MEL FREKANSI KEPSTRUM KATSAYILARINDAKİ DEĞİŞİMLERİN KONUŞMACI TANIMAYA ETKİSİ

Ömer ESKİDERE[*]
Figen ERTAŞ[**]

Özet: Konuşmacıya özgü bilgileri karakterize eden özniteliklerin çıkartılması, konuşmacı tanıma sisteminin performansı için hayati öneme sahiptir. Bu makalede, TIMIT ve NTIMIT veritabanları kullanılarak öznitelik vektörü oluşturma aşamalarının her biri için parametre değişiminin konuşmacı tanımaya etkisi incelenmekte ve tanımayı arttırıcı en iyi parametre değerleri bulunmaktadır. Bu veritabanları ile yapılacak diğer konuşmacı tanıma çalışmaları için, kaynak olabilecek optimum öznitelik değerleri belirlenmiştir. Bu sayede diğer araştırmacıların, en iyi parametreleri bulmak için tekrar deney yapmalarına gerek kalmayacaktır.

Anahtar Kelimeler: Mel frekansı kepstrum katsayıları, Konuşmacı tanıma, Gauss karışım modeli, TIMIT/NTIMIT veritabanları.

The Effects of Variabilities in Mel Frequency Cepstrum Coeffcients On Speaker Recognition

Abstract: Extraction of speaker-specific features which characterize the information towards identification of the correct speaker is vital importance. In this work TIMIT and NTIMIT databases are used. The effect of changing the feature vector elements to the speaker identification is analyzed and the best identifying elements are found. The best identifying feature vector elements may also be used for other speaker identification studies using the same databases. This way, any future work using these databases may not need to optimize the feature vectors towards identification.

Key Words: Mel frequency cepstrum coefficients, Speaker identification, Gaussian mixture model, TIMIT/NTIMIT databases.

1. GİRİŞ

Konuşmacının sesinden kendisini karakterize eden değerlerin çıkartılması işlemine öznitelik çıkartma işlemi adı verilir. Öznitelik çıkartma, bir konuşmacıyı sonradan tanımlayabilmek için ses sinyalinden elde edilmiş küçük bir veri topluluğu oluşturmaya yarayan bir işlemdir.

Wolf (1972) ideal özniteliklerin sahip olması gerektiği özellikleri tanımlamıştır. 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, özniteliklere ait istenen bu özelliklerin eş zamanlı olarak elde edilmesi çok zordur (Reynolds, 1992). Uygulamaya bağlı olarak bu öznitelik standartlarında kısmi değişimler oluşabilir.

[*] Uludağ Üniversitesi, Teknik Bilimler Meslek Yüksekokulu, Mekatronik Programı, 16059, Görükle, Bursa.
[**] Uludağ Üniversitesi, Mühendislik-Mimarlık Fakültesi, Elektronik Mühendisliği Bölümü, 16059, Görükle, Bursa.

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.

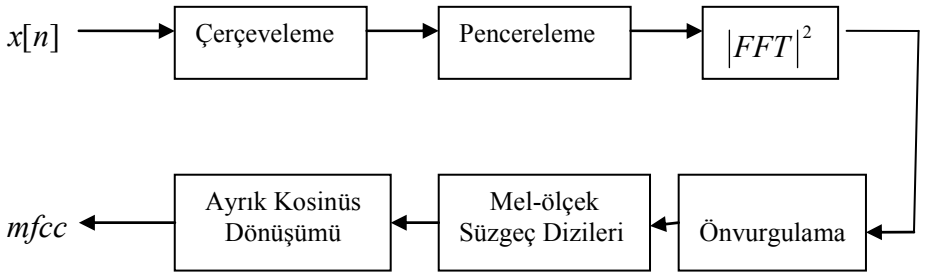
Konuşma spektrumunun öznitelik olarak kullanılma yöntemleri değişim göstermektedir. Yaygın spektrum gösterim yöntemleri; doğrusal öngörü katsayıları ve onun değişik dönüşümleri, süzgeç dizisi enerjileri ve onun kepstral gösterimleri sayılabilir. Doğrusal öngörü katsayıları (DÖK), konuşmada gürültü olması 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 kepstral 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 çalışmada bu işlem adım adım incelenip her bir öznitelik vektörü parametresinin konuşmacı tanıma üzerine etkisi araştırılmaktadır.

Deneylerde veritabanı olarak sıklıkla kullanılan TIMIT (Zue ve diğ., 1990) ve NTIMIT (Jankowski ve diğ., 1990) kullanılmıştır. TIMIT veritabanında Amerikan İngilizcesinin 8 ana lehçesine sahip bölgelerden seçilmiş 438 erkek, 192 kadın olmak üzere toplam 630 konuşmacıya ait 10'ar fonetik olarak zengin cümle bulunmaktadır. Konuşmalar sessiz ortamda mikrofon kullanılarak kaydedilmiş ve 16 kHz'de örneklenmiştir. NTIMIT ise TIMIT veritabanının telefon hattı üzerinden geçirilmesiyle elde edilmiştir.

2. MEL FREKANSI KEPSTRUM KATSAYILARI (MFCC)

Öznitelik vektörü olarak kullanılan kepstrum katsayıları elde edilirken, genellikle konuşmacı tanıma uygulamalarında MFCC kullanılır (Matsui ve diğ., 1995). Bunun nedeni, MFCC insan kulağının frekans seçiciliğini taklit ederek iyi bir şekilde konuşmacıları ayırt edici değerler elde edilmesidir. Ayrıca MFCC katsayıları değişimlerden, ses dalga yapısından çok daha az etkilenir. Reynolds (1992), tarafından önerilen MFCC vektörü çıkartımı blok diyagramı şekil 1'de görülmektedir.

Bazı çalışmalarda (Davis ve Mermelstein, 1980), önvurgulama 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 makalede incelenmektedir.

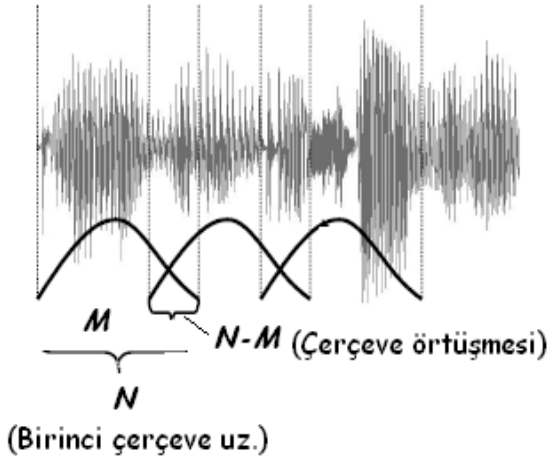


Şekil 1:
MFCC özniteliklerinin çıkarılma işleminin blok diyagramı

2.1. 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* örnek uzunluğunda konuşma parçalarına bölünür. İlk çerçeve *N* örnekten oluşurken sonraki çerçeve ilk çerçeveden *M* örnek sonra başlar ve böylece *N-M* örnek örtüşür (Rabiner ve Juang, 1993). Şekil 2 'de bir konuşma işareti üzerinde çerçeveleme işlemi görülmektedir.



