Journal of Engineering and Natural Sciences
Mühendislik ve Fen Bilimleri Dergisi

Sigma 27, 197-207, 2009

Research Article / Araştırma Makalesi

THE EFFECTS OF FILTER FREQUENCY SCALE VARIABILITY ON SPEAKER IDENTIFICATION PERFORMANCE

Ömer ESKİDERE*[1], Figen ERTAŞ[2]

[1]*Uludağ Üniversitesi, Teknik Bilimler Meslek Yüksekokulu, Mekatronik Programı, BURSA*
[2]*Uludağ Üniversitesi, Mühendislik-Mimarlık Fakültesi, Elektronik Mühendisliği Bölümü, BURSA*

Received/Geliş: 12.01.2009 Revised/Düzeltme: 17.08.2009 Accepted/Kabul: 28.09.2009

ABSTRACT

Extracting discriminatory feature vectors that contain speaker specific information is of crucial importance in speaker identification. Although the cepstrum coefficients on the Mel frequency scale are commonly used as feature vectors, it is demonstrated in this paper that linear and ERB frequency scales provide better results compared to the Mel scale. In the paper, ERB, Bark and linear scales are compared with Mel scale on the TIMIT and NTIMIT databases. On the TIMIT database, an identification rate of 100% is obtained with the linear frequency scale when the filter-bank is placed in 0-8 KHz range, and a rate of 98.81% is obtained with the ERB scale using 0-4 KHz filter-bank frequency range. On the NIMIT database, 73.51% identification rate is achieved with linear scale, resulting in 2.97% improvement over that of the Mel scale.
Keywords: Filter frequency scale, speaker identification, Gaussian mixture model, TIMIT/NTIMIT databases.

FİLTRE FREKANS ÖLÇEĞİ DEĞİŞİMLERİNİN KONUŞMACI TANIMAYA ETKİSİ

ÖZET

Kişileri birbirinden ayırt edici özellikleri taşıyan öznitelik vektörlerinin elde edilmesi, konuşmacı tanımanın en önemli kısmıdır. Öznitelik vektörü olarak her ne kadar Mel frekans ölçeğindeki kepstrum katsayıları yaygın olarak kullanılsa da, bu makalede görüleceği üzere doğrusal ve ERB frekans ölçekleri kullanılarak Mel ölçeğe kıyasla daha iyi sonuçlar elde edilmiştir. Bu makalede, TIMIT ve NTIMIT veritabanları için, Mel ölçeği ile ERB, Bark ve doğrusal ölçek karşılaştırılmıştır. TIMIT veritabanında süzgeç dizilerinin yerleştirildiği frekans bandı 0-8 kHz için doğrusal ölçekle %100, 0-4 kHz frekans bandı için ERB ölçekle %98.81 konuşmacı tanıma oranı elde edilmiştir. NTIMIT veritabanında doğrusal ölçekle %73.51 konuşmacı tanıma oranı elde edilip Mel ölçeğe kıyasla %2.97 tanıma artışı sağlanmıştır.
Anahtar Sözcükler: Süzgeç frekans ölçekleri, konuşmacı tanıma, gauss karışım modeli, TIMIT/NTIMIT veritabanı.

1. GİRİŞ

Konuşmacı tanıma sistemlerinin tasarımında en önemli noktalardan biri, kişiye ait konuşma karakteristiklerini temsil eden öznitelik vektörlerinin seçimidir. Parametre olarak uygun özniteliklerin seçimi tanıma oranını doğrudan etkiler. Konuşmacı tanıma sistemleri için şimdiye

*Corresponding Author/Sorumlu Yazar: e-mail/e-ileti: oeskidere@uludag.edu.tr, tel: (224) 294 23 68

kadar yapılan çalışmalarda, Mel frekansı kepstrum katsayıları (MFCC) en çok kullanılan öznitelik olmuştur. Bunun sebebi de MFCC parametrelerinin diğer öznitelik vektörü oluşturma yöntemlerine oranla daha iyi tanıma performansı sağlamasıdır [1].

İnsan algılama yapısı üzerinde yapılan psikofizyolojik ölçümler ile çeşitli frekans ölçekleri elde edilmiştir. Bu frekans ölçekleri insanın kulağının algılamada ayırt edici olduğu frekansları göstermektedir [2]. Öznitelik vektörleri oluşturulurken frekans ölçekleri ile kullanılan süzgeçlerin yeri ve bant genişlikleri ayarlanmaktadır. Süzgeç seçimi yapılırken kişiye ait konuşma özelliklerinin, en iyi biçimde bir vektör ile ifade edilmesi amaçlanır. Kullanılan bu süzgeçlerin konumu konuşmacı tanıma performansını doğrudan etkilemektedir [3, 4]. Süzgeçlerin konumu değişik frekans ölçekleri ile belirlenmektedir.

