sarikaya, R., J. H. L. Hansen, 2001. Analysis of the Root-Cepstrum for Acoustic Modeling and Fast Decoding in Speech Recognition, *Eurospeech-2001*, Denmark p. 2-4.
- Sarma, S. 1997. *A Segment-based Speaker Verification System* S.M. thesis, MIT Department of Electrical Engineering and Computer Science, p. 84-86.
- Sezer, O. G., A. Ercil., M. Keskinöz. 2005. Independent Component Based 3D Object Recognition using Support Vector Machines. *IEEE Signal Processing and Communications Applications Conference*, p. 99-102.
- Slaney, M. 1993. An efficient implementation of the Patterson-Holdsworth auditory filter bank. *Tech. Rep. 35*, Apple Computer, Inc.
- Slaney, M. 1998. *Auditory Toolbox: A MATLAB Toolbox for Auditory Modeling* Work Technical Report, Interval Research Corporation, p. 29-32.
- Shannon, B., J. Kuldip. and K. K. Paliwal. 2003. A Comparative Study of Filter Bank Spacing for Speech Recognition. *Microelectronic Engineering Research Conference*.
- Skowronski, M.D. Haris, J.G. 2004. Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition. *Journal of the Acoustical Society of America*. Vol. 116., No. 3, p. 1774-1780.

- Stevens, s. And J. Volkman. 1940. The Relation of Pitch to Frequency. American Journal of Psychology, vol. 53, p. 329.
- Soong, F. And A. Rosebberg. 1988. On the use of instantaneous and transitional spectral information in speaker recognition, ASSP, vol. ASSP-36, p. 871-879.
- Tancerel, L., R. Vesa, V.T. Ruoppila, R. Lefebvre 2000. Combined Speech and Audio Coding by Discrimination. Proc. IEEE Workshop on Speech Coding, p. 154-156.
- Tierney, J. 1980. A Study of LPC Analysis of Speech in Additive Noise. IEEE ASSP-28, p. 389-397.
- Teager, H. M. and S. M. Teager. 1989. Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract. in Speech Production and Speech Modelling, W.J. Hardcastle and A. Marchal, Eds., NATO Advanced Study Institute Series D, Vol. 55, Bonas, France.
- Tou, J. and R. Gonzalez. 1974. Pattern recognition Principles. Reading, Mass: Addison-Wesley Publishing Company.
- Tyagi, V., C. Wellekens, 2005. On Desensitizing the Mel-Cepstrum to Spurious Spectral Components for Robust Speech Recognition, in Acoustics, Speech, and Signal Processing, Proceedings, IEEE International Conference on, vol. 1, p. 529–532.
- Umesh, S., L. Cohen, D. Nelson. 1999. Fitting the Mel Scale. IEEE Transactions on Acoustics, Speech and Signal Processing., p. 217-220.
- Wildermoth, B. R. 2001. Text Independent Speaker Recognition Using Source Based Features. Master of Philosophy, Griffith University, Australia, p. 21-29.
- Woodward, J. D. 1997. Biometrics: Privacy's foe or Privacy's Friend. Proc. IEEE, Vol. 85, No. 9, p.1480-1492.
- Wolf, J. 1972. Efficient Acoustic Parameters for Speaker REcognition. Journal of the Acoustical Society of America, vol. 51, no. 6, p. 2044-2056.
- Quatieri, T. F., H. E. Thomas. and G. C. O'Leary. 1997. AM-FM Separation Using Auditory-Motivated Filters. IEEE Transactions on Speech and Audio Processing, vol. 5, no. 5. p. 465-480
- Quatieri T. F., Douglas A. Reynolds. and G. C. O'Leary. 2000. Estimation of Handset Nonlinearity with Application to Speaker Recognition. IEEE Transactions on Speech and Audio Processing, Vol. 8, no. 5. p. 567-585.
- Yapanel, Ü. 1997. Garbage Modeling Techniques for a Turkish Keyword Spotting System. MSc. Thesis, Boğaziçi University.

Zhu, Q., A. Alwan, 2000. On the use of variable frame rate in speech recognition. In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP 2000, Turkey, vol. 3, p. 1783–1786.

Zhou, G., J. Hansen. and J. F. Kaiser. 2001. A Nonlinear Feature Based Classification of Speech Under Stress. IEEE Transactions on Speech and audio Processing, vol. 9, no. 3. p. 300-313.