Şekil 2:
Bir konuşmanın çerçevelere bölünmesi

2.2. Pencereleme

Mel frekansı kepstrum 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:

$$w[k+1] = 0.54 - 0.46\cos\left(2\pi\frac{k}{N-1}\right) \qquad k = 0,\ldots\ldots\ldots\ldots,N-1 \tag{1}$$

Hanning:

$$w[k+1] = 0.5\left(1-\cos\left(2\pi\frac{k}{N-1}\right)\right), \qquad k = 0,\ldots\ldots\ldots\ldots,N-1 \tag{2}$$

Blackman:

$$w[k+1] = 0.42 - 0.5\cos\left(2\pi\frac{k}{N-1}\right) + 0.08\cos\left(4\pi\frac{k}{N-1}\right), \quad k = 0,\ldots\ldots\ldots\ldots,N-1 \tag{3}$$

Gauss:

$$w[k+1] = e^{-\frac{1}{2}\left(\alpha\frac{k-\frac{N}{2}}{N/2}\right)^2} \qquad 0 \le k \le N \qquad \text{ve} \quad \alpha \ge 2 \tag{4}$$

Dikdörtgen:

$$w[k+1] = 1, \qquad k = 0,\ldots\ldots\ldots\ldots,N-1 \tag{5}$$

Üçgen:

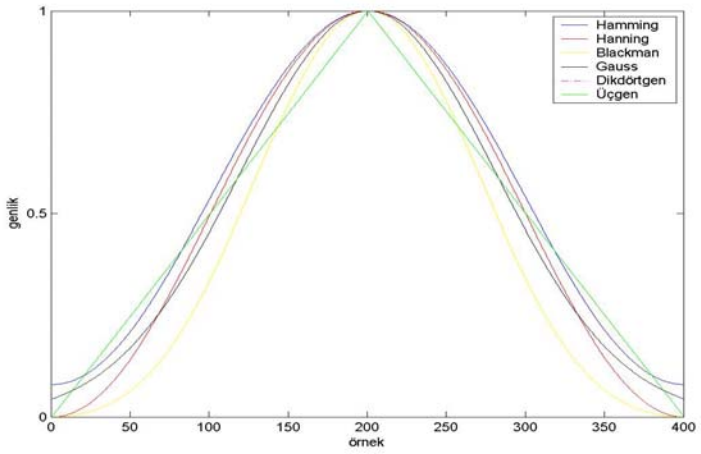
N tek için;

$$w[k] = \begin{cases} \dfrac{2k}{N+1},\ldots\ldots\ldots\ldots\ldots 1 \le k \le \dfrac{N+1}{2} \\ \dfrac{2(N-k+1)}{n+1},\ldots\ldots \dfrac{N+1}{2} \le k \le N \end{cases} \tag{6}$$

N çift için;

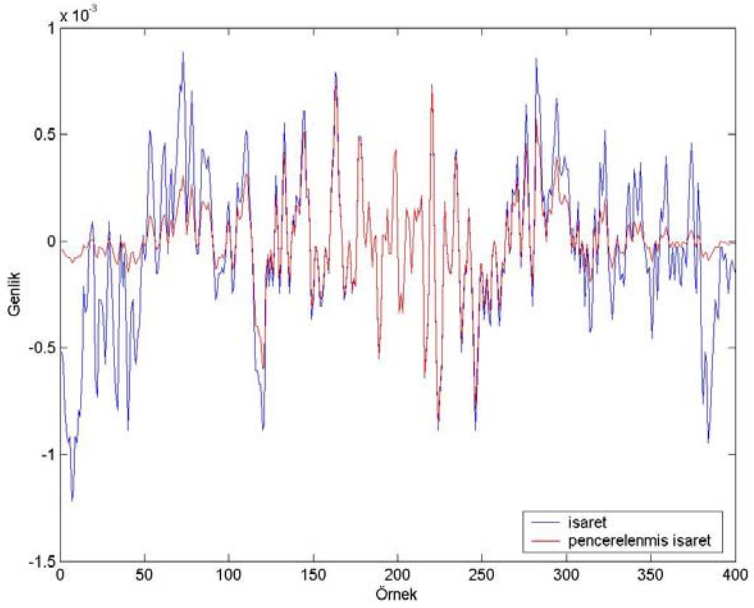
$$w[k] = \begin{cases} \dfrac{2k-1}{N},\ldots\ldots\ldots\ldots\ldots 1 \le k \le \dfrac{N}{2} \\ \dfrac{2(N-k+1)}{N},\ldots\ldots\ldots \dfrac{N}{2}+1 \le k \le N \end{cases} \qquad k = 0,\ldots\ldots\ldots\ldots,N-1 \tag{7}$$

Şekil 3'de, bu pencereleme fonksiyonlarının çerçeve süresi 400 örnek için eğrileri verilmektedir.



Şekil 3:
Pencereleme fonksiyonları

Şekil 4'de 25 msn'lik konuşma çerçevesi ve bu konuşma çerçevesinin hamming pencereleme uygulandıktan sonraki durumu görülmektedir. Şekil 4'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 4:
Konuşma çerçevesi ve hamming pencereden geçirilmiş hali

2.3. 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. Bir çerçevenin $\{x_0, x_1, ..., x_{N-1}\}$, ayrık fourier dönüşümü denklem 8'deki gibi tanımlanır.

$$X[k] = \sum_{n=0}^{N-1} x_n \cdot e^{-2\pi jkn/N}, \qquad k = 0,1,2,...,N-1 \tag{8}$$

Burada genellikle $X[k]$'ler kompleks sayılardır. Sonuç olarak elde edilen dizi $\{X[k]\}$: sıfır frekansı k=0 a karşılık gelip, pozitif frekanslar ($0 < f < f_s/2$), $1 \le k \le (N/2)-1$ değerlerine karşılık gelirken, negatif frekanslar ($-f_s/2 < f < 0$), $(N/2)+1 \le k \le N-1$'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 9'daki gibi tanımlanır.

$$X[k] = |X[k]| e^{j\angle X[k]} \tag{9}$$

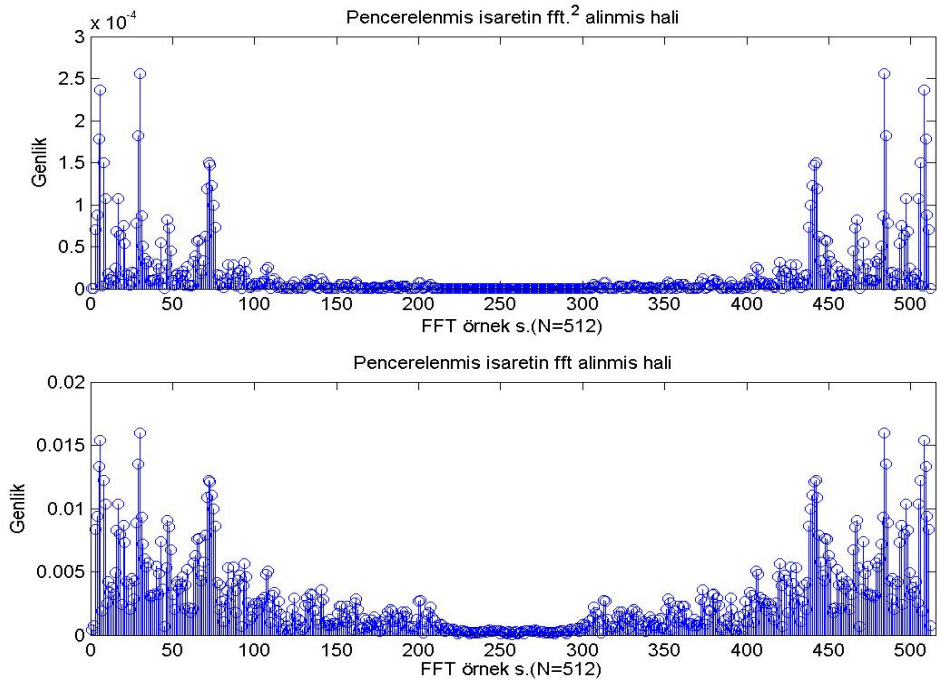
$$|X[k]| = \sqrt{X_{re}[k]^2 + X_{im}[k]^2} \tag{10}$$

$$\angle X[k] = \tan^{-1}\left(\frac{X_{im}[k]}{X_{re}[k]}\right) \tag{11}$$

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 $2^M \quad M \in N_+$ şeklinde başka bir değişle 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 5'de NTIMIT veritabanında 25 msn'lik pencerelenmiş konuşma parçasının $|FFT|^2$ ve $|FFT|$ alınmış hali bulunmaktadır. Şekilde işaretin *N*/2'ye göre simetrik olduğu görülmektedir.