Bu makalede Mel, Bark, ERB ve doğrusal frekans ölçekleri, mikrofon (TIMIT) ve telefon (NTIMIT) ortamlarından toplanan ses örnekleri için karşılaştırılmaktadır. Frekans ölçekleri karışım bileşen sayısı, örnekleme hızının düşürülmesi, süzgeçlerin 0-4 kHz aralığına sınırlandırılması ve kepstrum katsayı sayısı parametrelerine bağlı olarak incelenmektedir. Bu parametre değişimlerine bağlı olarak en ideal frekans ölçeği bulunmaktadır. Bant genişliği iyi ayarlanmış doğrusal frekans ölçeğinin, kişinin ayırt edici ses özelliklerini diğer frekans ölçeklerine göre daha iyi bulduğu gösterilmektedir.

2. ÖZNİTELİK VEKTÖRÜ OLUŞTURULMASI

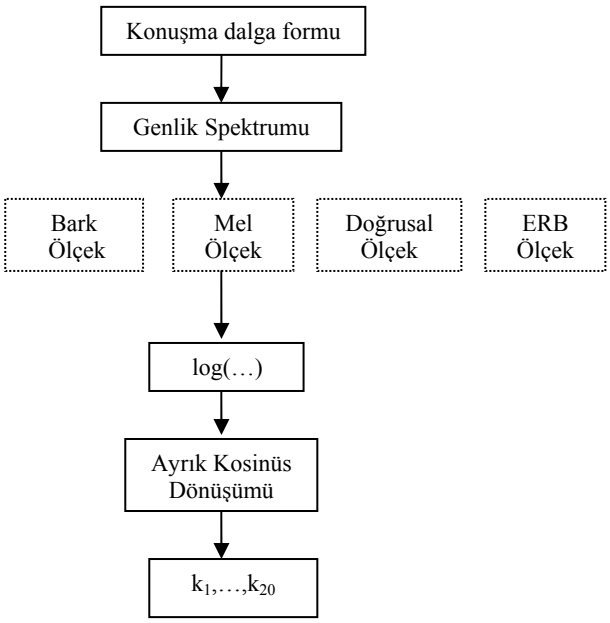
Her ne kadar konuşmacı tanımada konuşma özelliklerinin ayırt ediciliği pek fazla dikkate alınmasa da, konuşma spektrumunun konuşmacı tanımada etkili olduğu gözlenmiştir. Bu durum spektrumun kişinin ses yolu yapısını yansıtıp diğer kişilerin seslerine nazaran etkin fizyolojik bir ayırt edici faktör olması ile açıklanmaktadır [5].

Konuşmacılara ait ses örnekleri, ses değişimlerine karşı sabit kabul edilebilecek parçalara ayrılır. Genellikle 20–40 msn arasında değişen bu konuşma parçaları pencere fonksiyonlarından biri ile çarpılır ve konuşma parçasının orta kısmı vurgulanır. Elde edilen bu kısa süreli konuşma parçasının genlik spektrumu alınıp ön vurgulama uygulanır. Spektrum, sesin kısa süreli çerçeveler arası değişimine duyarlıdır. Spektrumu alınan işaret Şekil 1'de görülen frekans ölçeklerinden birine göre düzenlenmiş üçgen süzgeç dizilerinden geçirilip elde edilen işaretin logaritması alınır. En son olarak işarete ayrık kosinüs dönüşümü uygulanarak kepstrum katsayıları olarak bilinen öznitelik vektörleri elde edilir. Elde edilen bu öznitelikler konuşmacıların eğitim ve testinde kullanılır.

Üçgen süzgeç dizileri şu şekilde oluşturulmaktadır. Süzgeç sayısı *FS*, seçilen işaret bant genişliği [0, $f_s/2$] Hz ve f_s örnekleme frekansı olarak tanımlanır. Üçgen süzgeç dizilerinden biri l olsun, l ϵ [1, *FS*], bu süzgecin merkez frekansı f_{cl} olup alt ve üst bant geçiren frekansları ise; $f_{cl\text{-}1}$ ve f_{cl+1} olarak ifade edilir. Buna bağlı olarak f_{co}=0 ve $f_{cl}<f_s/2 \ \forall \ l$ olarak ifade edilir. Buna bağlı olarak süzgeç dizileri denklem 1'deki gibi ifade edilir.

$$F_l[k]=\begin{cases}((\frac{k}{N})f_s - f_{cl-1})/(f_{cl}-f_{cl-1}) & L_l \le k \le C_l \\ f_{cl+1}-(\frac{k}{N})f_s)/(f_{cl+1}-f_{cl}) & C_l \le k \le U_l\end{cases} \tag{1}$$

Burada $C_l = \frac{f_{cl}}{f_s} N$, $U_1 = \frac{f_{cl+1}}{f_s} N$ ve $L_l = \frac{f_{cl-1}}{f_s} N$ olup *l*'inci süzgecin merkez, üst ve alt frekanslarıdır [6]. Süzgeçlerin yerleştirildiği frekans ölçekleri ise aşağıdaki gibidir.



