Bir Konuşmacının Konuştuğu Dilin Belirlenmesi

[1]Kamil Öncü ŞEN **[3]Cemal KÖSE** **[3]Salih ARAS**

[1,3,4] Karadeniz Teknik Üniversitesi, Bilgisayar Mühendisliği, Trabzon

[1]e posta:oncusen@ktu.edu.tr [2]e posta: cemalkose@ktu.edu.tr
[3]e posta:saliharas@hotmail.com

ÖZET

Bu çalışmada, fonem, konuşma ve konuşulan sözcükler gibi bir dilin temel özeliklerinden yaralanarak, konuşmacının konuştuğu dildin belirlenmesine çalışılmıştır. Konuşulan dilin tanınması, konuşmacı tanıma, konuşmadan-metne ve metinden-konuşmaya dönüştürme konularıyla ilgili olarak ele alınabilir. Bu bağlamda, konuşma, ses tanınmasında, ve özellikle sesin işlenmesinde Hidden Markov Modeil (Saklı Markov Model) ile Yapay Sinir Ağları yöntemleri yaygın olarak kullanıldığı bilinmektedir. Bu çalışmada öncelikle fonem veri tabanı oluşturuldu Bu veri tabanını oluştururken İngilizce dili baz alındı ve bu dilin fonemleri kullanıldı. Bu fonem veri tabanını konuşmacının konuşmasıyla eşleşmesi sonucunda hangi fonemin konuşma içerisinde kaç sefer kullanılmasının tespit edilmesi ve bu tespite dayalı olarak konuşmacının hangi dili kullandığının bulunmasında kullanılmıştır. Sesin işlenmesi aşamasında Fast Fourier Transform (Hızlı Fourier Dönüşümü) , Forward Discrete Cosine Transform(İleri Ayrık Kosinüs Dönüşümü) ve Haar Wavelet Transform(Saç Dalgacık Dönüşümü) kullanılmıştır. Alınan ses kayıtlarından konuşmanın belirlenmesi için ise End-point Detection ve start-point Detection algoritmaları kullanılmıtır. Ses verilerinin eşlenmesi safhasında ise Cross Correlation(Çapraz Korelasyon) ve Dynamic Time Warping(Dinamik Zamanlı Çarpıtma) yöntemleri kullanılmıştır

