# BÜRÜNSEL ÖZELLİKLERİN KONUŞMACI TANIMA PERFORMANSINA ETKİSİ

*Ömer ESKİDERE*[*]
*Figen ERTAŞ*[**]

**Özet:** Bu makalede, **bürünsel** özniteliklerin gürültü içeren ortamlarda konuşmacı tanıma başarımına etkileri incelenmiştir. Bunun için, formant frekansı, sinyal enerjisi ve perde frekansı **bürünsel** özellikleri ve mel frekansı kepstrum katsayıları (MFCC) konuşma sinyalinden elde edilmiştir. Daha sonra her bir konuşmacı için özniteliklerin dağılımı Gauss karışım modeli ile modellenmiştir. Konuşmacı tanıma başarımı TIMIT ve NTIMIT veritabanları ile test edilmiştir. Gürültü ortamı NOISEX veritabanı kullanılarak oluşturulmuştur. Deneysel sonuçlar, enerjinin birinci türevi ve **formant** frekansları oranının ($F_3/F_2$), öznitelik vektörleriyle birlikte kullanılmasının konuşmacı tanıma hata oranını azalttığını göstermiştir. Ayrıca perde frekansının, gürültü ve telefon ortamının oluşturduğu bozulmalara karşı gürbüz bir öznitelik olduğu bulunmuştur.

**Anahtar Kelimeler:** Bürünsel özellikler, formant frekansı, enerji, perde frekansı, konuşmacı tanıma, Gauss karışım modeli.

### The Effect of Prosodic Features on Performance Speaker Identification

**Abstract:** In this paper, the effect of the prosodic features on the performance of the speaker identification system in the noisy environment is investigated. For this purpose, the prosodic features, formant frequency, signal energy and pitch frequency, and mel frequency cepstrum coefficients (MFCC) are extracted from the speech signal. And then the distribution of the features for each speaker is modeled by Gaussian Mixture Model (GMM). The speaker recognition is performed on the TIMIT and NTIMIT databases. The noisy environment is created using the NOISEX database. The experimental results showed that when first derivative of the energy and the ratio of the formant frequencies ($F_3/F_2$) are used in feature vector, the speaker identification error rate decreases. It is also founded particularly that the pitch frequency is the robust feature against noise and distortion in the phone lines.

**Key Words:** Prosodic features, formant frequency, energy, pitch frequency, speaker identification and Gaussian mixture model.

## 1. GİRİŞ

Konuşma işareti, konuşulan mesaj hakkında bilgi yanında konuşanın kimliği hakkında da bilgi taşır. Konuşma işareti konuşmacının psikolojik ve duygusal durumu, sağlığı ile sesin kaydedildiği ortam hakkında da bilgi içerir. Bu nedenle, farklı konuşmacıların konuşma sinyalleri arasında çok fazla değişiklik vardır ve daha da önemlisi aynı konuşmacının değişik zamanlarda kaydedilmiş konuşma sinyalleri arasında da farklılıklar bulunur (Reynolds, 1992).

Konuşanın kimliğini belirleme probleminde en önemli basamak, kişiyi diğerlerinden ayırt eden ses özelliklerinin çıkartılma aşamasıdır (Rose, 2001). Bürünsel özellikler, tek başlarına öznitelik olarak tanımlanabildiği gibi diğer öznitelikler ile birlikte de kullanılabilmektedirler. Bürünsel özellikler ile konuşmacıya özel vurgu, aksan ve ses perdesindeki yükselme ve alçalmalar ölçülür (Adami ve diğ., 2003, Shriberg, 2007). Genellikle sözcük süreleri, konuşmadaki durma süreleri ve sıklıkları, perde frekansı, formant frekansı ve enerji bürünsel özellik olarak kullanılmaktadır (Peskin ve diğ., 2003a).

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

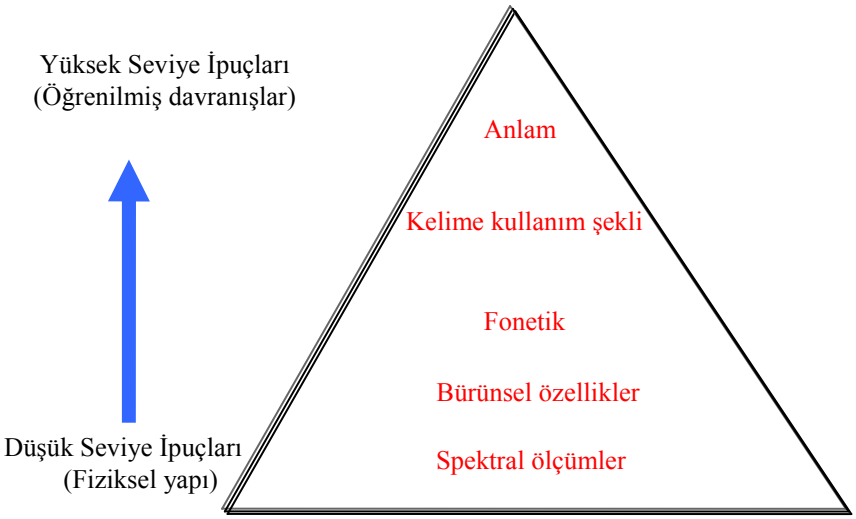
Bürünsel özellikler, konuşmacıya özel bilgilerin çıkartılmasında 1970'lerden itibaren metine bağımlı konuşmacı tanıma uygulamalarında kullanılmaya başlanmıştır (Atal, 1972). Özellikle perde frekansı, kişinin gırtlağının fiziksel özellikleri ile yakın ilişkili olduğundan sıklıkla konuşmacı tanıma çalışmalarında kullanılmıştır (Carey ve diğ., 1996, Chen ve Wang, 2004, Kinunen ve diğ., 2005). Formant frekansları ise konuşmacıdan çok konuşma ile ilgili özellikleri yansıttığından dolayı konuşmacı tanımada öznitelik olarak kullanılmamaktadır. Ancak Jankowski ve diğ. (1995), konuşmadaki ilk üç formant frekansının genlik ve frekans modülasyonunu öznitelik olarak Mel frekansı kepstrum katsayıları (MFCC) ile birlikte kullanmış ve NTIMIT veritabanı için tanıma başarımında artış elde etmiştir. Seddik ve diğ. (2004), Mezghani ve O'Shaughness (2005) ve Zeljkovic ve diğ. (2008), formant frekanslarını kullanarak, konuşmacı tanıma ve doğrulama başarımında artış elde etmiştir. Enerji ise bürünsel özellik olarak perde frekansı ile birlikte yaygın olarak kullanılmaktadır (Adami ve diğ., 2003, Reynolds ve diğ., 2003, Dehak ve diğ., 2007). Bu çalışmalarda logaritmik enerji, enerjinin türevi ve dağılımı bürünsel özellik olarak kullanılmaktadır.

Konuşma sinyalinden bürünsel özelliklerin çıkartılmasında pek çok çalışmada farklılık görülmektedir. Bir çalışmada her bir konuşma çerçevesinin perde frekansı çıkartılmaktadır (Reynolds 1995, Kinnunen ve diğ., 2005). Diğer birinde, otomatik konuşma tanıma kullanılarak fonem/hece sınırları belirlenmekte ve bürünsel özellikler çıkartılmaktadır (Shriberg ve diğ., 2005). Başka bir çalışmada ise doğrusal biçimlendirilmiş perde bölütlerinden (segment), dinamik perde frekansı değişimleri çıkartılmaktadır (Sönmez ve diğ., 1998, Reynolds ve diğ., 2003, Peksin ve diğ., 2003, Adami ve diğ., 2003). Ötümlü seslerin başlangıç veya bitiş noktaları bölütler olarak tanımlanmaktadır. Parçalara ayrılan konuşmalar, perde ve enerji değişimlerini göstermek üzere birkaç sınıfa bölünüp etiketlenmektedir. Bu etiketlenen kısımların *N*-gramı, bir konuşmacının karakteristiğini modellemede kullanılmaktadır (Adami ve Hermansky, 2003, Shriberg, 2007). Son yıllarda yapılan bir çalışmada, bürünsel özellik çıkartmada uygun hece dizilerinin belirlenmesinde otomatik konuşma tanıma yerine ötümlü sesin başlangıç noktalarını kullanan yöntem önerilmektedir. İki ötümlü sesin başlangıç noktaları arası hece benzeri bir bölge olarak tanımlanıp, perde frekansı ve enerji değişimleri her bir bölge için çıkartılmaktadır (Mary ve Yegnanarayana, 2008).

## 2. BÜRÜNSEL ÖZELLİKLER

Konuşanın kimliğinin belirlenmesinde, insan kulağının algı mekanizmasının anlaşılması önemli yer tutmaktadır. İnsanlar konuşanın kimliğini belirlemek için sözle ilgisi olmayan pek çok ipucu kullanmaktadır. Bu ipuçları iyi anlaşılmamakla 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. Örneğin kısa ses yolu, yüksek formant frekansı oluşturur. Ses tellerinin boyutlarındaki değişmeler ise ses tonundaki farklılıklar ile ilişkilidir (Rose, 2001). Bundan dolayı, doğuştan gelen bu özellikler bir konuşmacı için sabit olmakla beraber bazı sağlık durumlarından (burun boşluğunda değişikliğe neden olan nezle gibi) etkilenebilirler. Şekil 1'de konuşmanın taşıdığı bilgi seviyeleri ve ipuçları görülmektedir (Peskin ve diğ., 2003b).




*Şekil 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 diğ., 2004). Bu çalışmada düşük seviye bürünsel özelliklerden formant frekansları, enerji ve perde frekansının öznitelik olarak kullanıldığında konuşmacı tanımaya etkisi incelenecektir.