Şekil 5:
Pencerelenen konuşma çerçevesinin $|FFT|^2$ *ve* $|FFT|$ *alınmış hali*

2.4. Önvurgulama

Önvurgulama 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 önvurgulamaya ihtiyaç olmaz (Kinnunen, 2003). Denklem 12'de verilen 1. dereceden süzgeç ile işaret 6 dB/oktav iyileştirilir.

$$H(z) = 1 - \alpha z^{-1} \tag{12}$$

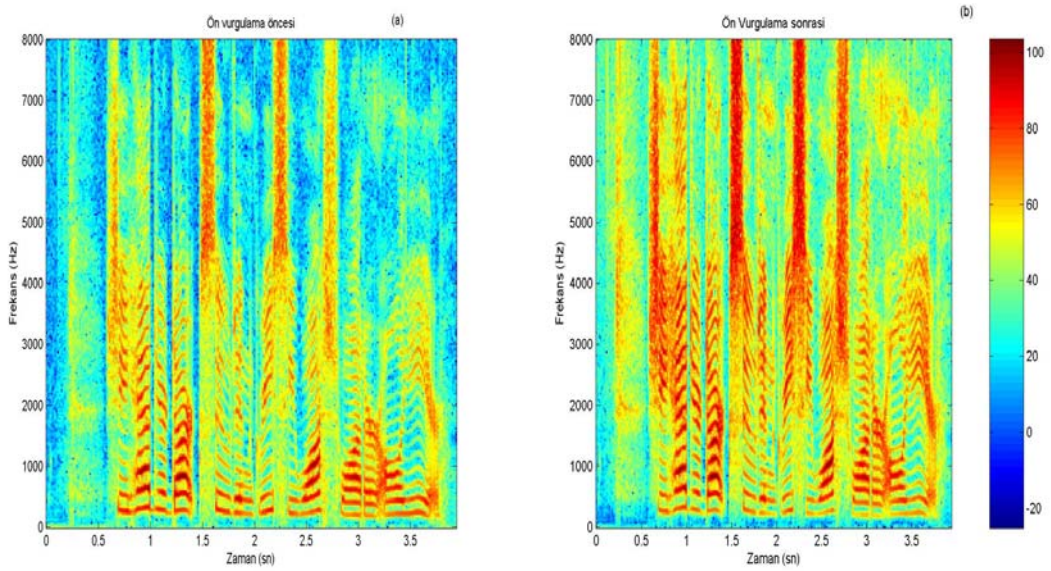
Burada α, önvurgulamanın derecesini yansıtıp genellikle 0.9 ile 1 arasında alınır (Wildermoth, 2001). Konuşma analizinde genellikle α değeri 0.95 alınmaktadır (Rabiner ve Juang, 1993). Süzgeç çıktısı y(n), fark denklemi olarak denklem 13'deki gibi ifade edilir.

$$y(n) = x(n) - \alpha \cdot x(n-1) \quad \text{n} = 0,1,2\ldots\ldots\text{N-1} \tag{13}$$

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

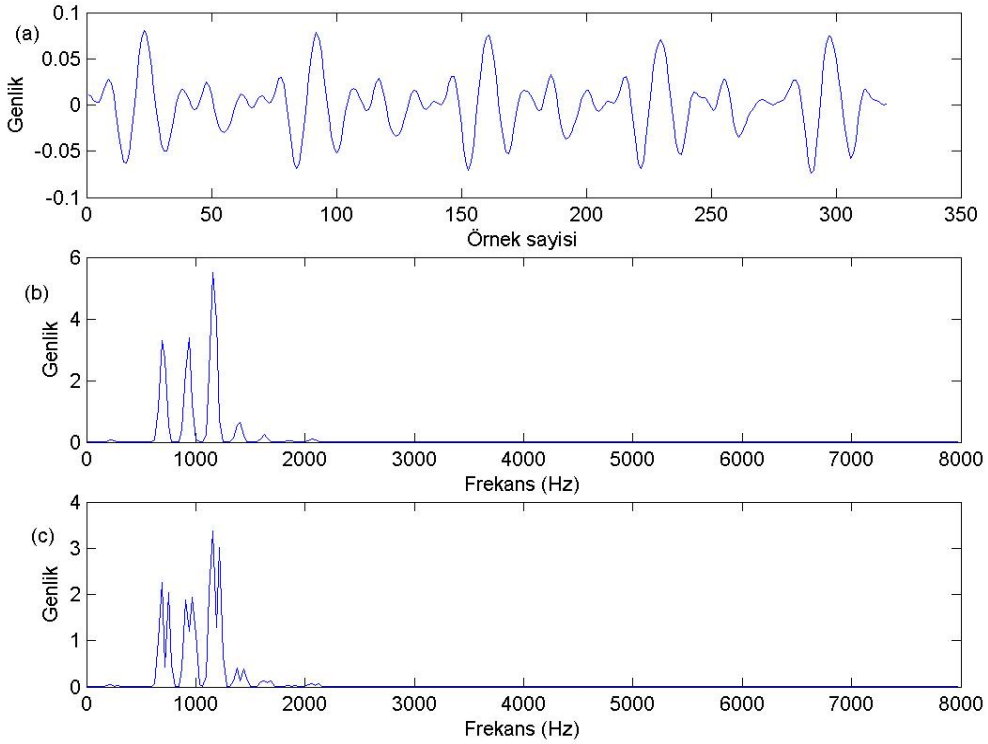
Bazı konuşmacı tanıma uygulamalarında işaretin çerçevelenmesi aşamasından önce önvurgulama uygulamak yerine güç spektrumu alındıktan sonra önvurgulama işlemi uygulanmaktadır (Reynolds, 1992; Reynolds ve Rose, 1995). Deneylerde birinci olarak işaret çerçevelenmeden önce önvurgulamanın etkisi incelenecek, ikinci olarak işaretin güç spektrumu alındıktan sonra vurgulama işlemi uygulanmasının etkisi incelenmektedir.

Bir cümle için yüksek frekanslı bileşenlerin güçlendirilmesi zaman-frekans eğrisinde (spektrogram) daha iyi belirlenebilmektedir. Şekil 6'da TIMIT veritabanına ait bir cümlenin önvurgulamadan ($\alpha = 0.95$) önce ve sonra zamana bağlı olarak frekansındaki değişimler görülmektedir. İşaret çerçevelenmeden önce önvurgulama işlemi uygulanmaktadır. Şekil 6 (b) den görüleceği üzere şekil 6 (a)'ya göre yüksek frekanslı bileşenler daha belirginleşmektedir.