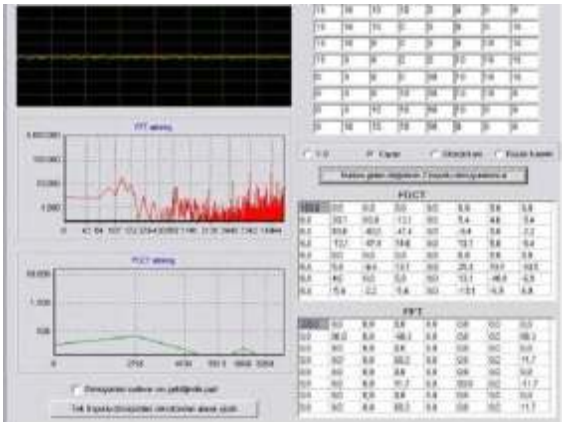
1.GİRİŞ

Konuşma insanoğlunun iletişim kurmasında ve varlığını sürdürmesinde kullandığı en önemli olgulardan biridir. İnsanoğlu yeryüzü üzerinde birçok dil kullanmaktadır. Kullanılan diller farklı olsa bile dillerin bu farklılığının anlaşılabileceği ortak yapılar mevcuttur. Bu çalışmada bu ortak yapılardan biri olan fonem olgusu kullanılmıştır. Fonem bir dilin özelliğini vurgulayan en temel sesbirimi olarak nitelenebilir. Böylece, fonemlerin bu temel özelliğinden yaralanarak konuşmacının konuştuğu dilin belirlenmesine çalışılmıştır. Dolayısıyla, öncelikle belirlenen dil için fonem veri tabanı oluşturuldu. Burada önemli olan oluşturulan veri tabanının dilin en çok kullanılan fonemlerini içeriyor olmasıdır. Böylece konuşma fonemleri çıkarılmış belirlenmiş dil ise eşlenen fonemlerin sayısı artacak ve tanıma işlemi daha kesin ve daha rahat yapılması sağlanmış olacaktır. Fonem veri tabanının hazır olmasından sonra yapılması gereken ilk iş sesi işlemeyi sağlayacak yapıların oluşturulmasıdır. Çünkü sesin ham şekilde kullanılması olanaksızdır. Bu projede öncelikle FFT kullanıldı fakat FFT'nin yavaş kaldığı görüldü, sonra FFT'den daha etkili ve hızlı olduğu belirtilen FDCT kullanıldı[1]. Sonuçlar daha iyi çıkmasına rağmen yeterli görülmeyerek son zamanlarda kullanımı sıklaşan Haar Wavelet Transform'u çalışmaya dahil edildi. Bu farklı dönüşümlerin kullanılmasının sebebi doğruluk oranını en iyi şekilde yakalayabilmek ve yöntemlerin gözlemlerini daha iyi elde edebilmek içindir. Frekans alanına geçiş yapan sesi şimdi karşılaştırma işlemlerine tabi tutarak belirli fonemin kullanım sayısı tespitini yapabilmek için iki yöntemden yararlanılmıştır. Bunlar sırası ile Cross Correlation ve Dynamic Time Warping. Karşılaştırma işlemi başlatılmadan önce alınan ses kayıtlarınından konuşma ve gürültü kısmı ayırtedilmesi eşlenmenin çok daha başarılı olacagı düşünülerek End-point Detection ve Start-point detection yapıları yardımıyla konuşma kısımları kayıtta tespit edilmiştir. Çalışmada ilk olarak Cross Correlation yöntemi kullanılmıştır, fakat bu yapı zamanda kaymış verilerin eşlenmesinde yeterli olmadığı görülmüştür. Bu yöntem zamanda veri kaymasını tespit edebiliyor fakat verilerin eşlenmesinin oranını bu kayma yüzünden doğru yapamıyor. Buna karşın yöntem kayma olmayan verilerin tespitini yüksek oranda belirleyebildiği gözlenmiştir. Verilerdeki kaymayı fark edebilen ve

onu tolore ederek başarılı bir şekilde eşleyerek eşlenme oranını büyük ölçüde doğru tespit edebilen bir yöntem olan Dynamic Time Warping(Dinamik Zamanlı Çarpıtma) yöntemi kullanıldı. Bu iki yöntemde kullanıcıya opsiyon olarak sunuluyor ve istenilen yöntemle karşılaştırılma yapılabilmesine olanak sunulmuştur. kullanıcın iki farklı yöntemi kullanmasına olanak sağlanmış ve kullanıcının geliştireceği sistemin performansını sağlanan bu yöntemlere dayalı olarak ölçmesi hedeflenmiştir.