EK 1 TERİMLER SÖZLÜĞÜ

Apex: Basilar membranın orta kulaktan uzak ucu

Basilar Membran: İç kulaktaki sıvımsı yapı

Base: Basilar membranın orta kulağa yakın ucu

Cochlea: Kulak salyangozu

Diagonal: Köşegen

Frame: Çerçeve

F-ratio: F-oranı

Formant frequency: Ses tellerinden geçen işaretin ses yolunda meydana getirdiği rezonansların frekansı

Fundamental frequency: Temel frekans

Glottal : Gırtlak

Glottis : Gırtlak dili

Identification: Tanıma, kimliklendirme, tanıma

Iteration: Özyineleme

Kovaryans: Ortak değişinti matrisi

Mixture: Karışım

Phoneme : Ses birimi, konuşmadaki temel sesler

Phonetic: Sesçil

Pitch frequency: Perde frekansı, algılanan temel frekans, ses tellerinin saniyedeki titreşim sayısı

Prosodic Features: Bürünsel öznitelikler, konuşmadaki şiddet, temel frekans değişimleri ve zamanlama özellikleri

Singularity: Tekillikler

Scala tympani: Kohleanın altındaki spiral yapı

Perilymph: Basilar membranda kemiksi yapının çevresindeki boşlular

Spektrum: Bir sinyal içindeki tüm frekansların gösterimi

Unvoiced : Nefessiz sesler ses telleri titreştirilmeden üretilir.

Voiced : Nefesli sesler, ses tellerinin titreşimi ile üretilen sesler

EK 2. GKM PARAMETRE KESTİRİMİ

GKM parametrelerinin maksimum benzerlik denklemlerinin kestirimi için Baum'un yaklaşım fonksiyonu kullanılmaktadır (Baum ve ark 1970). Bu parametre kestirim denklemleri, özyinelemeli parametre kestirim prosedürünün temelini oluşturmaktadır.

GKM konuşmacı modeli, konuşmacıların özelliklerinin temsil edildiği M akustik sınıf içerir ve $X = \{\bar{x}_1, \dots, \bar{x}_T\}$ gözlem vektörleri dizisidir. $I = \{i_1, \dots, i_T\}$, ($i_t \in [1, M]$) X tarafından üretilen akustik sınıfların belirli bir dizisi veya durumudur. X e bağlı olarak benzerlik fonksiyonu denklem E.1 deki gibi yazılabilir.

$$p(X|\lambda) = \sum_I p(X, I|\lambda) \quad (\text{E.1})$$

Burada;

$$p(X, I|\lambda) = \prod_{t=1}^T p_{i_t} b_{i_t}(\bar{x}_t) \quad (\text{E.2})$$

X ve I nin ek pdf'i ve \sum_I notasyonu, dizilerdeki tüm mümkün durumların toplamını göstermektedir. X ve I nin ek pdf'inin temel şekli gözlemlerin ve akustik sınıfların birbirinden bağımsız olmasının bir sonucudur. λ ile verilen modelin benzerlik fonksiyonunu arttırmak için yeni model parametreleri ($\bar{\lambda}$) bulunur. Bulunan yeni model λ modeline bağlı olarak denklem E.3'ü sağlamalıdır.

$$p(X|\bar{\lambda}) \geq p(X|\lambda) \quad (\text{E.3})$$

Bu maksimumlaştırma işlemi Baum'un (1970) yaklaşım fonksiyonu metodu ile elde edilir. Yaklaşım fonksiyonu, bir ortalamada gözlenen verinin benzerlik fonksiyonunun maksimum olduğu özyineleme artışı sağlar. Bu fonksiyon E.4'deki gibi ifade edilir.

$$\Theta(\lambda, \bar{\lambda}) = \sum_I p(X, I|\lambda) \log p(X, I|\bar{\lambda}) \quad (\text{E.4})$$

$\bar{\lambda}$ modeli, $\Theta(\lambda, \bar{\lambda})$ maksimum olduğu yerde, $-p(X|\bar{\lambda}) \geq p(X|\lambda)$ gözlenen benzerlik verisinde bir artış sonucunu verir.

Sonraki adımda E.2’de verilen denklem ($\lambda = \bar{\lambda}$) durumunda denklem E.4’de yerine yazılırsa denklem E.5 elde edilir.

$$\Theta(\lambda, \bar{\lambda}) = \sum_I p(X, I | \lambda) \sum_{i=1}^T \log[\bar{p}_i \bar{b}_i(\bar{x}_i)] \quad (\text{E.5})$$

Burada \bar{p}_i yeni karışım ağırlığı ve $\bar{b}_i(\bar{x})$ yeni yoğunluk bileşeni olup yeni ortalama ve ortak değişinti model parametreleri kullanılarak elde edilir. Yeni model parametrelerinin saklı durum değişkenlerine (i_t) bağımlılığından kurtulmak için denklem E.6’da verilen bir fonksiyon tanımlanır.

$$\eta_t(i, I) = \begin{cases} 1 \dots i_t = i \\ 0 \dots \text{diğer} \end{cases} \quad (\text{E.6})$$