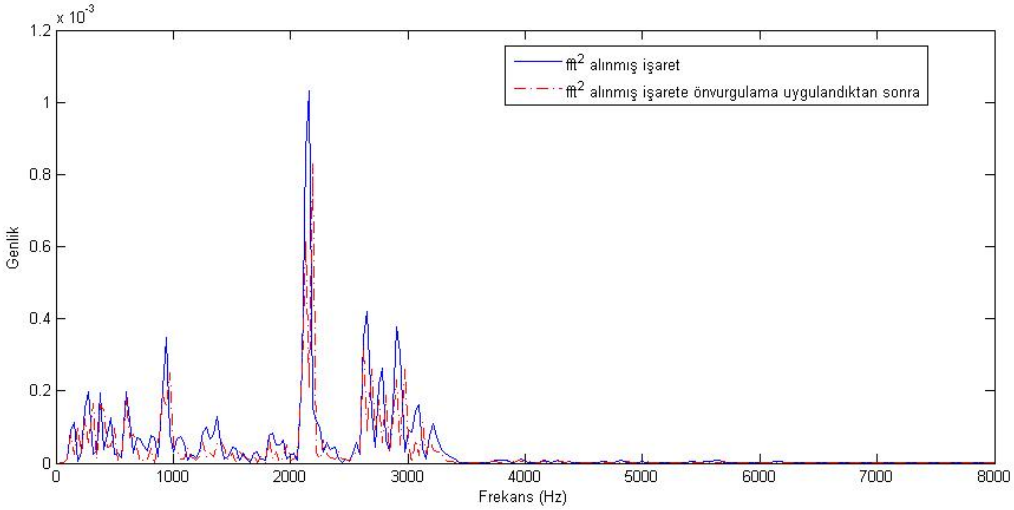


Şekil 6:
Bir cümlenin birinci dereceden süzgeçten ($\alpha = 0.95$) (a) geçirilmeden (b) geçirildikten sonra zaman-frekans değişimi

Şekil 7'de ise güç spektrumu alınmış işarete ön vurgulanma uygulanması durumunda değişim görülmektedir. Şekil 7 (b) ile 7 (c) karşılaştırıldığında işaretin düşük frekanslı bileşenlerin genliğinin zayıflatıldığı görülmektedir. Şekil 8'de işaretin ön vurgulamasız 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 7:
(a) Yirmi msn uzunluğunda bir konuşma parçası (b) bu konuşma parçasının $|FFT|^2$ spektrumu (c) spektrumu alınmış işaretin ön vurgulanmış hali



Şekil 8:
İşaretin $|FFT|^2$ alınmış hali üzerindeki ön vurgulamanın etkisi

2.5. Süzgeç Dizileri

Mel ölçek kepstrum 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 ϵ [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 $f_{co}=0$ ve $f_{cl}<f_s/2 \ \forall \ l$ olarak ifade edilir. Süzgeç dizileri, denklem 14'deki gibi ifade edilir.

$$F_l[k]=\begin{cases}((\frac{k}{N})f_s-f_{cl-1})/(f_{cl}-f_{cl-1}) & L_l \le k \le C_l \\ f_{cl+1}-(\frac{k}{N})f_s)/(f_{cl+1}-f_{cl}) & C_l \le k \le U_l\end{cases} \tag{14}$$

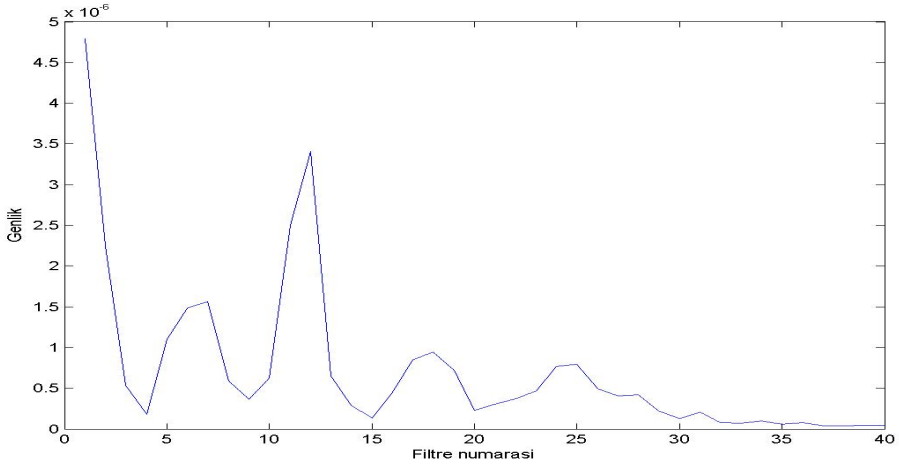
Burada $C_l=\frac{f_{cl}}{f_s}N,\ U_1=\frac{f_{cl+1}}{f_s}N$ ve $L_l=\frac{f_{cl-1}}{f_s}N$ olup l'inci süzgecin merkez, üst ve alt frekanslarıdır (Reynolds, 1992). Kullanılan üçgen süzgeç dizilerinin merkez frekansları, Mel ölçeğinde eşit olarak yerleştirilir.

Davis ve Mermelstein (1980) tarafından tanımlanan ilk 10 süzgecin merkez frekansları doğrusal olarak, sonraki 10 süzgeç ise logaritmik olarak yerleştirilmiştir. Tüm süzgeçler eşit genliğe sahiptir.

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.

Ön vurgulanan 1x512 boyutundaki konuşma işareti, 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 9'da görüldüğü gibi 1x40 boyutunda çıkış değeri elde edilir ve her süzgeç çıkışı bir değer ile temsil edilmektedir.



Şekil 9:
İşaretin süzgeç dizisinden geçirildikten sonraki durumu

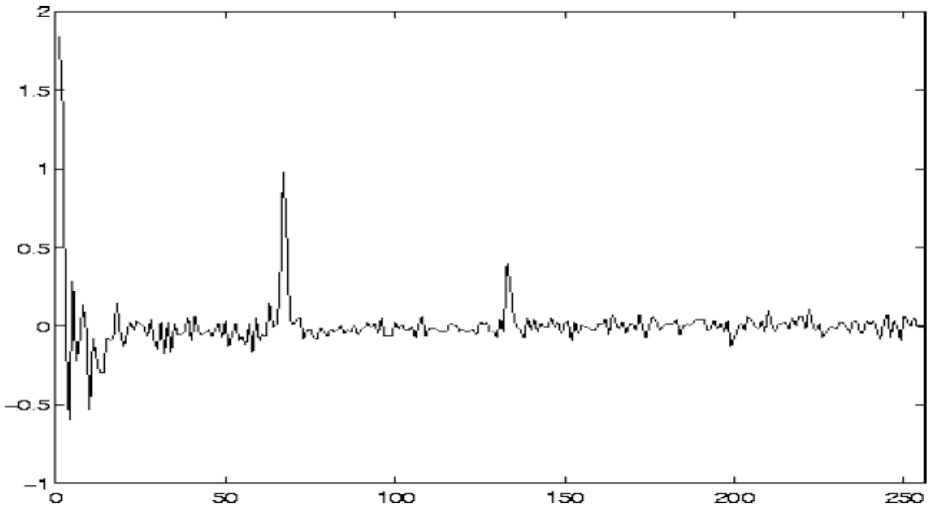
2.6. Logaritma Alma

Konuşma işaretinin süzgeç dizisinden geçirildikten sonra logaritması alınır. Spektrum'un logaritması alınmasının nedeni şu şekilde açıklanabilir. Konuşma işareti, $\left|S(e^{jw})\right|=\left|X(e^{jw})\right|\left|F(e^{jw})\right|$ 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 $\log|S(e^{jw})| = \log|X(e^{jw})| + \log|F(e^{jw})|$ 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 15 ile gösterilen işlem sonunda elde edilen katsayılar kepstrum katsayıları olarak adlandırılır.

$$ceps = FFT^{-1}(\log(|FFT(hamm.(512)\cdot * x(n))|)) \quad (15)$$

Burada $x(n)$, çerçevelenmiş konuşma parçasına karşılık gelmektedir. Şekil 10'da bir konuşma parçasının denklem 15 uygulandıktan sonra elde edilen şekil görülmektedir.