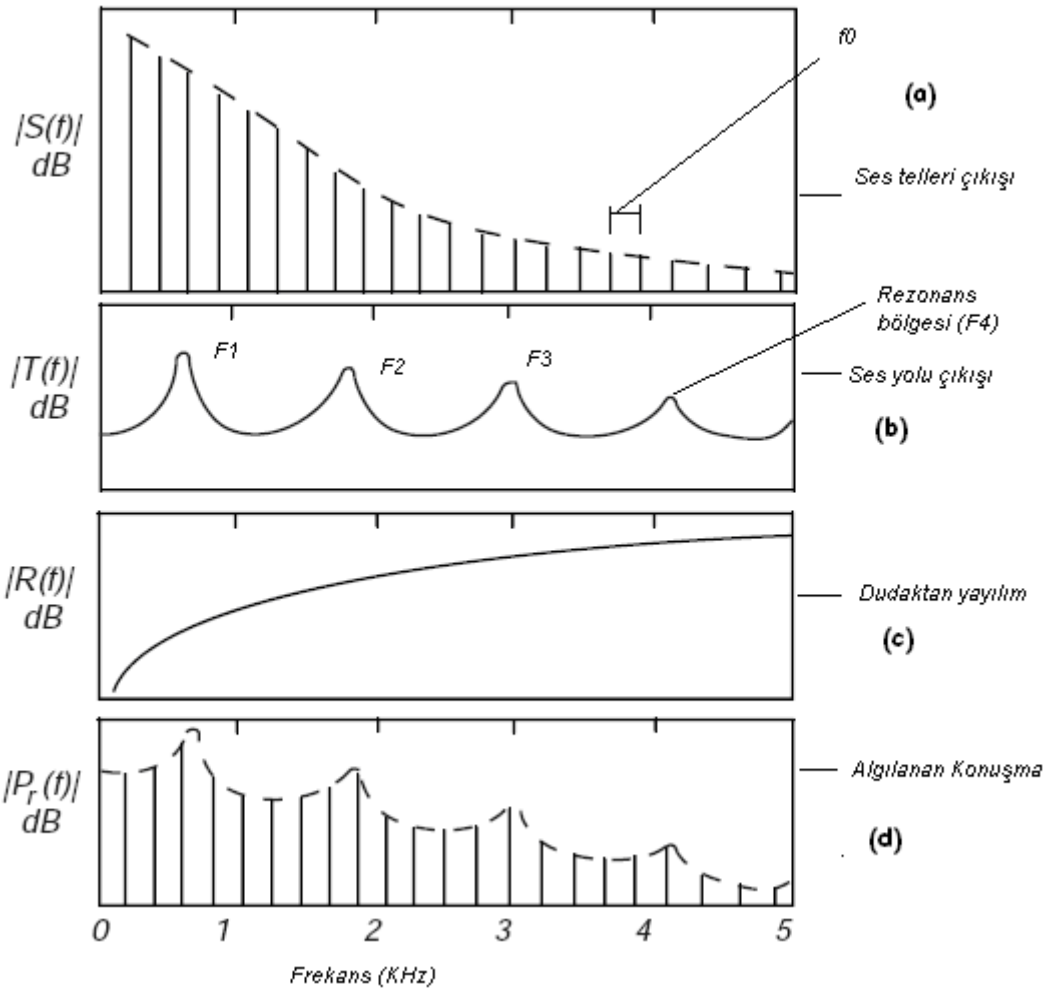
### 2.1. Formant Frekansları

Konuşma işareti $P_r(t)$, gırtlaktan çıkan dalga $S(t)$, $T(f)$ ve $R(f)$ kaskat süzgeçlerinden geçirilerek denklem 1'deki gibi üretilir (Park, 2002).

$$P_r(f) = S(f) \cdot T(f) \cdot R(f) \tag{1}$$

Burada $P_r(f)$ algılanan konuşma işaretinin spektrumunu, $S(f)$ boğaz kaynak spektrumu, $T(f)$ ses yolu transfer fonksiyonu ve $R(f)$ yayılım karakteristiğini ifade etmektedir. Şekil 2'de bu ifadelerin spektral gösterimi görülmektedir. Formant frekansları şekilde de görüleceği üzere ses yolu transfer fonksiyonu $T(f)$ ile doğrudan ilişkilidir.

Akustik teoride ses yolu, değişik 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. Ötümlü 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 (Rose, 2001).



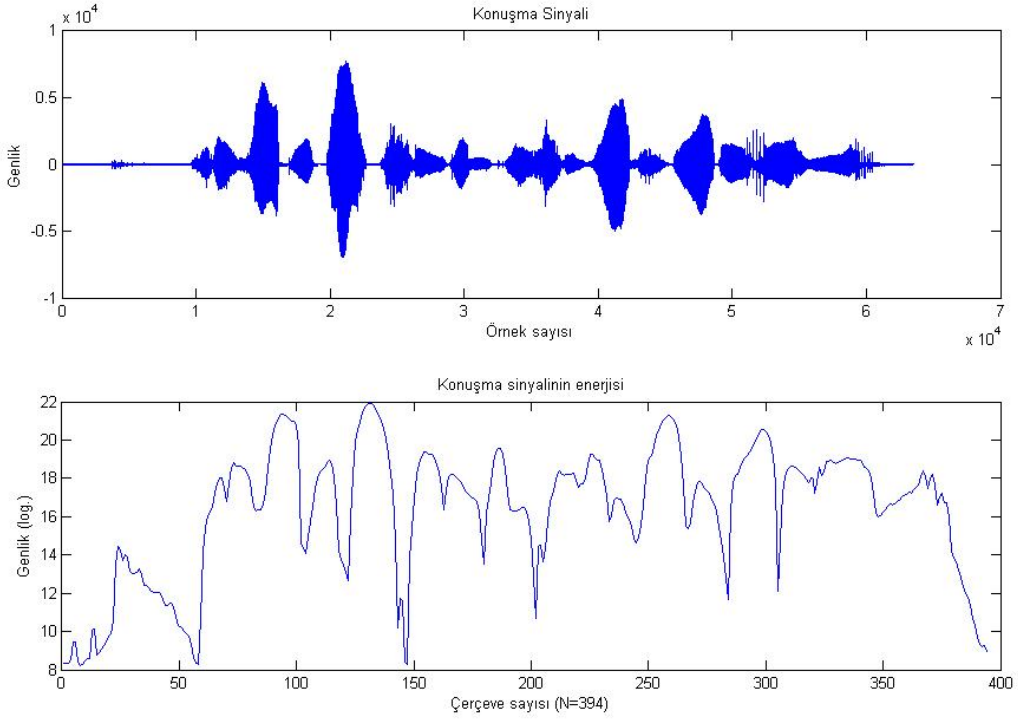

*Şekil 2:*
*Bir konuşma işareti için (a) ses telleri çıkışı, (b) ses yolu çıkışı, (c) dudaktan yayılım karakteristiği, (d) algılanan konuşma işareti (Park, 2002).*

**2.2. Enerji**

Bir işaretin enerjisi, zaman alanında işaretin genliklerinin karelerinin toplamı olarak ifade edilmektedir. Çerçevelenen konuşma sinyalinin enerjisi denklem 2'de elde edilir.

$$E = \sum_{i=1}^{N} x(i)^2 \tag{2}$$

Burada $N$ çerçeve uzunluğu olup, $x(i)$ konuşma işaretinin i. örnek değerini göstermektedir. Şekil 3'de bir konuşma sinyali ve logaritmik olarak enerji değeri görülmektedir.




*Şekil 3:*
*Konuşma sinyali ve N=394 için sinyalin enerjisi*

### 2.3. 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. Teager enerji operatörü ile ses yolundaki hava akışının doğrusal olmadığı kabul edildiği durumlarda anlık enerji değişimleri bulunur. Şekil 4'de ses yolunda hava akışı ve girdap oluşumu görülmektedir.

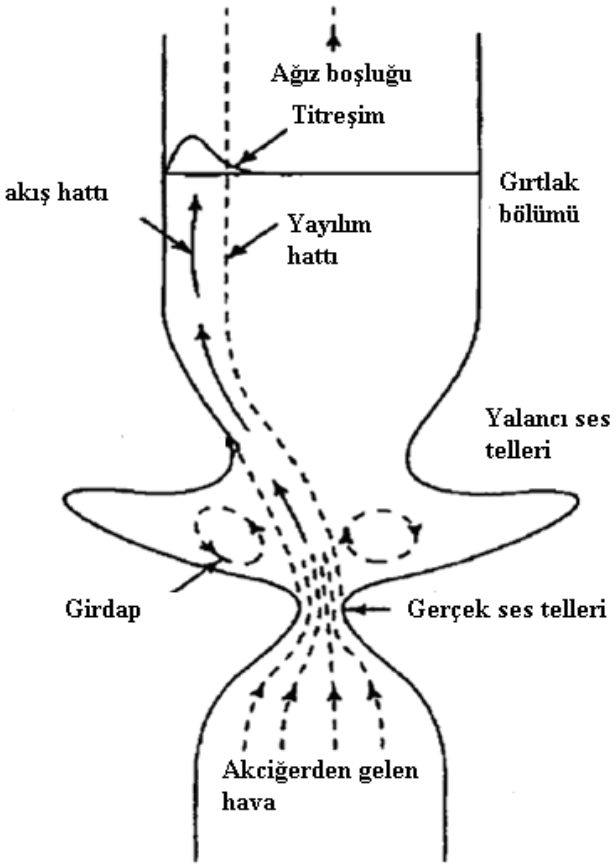
Teager, ses üretimi esnasında meydana gelen girdap-ses akışı etkileşmesinden dolayı akışın doğrusal olmadığını söylemiştir. Bu teori, akış mekanizması ve ses basıncı izlenerek desteklenmiştir (Hansen ve diğ., 1998). Kişinin içinde bulunduğu durum (stres, heyecan v.b.), fiziksel değişikliklere yol açtığı ve bu durumun ses yolunda girdap akış etkileşmesine neden olduğu varsayılmaktadır (Plumpe ve diğ., 1999). Şekil 4'de görülen doğrusal olmayan girdap-hava akışı etkileşiminin anlık enerji değişimi, Teager tarafından, Teager enerji operatörü olarak denklem 3'deki gibi ifade edilmektedir (Zhou ve diğ., 2001).

$$\Psi_c\left[x(t)\right] = \left(\frac{d}{dt}x(t)\right)^2 - x(t)\left(\frac{d^2}{dt^2}x(t)\right) \tag{3}$$

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 4'deki gibi tanımlanmaktadır (Hamila ve diğ., 1999).