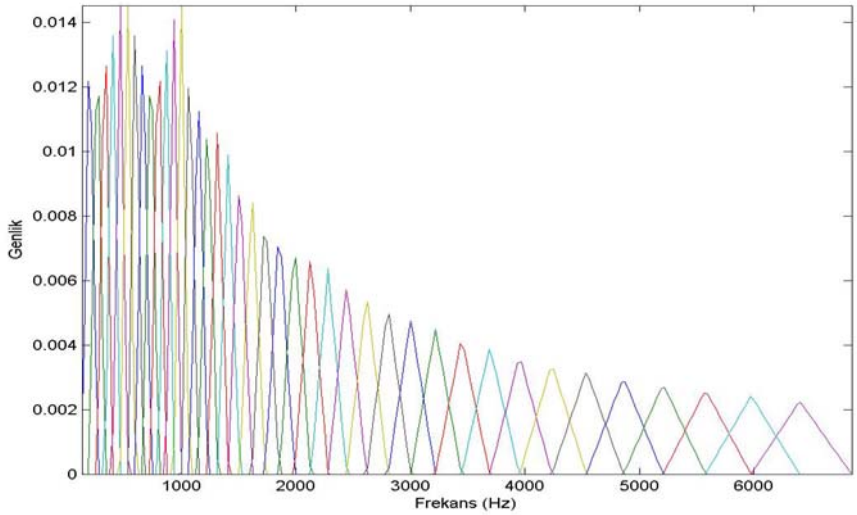
Şekil 1. Öznitelik vektörü oluşturma blok diyagramı

2.1. Mel Ölçek

Kulak tarafından algılanan frekansları ifade eden Mel değerleri Steven ve Volkman [2], tarafından tespit edilmiştir [7]. Bu Mel değerleri O'Shaughnessy [8], Fant [9] ve Slaney [10] tarafından tanımlanan Mel ölçekleri ile yaygın olarak ifade edilmektedir. Denklem 2, Hz'den Mel ölçeğe dönüşüm için kullanılan ifadedir.

$$Mel(f) = a\log(1+\frac{f}{b}) \tag{2}$$

Burada f, Hz olarak frekansı göstermekte olup $Mel(f)$ ise mel ölçekte frekansı göstermektedir. O'Shaughnessy [8], a=2595 ve b=700, Fant [9], a=1000/log2 ve b=1000 olarak tanımlamaktadır. Slaney [10], 1000 Hz altı 66.6 Hz bant genişliğinde doğrusal, 1000 Hz üstü logaritmik olarak tanımlamaktadır. Şekil 2'de Slaney [10] tarafından tanımlanan Mel ölçekte dizilmiş üçgen süzgeç dizileri görülmektedir.



Şekil 2. Mel ölçekte dizilmiş süzgeç dizileri

2.2. 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'deki formülle açıklanabilir [11].

$$Bark(f) = 13\arctan\left(\frac{0.76f}{1000}\right) + 3.5\arctan\left(\frac{f^2}{7500^2}\right) \tag{3}$$

Buradaki frekans ölçeğinin birimi kritik bant genişliği oranı ya da bark olarak adlandırılır. Yukarıda belirtilen bark ölçeği formülleriyle süzgeç oluşturmak için kritik bant genişliklerinin belirlenmesi gerekir. Kritik bant genişlikleri denklem 4'deki gibi belirlenir.

$$BWkritik=25+75[1+14(f/1000)^2]^{0.69} \tag{4}$$

2.3. 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 [12], deneysel ölçümlerle insan işitsel süzgeçlerinin ERB'si ile süzgeçlerin merkez frekansları arasındaki bağıntıyı denklem 5'deki gibi tanımlamaktadır.

$$ERB(f) = 0.108f + 24.7 \tag{5}$$

Bu denklemde *f*'in birimi Hz dir. Aynı şekilde işaret bant genişliği boyunca istenilen sayıda süzgeç ERB ölçeğinde eşit aralıklı olarak yerleştirilir [13]. ERB ölçeğine göre ayarlanan *i.* süzgeç dizisinin merkez frekansı ifadesi denklem 6'daki gibidir.

$$cf = -(E \cdot mBW) + \exp((i) \cdot (-\log(fs/2 + E \cdot mBW) + \log(lfr + E \cdot mBW))/nc) \cdot (fs/2 + E \cdot mBW) \tag{6}$$

Burada *cf* merkez frekansı, *E* asimptotik süzgeç kalite faktörü, *mBW* minimum bant genişliği, *lfr* en düşük frekans ve *nc* süzgeç sayısıdır. Moore ve Glasberg [14], asimptotik süzgeç kalite faktörünü, 9.26449 ve minimum bant genişliğini, 24.7 olarak tanımlamaktadır [10].