Bu fonksiyon denklem E.5’de yerine yazılırsa denklem E.7 elde edilir.

$$\Theta(\lambda, \bar{\lambda}) = \sum_{i=1}^T \sum_I p(X, I | \lambda) \sum_{i=1}^M \log[\bar{p}_i \bar{b}_i(\bar{x}_i)] \eta_t(i, I) \quad (\text{E.7})$$

Bu denklem tekrar düzenlenirse denklem E.8 ve E.9 elde edilir.

$$\Theta(\lambda, \bar{\lambda}) = \sum_{i=1}^T \sum_{i=1}^M \log[\bar{p}_i \bar{b}_i(\bar{x}_i)] \gamma_t(i) \quad (\text{E.8})$$

$$\gamma_t(i) = \sum_I \eta_t(i, I) p(X, I | \lambda) \quad (\text{E.9})$$

Denklem E.9, denklem E.10’daki gibi gösterilebilir.

$$\gamma_t(i) = p(X | \lambda) p(i_t = i | \bar{x}_t, \lambda) \quad (\text{E.10})$$

Burada

$$p(i_t = i | \bar{x}_t, \lambda) = \frac{p_i b_i(\bar{x}_t)}{\sum_{k=1}^M p_k b_k(\bar{x}_t)} \quad (\text{E.11})$$

i. durum bir sonsal olasılığa karşılık gelmektedir.

Benzerlik fonksiyonunun artışıyla denklem E.8 maksimumlaştırılması ile yeni model parametreleri $\bar{\lambda} = \{\bar{p}_i, \bar{\mu}_i, \bar{\Sigma}_i\}$ kestirilir. $\Theta(\lambda, \bar{\lambda})$ parametrelerin fonksiyonu konkav şeklinde olup fonksiyonun kritik değerleri bulunarak maksimumlaştırılır. Yani $\bar{\lambda}$ model parametreleri $\partial\Theta(\lambda, \bar{\lambda})/\partial\bar{\lambda} = 0$ alınarak bulunur.

Karışım ağırlıkları

Karışım ağırlıkları $\Theta(\lambda, \bar{\lambda})$ maksimum olduğu noktada elde edilir. p_i karışım ağırlığı, $\sum_{i=1}^M \bar{p}_i = 1$ aralığında sınırlıdır. Denklem E.12'de karışım ağırlıkları ifadesi verilmektedir.

$$\bar{p}_i = \frac{\sum_{t=1}^T \gamma_t(i)}{\sum_{t=1}^T \sum_{k=1}^M \gamma_t(k)} \quad (\text{E.12})$$

Bu ifade tekrar düzenlenirse denklem E.11'de verilen $p(i_t = i | \bar{x}_t, \lambda)$ ifade kullanılırsa denklem E.13 elde edilmektedir.

$$\bar{p}_i = \frac{1}{T} \sum_{t=1}^T p(i_t = i | \bar{x}_t, \lambda) \quad (\text{E.13})$$

Bileşen yoğunluklarının ortalamaları

Denklem E.8'in gradyentinin alınması ile yoğunluk ortalama vektörü $\bar{\mu}_i$ denklem E.14'deki gibi elde edilir.

$$\begin{aligned} \frac{\partial\Theta(\lambda, \bar{\lambda})}{\partial\bar{\mu}_i} &= \frac{\partial}{\partial\bar{\mu}_i} \sum_{t=1}^T \gamma_t(i) \log \bar{b}_i(\bar{x}_t) \\ &= \sum_{t=1}^T \gamma_t(i) \frac{\partial}{\partial\bar{\mu}_i} \left[-\frac{1}{2} (\bar{x}_t - \bar{\mu}_i)' \Sigma^{-1} (\bar{x}_t - \bar{\mu}_i) \right] \end{aligned} \quad (\text{E.14})$$

Matrisler için bazı farklılaştırma kuralları kullanılırsa $D \times 1$ boyutunda bir \bar{a} vektörü ve $D \times D$ boyutunda bir C matrisi

$$\frac{\partial}{\partial\bar{a}} \bar{a}' C \bar{a} = 2C\bar{a}$$

Kuralı denklem E.14 uygulanıp bu denklem 0 a eşitlenirse $\bar{\mu}_i$ elde edilir.

$$\bar{\mu}_i = \frac{\sum_{t=1}^T \gamma_t(i) \bar{x}_t}{\sum_{t=1}^T \gamma_t(i)} \quad (\text{E.15})$$

Denklem E.10'da belirtilen $\gamma_t(i)$ ifadesi denklem E.15'de yerine yazılırsa denklem E.16 ifadesi elde edilir.