$$\Psi\left[x(n)\right] = x^2(n) - x(n+1)\cdot x(n-1) \tag{4}$$

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



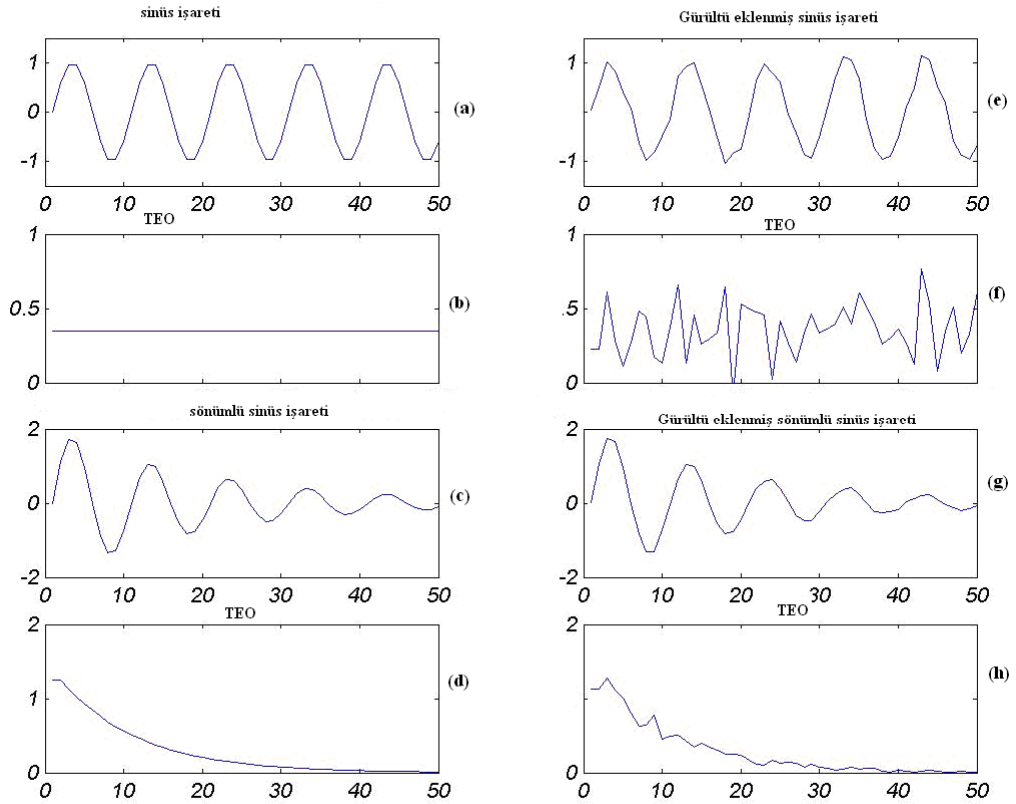

*Şekil 4:*
*Ses yolunda girdap-hava akış etkileşimi (Zhou ve diğ., 2001).*

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

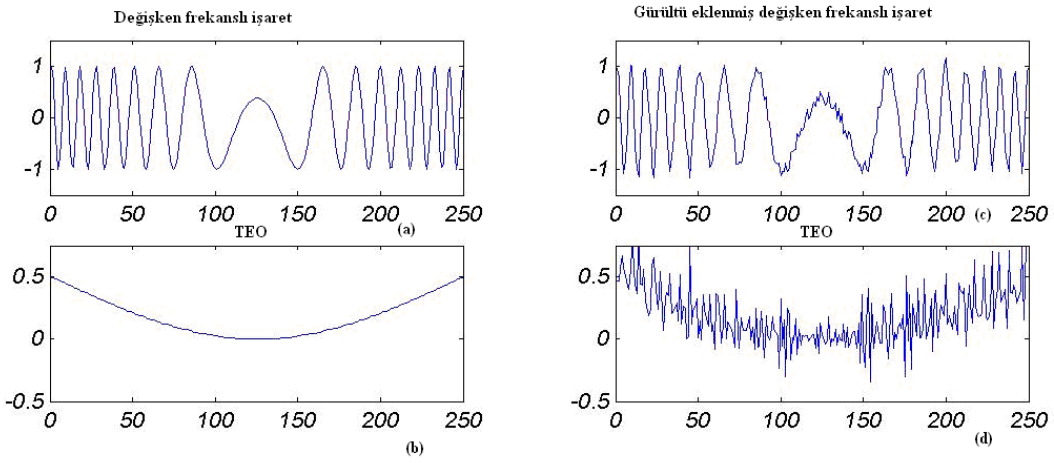
Aynı sinüs işaretine SNR = 40 dB gürültü eklendiğinde Şekil 5 (e), bu işarete TEO uygulanması durumunda Şekil 5 (f) elde edilmektedir. Şekil 5 (g) ve (h)'de gürültü eklenmiş sönümlü dalga ve TEO ile genlik değişimi izlenmiş işaret görülmektedir.

Şekil 6'da 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.

Şekil 6 (a)'da değişken frekanslı bir sinüs işareti, Şekil 6 (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 frekansının artması ile işaretin değeri de 0,5'e doğru artmaktadır. Şekil 6 (c)'de 40 dB gürültü eklenmiş değişken frekanslı sinüs işareti ve Şekil 6 (d)'de bu işarete TEO uygulanmış hali görülmektedir.




*Şekil 5:*
*Bir sinüs işareti ve bu işaretin TEO ile genliğinin izlenmesi (a, b) Sönümlü bir sinüs işareti ve bu işaretin TEO ile genliğinin izlenmesi (c, d) Gürültü eklenmiş bir sinüs işareti ve bu işaretin TEO ile genliğinin izlenmesi (e, f) Gürültü eklenmiş sönümlü sinüs işareti ve bu işaretin TEO ile genliğinin izlenmesi (g, h).*




*Şekil 6:*
*Değişken frekanslı bir sinüs işareti ve bu işaretin TEO ile frekansının izlenmesi (a, b) Gürültü eklenmiş değişken frekanslı bir sinüs işareti ve bu işaretin TEO ile frekansının izlenmesi (c,d)*

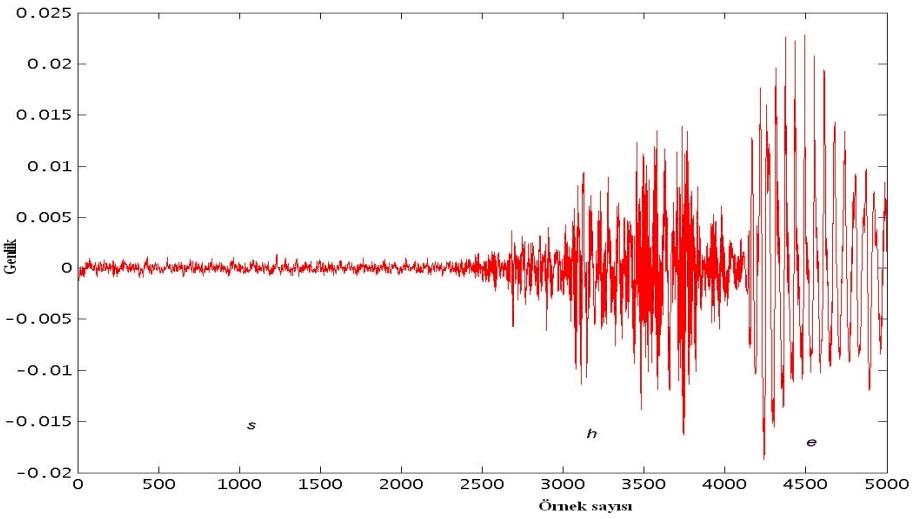
### 2.4. Perde Frekansı

Boğazda bulunan ses telleri, periyodik darbeler oluşturur ve bu darbelerin frekanslarına temel frekans adı verilir. Algılanan temel frekans perde frekansı olarak adlandırılmaktadır (Atal, 1974). Temel frekansın daha iyi anlaşılabilmesi için insan ses üretim mekanizmasının bilinmesi gerekmektedir. Ses üretiminde akciğerler, hava üreten bir enerji kaynağı gibi davranır. Kişinin konuşması ile birlikte hava akciğerlerden 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 ötümlü veya ötümsüz olacağını belirler. Ötümlü sesler için ses telleri hızlıca açılıp kapanarak havayı modüle eder.

Ses tellerinin titreşim hızı, tellerin gerginliğine ve kütlesine bağlıdır. Ötümsüz sesler için ses telleri periyodik olmayan akış olacak şekilde pozisyon alır ve konuşma içerisinde periyodik olmayan bileşenler oluşur. Şekil 7'de bir bayan konuşmacı tarafından söylenen "she" sözcüğüne ait ses sinyali gösterilmektedir. Şekilden görüleceği üzere konuşmanın ötümlü sese karşılık gelen bölgeleri periyodik iken, ötümsüz ses bölgeleri gürültü benzeri bir yapıya sahiptir. İşaretin periyodik bölgeleri konuşmanın perde frekansının ( $f_0$ ) hesaplanmasına yardımcı olur.

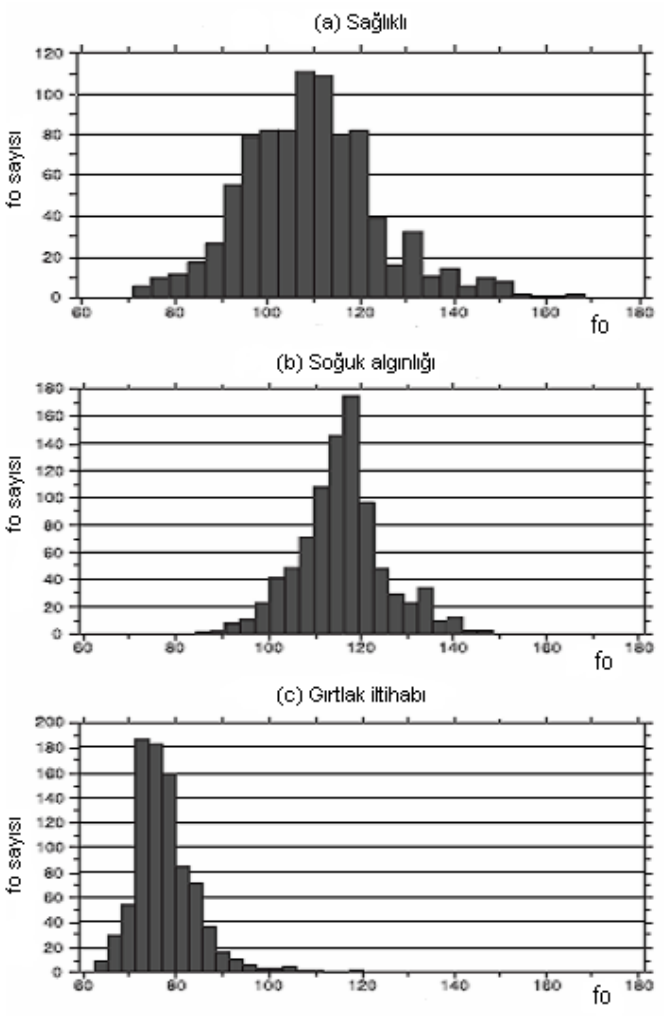
Ötümlü sesler için $f_0$, erkeklerde 50-250 Hz arasında değişirken, bayanlarda 120-400 Hz ve çocuklarda 150-450 Hz arasında değişir. Geniş değişim aralığı ve diğer faktörler $f_0$'ın % 100 doğrulukta ayırt edilmesini zorlaştırmaktadır.