2.4. Doğrusal Ölçek

Doğrusal frekans ölçeği ile tüm frekans bölgesinin konuşmacının algılanmasında eşit etkiye sahip olduğu varsayılıp buna göre süzgeçlerin merkez frekansları eşit aralıklarla ve sabit bant genişliği ile konuşmacı frekans bandına yerleştirilir. TIMIT veritabanı için 0-8000 Hz, NTIMIT veritabanı için 300-3400 Hz frekans aralığına, 66,6 Hz bant genişliğinde üçgen süzgeçler, % 50 örtüşme uygulanarak düzgün aralıklarla yerleştirilmektedir. Şekil 3'de 0-8000 Hz aralığında maksimum değerine normalize edilmiş Mel, doğrusal, Bark, ERB ölçekleri görülmektedir.

3. GAUSS KARIŞIM MODELİ

Elde edilen öznitelik vektörleri Gauss karışım modeli kullanılarak modellenmektedir. Gauss karışım modeli, *M* adet Gauss yoğunluğun ağırlıklı toplamı olarak denklem 7'deki gibi gösterilmektedir [5].

$$p(x/\lambda) = \sum_{i=1}^{M} p_i b_i(x) \tag{7}$$

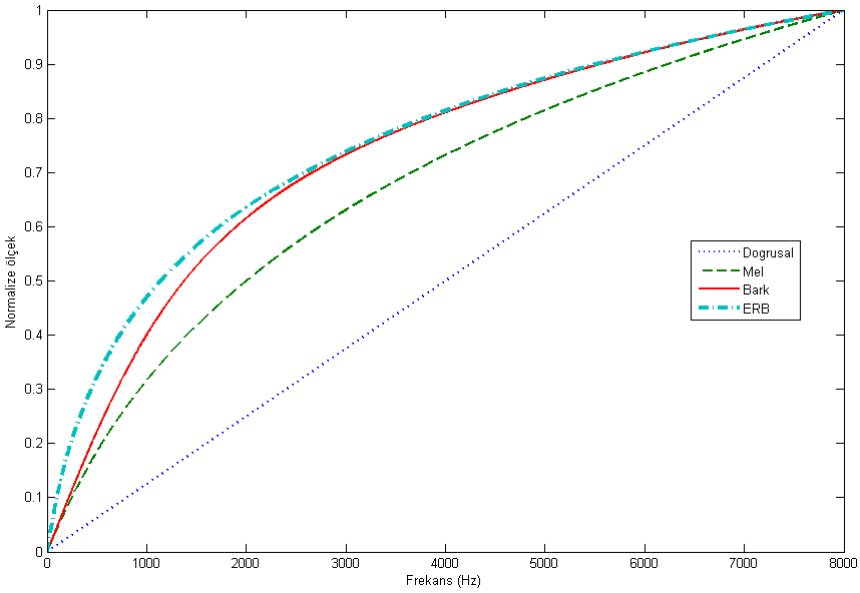
Burada x, D boyutlu rastsal vektörü; $b_i(x)$, i=1,2,3...*M,* Gauss yoğunluk bileşenlerini ve p_i de karışım ağırlığını göstermektedir. Gauss karışım modelinde her bileşenin ortalama vektörü, ortak değişinti matrisi ve karışım ağırlık değerleri ile denklem 8'deki gibi gösterilir.

$$\lambda = \{p_i, \mu_i \sum_i\} \qquad \text{i=1,2 ....,}M \tag{8}$$

Burada $\sum_{i=1}^{M} p_i = 1$ olup μ_i ortalama vektör ve $\sum_i$ ortak değişinti matrisini ifade etmektedir. Gauss karışım modelindeki bileşenlere ait parametrelerin tahmini için maksimum benzerlik tahmin yöntemi kullanılır. Bu yöntemde amaç eğitim verilerinden p(x/λ)'yı en büyük yapacak model parametrelerini bulmaktır. *T* adet vektörden oluşan eğitim dizisi *X* ile gösterilsin: X={$x_1, x_2, \ldots\ldots x_T$ }. Bu *X* dizisi için Gauss karışım olasılığı denklem 9'daki gibi yazılabilir.

$$\text{p(X/}\lambda\text{)} = \prod_{t=1}^{T} p(x_t/\lambda) \tag{9}$$

Bu ifade, λ parametrelerinin doğrusal olmayan bir işlevidir ve direkt olarak en büyük yapılması mümkün değildir. Beklentinin maksimumlaştırılması (BM) algoritması kullanılarak λ parametrelerine göre denklem 9 en büyük yapılır [6].