Şekil 10:
Konuşma parçasına denklem 15 uygulanması durumunda elde edilen kepstrum katsayıları

Kepstrum katsayılarına ait şekil 10'dan 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 katsayı (70. örnek civarı) perde periyodu hakkında bilgi vermektedir.

Konuşma spektrumu x, sıfıra yakınsadığı durumlarda $\log(x)$ eksi sonsuza yönelir. Logaritma fonksiyonu x'in küçük değerlerine karşı çok hassastır. Spektrumda düşük güce sahip yerler (SNR'ın düşük olduğu) en hassas kısımlardır. Spektrumun küçük değerleri için $\log(x)$ yerine $\log(x+c)$ kullanılır (Hunt, 1999). Burada c küçük bir sabittir. Kinnunen (2003), konuşmacı tanıma deneylerinde $\log(x)$ değerine 1 sabitini eklemektedir.

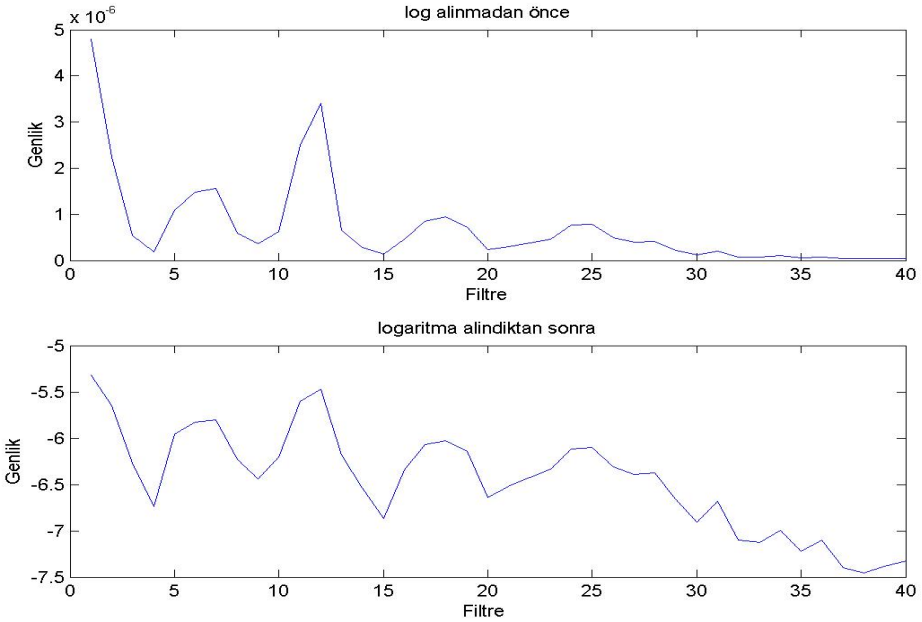
Konuşma spektrumu x, sıfıra yakınsadığında oluşan problemlerden dolayı konuşma güç spektrumunun logaritma fonksiyonu yerine güç $(.)^{\gamma}$ veya kök $(.)^{1/\gamma}$ fonksiyonu gösterimi önerilmiştir (Lim, 1979). Ses şiddeti ve algılanan duyma düzeyi arasındaki doğrusal olmayan bir ilişkinin varlığına dayanılarak bu ilişkinin modellemesi yapılır. Algılanan sesin düzeyi sesin şiddetinin küp köküne eşittir. Spektrumun küp kökünü alma gürültü içeren konuşma tanıma deneylerinde (Alexandre ve Lockwood, 1993; Chu ve diğ., 2003) logaritma fonksiyonuna göre daha iyi sonuçlar elde edilirken, temiz konuşma için düşük başarım elde edilmiştir. Sarıkaya ve diğ. (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 vektör nicemleme 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.

l. süzgec için logaritmik enerji çıkışı denklem 16'da görüldüğü gibi $mfb(l)$ olarak ifade edilir.

$$mfb(l) = \log\left(\frac{1}{A_l}\sum_{k=L_1}^{U_1} F_l[k]X[k]\right) \tag{16}$$

Burada A_l süzgeçlerin bant genişliğine bağlı olarak kullanılan normalizasyon katsayısı olup $A_l = \sum_{k=L_1}^{U_1} F_l[k]$ olarak tanımlanır. Sonuç olarak elde edilen vektöre Mel-süzgeç dizisi vektörü denir. Logaritma alarak, dinamik sıkıştırma yapılıp, öznitelik vektörleri, dinamik değişimlere karşı daha az hassas olmaktadır (Claudio, 1999). Şekil 11'de işaretin logaritması alındığında işaretteki değişimler görülmektedir.



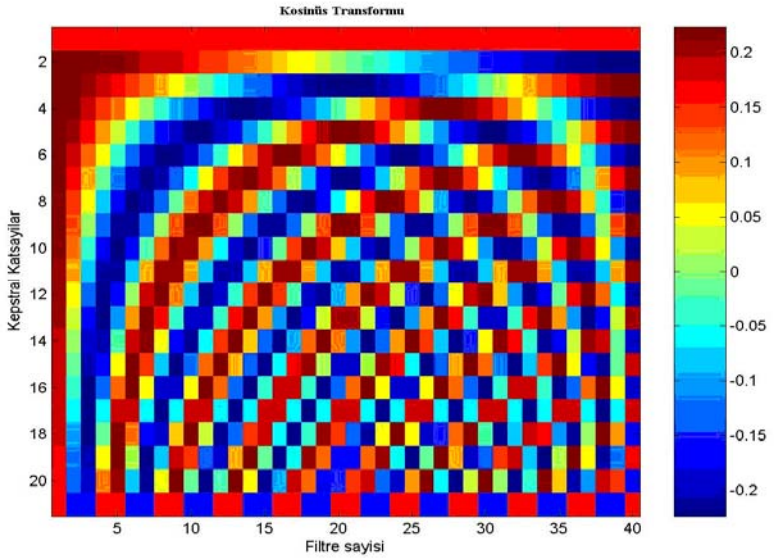
Şekil 11:
İşaretin süzgeç çıkışı ve logaritmalı hali

2.7. Ayrık Kosinüs Dönüşümü (AKD)

MFCC elde edilmesinde en son olarak kepstrum katsayıları hesaplanır. Kepstral gösterimi ile kayıt ve iletim ortamından dolayı oluşan spektral şekil değişimleri kaldırılır. Ayrıca kepstral katsayılar, yüksek derecede istatiksel bağımsızlık gösterip genlik spektrum gösteriminden daha yüksek tanıma oranı verirler. Gerçel kepstrum, logaritmik genlik spektrumun ters fourier dönüşümü olarak tanımlanıp, gerçel işaretler için kosinüs dönüşümü kullanılarak hesaplanır. Mel frekansı kepstrum katsayıları, *MFCC(i)*, süzgeç çıkışlarından denklem 17'deki gibi hesaplanır.

$$MFCC(i) = \frac{1}{FS}\sum_{l=1}^{FS} mfb(l)\cos\left(i\left(l-\frac{1}{2}\right)\frac{\pi}{FS}\right), \qquad i = 1,...,FS-1. \tag{17}$$

Mel ölçek süzgeç sayısı 40, kepstrum katsayı sayısı 21 için ayrık kosinüs dönüşümü şekil 12'de görülmektedir. Şekilde verilen renk ölçeğine göre matrisin aldığı değerler görülmektedir.



Şekil 12:
Ayrık kosinüs dönüşümü

2.8. Gauss Karışım Modeli (GKM)

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 yansıtıp, konuşmacı kimlik modellenmesinde kullanılır (Reynolds, 1992). Ayrıca Gauss karışım yoğunluğu, bir konuşmacıdan alınan sözcüklerle gözlemlerin uzun süreli dağılımında düzgün bir yaklaşım sağlamaktadır (Bhattacharyya ve diğ., 2001).