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 (10-20 msn), konuşma işaretinin $f_0$'ı ve spektral karakteristiğin genliği değişmektedir. Çok hızlı değişen konuşmada, ses işareti 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 ötümlü ve ötümsüz 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).




*Şekil 7:*
*NTIMIT veritabanından alınmış bir bayana ait "She" sözcüğünün ses sinyali.*

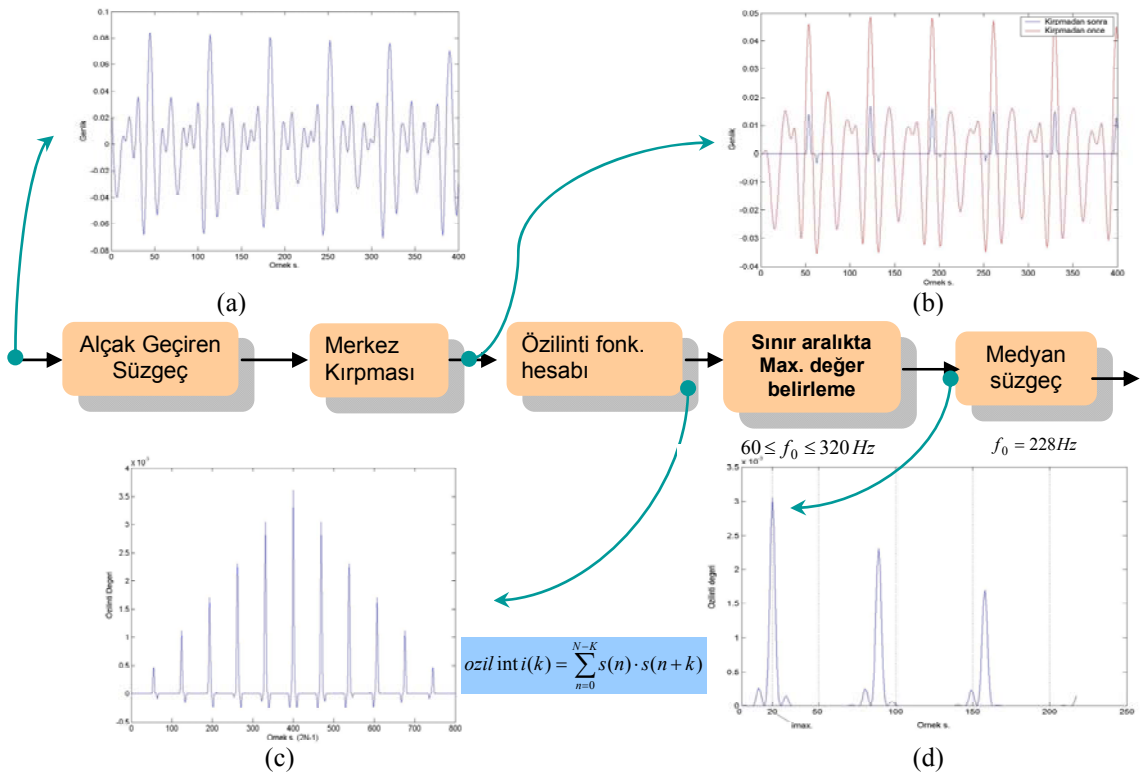
Perde frekansı kişinin içinde bulunduğu (stres, kızgınlık, üzüntü, sevinç v.b.) ruh haline bağlı olarak değişebilir. Perde frekansı soğuk algınlığı, gırtlak iltihaplanması gibi durumlardan etkilenir. Şekil 8'de değişik sağlık koşullarında bir kişinin $f_0$ histogramı görülmektedir (Rose, 2001).



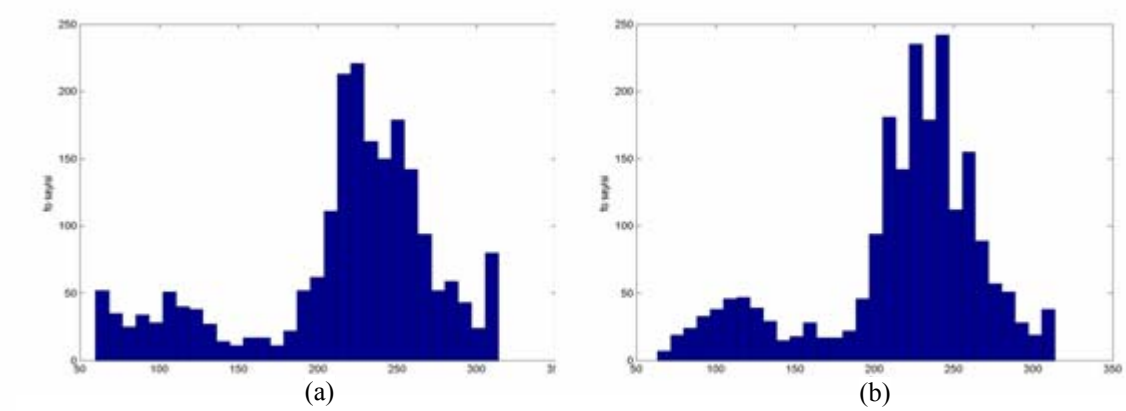

*Şekil 8:*
*Değişik sağlık koşullarında perde frekansının histogramı (Rose, 2001).*

Şekil 8'den görüleceği üzere sağlıklı bir kişinin perde frekansı 70–170 Hz arasında iken, aynı kişi soğuk algınlığına yakalanması durumunda perde frekans dağılımı 85–150 Hz, gırtlak iltihaplanması oluştuğunda ise 65–110 Hz arasında dağılmaktadır. Sağlık koşullarına bağlı olarak kişinin perde frekansı önemli oranda değişim göstermektedir.

Konuşmacı tanıma deneylerinde perde frekansı elde edilmesinde Şekil 9'da verilen özilinti yöntemi kullanılmıştır. Şekil 9'da görülen perde frekansı izleme algoritması üç temel adımdan oluşur: ön işleme, perde frekansı kestirimi ve son işlem. Ön işleme aşamasında çerçevelere ayrılan konuşma işareti kesim frekansı 900 Hz olan alçak geçiren süzgeçten geçirilip, periyodikliği arttırıcı merkez kırpması uygulanır. Perde frekansı kestiriminde özilinti fonksiyonu kullanılır. Son adımda konuşmacıların perde 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 10'da görülmektedir.




*Şekil 9:*
*Özilinti Yöntemi ile perde frekansı bulunması (a) konuşma işareti (b)merkez kırpması uygulanmış konuşma işareti (c) özilinti fonksiyonu ile perde frekansı kestirimi (d) istenen aralıkta perde frekansının bulunması.*




*Şekil 10:*
*Özilinti yöntemi ile elde edilen perde frekansının (a) medyan süzgeç öncesi (b) sonrası dağılımı*

## 3. DENEYSEL ÇALIŞMA

Bu çalışmada TIMIT ve NTIMIT veritabanları ile bürünsel özellikler ve MFCC öznitelikleri kullanılarak konuşmacı tanıma üzerine çeşitli çalışmalar yapılmıştır.

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. TIMIT veritabanında konuşmalar sessiz bir ortamda mikrofon üzerinden kaydedilmiştir. Bu veritabanı ile yapılan konuşmacı tanıma deneylerinde % 100'e yakın sonuçlar elde edilmiştir (Reynolds ve diğ., 1995). NTIMIT veritabanı, TIMIT veritabanındaki konuşmaların telefondan hattından geçirilmesi ile elde edilmiştir. NTIMIT veritabanında konuşmalar telefon hattından iletildiğinden, bu

veritabanı telefon ahizesi ve iletim hattının etkilerini içermektedir. Konuşma işareti 16 kHz'de örneklenmiş olup kullanışlı bant genişliği telefon bant genişliği ile sınırlıdır. Bu veritabanı ile MFCC öznitelikleri kullanılarak yapılan deneylerde, 168 kişilik test dizini için %76 tanıma başarımı elde edilmektedir. (Jankowski ve diğ., 1995).

Sarma ve Zue (1997), TIMIT ve NTIMIT veritabanlarını kullanarak, Mel frekansı kepstrum katsayıları (MFCC) ile birlikte enerji, sözcük süresi ve perde frekansı bürünsel özelliklerini konuşmacı doğrulama başarımlarını araştırmıştır. Bu çalışmada öznitelikler 6 geniş fonetik sınıfa ayrılıp Gauss karışımları ile modellenmiş ve NTIMIT için perde frekansı en önemli öznitelik olarak gösterilmiştir. Park (2002), formant frekansları ve perde frekansını MFCC'lere alternatif olarak araştırmış, $F_1$, $F_2$, $F_3$ ve $F_4$ formant frekanslarını öznitelik olarak kullanarak TIMIT için % 64.6, NTIMIT için % 4.8 konuşmacı tanıma başarımı elde etmiştir. Aynı çalışmada perde frekansı öznitelik olarak kullanılarak TIMIT için % 45.5, NTIMIT içinse % 39.1 konuşmacı tanıma başarımı elde edilmiştir. Ancak bürünsel özellikler MFCC öznitelikleriyle birlikte kullanılmamıştır. Jankowski ve diğ. (1995), perde frekansının salınımlar arası uzaklığı (pitch jitter) ve ikinci darbelerin (secondary pulse) yerini öznitelik olarak kullanmaktadır. NTIMIT veritabanında 168 kişi ile yapılan deney sonucunda MFCC öznitelikleri ile konuşmacı tanıma başarımı % 76, MFCC özniteliklere perde salınımları ve ikinci darbe parametreleri eklenmesi ile tanıma başarımı % 78.7'ye çıkartılmaktadır. Literatürdeki bu çalışmalardan farklı olarak bu makalede MFCC öznitelikleriyle birlikte enerji değişimi ve formant frekans oranı ($F_3/F_2$) öznitelik olarak kullanılmakta ve gürültülü ortamlarda konuşmacı tanıma başarımı araştırılmaktadır.