Şekil 3. Normalize edilmiş ERB, Mel, Bark, ve doğrusal frekans ölçekleri

4. DENEYSEL ÇALIŞMA

Yapılan deneylerde TIMIT ve NTIMIT veritabanlarına ait ses örnekleri kullanılmaktadır. TIMIT veri tabanı toplam 630 kişinin her birinin söylediği 10'ar adet cümleden oluşmaktadır. Konuşma işareti 16 kHz örnekleme frekansı ile kaydedilmiştir. NTIMIT veritabanı, TIMIT veritabanındaki cümlelerin karbondan yapılma telefon ahizesi üzerinden bir yerel veya uzun mesafe merkez ofise iletilmiş ve aynı hat üzerinden tekrar kayıt için geri alınmış halidir. Deneylerde TIMIT veritabanının tamamı ve her iki veritabanın 168 konuşmacıdan oluşan test dizini kullanılmaktadır.

Konuşmacılar 32 adet Gauss karışımı ile modellenmektedir. BM algoritması model başlangıç değeri, k-ortalama algoritması ile kestirilip, minimum değişinti sınırı 0.01 alınmaktadır. Model 15 özyinelemede istenen değere yakınsamaktadır. Konuşmalar test edilirken test sözcüklerine ait değerler, hafızadaki her bir konuşmacı modele uygulanır ve maksimum olasılıklı modele ait kişiye eşleştirilir. Eğitim için yaklaşık toplam 24 saniye uzunluğunda (2 sa, 3 si ve 3 sx) cümleleri, test için ise kalan 3 saniye uzunluğunda yaklaşık 1 cümle kullanılmıştır.

TIMIT veritabanındaki her bir konuşmacının analizinde; konuşmalar 10 msn örtüşme ile 20 msn uzunluğunda kısa süreli çerçevelere ayrılıp Hamming pencereden geçirilir. Elde edilen işaretin genlik spektrumu alınıp ayarlanan frekans ölçeklerine bağlı olarak elde edilen süzgeç dizilerinden geçirilir. Üçgen süzgeç dizileri kullanılacak olan Mel, Bark, ERB ve doğrusal ölçeğe bağlı olarak yerleştirilir. Süzgeç çıkışlarının log enerjileri alınıp ayrık kosinüs dönüşümü uygulandıktan sonra öznitelik vektörleri elde edilmektedir. 0. öznitelik vektörü ortalama enerjiyi gösterdiğinden alınmamaktadır. Konuşmanın her bir çerçevesi 24 kepstrum katsayısı ile ifade edilir. Bu şartlarda aşağıdaki deneyler yapılmaktadır.

1. İki değişik konuşmacı grubu için frekans ölçekleri değişimine göre doğru konuşmacı tanıma oranları incelenecektir. Konuşmacı grupları, 168 kişiden oluşan test dizini ve 630 kişiden oluşan TIMIT veritabanının tamamıdır. Çizelge 1'de bu iki konuşmacı grubu için Bölüm 2'de tanımlanan frekans ölçeklerinde süzgeçlerin yerleştirilmesi ile elde edilen konuşmacı tanıma oranları görülmektedir.

Çizelge 1. Değişik süzgeç ölçekleri için konuşmacı tanıma oranları (%)

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

Süzgeç aralığı 0-8 kHz, kepstrum katsayı sayısı 24, örnekleme frekansı 16 kHz, karışım bileşen sayısı 32, TIMIT veritabanı

Çizelge 1'den görüleceği üzere konuşmacı sayısı 168 kişi için doğrusal ve ERB frekans ölçekleri kullanılarak %100 lük konuşmacı tanıma oranı elde edilmektedir. Veritabanının tamamı ile yapılan deneyde doğrusal frekans ölçeği ile test edilen konuşmacı grubu için %100, Mel ölçeği için %99.4 tanıma oranı elde edilmektedir.

2. TIMIT veritabanında 168 konuşmacı için, karışım bileşen sayısı değişimine bağlı olarak, frekans ölçeklerinin değişiminin tanıma üzerine etkisi incelenecektir. Konuşmacıların ses örneklerinin örnekleme hızı 16 kHz'den 8 kHz'e düşürüldüğünde Çizelge 2'deki sonuçlar elde edilmektedir.