2. Ses İşleme Teknikleri

Ses, fonem veya konuşma tanımasında iki farklı verinin işlenmesi gerekmektedir. Bunlardan biri fonemlerin özel olarak alınması ve wav dosyası olarak kaydedilmesi diğeri ise gerçek zamanlı mikrofondan alınan ses verilerin işlenmesidir. Burada karşılaştırılmaya hazırlanacak ilk verinin bir wav dosyası olduğunun bilinmesi önemlidir. İlk etapta otomatik bir sistem kurmak amaçlanmamıştır. Dolayısıyla, manüel bir alt yapı oluşturularak, kullanıcının istediği veriyi istediği bir yöntemle işleyebilmesi sağlanmıştır. Böylece, kullanıcıya izlenecek yöntemler üzerinde kontrol yeteneği sağlanarak, kullanıcının işlemleri her bir adımda izlemesi hedeflenmiştir. Burada, amaç kullandığımız yöntemlerin gözlenebilirliğini sağlamaktır. Bu yüzden hangi fonemin konuşmada aranacağına kullanıcıya sunulan seçenekle kullanıcının belirlemesi isteniyor. Sistem ilk olarak gürültülü ortamlarda da test edilmiş fakat burada bazı sorunlarla karşılaşılmıştır. Örneğin mikrofondan gelen sesin gürültülü olması ve bu gürültünün eşlenmenin oranını çok düşürmesine neden olduğu görülmüştür. Bu nedenle, daha az gürültülü bir ortamda, yani gürültünün minimum olduğu bir özel ortamda, kaydedilmiş veri üzerinde işlem yapılmasına karar verilmiştir. Bunun için uygun ortamda veriler hazırlandı. Ses işleme yapılacak veriler belirlendikten sonra hangi dönüşümün kullanılacağını belirlemek gerekiyor ki burada üç metot tercih edildi FFT, FDCT ve Haar Wavelet dönüşümleridir. Bu üç dönüşümden FFT'nin ve FDCT'nin gerçek zamanlı dönüştürülmesi şekil 1.a da verilmiştir. Burada kullanılan dönüşümlerin doğruluğunu kontrol için hazır veri kullanılmasına da olanak sunulmuştur. Burada bazı özel matrisler verilerek ki bunların sonuç değerleri belirlidir kullanılan fonksiyonların doğruluğu teyit edilmiştir. FFT, FDCT frekans boyutuna geçmede kullanılan yöntemlerdir. Alınan ses verisi FFT ve FDCT ile frekans boyutuna aktarıldığı tek boyut dönüşümü ve edit kutularına girilen değerlerin iki boyutlu dönüşümlerinin alınmasından oluşmaktadır. Burada, 22050kHz, 16 bit stereo özelliği seçilmiştir. FFT'nin örnek sayısı ise 1024 olarak seçilmiştir. FFT kodunda kullanılacak katsayılar (W) aynı olduğu için (girilen değerlere göre değişmediğinden) başta bir döngü yardımıyla hesaplanarak 2048'lik bir dizide tutulmaktadır. Bu sayede, katsayılar her defasında bir daha hesaplanmamakta, dizideki değerler kullanılmaktadır. FFT'nin 1024'lük seçilmesinin sebebi ise frekans aralığını daha iyi belirleyebilmektir. FDCT'te ise, butterfly(kelebek) yapısı uygulanmıştır. İki boyutlu dönüşümlerde ise tek boyutlu dönüşümlerden yararlanılmıştır. Burada, matrisler, ilk önce FFT veya FDCT dönüşümlerinden geçirilmiş, sonra matrisin transpozesi alınmış tekrar dönüşümden geçirilmiş ve tekrar transpozesi alınarak 2 boyutlu dönüşüm tamamlanmıştır. Kullanıcı için de bazı hazır matris seçenekleri sunulmuştur (1-0, kayan, ortada bir kare, ve küçük kare). FFT'nin hesaplanmasında önce 16'lık ağ yapısı kullanılmış, sonra bunun iyi sonuç vermediği görülmüş ve formül üzerinde değişikliğe gidilerek 1024'lük FFT uygulanmıştır. FFT'nin genliğinin hesaplanmasında da yaklaşık aynı sonuçları verdiğinden sadece Reel kısmın mutlak değeri alınmıştır. Hızlandırma çalışmaları esnasında sonuçların imajiner kısımlarının kullanılmadığından dolayı hesaplama işlemleri kaldırılmış ve simetri özelliğinden dolayı sadece bir yarısında hesaplamalar yapılmıştır. Bu şekilde FFT ve FDCT hesaplaması tamamlanmıştır. Haar Wavelet dönüşümünün kullanılmasının nedeni zamanda çözünürlüğünün yüksek olması ve bunun yanında işlem yükünün fazla olmamasından ötürü hızlı olmasıdır. Diğer wavelet yöntemlerinin Haar Wavelet yönteminden daha iyi çözünürlüğe sahip olduğu bilinmektedir fakat daha ağardırlar ve bu yüzden gerçek zamanlı çalışmalarda genelde Haar Wavelet dönüşümü kullanılmaktadır[2]. Haar Wavelet dönüşümünün algoritması şu şekilde özetlenebilir. Bu yöntem alınan veri dizisinden iki ve ikinin katları(2,4,6,8, ...) şeklinde örneklerle yeni veri serisi oluşturur. Bu veri serisinin ve katsayı dizisinin yenilenmesini sağlar. Yeni seri, eski serinin örneklenme penceresinin ortalaması olur. Katsayılar örnekleme penceresinin ortalaması olarak temsil edilir. Bu işlemin ardından örneğin 1024'lük bir örnekleme verisi 512 veriyle temsil edilebilir. Bu elde edilen yeni veri katsayılarıyla Haar Wavelet dönüşümünün formülüzasyonu alınan verilere kullanılarak veriler frekans, genlik ve zaman boyutlarında hesaplanmaları gerçeklenmiş olur. Veriler sonuç olarak 3 boyutlu dizide elde edilir. Haar Wavelet dönüşümüne örnek bir veri grafiği şekil 2. 'de görülmektedir. Bu yöntemler yardımı ile ses verileri işlenmiş ve karşılaştırmaya hazırlanmış oldu.