Konuşmacı tanımada eğitim ve test gürültüsüz ortamda yapılırsa yüksek tanıma başarımları elde edilebilir. Konuşmacı tanıma sisteminin eğitiminde mikrofon veya telefon hattı üzerinden alınan ses kayıtları kullanılmaktadır. Sessiz bir ortamda mikrofon ile kayıt yapılması tanıma başarımını yükseltirken, telefon ortamı ile kayıtta, telefon iletim hattı etkisinden dolayı tanıma başarımı düşmektedir. Bu duruma ek olarak konuşmacı tanıma sistemlerinde aday konuşmacının test edildiği ortam gürültü içerebilir. Bu ortam gürültüleri insan konuşması, araba, fabrika gürültüsü gibi gürültüler olabilir. Deneylerde bu ortam gürültüleri NOISEX (Varga ve diğ., 1992) veritabanı kullanılarak elde edilmiştir. Bu veritabanındaki gürültüler, değişik işaret gürültü oranlarında (SNR) konuşmacıların test cümlelerine eklenmiştir.

TIMIT veritabanı için özniteliklerin elde edilmesinde 10 msn'lik kısmı örtüşen 20 msn'lik çerçeveler alınıp, çerçevelere Hamming pencereleme uygulanmaktadır. Pencerelenen işaretin 512 örnek FFT'si alınıp, Slaney (1998) tarafından tanımlanan Mel ölçekte, üçgen süzgeç dizilerinden geçirilir. Süzgeçten geçirilen işaretin logaritması alınıp ayrık kosinüs dönüşümü alınır. NTIMIT veritabanı için ise işaret 25 msn'lik çerçevelere ayrılıp 10 msn de bir çerçeveler yenilenir. Aliaa ve diğ. (2004) tarafından tanımlanan algoritma kullanılarak konuşmadan sessiz kısımlar atılır. Süzgeç dizileri 300–3380 Hz arasına ve 70 Hz aralıkla % 50 örtüşmeli olarak yerleştirilir. Her bir çerçeveye karşılık olarak TIMIT veritabanı için 24, NTIMIT veritabanı için 20 boyutlu öznitelik vektörleri kullanılmaktadır. Yapılan tüm deneylerde parametreler bu şekilde kullanılmaktadır.

Deneylerde konuşmacıların modellenmesinde Gauss karışım modeli (GKM) kullanılmaktadır. Bir konuşmacıya ait 10 cümlenin 8'i (2 Sa, 5 Si, 3 Sx cümleleri) eğitim, kalan 2 cümlenin her biri test için kullanılmıştır. Bu işlem 168 kişi için tekrar edilmektedir. Gauss karışım modelinde karışım sayısı 32 alınıp Beklentinin Maksimumlaştırılması (BM) algoritması ile eğitilmektedir. Modelde minimum değişinti sınırı 0.01 alınmaktadır. Model parametreleri 15 özyinelemede istenen değere yakınsamaktadır (Reynolds ve diğ., 1995).

Deneysel çalışmada bürünsel özellik olarak formant frekansları, enerji ve perde frekansının konuşmacı tanımaya etkisi incelenmiştir. Bürünsel özellikler MFCC öznitelik vektörüne çeşitli şekillerde eklenerek TIMIT/NTIMIT üzerinde konuşmacı tanıma yapılmıştır.

### 3.1. Formant Frekansları

Dört farklı ses verisi üzerinde formant frekanslarının konuşmacı tanımaya etkisi incelenmiştir. Gürültüsüz mikrofon ortamına karşılık TIMIT veritabanı, telefon ortamına karşılık olarak NTIMIT veritabanı kullanılmaktadır. TIMIT ve NTIMIT veritabanlarındaki ses kayıtlarına, SNR 20 dB olacak şekilde NOISEX veritabanından beyaz gürültü eklenmiştir. Ses sinyalinin formantlarının yerlerinin bulunmasında 19. dereceden doğrusal öngörü katsayıları (DÖK) kullanılmaktadır (Jankowski ve diğ.,

1995). DÖK köklerinden genlik ve frekanslar kullanılarak formantlar seçilir. Formant frekansları olarak logaritma 2 tabanında $F_1, F_2$ ve $F_3$ kullanılmıştır. Logaritma alarak, dinamik sıkıştırma yapılıp, öznitelik vektörleri, dinamik değişimlere karşı daha az hassas olmaktadır (Claudio ve diğ. 1999).

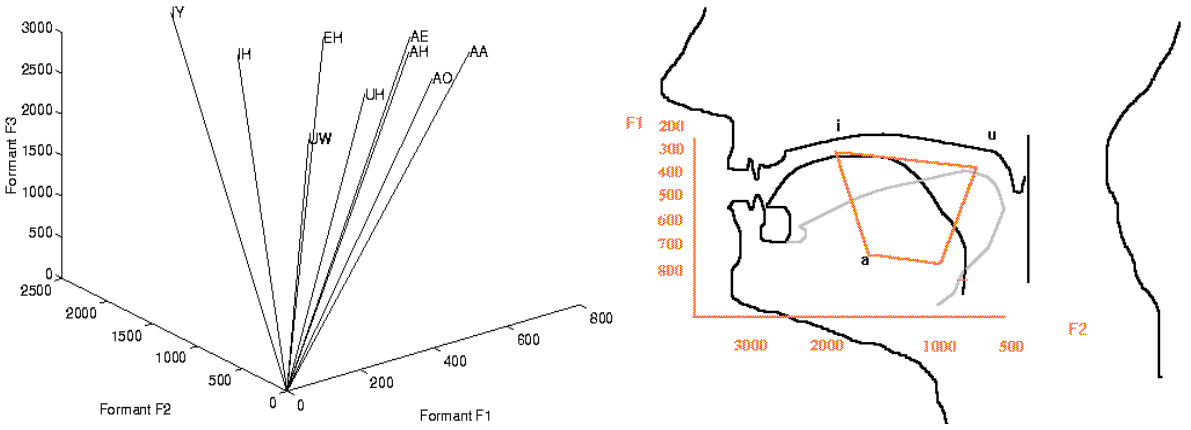
Konuşmacı tanımada öznitelik olarak MFCC ve formant frekansları kullanılmıştır. Formant frekanslarının ve formant frekansları oranlarının ($F_3/F_1$, $F_3/F_2$, $F_2/F_1$), MFCC parametrelerine eklenmesinin konuşmacı tanıma başarımına etkisi incelenmiştir. Elde edilen tanıma oranları Tablo I'de verilmektedir.

**Tablo I. Formant frekansları için tanıma oranları (%)**

| Öznitelik vektörleri | Konuşmacı tanıma oranı (%) | | | |
|---|---|---|---|---|
| | TIMIT | TIMIT+beyaz gürültü | NTIMIT | NTIMIT+beyaz gürültü |
| F1, F2, F3 | 47 | 16.4 | 13.7 | 3.6 |
| MFCC+ F1, F2, F3 | 99.4 | 47 | 64.6 | 25.3 |
| MFCC+F3 | 99.4 | **56.2** | 69.9 | 32.4 |
| MFCC+F2 | **99.7** | 53.3 | 70.8 | 33.9 |
| MFCC+F1 | 99.4 | 54.1 | 71.1 | 30.1 |
| MFCC+F3/F1 | 99.4 | 52.4 | 70.2 | 31.3 |
| MFCC+F3/F2 | 99.4 | **56.8** | 71.4 | **42** |
| MFCC+F2/F1 | 99.1 | 53.9 | 72.02 | 34.8 |
| MFCC | 99.4 | 55.9 | **73.8** | 40.5 |

Tablo I'den görüleceği üzere öznitelik olarak yalnız formant frekansları kullanılması durumunda konuşmacı tanıma oranı oldukça düşük çıkmaktadır. MFCC ile birlikte formant frekansları kullanılması durumunda, gürültü eklenmemiş TIMIT veritabanı için $F_2$ tanıma başarımını arttırmaktadır. TIMIT ve NTIMIT veritabanlarına gürültü eklendiği durum için $F_3/F_2$, MFCC ile birlikte kullanıldığında tanıma başarımını arttırmaktadır. $F_1$ daha çok konuşmaya bağlıdır ve konuşmacı tanıma başarımını düşürmektedir.

Formant frekansları, konuşulan ses, özellikle ötümlü sesler hakkında bilgi vermektedir. Şekil 11 (a)'da görüleceği üzere bir kişiye ait ötümlü sesin $F_1$, $F_2$ ve $F_3$ değeri, rastgele değişen bir değer olmayıp belirli aralıklarda değerler almaktadır. Bir ötümlü ses için ses yolunun belirli bir şekil alması ile birlikte, $F_1$ dilin yüksekliğine, $F_2$ dilin önde veya arkada olmasına bağlı olarak değişen formant frekanslarıdır. Formant frekansları daha çok ötümlü/ötümsüz seslere bağlı olarak değiştiği için konuşmacı tanımada fazla kullanılmamaktadır.