Çizelge 2. Karışım bileşen 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 0-8 kHz, kepstrum katsayı sayısı 24, örnekleme frekansı 8 kHz, TIMIT veritabanı

Çizelge 2'den görüleceği üzere değişik karışım bileşen sayıları için en yüksek tanıma oranı doğrusal frekans ölçeğinde elde edilmektedir. Doğrusal frekans ölçeği diğer frekans ölçeklerine nazaran daha gürbüz davranmaktadır.

3. TIMIT veritabanı için filtre dizilerine bant sınırlama uygulanması durumunda tanıma oranı değişimi gözlenecektir. Süzgeç dizileri, 0-4 kHz frekans aralığında hazırlanıp ses işaretine ön vurgulama uygulanmasına bağlı olarak konuşmacı tanıma performansı ölçülecektir. Örnekleme frekansı 16 kHz için elde edilen sonuçlar Çizelge 3'de görülmektedir.

Çizelge 3. Süzgeç aralığı 0-4 kHz için değişik frekans ölçekleri için 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 0-4 kHz, kepstrum katsayı sayısı 20, örnekleme frekansı 16 kHz, Konuşmacı sayısı 168, TIMIT veritabanı

Çizelge 3'den görüleceği üzere süzgeçler 0-4 kHz aralığında yerleştirildiğinde Mel ölçeğinde en yüksek sonuç 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 % 2.08 daha iyi tanıma sağlamaktadır.

4. TIMIT veritabanında üçgen süzgeç dizileri bant sınırlamalı (0-4 kHz) ve bant sınırlamasız (0-8 kHz) frekans aralığında yerleştirilmektedir. Doğrusal, ERB, Mel, Bark frekans ölçekleri için kepstrum katsayıları 9, 12, 15, 18, 20, 22 ve 24 olması durumunda elde edilen konuşmacı tanıma oranları Çizelge 4'deki gibidir.

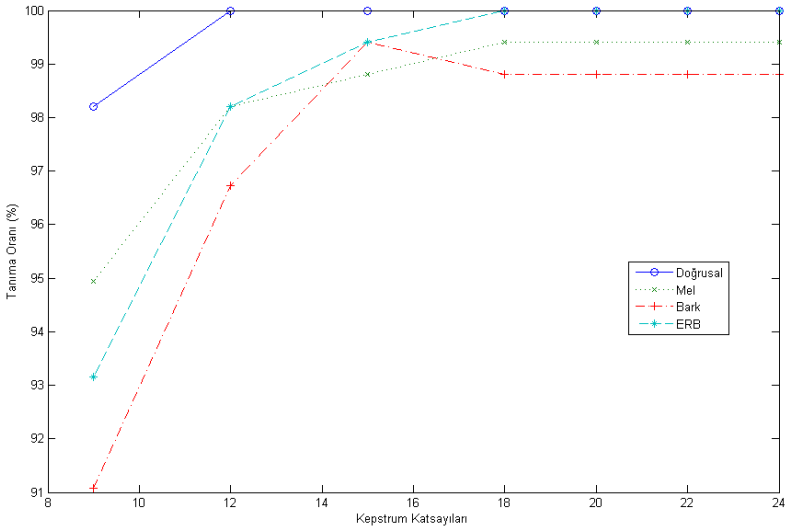
Çizelge 4. Dört değişik frekans ölçeği için konuşmacı tanıma oranları (%)

Kepstrum katsayıları	Doğrusal ölçek		Mel ölçek		Bark ölçek		ERB ölçek	
	0-8 kHz	0-4 kHz	0-8 kHz	0-4 kHz	0-8 kHz	0-4 kHz	0-8 kHz	0-4 kHz
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

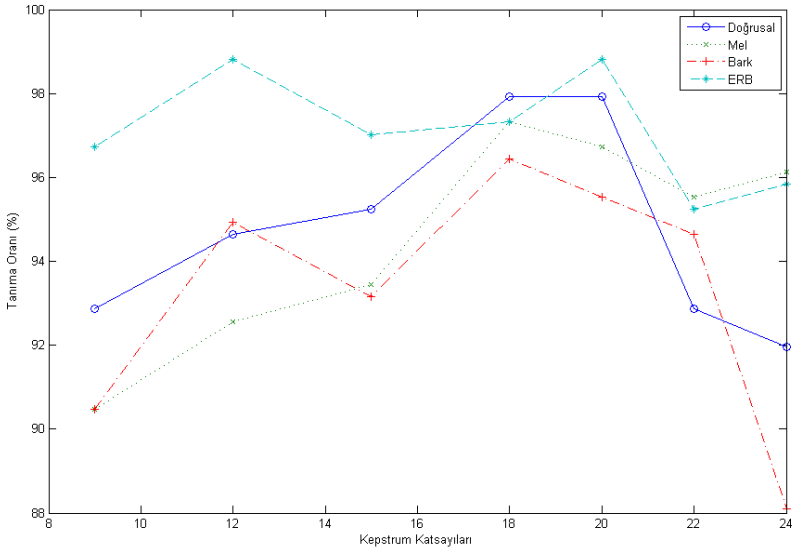
Örnekleme frekansı 16 kHz, karışım bileşen sayısı 32, TIMIT veritabanı Mel, Bark ölçek ön vurgulamalı, Doğrusal ve ERB ölçek ön vurgulamasız, konuşmacı sayısı 168