Şekil 1. Program gerçek zamanlı sesin alımını ve onu FFT ve FDCT dönüşümlerini grafiksel ortamda sunumunu gerçekleştiriyor

Şekil 2. Haar Wavetlet dönüşümünden geçirilmiş veri

3. Konuşmanın Kayıt İçinde Belirlenmesi

Konuşma kısmının kayıt içinde tespitedilmesi sisteme olan katkısı iki şekilde açıklanabilir. Birincisi fazla işlem yapılması engellenerek sistem hızlandırılmış olur ikinci olarak ise sadece konuşma kısımlarının karşılaştırmada kullanılması karşılaştırma işleminin doğruluk oranını olumlu şekilde artırması olarak açıklanabilir[6]. Konuşmanın belirlenmesinde öncelikle Threshold(eşik degeri) ve Enerji düzeyleri belirlendiki bu değerler yardımı ile arka plan gürültüsünden eleme yapılabilsin. Zero Crossing Rate ve Time feature Energy[5] değerlerinin belirlenmesi sağlandı. Bu aşamanın ardından bulunan degerlerin yardımıyla alınan pencerede konuşmanın başlangıç ve bitiş degerleri belirlenmesi yapıldı. Şimdi sırai ile bahsi gecen yapıların formülüzasyonunu verelim.

a) Zero Crossing Rate:

$$ZCR = 0{,}5\sum_{i=1}^{N}(s(i+1)-s(i)) \qquad \text{Formül 1}$$

Burada N pencerenin eleman sayısıdır(bu projede 512 olarak seçilmiştir) ve s(i) ses işaretidir. Bu ana formül yardımıyla Z_N(pencere ZCR degri), Z_f(ileri yön ZCR), Z_b(geri yön ZCR) değerlerinin hesaplanması aşağıdaki gibidir;

$Z_f = \{Z_1 + Z_2 / 2\}$ eğerki $0{,}5 < Z_1/Z_2 \leq 2$ aksi taktirde $Z_f = \min(Z_1/Z_2)$ olarak degerlendirilir.

$Z_b = \{Z_k + Z_{k-1} / 2\}$ eğerki $0{,}5 < Z_k/Z_{k-1} \leq 2$ aksi taktirde $Z_b = \min(Z_k/Z_{k-1})$ olarak degerlendirilir

$Z_N = \{Z_f + Z_b / 2\}$ eğerki $0{,}5 < Z_f/Z_b \leq 2$ aksi taktirde $Z_N = \min(Z_f/Z_b)$ olarak degerlendirilir.

b) Enerji seviyelerinin tespiti:

Enerji seviyesi hesaplanmasında aşağıdaki formül 2 kullanılır,

$$E_k = \sum_{n=((k-1)F/2)+1}^{(k+1)F/2} (s_n)^2$$ Fomül 2

k= 1,2...K

F= Her pencerenin elemen sayısı

K= Toplam pencere sayısı

Bu formül ana denklemdir ve E_f`nin(ileri yön enerji) ve $E_b$`nin(geri yön enerji) hesablanmasında da kullanılacaktır. Şimdi E_f ve E_b nin hesaplanması formül2 yardımıyla aşağıdaki gibidir;

$E_f = \{E_1 + E_2 / 2\}$ egerki $0{,}5 < E_1/E_2 \leq 2$ aksi taktirde $E_f = \min(E_1, E_2)$ değerini alır.

$E_b = \{E_k + E_{k-1} / 2\}$ eğerki $0{,}5 < E_k/E_{k-1} \leq 2$ aksi taktirde $E_b = \min(E_k, E_{k-1})$ değerini alır.

$E_N = \{E_f + E_b / 2\}$ eğerki $0{,}5 < E_f/E_b \leq 2$ aksi taktirde $E_N = \min(E_f, E_b)$ değerini alır ve bu tüm pencerenin enerjisini temsil eder.

c) Konuşmanın başlangıç ve bitiş yerinin tespiti:

Belirlenmiş ZCR ve enerji seviyeleri yardımıyla konuşmanın başlangıç ve bitiş yerinin tespiti P_{f3}, P_{b3}, P_{f2}, P_{b2} değerleri ile hesaplanır. Bu degerlerin elde edilme yöntemleri aşağıda verilmiştir;

$P_{f3} = \text{argmin}\{E_k > T_E, k=1,2,3..K\}$ ve $T_E = C_E * E_N$ olarak hesaplanır.(C_E denemeler sonucunda elde edililen katsyıdır)

$P_{b3} = \text{argmax}\{E_k > T_E, k=K,K-1,...,1\}$ geriye doğru olarak hesaplanır.

$P_{f3} - P_{b3} < t_{min}$ olduğu taktirde bu aralıkta konuşma yok denilebilir ve pencere işleme sokulmayabilir.

Enerji katsayılarıyla belirlenen aralık daha ayrıntılı ve kesin olarak ZCR katsayılarıyla hesaplanması aşağıdaki şekilde gerceklenir;

$P_{f2} = \text{argmin}\{Z_k > T_{Zf}, k=P_{f3}, P_{f3-1},...,1\}$ ve $T_{Zf} = C_{Zf} * Z_N$

hesaplanır.(C_{Zf} denemeler sonucunda elde edilir)

$P_{b2} = \text{argmax}\{Z_k > T_{Zb}, k=P_{b3}, P_{b3+1},...,K\}$ bulunur ve $T_{Zb} = C_{Zb} * Z_N$ hesaplanır.(C_{Zb} denemeler sonucunda bulunur)

Sonuç olarak elde edilen konuşma aralığı P_{f2} ve P_{b2} olarak belirlenmiş olur.

4. Ses Verilerinin Karsılaştırılması

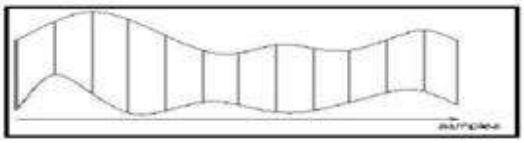
İlk olarak ses verisi işlenerek frekans bölgesine aktarıldı. İkinci adımda bu işlenmiş ses verilerini karşılaştırarak aranan fonemin konuşma içinde varlığını ve eğer varsa kaç kez bu konuşmada geçtiği belirlenmesine geçildi. Bunun için iki farklı yöntem kullanılmasına karar verilmiştir. Bu yöntemler sırası ile Cross Correlation (Çapraz Korelasyon) ve Dynamic Time Warping (Dinamik Zaman Ç). Cross Correlation yöntemi iki sinyalin benzerliğini ölçmede kullanılan yöntemlerden biridir ve bazen cross covariance (Çapraz Kovaryans) olarakta bilinir. Bu yöntemin formülü [3] formül 3'de verilmiştir:

$$(f \star g)(x) \stackrel{\text{def}}{=} \int f^*(t) g(x+t)\, dt$$

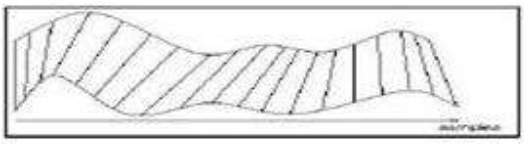
Formül 3. Formüldeki integral (t)ye uygun değerler çerçevesinde işlem yapıyor.

Formü1'deki eşitlik giriş verilerine kaydırılarak tüm verilere uygulanarak karşılaştırma gerçeklenmiştir. Buradaki bu yapının eksikliği kaymalara karşı yetersiz olmasıdır. Burada, verilerde kayma(zaman ekseninde karşılaştırılan verinin geç başlaması veya içinde bazı gürültüler barındırması gibi) olduğunun farkına varılmış fakat sonuç olarak eşleşme

yüzdesi düşük çıkması engellenememiştir. Bu problem Dynamic Time Warping(Dinamik Zaman Çarpıtma) ile aşılmaya çalışılmıştır. Bu yapı kayma varsa bile bu kaymadan etkilenmeden işlem yaparak eşlemenin oranını çok daha doğru şekilde üretebilmektedir. Bu yöntemin diğer yöntemden farkı grafiksel olarak gösterimi şekil 3(a) ve şekil3(b)'de verilmiştir..

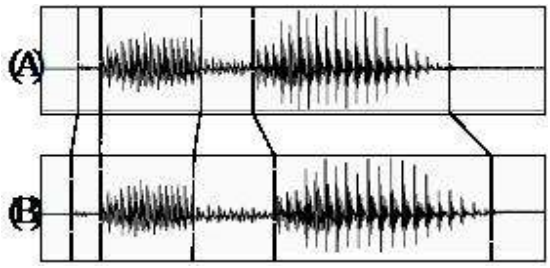


Şekil 4.(a) Cross Correlation eşleme



Şekil 3.(b) DTW ile eşleme

Genel olarak DTW, kesin kısıtlamalar altında iki farklı seri arasındaki benzerliğin hesaplanmasını sağlamktadır. Örneğin Şekil 4'te ki gibi iki ses işareti verilsin



Şekil 4. Farklılıkları veya kayma olan A ve B gibi iki ses işaretinin eşlenmesi

Bu iki ses işaretinin birbiriyle DTW yöntemi ile karşılaştırılmasının algoritması[4] aşağıda verilmiştir;

1. Başlangıç (zaman 1 ilk zaman çerçevesi)

   D(1,1)=d(1,1)

2. Özyineleme

$$D(x,y) = \min_{(x',y')} \left[D(x',y') + \zeta((x',y'),(x,y)) \right]$$

$$\zeta((x',y'),(x,y)) = \sum_{l=1}^{L} d(x_l, y_l) m(l)$$

$l =$ index along valid path from (x', y') to (x, y), from 1 to L

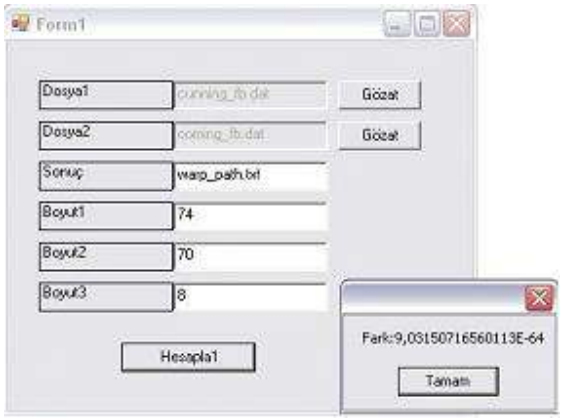
3. Sonlandırma

$$result = \frac{D(x_T, y_T)}{M = \sum_{k=1}^{T} m(k)}$$

(M bazen böylede tanımlana bilir:

T_x, T_x+T_y,veya $(T_x{}^2 + T_y{}^2)^{½}$)

Bu algoritmanın uygulanması ile DTW gerçeklenir. Böylece ikinci yöntemle de veride herhangi bir kayma varsa bile onu en yüksek oranda eşleyebilecek yapıyı da gerçeklenmiş olur.