Bir Gauss karışım yoğunluğu, *M* bileşenli yoğunlukların toplamının ağırlıklandırılması olup denklem 18'deki gibi ifade edilir.

$$p(\vec{x}/\lambda)=\sum_{i=1}^{M} w_i b_i(\vec{x}) \tag{18}$$

Burada $\vec{x}$, *D* boyutlu rastgele değişen vektör, $b_i(\vec{x})$, bileşen yoğunlukları $(i=1,.....,M)$ ve w_i, karışım ağırlıklarıdır. Her bir bileşen için *D* boyutlu Gauss fonksiyonu denklem 19'da görülmektedir.

$$b_i(\vec{x})=\frac{1}{(2\pi)^{D/2}\left|\sum_i\right|^{1/2}}\exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)'\sum_i{}^{-1}(\vec{x}-\vec{\mu}_i)\right\}, \tag{19}$$

Burada $\vec{\mu}_i$ ortalama vektör ve $\sum_i$ 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 20'deki gibi ifade edilmektedir.

$$\lambda=\{w_i,\vec{\mu}_i,\sum_i\} \qquad i=1,...,M \tag{20}$$

3. DENEYLER

Öznitelik vektörü oluşturma sonucu ortaya çıkan ve konuşmacıyı tanımlayıcı özelliği olan öznitelik vektörleri, sınıflandırma aşamasında konuşmacı için bir model oluşturmakta kullanılır. Konuşmacıların modellenmesinde Gauss karışım modeli kullanılmaktadır. Sınıflandırma için kullanılan tekniklerin dayandığı ortak durum, aday konuşmacı ile referans alınan modellerin benzerliğin hesaplanmasıdır.

Konuşmacı tanıma sisteminin en iyi başarıyı vermesi için öznitelik vektörlerin oluşturulduğu tüm adımların, konuşmacı tanıma üzerine etkisi incelenmektedir. Bu adımların her birinde sabit olarak şu ayarlamalar yapılmıştır. TIMIT ve NTIMIT veritabanlarının test dizinindeki 168 kişinin her birine ait 10 cümleden 8'i (yaklaşık 24 saniye) eğitim için, kalan 2 cümle (yaklaşık 3'er saniye) ise ayrı ayrı kullanılarak toplam 336 test yapılmıştır.

Beklentinin maksimumlaştırılması (BM) parametre kestirimi, benzerlik fonksiyonunun maksimum olduğu model parametre değerlerinin bulunmasıdır. BM algoritmasının temelindeki fikir, ilk model başlangıcının yeni model $\overline{\lambda}$, $P(X|\overline{\lambda}) \geq p(X|\lambda)$ olarak kestirilmesidir (Dempster, 1977). Eski model yerine yeni model yerleştirilir ve bu işlem, yakınsama süreci eşik değerine ulaşılana kadar devam edilir. Deneylerde eğitim için BM algoritması kullanılıp, Gauss karışım sayısı 32 alınmaktadır.

BM algoritması, her bir özyineleme benzerlik fonksiyonunun artışını sağlar. Özyineleme sayısı, pratik anlamda benzerlik fonksiyonunun yeterli oranda yakınsayıp yakınsamadığını bulmak için gereklidir. Özyineleme sayısının 15 alınması yeterli yakınsamayı sağlamaktadır. Modelin eğitimi esnasında oluşan model değişinti değerlerinin sıfıra yönelmesini önlemek için sabit değişinti sınırlaması uygulanır. Değişinti sınırlaması olarak $\sigma^2_{\min}$ =0.01 değeri kullanılmaktadır (Reynolds, 1992).

Model başlangıç değerleri için ilk olarak Linde Buzo Gray (LBG) algoritması (Linde ve diğ., 1980) kullanılıp elde edilen değerler k-ortalama algoritması ile model başlangıç değerleri olarak belirlenmektedir. 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).

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 alınır. Her bir çerçeveye karşılık olarak TIMIT veritabanı için 24, NTIMIT veritabanı için 20 boyutlu öznitelik vektörleri kullanılmaktadır. Bu şartlarda şekil 1'de belirtilen öznitelik vektörü elde etme adımlarının her biri değiştirilerek konuşmacı tanıma üzerine etkileri incelenmektedir.

3.1. Çerçeveleme

En ideal çerçeveleme süresi veritabanlarına ve kullanılan yöntemlere bağlı olarak değişmektedir. Yukarıda öznitelik vektörü elde edilmesinde kullanılan parametreler için çerçeve sürelerine bağlı olarak elde edilen konuşmacı tanıma oranları Tablo I'deki gibidir. Deneyde TIMIT ve NTIMIT veritabanları için hamming pencereleme fonksiyonu kullanılmaktadır.

Tablo I. Çerçeveleme sürelerinin konuşmacı tanımaya etkisi (%)

Veritabanları	Çerçeveleme süreleri (msn.)			
	30	25	20	15
TIMIT	**99.4**	**99.4**	**99.4**	**99.4**
NTIMIT	67.9	67.9	**69.9**	68.1

Tablo I'den görüleceği üzere TIMIT veritabanı için çerçeveleme süresi değişimi konuşmacı tanıma başarımını değiştirmez iken, NTIMIT veritabanı için 20 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 diğ., 2001), ardışıl çerçeve parametreleri arasında öklit uzaklığı hesabının ölçülmesi (Zhu ve Alwan, 2000) gibi değişik metotlar ile uyarlamalı çerçeve uzunluğu kullanılabilir.

3.2. Pencereleme

Konuşmacı tanıma sisteminde başarımı 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 bir önceki deneyde en yüksek sonuç alınan değer, 20 msn, alınmıştır. Pencereleme fonksiyonlarına bağlı olarak elde edilen konuşmacı tanıma oranları tablo II'deki gibidir.

Tablo II. Pencereleme fonksiyonlarına bağlı olarak konuşmacı tanıma oranları (%)

Pencereleme fonk.	Veritabanları	
	TIMIT	NTIMIT
Hamming	99.4	**69.9**
Hanning	99.4	69.3
Blackman	**99.7**	68.4
Gauss	**99.7**	67.6
Dikdörtgen	99.1	64.9
Üçgen	99.4	67.6

Tablo II'den görüleceği üzere TIMIT veritabanı için Gauss ve Blackman pencereleme fonksiyonları, NTIMIT veritabanı için ise Hamming pencereleme fonksiyonu kullanılarak en yüksek konuşmacı tanıma başarımı elde edilmiştir. En düşük tanıma başarımı pencereleme uygulanmama durumuna karşılık gelen dikdörtgen pencereleme ile elde edilmiştir.