Çizelge 4'den görüleceği 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 bağlı olarak değişimi Şekil 4'de daha ayrıntılı görülmektedir**.**

Süzgeçlerin yerleştirildiği bant aralığı 0-8 kHz için, doğrusal ve ERB ölçekleri kepstrum katsayısı 18 ve üzeri olması durumunda % 100 lük konuşmacı tanıma elde edilmektedir. Bant aralığı 0-4 kHz için doğrusal, Mel, Bark, ERB frekans ölçeklerinde değişik kepstrum katsayıları için konuşmacı tanıma oranları Şekil 5'de görülmektedir. Süzgeçlerin yerleştirildiği bant aralığı 0-4 kHz için ERB ölçeğinde kepstrum katsayılarının 12 ve 20 olduğu durumlarda en yüksek (% 98.81) konuşmacı tanıma oranı elde edilmiştir.



Şekil 4. Değişik frekans ölçeklerinin kepstrum katsayıları değişimlerine bağlı olarak karşılaştırılması (0-8 kHz)



Şekil 5. Değişik frekans ölçeklerinin kepstrum katsayı değişimlerine bağlı olarak karşılaştırılması (0-4 kHz)

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 uygulanmaktadır. Her bir çerçeve için 20 kepstrum katsayısı kullanılıp, konuşma işaretine ön vurgulama uygulanmayıp, Gauss karışım bileşen sayısı 32 alınmaktadır. Her bir konuşmacı sekiz cümle kullanılarak eğitilmekte, 1 cümle kullanılarak test edilmektedir. Çizelge 5'de NTIMIT veritabanı için değişik frekans ölçeklerinde konuşmacı tanıma oranları görülmektedir.

Çizelge 5. Değişik frekans ölçekleri için konuşmacı tanıma oranları (%)

Konuşmacı sayısı	**Ölçek çeşidi**			
	Doğrusal	**Mel**	**Bark**	**ERB**
168	**70.24**	69.05	58.33	68.45

Kepstrum katsayı sayısı 20, ön vurgulama yok, NTIMIT veritabanı

Çizelge 5'den görüleceği üzere doğrusal frekans ölçeği ile % 70.24 konuşmacı tanıma oranı elde edilmiştir. Mel ölçeği kullanıldığında konuşmacı tanıma oranı % 69.05 olmaktadır.

6. NTIMIT veritabanı için konuşmadan sessiz kısımların atılması durumunda üçgen süzgeç dizilerinin yerleştirildiği frekans ölçeği değişiminin konuşmacı tanımaya etkisi incelenecektir. TIMIT veritabanında konuşmadan sessiz kısımların atılması tanıma oranını değiştirmemektedir. Konuşmada sesli sessiz ayırımında Aliaa ve diğ. [15], tarafından belirtilen eşik değeri kullanılmaktadır. Konuşmadaki eşik değerinin altındaki sessiz çerçevelere karşılık gelen kısımlar atılmakta ve buna bağlı olarak öznitelik vektörleri

üretilmektedir. Konuşmacıların öznitelik vektörleri üretilirken doğrusal, Mel, Bark ve ERB ölçekte süzgeçler 300-3400 Hz arasına yerleştirilir. 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 1 cümle kullanılmaktadır. Bu durumda elde edilen tanıma oranları Çizelge 6'da görülmektedir.

Çizelge 6. Konuşmadan sessiz kısımların atılmasına bağlı olarak dört değişik frekans ölçeği için konuşmacı tanıma oranları (%)

Konuşmacı sayısı	Doğrusal	Mel	Bark	ERB
168	**73.51**	70.54	60.42	69.94

Kepstrum katsayı sayısı 20, ön vurgulama yok, NTIMIT veritabanı

Çizelge 6'dan görüleceği üzere konuşmadan sessiz kısımlar atıldığında, doğrusal ölçek için konuşmacı tanıma oranı 70.24'ten % 73.51'e çıkmaktadır. Mel ölçek için tanıma oranı % 69.05'den % 70.54'e çıkmaktadır. Dört frekans ölçeği içinde en iyi tanıma oranı doğrusal ölçek ile elde edilmektedir.