(a) (b)

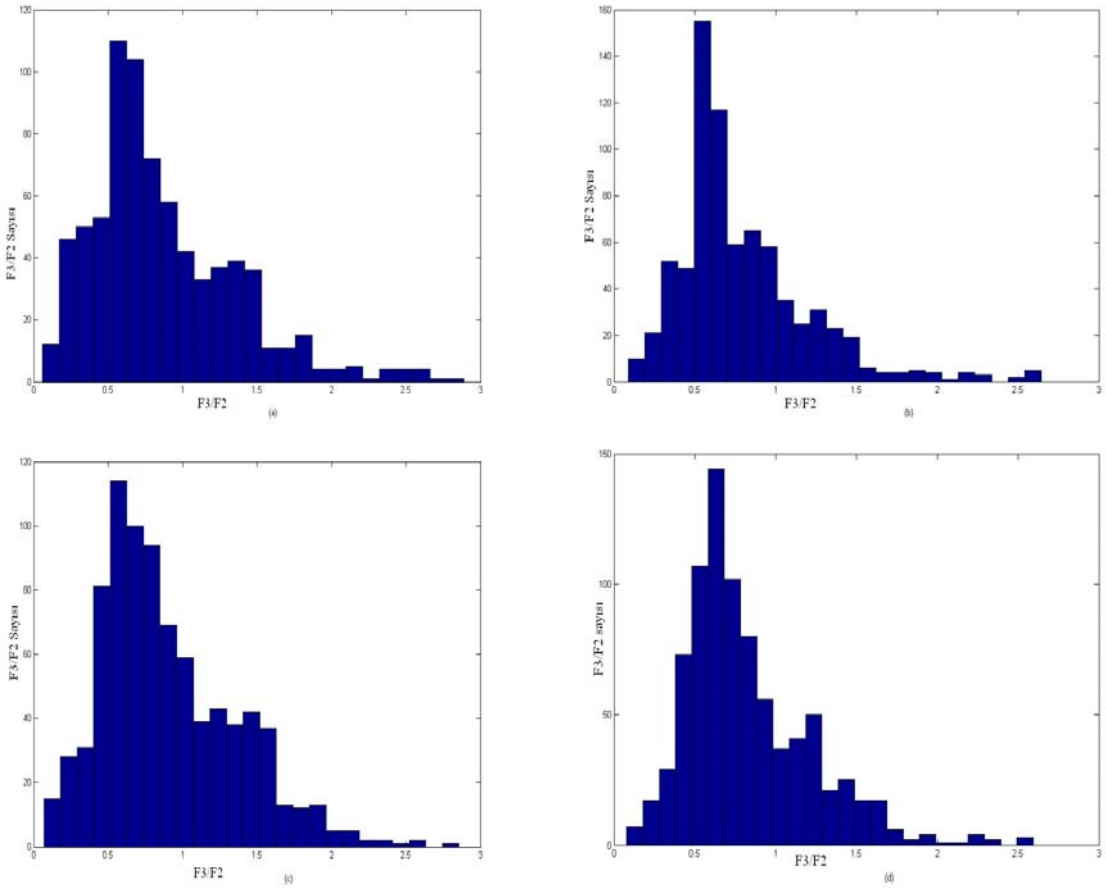
*Şekil 11:*
*(a)İngilizcedeki ötümlü seslerin ortalama $F_1$, $F_2$ ve $F_3$ formant frekansları (b) ötümlü sesler için $F_1$ ve $F_2$ formant frekanslarının oluşumu.*

Ancak formant frekanslarının üretildiği ses yolunun şekli (ses yolu uzunluğu, dil pozisyonu, dudak şekli v.b.) konuşmacılar için ayırt edici özelliklerde taşımaktadır. Bu özelliklerin çıkartılmasında doğrudan formant frekanslarını kullanmak yerine, formant frekansları oranlarının alınması ($F_3/F_2$) gürültülü ortamda tanıma başarımını arttırmaktadır (Sankar ve diğ. 1997). $F_3$, $F_1$ ve $F_2$'ye göre daha fazla konuşmacı bağımlı özellik göstermektedir.

Jankowski ve diğ. (1995), formant frekanslarından kepstrum katsayıları elde etmektedir. Bu çalışmada formant frekanslarının frekans ve genlik modülasyonu alınıp, teager enerji operatöründen geçirilerek kepstrum katsayıları elde edilmekte ve NTIMIT veritabanı için tanıma başarımı % 76'dan % 77.2 ye artmaktadır. Formant frekansları doğrudan olmasa da çeşitli yöntemlerle konuşmacı tanıma başarımını arttırmakta kullanılabilmektedir.

Temiz bir ünlü sözcük için formant frekanslarında 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). Formantlarda gürültü etkisi ile dalgalanma olmamaktadır. Bu durum gürültülü ortamlarda MFCC öznitelikleriyle formantların birlikte kullanılma durumunda tanıma başarımını arttırmasının nedenini açıklamaktadır.

Şekil 12'de TIMIT veritabanında iki konuşmacı tarafından ortak olarak söylenen Sa1 ve Sa2 cümlelerine, 20 dB beyaz gürültü eklenip ve eklenmediği durumlarda, $F_3/F_2$ formant frekanslarının oranının histogramı görülmektedir. Konuşmalara gürültü eklenmesi ile iki konuşmacı içinde $F_3/F_2$ dağılımında değişiklikler oluşmaktadır. Ayrıca Şekil 12'deki gürültüsüz ve gürültü içeren cümleler kendi aralarında karşılaştırıldığında, kullanılan cümleler aynı olmasına rağmen $F_3/F_2$ dağılımında farklılıklar oluşmaktadır.




*Şekil 12:*
*Birinci konuşmacının (a) gürültüsüz (b) gürültü eklenmiş cümleleri için F3/F2 oranının dağılımı. İkinci konuşmacının (c) gürültüsüz ve (d) gürültülü cümleleri için F3/F2 oranının dağılımı.*

### 3.2. Enerji

Bürünsel özelliklerden Enerji (*E*)'de konuşmacı tanımada yaygın olarak kullanılmaktadır. Enerji kullanımında perde frekansına bağlı olarak enerjinin türevi konuşma parçalarında öznitelik olarak alınmaktadır (Adami ve diğ., 2003, Shriberg ve diğ., 2005). Teager enerji (*TE*), yüksek çözünürlüklü enerji kestirimi için konuşmacı tanımada kullanılmaktadır (Jankowski ve diğ., 1995). Ayrıca konuşma tanıma (Jabloun ve Çetin 1999) ve kişinin sesinden duygu durumunun (kızgın, stresli, normal v.b.) belirlenmesinde Teager enerji kullanılmaktadır (Nwe ve diğ., 2003).

Bu çalışmada enerjinin ve Teager enerjinin konuşmacı tanıma başarımına etkisi incelenecektir. Deneylerde Bölüm 3'de verilen parametreler kullanılmaktadır. Enerji değeri olarak, logaritma 2 tabanında denklem 2'de belirtilen enerji ve denklem 4'de belirtilen Teager enerji kullanılmaktadır. Ayrıca enerji ve teager enerjinin birinci derece türevleri alınıp özniteliklere eklenmektedir. Deneylerin test aşamasında TIMIT ve NTIMIT veritabanlarına 20 dB beyaz gürültü eklenmiştir. Elde edilen sonuçlar Tablo II'de verilmektedir.

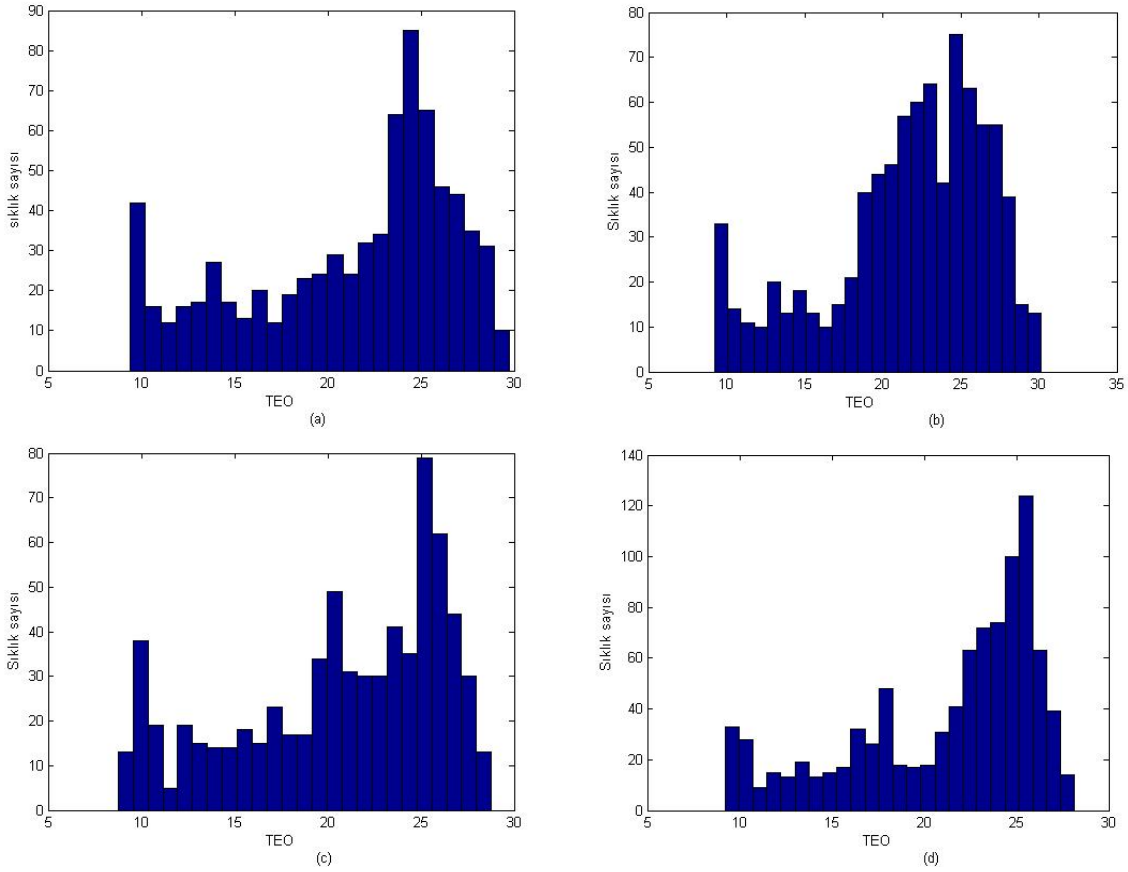
Tablo II'den görüleceği üzere özniteliklere enerjinin birinci türevinin eklenmesi TIMIT veritabanında tanıma başarımını arttırmaktadır. TIMIT ve NTIMIT veritabanına beyaz gürültü eklendiği durumda, özniteliklere enerjinin eklenmesi konuşmacı tanıma başarımında azalmaya neden olmaktadır.