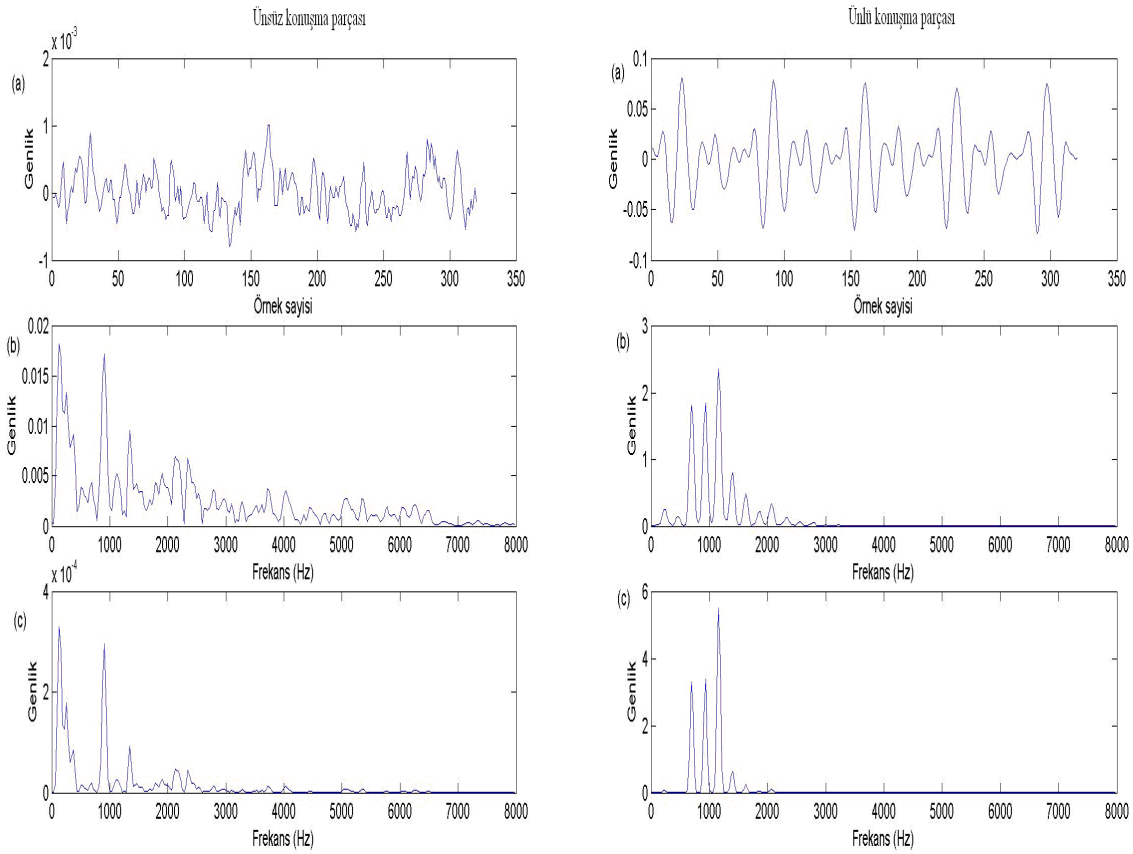
3.3. Hızlı Fourier Dönüşümü

Bir önceki deneyde belirtilen öznitelik vektörü üretim, eğitim ve test şartlarında, konuşma işaretinin genlik ve güç spektrumu konuşmacı tanımaya etkisi incelenecektir. Her iki veritabanı içinde çerçeveleme süresi 20 msn alınıp, TIMIT veritabanı deneyleri için Gauss, NTIMIT veritabanı deneyleri için hamming pencereleme kullanılmaktadır. FFT kuvvetlerinin konuşmacı tanımaya etkisi tablo III'de görülmektedir.

Tablo III. FFT kuvvetlerinin konuşmacı tanımaya etkisi (%)

Veritabanları	FFT kuvveti	
	$\lvert FFT \rvert$	$\lvert FFT \rvert^2$
TIMIT	99.4	**99.7**
NTIMIT	66.1	**69.9**

Tablo III'den görüleceği üzere TIMIT ve NTIMIT veritabanı için konuşma işaretinin güç spektrumunu almak, genlik spektrumu kullanılmasına göre daha yüksek konuşmacı tanıma başarımı elde edilmesini sağlamaktadır. Çünkü $\lvert FFT \rvert^2$ 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. Şekil 13'de, 20 msn. uzunluğunda ünsüz ve ünlü konuşma parçalarının $\lvert FFT \rvert$ ve $\lvert FFT \rvert^2$ alınmış hali görülmektedir. Şekillerden görüleceği üzere işaretteki düşük genlikli gürültü bileşenlerinin etkinliği azaltılmaktadır.



Şekil 13:

Yirmi msn uzunluğunda (a) konuşma parçası (b) bu konuşma parçasının $|FFT|$ *(c)* $|FFT|^2$ *alınmış hali*

3.4. Önvurgulama

TIMIT ve NTIMIT veritabanları kullanılarak önvurgulamanın konuşmacı tanımaya etkisi incelenecektir. Önvurgulama süzgeci olarak denklem 13 kullanılıp ve $\alpha = 0.95$ alınmaktadır. Konuşma işaretine tablo IV'de görülen şekillerde ön vurgulama uygulanmaktadır. Deneyde her iki veritabanı içinde çerçeveleme süresi 20 msn alınıp TIMIT veritabanı deneyleri için Gauss, NTIMIT veritabanı deneyleri için hamming pencereleme kullanılmaktadır. Konuşma işaretinin güç spektrumu alınmaktadır. Bir önceki deneyde verilen konuşmacı tanıma sistemi parametrelerine bağlı olarak elde edilen konuşmacı tanıma oranları tablo IV'de görülmektedir.

Tablo IV. Önvurgulamanın konuşmacı tanıma üzerine etkisi (%)

Önvurgulama uygulama şekilleri	Veritabanları	
	TIMIT	NTIMIT
Çerçevelemeden önce	99.4	**70.2**
Çerçevelemeden sonra	99.4	60.1
Pencerelemeden sonra	99.4	69.1
Güç spektrumu alındıktan sonra	**99.7**	67.3
Önvurgulama yok	**99.7**	69.9

Tablo IV'den görüleceği üzere TIMIT veritabanı için önvurgulama uygulanmadığı durumda ve güç spektrumu alındıktan sonra önvurgulama uygulandığında en yüksek konuşmacı tanıma başarı-

mına ulaşılmıştır. NTIMIT veritabanı için önvurgulamanın çerçevelemeden önce uygulandığı durumda, en yüksek konuşmacı tanıma başarımı elde edilmiştir.

3.5. Mel Ölçekte Dizilmiş Süzgeç Dizileri

Davis ve Mermelstein (1980) ve Slaney (1998) tarafından tanımlanan Mel ölçek süzgeç dizilerinin konuşma tanımadaki başarımları karşılaştırılacaktır. TIMIT veritabanında (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. NTIMIT 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.

Deneyde konuşmaların çerçeveleme süresi 20 msn alınıp TIMIT veritabanı deneyleri için Gauss, NTIMIT veritabanı deneyleri için hamming pencereleme uygulanmaktadır. Pencerelenen sesin güç spektrumu alınıp Davis ve Mermelstein (1980) ve Slaney (1998) tarafından tanımlanan 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. Ön vurgulama, TIMIT veritabanı için güç spektrumu alındıktan sonra, NTIMIT veritabanı için çerçevelemeden önce uygulanmaktadır. Bu parametrelere bağlı olarak elde edilen konuşmacı tanıma oranları tablo V'de görülmektedir.

Tablo V. İki farklı Mel ölçek için konuşmacı tanıma oranları (%)

Veritabanları	Mel Ölçek	
	Davis ve Mermelstein (1980)	Slaney (1998)
TIMIT	99.4	**99.7**
NTIMIT	67.9	**70.2**

Tablo V'den görüleceği üzere TIMIT ve NTIMIT veritabanında Slaney (1998)'in önerdiği Mel ölçekte dizilmiş süzgeç dizileri, Davis ve Mermelstein (1980) tarafından tanımlanan Mel ölçeğe göre daha iyi başarım sağlamaktadır. Slaney (1998)'in önerdiği Mel ölçekteki süzgeç dizlerinin daha iyi olmasının temel nedeni, süzgeçlerin bant genişliklerinin daha dar olması ve bu şekilde orta ve yüksek frekans bandının daha iyi modellenmesidir.

3.6. Logaritma Alma

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 başarım artışı elde edilen süzgeç çıkışının (1/3) ve (0.008) kuvvetlerinin alınmasının her iki veritabanı için sonuçları incelenecektir (Alexandre ve Lockwood, 1993; Chu ve diğ., 2003; Sarıkaya ve diğ., 2001).