$$\bar{\mu}_i = \frac{\sum_{t=1}^T p(i|T=i|\bar{x}_t, \lambda) \bar{x}_t}{\sum_{t=1}^T p(i=i|\bar{x}_t, \lambda)} \quad (\text{E.16})$$

Bileşen yoğunluklarının ortak değişinti matrisleri

Ortalama vektörü maksimumlaştırılmasında $\Theta(\lambda, \bar{\lambda})$ maksimumlaştırılması ile elde edildiğinden tüm ortak değişinti matrisleri elemanları eş zamanlı olarak denklem E.8'in gradyenti 0'a eşitlenerek $\bar{\Sigma}_i$ matrisi, denklem E.17'den elde edilir.

$$\begin{aligned} \frac{\partial \Theta(\lambda, \bar{\lambda})}{\partial \bar{\Sigma}_i} &= \frac{\partial}{\partial \bar{\Sigma}_i} \sum_{t=1}^T \gamma_t(i) \log \bar{b}_i(\bar{x}_t) \\ &= \sum_{t=1}^T \gamma_t(i) \frac{\partial}{\partial \bar{\Sigma}_i} \left[-\frac{1}{2} (\bar{x}_t - \bar{\mu}_i)' \bar{\Sigma}_i^{-1} (\bar{x}_t - \bar{\mu}_i) - \frac{1}{2} \log |\bar{\Sigma}_i| \right] \end{aligned} \quad (\text{E.17})$$

Tekrar matrisler için bazı farklılaştırma kuralları kullanılırsa $D \times 1$ boyutunda bir \bar{a} vektörü ve $D \times D$ boyutunda bir C matrisi;

$$\frac{\partial}{\partial C} \bar{a}' C^{-1} \bar{a} = -\bar{a} \bar{a}' C^{-1} (C^{-1})' \text{ ve } \frac{\partial}{\partial C} \log |C| = C^{-1}$$

Bu kurallar denklem E.17'ye uygulanırsa ve denklem 0'a eşitlenirse $\bar{\Sigma}_i$ denklem E.18'deki gibi elde edilir.

$$\bar{\Sigma}_i = \frac{\sum_{t=1}^T \gamma_t(i) (\bar{x}_t - \bar{\mu}_i) (\bar{x}_t - \bar{\mu}_i)'}{\sum_{t=1}^T \gamma_t(i)} \quad (\text{E.18})$$

$\gamma_t(i)$ ifadesi denklem E.18'de yerine yazılırsa ortak deęişinti kestirim denklemini E.19'daki gibi ifade edilmektedir.

$$\bar{\Sigma}_i = \frac{\sum_{t=1}^T p(i_t = i | \bar{x}_t, \lambda) \bar{x}_t \bar{x}_t'}{\sum_{t=1}^T p(i_t = i | \bar{x}_t, \lambda)} - \bar{\mu}_i \bar{\mu}_i' \quad (\text{E.19})$$

Köşegen ortak deęişinti matrisleri kullanılması durumunda sadece köşegen elemanlar veya deęişintilerin kestirilmesine gerek vardır. bu durumda denklem E.20 kullanılabilir.

$$\bar{\sigma}^2_i = \frac{\sum_{t=1}^T p(i_t = i | \bar{x}_t, \lambda) \bar{x}_t^2}{\sum_{t=1}^T p(i_t = i | \bar{x}_t, \lambda)} - \bar{\mu}_i^2 \quad (\text{E.20})$$

TEŐEKKÖR

Tezle ilgili arařtırmalarımnda gösterdiđi yardım ve bu alıřmanın geliřtirilmesindeki ynlendirici katkılarından dolayı danıřmanım Yrd. Do. Dr. Figen ERTAŐ' a, kritik noktalardaki yardımından dolayı Prof. Dr. Tuncay ERTAŐ'a ve engin sabrından dolayı eřime teőekkr ederim.

ÖZGEÇMİŞ

1975 yılında İstanbul'da doğdu. İlköğretim ve lise öğrenimini İstanbul'da tamamladı. 1993 yılında Uludağ Üniversitesi Mühendislik Mimarlık Fakültesi Elektronik Mühendisliği Bölümüne girmeye hak kazandı. 1998 yılında Uludağ Üniversitesi Fen Bilimleri Enstitüsü Elektronik Mühendisliği bölümünde Yüksek lisansa başladı. 2001 yılında Uludağ Üniversitesi Fen Bilimleri Enstitüsü Elektronik Mühendisliği bölümünde Doktora öğrenimine başladı.

1997 yılında Uludağ Üniversitesi Teknik bilimler MYO'da Uzman olarak göreve başladı. 2001 yılından itibaren aynı bölümde Öğr. Gör. olarak çalışmaktadır.