Şekil 5. DTW için özel proğramın ekran görüntüsü

Sekil 6. Karsılaştırmaları yapan ana programın ekran görüntüsü

Sekil 5'te DTW ekran görüntüsü ve bir özel sonuç verilmiştir. Burada karsılaştırma yapılmış ve sonuçlar belirlenmiş ve sonuçlar bir dosyasında saklanmıştır. Burada, Şekil 5'da, bir örnek uygulama olarak, aralarında gecikme olan iki işaret arasındaki fark %9 olarak bulunmuştur. Sekil 6 de ise manüel seçenekler sunarak gözlem yapabilme olanağı sunan ana programın örnek ekran çıktısı verilmiştir.

Sonuç

Oluşturulmuş sistemle ses işaretinin işlenmesi ve belirli fonem kayıtlarının bir konuşmanın içerisinde geçip geçmediği eşlenmesi yaparak konuşulan dili belirleyen bir uygulamanın sonucu sunulmuştur. Bu çalışmada verilerin eşlemesinde eğer kayma yoksa %99 oranında başarı elde edilmiştir. Verilerde kayma veya gürültü var ise bu bozuklukların değerine göre eşlenme %40'lara kadar düştüğü tespit edilmiştir. Böylece çalışmada %40 ile %98 arasında değişen eşlenme oranları elde edilmiştir. Eşlenmenin %50'nin altında olduğu değerler kabul edilmemiştir. Böylece eşlenme sayısının hesaplanmasında eşik değer olarak %50 belirlenmiştir. Konuşma içinde geçen belirli fonemlerin kullanılma sayısının tespit edilmesi eşik ölçütüne bağlı olarak yapılmıştır. Önerilen bu yöntemin bir konuşmacının konuştuğu dilin belirlenmesine olanak sunup sunmayacağı test edilmektedir. Bu aşamada mevcut sistem manüel çalıştırılmaktadır. Böylece kullanıcıların uyguladığı yöntemleri adım adım izlemesine olanak sunulmuştur. Sistem istenildiği takdirde tam otomatik olarak çalışabilecek duruma getirilebilir. Ancak sistemin tam

otomatik olabilmesi için sistemin belirleme ve karar olgunlaştırılması gerekmektedir. Böylece bütünleşik bir sistemle konuşulan bir dil otomatik algılanabilir.

Kaynaklar

1. Rıfat Yazıcı ve Cemal Köse, Kısıtlamasız Türkce Söz Sentezi, 1. Sinyal işleme ve uygulamaları kurultayı, 1993,Boğazici Üniversitesi
2. Srinath Hosur and Ahmed H. Tewfik, "Wavelet Transform Domain Adaptive FIR Filtering", IEEE Transactions on Signal Processing, Vol. 45, No. 3, pp.617-630, March 1997
3. Gary L. Pavlis, Peng Wang, Frank Vernon, Generalized Cross Correlation: Receiver and Source Array Processing
4. John-Paul Hosom, Hidden Markov Models for Speech Recognition, Oregon Health & Science University OGI School of Science & Engineering, April 5 2006
5. Yiying Zhang, Xiaoyan Zhu, Yu Hao and Yupin Luo, "A Robust And Fast Endpoint Detection Algorithm For Isolated Word Recognition" 1997 IEEE International Conference on Intelligent Processing Systems
6. Hidefumi KOBATAKE, Katsuhisa TAWA and Akiraısh "Speech/nonspeech discrimination for speech recognation system under real life noise environments