Deneyde her iki veritabanı içinde çerçeveleme süresi 20 msn alınmaktadır. TIMIT veritabanı deneyleri için Gauss, NTIMIT veritabanı deneyleri için hamming pencereleme uygulanmaktadır. Pencerelenen sesin güç spektrumu alınıp Slaney (1998) tarafından tanımlanan Mel ölçekte yerleştirilmiş üçgen süzgeç dizilerinden geçirilmiştir. Ön vurgulama, TIMIT veritabanı için güç spektrumu alındıktan sonra, NTIMIT veritabanı için çerçevelemeden önce uygulanmaktadır. Tablo VI'da bu parametrelere bağlı olarak elde edilen konuşmacı tanıma oranları görülmektedir.

Tablo VI. Süzgeç çıkışlarının logaritması ve kuvvetleri alınmasının tanımaya etkisi (%)

	Veritabanları	
	TIMIT	**NTIMIT**
$\log(x)$	**99.7**	**70.2**
$x^{1/3}$	86.6	15.2
$x^{0.008}$	35.1	13.1
x	13.7	3.3

Tablo VI'dan 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.

4. SONUÇLAR

Bu makalede, TIMIT ve NTIMIT veritabanlarının kullanıldığı bir konuşmacı modeli, MFCC parametre değişimlerine karşı incelenip, en yüksek tanıma oranını verecek ideal parametreler belirlenmiştir. TIMIT veritabanı, gürültü olmayan ortamda mikrofon ile veriler toplandığından dolayı temiz bir veritabanı olarak tanımlanmaktadır. NTIMIT veritabanı, konuşmalar telefon ortamından iletildiğinden dolayı telefon ahizesi ve iletim hattının etkilerini içermektedir. TIMIT ve NTIMIT veritabanının bu farklılıklarından dolayı ideal parametreler, bu veritabanlarına bağlı olarak farklılık arz etmektedir. Reynolds (1992), tarafından önerilen şekil 1'deki öznitelik vektörü elde etme blok diyagramında, önvurgulamanın yeri güç spektrumu alındıktan sonra gösterilmektedir. Tablo IV'den görüleceği üzere NTIMIT veritabanı için en yüksek başarım, önvurgulama çerçevelemeden sonra yapıldığı durumda gerçekleşmektedir. Dolayısıyla şekil 1'deki bazı parametreler veritabanlarına bağlı olarak değişmektedir.

Deneylerden elde edilen sonuçlardan konuşmacı tanıma etkisini arttırıcı, ideal Mel frekansı kepstrum katsayı parametreleri şunlardır. TIMIT veritabanı için öznitelik vektörleri çıkartılırken çerçeve süresi değişimi tanıma başarımını değiştirmemektedir. Çerçevelere Gauss veya blackman pencereleme fonksiyonu kullanılması, çerçevelerin genlik spektrumu yerine güç spektrumu alınması, Slaney'in (1998) önerdiği Mel ölçek kullanılması ve önvurgulamanın uygulanmadığı veya güç spektrumu alındıktan sonra uygulanması durumlarında en iyi konuşmacı tanıma başarımı elde edilmektedir.

NTIMIT veritabanında öznitelik vektörü elde edilmesinde; konuşma 20 msn'lik çerçevelere ayrılması, hamming pencereleme fonksiyonu kullanılması, çerçevelerin genlik spektrumu yerine güç spektrumu alınması, Slaney'in (1998) önerdiği Mel ölçek kullanılması ve işarete çerçevelemeden önce önvurgulama uygulanması durumlarında en iyi konuşmacı tanıma başarımı elde edilmiştir.

5. KAYNAKLAR

1. 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.
2. 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.
3. Besacier, L. J.F. Bonastre, (1998) Frame Pruning for Automatic Speaker Verification, Proc. EUSIPCO'98, Greece, September 8-11, Vol.1, p. 367-370.
4. 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.

5. Claudio, B. and L. P. Ricotti (1999) Speech Recognition Theory and C++ Implementation. John WILEY&Sons, Ltd, p. 125-137.
6. Davis, S. B. and P. Mermelstein (1980) Comparison of parametric repesentations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-28, p. 389-397.
7. Dempster, A. N. Laird and D. Rubin (1977) Maximum likelihood from incomplete data via the EM alorithm, Journal of the Royal Statistical Society, vol. 39, p. 1-38.
8. Furui, S. (1989) Digital Speech Processing, Synthesis, and Recognition. M. Dekker Inc.
9. Ganchev, T. (2005) Speaker Recognition , PhD thesis, Dept. of Electrical and Computer Engineering, University of Patras, Greece. p. 61-82.
10. Huang, X., Acero, A., Hon, H.-W., (2001) Spoken Language Processing: a Guide to Theory, Algorithm, and System Development. Prentice-Hall, New Jersey.
11. Hunt, M. J., (1999) Spectral Signal Processing for ASR. IEEE ASRU Workshop, Colorado, Keystone, U.S.A.
12. Jankowski, C., Kalyanswamy, A., Bason, S. and Spitz, J. (1990) NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database, IEEE International Conference on Acoustic, Speech and Signal Processing , p. 109-112
13. Karpov, E. (2003) Real-Time Speaker Identification, Master thesis, University of Joensuu, Department of Computer Science p. 17-26.
14. Kinnunen, T. (2003) Spectral Features for Automatic Text-independent Speaker Recognition, Ph.D. thesis, University of Joensuu, Department of Computer Science p. 49-115.
15. Lim, J. S. (1979) Spectral Root Homomorphic Deconvolution system, IEEE Trans. on ASSP, Vol. ASSP-27, No. 3.
16. Linde, Y., A. Buzo., R. M. Gray. (1980) An Algorithm for Vector Quantization, IEEE Trans. Communications, Vol. 28, No. 1, p. 84-95.
17. Lincoln, M. (1999) Characterization of Speakers for Improved Automatic Speech Recognition. Thesis Doctor of Philosophyin the School of Information Systems, University of East Anglia, Norwich. p. 18-23.
18. Matsui, T. and S. Furui (1995) Speaker Recognition Tecnology. NNT Review, Vol. 7, No. 2, p. 40-48.
19. Rabiner, L. R. and B. H. Juang (1993) Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs.
20. Reynolds, D.A. (1992) A Gaussian Mixture Modeling Approach to Text Independent Speaker Identification. Ph.D. thesis, Georgia Inst. of Technology.
21. Reynolds D. A., and Rose, R.C. (1995) Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models, IEEE Trans. Speech Audio Proc., 3, (1), p. 72–83.
22. Sarıkaya, 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.
23. Sarma, S. (1997) A Segment-based Speaker Verification System S.M. thesis, MIT Department of Electrical Engineering and Computer Science, p. 84-86.
24. Slaney, M. (1998) Auditory Toolbox: A MATLAB Toolbox for Auditory Modeling Work Technical Report, Interval Research Corproation, p. 29-32.
25. Wildermoth, B. R. (2001) Text Independent Speaker Recognition Using Source Based Features. Master of Philosophy, Griffith University, Australia, p. 21-29.
26. Wolf, J. (1972) Efficient Acoustic Parameters for Speaker Recognition. Journal of the Acoustical Society of America, vol. 51, no. 6, p. 2044-2056.
27. 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.
28. Zue, V., Seneff, S. And Glass, J., (1990) Speech Database Development at MIT: TIMIT and beyond, Speech Communication, p. 351-356

Makale 13.01.2009 tarihinde alınmış, 29.05.2009 ve 04.08.2009 tarihlerinde düzeltilmiş, 05.08.2009 tarihinde kabul edilmiştir.
İletişim Yazarı: Ö. Eskidere (oeskidere@uludag.edu.tr).