Konuşma işaretinin enerjisi, konuşma yüksek veya alçak olduğunda değişeceği için konuşmacı tanıma için uygun bir özellik olarak gözükmemektedir. Enerji ve teager enerjinin birinci derece türevleri konuşmacıya özgü özellikler taşıdığından dolayı tanıma başarımını arttırmaktadır. Ancak iletim hattı ve gürültü içeren ortamlarda enerji dağılımı büyük değişimlere uğradığından dolayı tanıma başarımını azaltmaktadır.

**Tablo II. Enerji ve birinci derece türevlerinin konuşmacı tanımaya etkisi (%)**

| Öznitelik vektörü | Konuşmacı Tanıma oranı | | | |
|---|---|---|---|---|
| | TIMIT | TIMIT+beyaz gürültü | NTIMIT | NTIMIT +beyaz gürültü |
| **MFCC +*E*** | 99.4 | 53.3 | 62.5 | 33.6 |
| **MFCC +*TE*** | 99.4 | 51.5 | 66.1 | 29.2 |
| **MFCC + Δ (*E*)** | **99.7** | 53.3 | 71.4 | 36.9 |
| **MFCC +Δ*(TE)*** | **99.7** | 53 | 71.1 | 33.9 |
| **MFCC** | 99.4 | **55.9** | **73.8** | **40.5** |

Teager enerji konuşmadaki ani değişimleri daha fazla kuvvetlendirmektedir. Şekil 13'de TIMIT veritabanından 4 farklı konuşmacının aynı cümleleri söylemesi ile elde edilen Teager enerji dağılımları görülmektedir. Tüm konuşmacılar için 25 değeri civarında Teager enerji dağılımı tepe noktaya ulaşmaktadır. Konuşma yüksek veya alçak olmasına bağlı olarak Teager enerjinin dağılımı değişmektedir.




*Şekil 13:*
*TIMIT veritabanından dört farklı kişiye ait ortak cümleler için Tager enerji histogramları*

### 3.3. Perde Frekansı

Deneysel çalışmada, değişik işaret gürültü oranları için öznitelikler ile birlikte perde frekansının kullanılmasının tanıma başarımına etkisi incelenmiştir. Perde frekansı özilinti yöntemi ile elde edilmiş ve deneylerde Bölüm 3'de verilen parametreler kullanılmıştır. Konuşmacıların eğitiminde gürültü eklenmeyip, test aşamasında test cümlelerine 30, 20 ve 10 dB işaret gürültü oranlarında Gauss gürültüsü eklenmektedir. TIMIT ve NTIMIT veritabanları için elde edilen tanıma başarımları Tablo III'de verilmektedir.

**Tablo III. Değişik işaret gürültü oranları için perde frekansının konuşmacı tanımaya etkisi (%)**

| Test Ortamı | | TIMIT | | NTIMIT | |
|---|---|---|---|---|---|
| | | MFCC | MFCC+$f_0$ | MFCC | MFCC+$f_0$ |
| Temiz | | **99.4** | **99.4** | 73.8 | **80.6** |
| Gauss gürültüsü | 30 dB | 96.4 | **96.4** | 71.4 | **79.5** |
| | 20 dB | 54.2 | **55.9** | 42.6 | **55.1** |
| | 10 dB | 8.0 | **12.2** | 8.3 | **10.7** |

Tablo III'den görüleceği üzere MFCC katsayılarına perde frekansı ile birlikte kullanılması durumunda, test cümlelerine gürültü eklenmediği durum için TIMIT veritabanı için tanıma başarımında değişme olmamakta, NTIMIT veritabanı için ise tanıma başarımı 6.8 puan artmaktadır. MFCC katsayılarına perde frekansının eklenmesi, her iki veritabanında da tüm işaret gürültü oranları için tanıma başarımını artırmaktadır.

Bir sonraki adımda, değişik gürültü tipleri için perde frekansının konuşmacı tanıma başarımına etkisi araştırılmıştır. Konuşmacıların eğitiminde gürültü eklenmeyip, test aşamasında NOISEX veritabanından beyaz, konuşma, fabrika ve araba gürültüleri, 20 dB işaret gürültü oranında test cümlelerine

eklenmektedir. Bu şekilde değişik gürültü ortamlarında bürünsel özelliklerin tanıma performansı ölçülmektedir. Değişik gürültüler için perde frekansının konuşmacı tanıma başarımları Tablo IV'de verilmektedir.

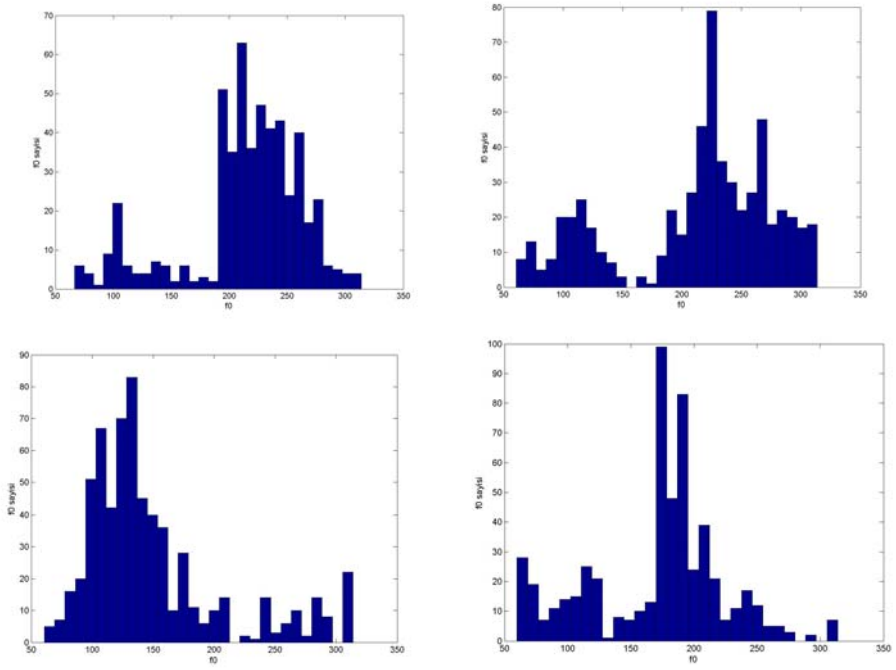
**Tablo IV. Değişik gürültüler için perde frekansının konuşmacı tanıma başarımları (%)**

| Test Ortamı (SNR=20dB) | TIMIT | | NTIMIT | |
|---|---|---|---|---|
| | MFCC | MFCC+$f_0$ | MFCC | MFCC+$f_0$ |
| **Beyaz gürültü** | 55.9 | **60.4** | 40.5 | **48.8** |
| **Konuşma gürültüsü** | 58.0 | **61.6** | 39.9 | **50.3** |
| **Fabrika gürültüsü** | 57.4 | **62.5** | 38.4 | **46.1** |
| **Araba gürültüsü** | 62.2 | **64.9** | 40.5 | **48.5** |

Tablo IV'den görüleceği üzere MFCC katsayılarına perde frekansı eklenmesi ile dört farklı gürültü için TIMIT veritabanında ortalama 4 puan, NTIMIT veritabanında ise ortalama 8.6 puanlık artış sağlanmaktadır.

Sonuç olarak perde frekansı konuşmacıları birbirinden ayırmada etkilidir (Reynolds ve diğ., 2003, Adami ve diğ., 2003, Shriberg 2007). Perde frekansı, iletişim ortamı özellikleri ve gürültüden, spektral katsayılara göre daha az etkilenir (Arcienega ve Drygajlo, 2001). Bu nedenle konuşmacı tanımada öznitelik olarak kullanılmaktadır. Şekil 14'de NTIMIT veritabanında tüm konuşmacılar tarafından ortak olarak söylenen Sa1 cümlesinin 4 farklı konuşmacı için perde frekansının histogramı görülmektedir. Şekil 14'den görüleceği üzere her bir konuşmacının söylediği cümleler aynı olmasına rağmen perde frekansı dağılımında büyük farklılıklar oluşmaktadır. Bu farklılıkların kişiyi ayırt edici özellik olarak kullanılması sonucu yukarıda yapılan deneylerden de görüleceği üzere tanıma başarımında önemli oranda artış sağlanmaktadır.

Son olarak, konuşmacı tanıma başarımlarını yükselten bürünsel özellikler, MFCC özniteliklere eklenerek kullanılmasının konuşmacı tanıma başarımına etkisi incelenmiştir. Bölüm 3.1, 3.2 ve 3.3'de yapılan deneylerde $F_2$, $F_3$, $F_3/F_2$, $\Delta E$ , $\Delta TE$ ve $f_0$ bürünsel özellikleri farklı durumlar için tanıma başarımını arttırmaktadır. Deneylerde TIMIT ve NTIMIT veritabanları test cümlelerine 20 dB beyaz gürültü eklenmiştir. MFCC katsayıları ve perde frekansı ile birlikte deneylerde başarımı arttıran bürünsel özelliklerin konuşmacı tanıma başarımları Tablo V'de verilmektedir.



*Şekil 14:*
*Dört farklı konuşmacının aynı cümleyi söylemesi ile elde edilen perde frekanslarının histogramları*

**Tablo V. Perde frekansının konuşmacı tanımaya etkisi (%)**

| Öznitelik vektörleri | Konuşmacı tanıma oranları | | | |
|---|---|---|---|---|
| | TIMIT | TIMIT+beyaz gürültü | NTIMIT | NTIMIT +beyaz gürültü |
| MFCC+$f_0$+$F_3$ | 99.4 | 58.6 | 76.8 | 41.4 |
| MFCC+$f_0$+$F_2$ | 99.4 | 57.7 | 75.9 | 45.8 |
| MFCC+$f_0$+$F_3/F_2$ | 99.4 | 58.3 | 78.9 | **50.9** |
| MFCC+$f_0$+$\Delta E$ | **100** | **62.5** | 77.1 | 45.5 |
| MFCC+$f_0$+$\Delta TE$ | 99.7 | 60.7 | 77.1 | 43.1 |
| MFCC+$f_0$+$\Delta f_0$ | 99.7 | 61 | 80.3 | 48.5 |
| MFCC+$f_0$ | 99.4 | 60.4 | **80.6** | 48.8 |
| MFCC | 99.4 | 55.9 | 73.8 | 40.5 |

Tablo V'den görüleceği üzere TIMIT veritabanında test cümlelerine gürültü eklenmediği ve eklendiği durumlarda, MFCC parametrelerine perde frekansı ve enerjinin birinci türevinin eklenmesi tanıma başarımlarını arttırmaktadır. NTIMIT veritabanında konuşmalara gürültü eklendiği durumda, MFCC katsayılarına perde frekansı ve $F_3/F_2$ eklenmesi tanıma başarımlarını arttırmaktadır.