5. SONUÇLAR

Bu çalışmada öznitelik vektörü elde edilmesinde kullanılan süzgeçlerin yerleştirildiği frekans ölçekleri, metinden bağımsız Gauss karışım modeli kullanılarak, konuşmacı tanıma oranları karşılaştırılmıştır. Bilinenin aksine bant genişliği iyi ayarlanmış doğrusal frekans ölçeği kişinin ayırt edici ses özelliklerini Mel frekans ölçeğinden daha iyi yakalamaktadır.

TIMIT veritabanı ile frekans bandı 0-8 kHz için doğrusal ölçek ile % 100 tanıma oranı elde edilmiştir. TIMIT veritabanındaki konuşmalara 0-4 kHz bant sınırlaması uygulandığında, ERB frekans ölçeğinin konuşmacı tanımada en iyi performansı gösterdiği görülmektedir. Reynolds ve diğ. [16], Mel ölçeğini kullanarak bant sınırlamalı durumda % 95.2 tanıma oranı elde etmiştir. Yaptığımız deneylerde bant sınırlamalı durumda ERB ölçek ile % 98.81 tanıma oranı elde edilip, Mel ölçeğe nazaran tanıma oranında % 3.61 iyileşme sağlanmıştır. Bant sınırlamalı durumda frekans ölçekleri tanıma oranlarına bağlı olarak ERB, doğrusal, Mel ve Bark şeklinde sıralanmaktadır.

NTIMIT veritabanında konuşmalar telefon hattından elde edildiğinden dolayı, TIMIT veritabanına nazaran tanıma oranı % 26.49 daha düşük olup en yüksek tanıma oranı doğrusal ölçekte % 73.51 olarak elde edilmiştir. Bu sonuç Mel ölçeğe kıyasla % 2.97 tanıma artışı sağlamaktadır. NTIMIT veritabanı için tanıma oranına göre süzgeç dizilerinin yerleştirildiği frekans ölçekleri; doğrusal, Mel, ERB ve Bark olarak sıralanmaktadır.

REFERENCES / KAYNAKLAR

[1] Liu, Li., J. He and Palm G., "Signal Modeling for Speaker Identifiction". Proc. Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP-96), Vol. 2, 1996, pp. 665-668.

[2] Stevens, S. and J. Volkman, "The Relation of Pitch to Frequency". American Journal of Psychology, vol. 53, p. 329, 1940.

[3] Kinnunen, T. "Spectral Features for Automatic Text-independent Speaker Recognition", Ph.Lic. thesis, University of Joensuu, Department of Computer Science p. 49-115, 2003.

[4] Ganchev, T. "Speaker Recognition", Ph.D. thesis, Dept. of Electrical and Computer Engineering, University of Patras, Greece. p. 61-82. 2005.

[5] Reynolds D. A., and Rose, R. C., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. Speech Audio Proc., 3, (1), pp. 72–83, 1995.

[6] Reynolds, D. A., "A Gaussian Mixture Modeling Approach to Text Independent Speaker Identification", Ph.D. Thesis, Georgia Institute of Technology, 1992.

[7] Umesh, S., L. Cohen and Nelson D., "Fitting the Mel Scale". Proc. Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP-99), Vol. 1, 1999, pp. 217–220.

[8] O'Shaughnessy, D., "Speech Communication Human and Machine". Addison Wesley, New York, 1987.

[9] Fant, G., "Acoustic Theory of Speech Production". Mouton & Co., The Hauge, 1960.

[10] Slaney, M., "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank", Tech. Rep. 35, Apple Computer, Inc., 1993.

[11] Picone, J., "Fundamentals of Speech Recognition: a Short Course". Institute for Signal and Information Processing, pp. 68-69, 1996.

[12] Moore , B. C. J. and B. Glasberg R., "Suggested Formula for Calculating Auditory Filter Bandwidths and Exicatation Patterns", J. Acoust. Soc. Am., 74, p. 750-753, 1983.

[13] Ertaş, F., "Ses İşaretlerine Karşı Basilar Membran Hareketinin Yazılım Benzetimi", S.D.Ü. Fen Bilimleri Dergisi 6:1, s. 86-93, 2002.

[14] Glasberg, B. R. and Moore B. C. J., "Derivation of Auditory Filter Shapes From Notched-Noise Data", Hearing Research, vol. 47, pp. 103–108, 1990.

[15] Aliaa, A. Y., Ebada A. S. and El Behaidy W. H., "Development of Automatic Speaker Identification System", 21^{st} National Radio Science Conf., 2004.

[16] Reynolds D. A., Zissman M. A., Quatieri T. F., et. al., "The Effects of Telephone Transmission Degradations on Speaker Recognition Performance", ICASSP (Detroit), May 9-12, 1995, 329-331.