## 4. SONUÇLAR

Bu makalede bürünsel özelliklerden formant frekansları, enerji ve perde frekansının konuşmacı tanıma başarımına etkisi incelenmiştir. Bürünsel özellikler, genellikle tek kullanılmayıp, MFCC katsayıları ile birlikte tamamlayıcı özellik olarak kullanılmaktadır.

Formant frekansları (özellikle $F_1, F_2$) konuşmacıdan daha çok konuşulan sözcüğe bağlı olarak değişmektedir (Rabiner ve Juang, 1993). NIMIT veritabanı için telefon ahizesinin doğrusal olmayan etkisi, sahte formant frekansları oluşturmakta, bu da tanıma başarımını azaltmaktadır (Jankowski ve diğ., 1994). Bununla birlikte formant frekanslarının oranı ($F_3/F_2$), konuşmacı tanıma başarımını arttırmaktadır. Enerji konuşmanın yüksek ve alçak olmasına göre değişen bir bürünsel özellik olup, enerjinin birinci türevi, MFCC ve perde frekansı ile birlikte kullanıldığında TIMIT veritabanı için tanıma başarımında iyileşme olmaktadır. MFCC parametrelerine perde frekansı eklenmesi, telefon hattı etkisi ve gürültü içeren ortamlarda konuşmacı tanıma başarımını önemli oranda iyileştirmektedir. TIMIT ve NTIMIT veritabanlarında gürültü ve iletişim ortamının (telefon v.b.) doğrusal olmayan etkisinden (Reynolds ve diğ., 1995) perde frekansı daha az etkilenmekte (Arcienega ve Drygajlo, 2001), ve bu nedenle daha iyi konuşmacı tanıma başarımı elde edilmektedir.

Formant frekansları, enerji ve Teager enerji konuşmacı tanıma başarımını azaltmaktadır. Bununla birlikte gürültü içeren ortamlarda, enerjinin birinci türevi, $F_3/F_2$ formant frekanslarının oranı ve perde frekansı ise telefon ortamında ve gürültü içeren ortamlarda konuşmacı tanıma başarımında artış sağlamaktadır.

## 5. KAYNAKLAR


1. Adami, A.G., Mihaescu, R., Reynolds, D.A., Godfrey, J.J., (2003) Modeling prosodic dynamics for speaker recognition. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process., Vol. 4, Hong Kong, China. pp. 788–791.
2. Adami, A.G., Hermansky, H., (2003) Segmentation of speech for speaker and language recognition. In: Proc. EUROSPEECH, Geneva. pp. 841–844.
3. Aliaa, A. Y., A. S. Ebada and W. H. El Behaidy. (2004) Development of Automatic Speaker Identification System, $21^{st}$ National Radio Science Conference.
4. Arcienega, M., A. Drygajlo. (2001) Pitch-dependent GMMs for Text-Independent. Speaker Recognition Systems. Eurospeech'01, Scandinavia, p. 2821-2824.
5. 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.

6. Atal B.. (1972) Automatic speaker recognition based on pitch contours. Journal of the Acoustic Society of America, 52(6):1687–1697, 1972.
7. Carey M.J., E.S. Parris, H. Lloyd-Thomas, and S. Bennett. (1996) Robust prosodic features for speaker identification. In Proc. Int. Conf. on Spoken Language Processing (ICSLP 1996), USA, p. 1800–1803.
8. Chen shi-han and Hsiao-chuan wang (2004) Improvement of Speaker Recognition by combining residual and prosodic features with acoustic features acoustics, speech, and signal processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference volume: 1, p. 93-96.
9. Claudio, B. and L. P. Ricotti. (1999) Speech Recognition Theory and C++ Implementation. John WILEY&Sons, Ltd, p. 125-137.
10. Dehak, N., P. Kenny, and P. Dumouchel, (2007) Continuous prosodic features and formant modeling with joint factor analysis for speaker verification, in Proc. Interspeech, Antwerp.
11. Duman, F. O. Eroğul., Z. Telatar., S. Yetkin. (2005) Uyku İğciklerinin Kısa ve Uzun Dönemli Karma Analizi. SIU, Kayseri.
12. Ertaş, F. (2001) Feature Selection and Classification Techniques for Recognition. Journal of Engineering Sciences, No. 1, Pamukkale, p. 47-54.
13. Fant, G. (1960) Acoustic Theory of Speech Production. Mouton & Co., The Hauge.
14. 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.
15. 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.
16. Jabloun, F. A.E. Cetin, (1999) The Teager energy based feature parameters for robust speech recognition in car noise. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 273–276.
17. 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.
18. 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.
19. Kasi, K. (2002) Yet Another Algorithm for Pitch Tracking, Master thesis, Old Dominion University, p. 9-13.
20. Kinnunen, T., Gonzalez-Hautamaki, R. (2005) Long-Term F0 Modeling for Text-Independent Speaker Recognition. In: Proceedings of the 10th International Conference Speech and Computer (SPECOM), Patras, Greece, p. 567–570.
21. Mary L, B. Yegnanarayana. (2008) Extraction and representation of prosodic features for language and speaker recognition. Speech Communication, Volume 50, Issue 10, p. 782-796.
22. Mezghani, A., and O'Shaughnessy, D., (2005) Speaker Verification Using a New Representation Based on a Combination of MFCC and Formants, IEEE Canadian Conference on Electrical and Computer Engineering, Saskatoon, SK, p. 1461-1464.
23. Nwe, T.L. S.W. Foo, and L.C. De Silva. (2003) Detection of stres and emotion in speech using traditional and FFT based log energy features. In Fourth Pacific Rim Conference on Multimedia, Information, Communications and Signal Processing, volume 3, pages 1619–1623.
24. O'shaughnessy, D. (1987) Speech Communication Human and Machine. Addison Wesley, New York.
25. 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.
26. Peskin, Barbara et al. (2003 a) Using Prosodic And Conversational Features for High-Performance Speaker Recognition. Report from JHU WS'02", IEEE Trans. Speech Audio Processing, p. 792-796.
27. Peskin, B., A. Adami., Q. Jin., D. Klusácek., J. S. Abramson., R. Mihaescu., J. J. Godfrey, D. A. Jones and B. Xiang. (2003 b) The Super 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.
28. Plumpe, M. D., T. F. Quatieri. and D. A. Reynolds. (1999) Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification. IEEE Transactions on Speech and Audio Processing, vol. 7, no. 5.
29. Rabiner, L. R. and B. H. Juang. (1993) Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs.

30. Reynolds, D.A. (1992) A Gaussian Mixture Modeling Approach to Text Independent Speaker Identification. Ph.D. thesis, Georgia Inst. of Technology.
31. Reynolds D. A., Zissman M. A., Quatieri T. F., O'Leary G. C., Carlson B. A. (1995) The Effects of Telephone Transmission Degradations on Speaker Recognition Performance, ICASSP (Detroit), May 9-12. p. 329-331.
32. Reynolds D.A., et al. (2003) The Super SID Project: Exploiting High-Level Information for High-Accuracy Speaker Recognition. in Proc. ICASS, p. 784–787.
33. 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.
34. Rose, P. (2001) Forensic Speaker Identification, Taylor & Francis Forensic Science Series, ISBN 0-415-27182-7, p. 225-280.
35. Sankar k. Pal and Dwijesh D. M. (1997) Fuzzy Sets and Decision making Approaches in Vowel and Speaker Recognition IEEE Transactions on Systems, Man, and Cybernetics, pp. 625-629.
36. Sarma, S. and V. Zue, (1997) A Segment-based speaker verification system using SUMMIT, in Proc. Eurospeech, Rhodes, pp. 843-846.
37. Seddik Hassen, AmeI B. S. Rahmouni and Mounir Sayadi (2004) Text Independent Speaker Recognition based on the Attack State Formants and Neural Network Classification IEEE International Conference on Industrial Technology Volume: 3, p. 1649- 1653.
38. Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A., (2005) Modeling prosody for speaker recognition. Speech Comm. 46, 455–472.
39. Shriberg, E. (2007) Higher-Level Features in Speaker Recognition, in C. Müller, editor, Speaker Classification I, vol. 4343 of Lecture Notes in Computer Science/AI. Springer, Berlin.
40. Sonmez, M.K., Shriberg, E., Heck, L., Weintraub, M., (1998) Modeling dynamic prosodic variation for speaker variation. In: Proc. Int. Conf. Spoken Language Process., Vol. 7, Sydney, Australia. pp. 3189–3192.
41. Stevens, S. and J. Volkman (1940) The Relation of Pitch to Frequency. American Journal of Psychology, vol. 53, p. 329.
42. Slaney, M. (1998) Auditory Toolbox: A MATLAB Toolbox for Auditory Modeling Work Technical Report, Interval Research Corporation, p. 29-32.
43. Teager, H. M. and S. M. Teager. (1989) Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract. in Speech Production and Speech Modeling, W.J. Hardcastle and A. Marchal, Eds., NATO Advanced Study Institute Series D, Vol. 55, Bonas, France.
44. 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.
45. Umesh, S., L. Cohen and D. Nelson. (1999) Fitting the Mel Scale. IEEE Transactions on Acoustics, Speech and Signal Processing., p. 217-220.
46. Varga, A.P., Steeneken, H.J.M., Tomlinson, M., Jones, D., (1992) The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical Report, Speech Research Unit, Defence Research Agency, Malvern, UK.
47. Zeljkovic, I. P. Haffner, B. Amento, and J. Wilpon, (2008) GMM/SVM n-best speaker identification under mismatch channel conditions, in ICASSP, Las Vegas, USA, pp. 4129–4132.
48. 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.

Makale 13.01.2009 tarihinde alınmış, 25.06.2009, 27.08.2009 ve 26.10.2009 tarihlerinde düzeltilmiş, 27.10.2009 tarihinde kabul edilmiştir. İletişim Yazarı: Ö. Eskidere (oeskidere@uludag.edu.tr).