İNSAN İLE BİLGİSAYAR ARASINDA SESLİ İLETİŞİMİN İYİLEŞTİRİLMESİ

Hasan Feyzi Özustaoğlu[1], Prof. Dr. Arif Nacaroğlu[2]

[1, 2]Elektrik-Elektronik Mühendisliği Bölümü, Gaziantep Üniversitesi
Şahinbey, Gaziantep.
[1]*e-posta: hasanfeyzi77@hotmail.com* [2]*e-posta : arif1@gantep.edu.tr*

Anahtar sözcükler: sinyal işleme, ses tanıma

ABSTRACT

Reliable speech recognition is a hard problem, requiring a combination of many techniques; however modern methods have been able to achieve an impressive degree of accuracy but the improved algorithms still have some problems. This paper attempts to examine one of those techniques, and to apply them to build a sample voice recognition system. Pattern recognition method is used to recognize isolated words, spoken individually. The speech data is loaded in to the computer using data acquisition card. This data is processed by software prepared in MATLAB. Linear Predictive Coding and Dynamic Time Warping method is used in the program algorithm.

1. GİRİŞ

Konuşma insanların arasında bilgi alışverişi için en etkili ve doğal bir yoldur. İnsanlar ve makineler arası iletişimi sağlamak için ve akıllı bilgisayarlar elde etmek için ise makinelerin "duyabilir", "anlayabilir" ve buna gore "davranabilir" olması önem kazanmaktadır. Bu durumda bilgisayar-insan iletişiminin sağlanabilmesi için "konuşma tanıma" bir bilgisayara gerekli duruma gelmektedir. Ses tanıma sistemleri insanlar arası sesli iletişim sürecinde dinleyicinin yaptığı işlevleri yapay olarak gerçekleştirmeye çalışır.

Bilindiği üzere bilgisayara veri girişi amacıyla geleneksel yöntem olan klavye, fare ve tablet gibi cihazlar kullanılmaktadır. Ses tanıma sistemlerinin kullanılması ile İnsan - Bilgisayar iletişimi ve veri girişi açısından, kullanıcının alışkın olduğu en yaygın iletişim aracını, yani doğal konuşma dilini kullanması bilgisayar kullanımını kolaylaştıracaktır.

Güvenilir konuşma tanıma, bugün modern metodların da etkileyici bir doğruluk payı elde etmelerine rağmen halen bazı problemleri olan ve bir çok tekniği içeren zor bir problemdir.

Bu bildiride ses tanıma ile ilgili bu tekniklerden birisi ele alınmış ve örnek bir ses tanıma sistemi geliştirilmiştir. Konuşma sesleri ses kartı kullanılarak bilgisayara aktarılmış ve bu veri MATLAB'da hazırlanmış program yardımı ile işlenmiştir. Program Linear Predictive Coding ve Dynamic Time Warping yöntemleri kullanılarak hazırlanmıştır.

2. SES TANIMA SÜRECİNDE KULLANILAN TEKNİKLER

Ses tanıma süreci, sesin kaydedilmesi ile başlar, sesin işlenmesi, öz niteliklerinin çıkarılıp kaydedilmesi , karşılaştırma ve eşleştirme yapılarak sesin tanınması ile son bulur.

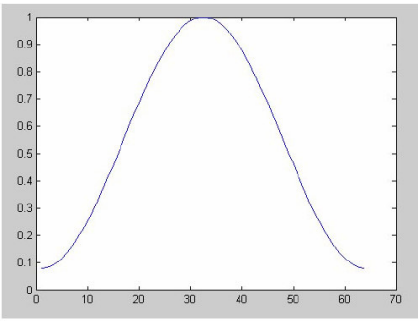
Bu süreç boyunca ses çeşitli aşamalardan geçer. İlk olarak sesin sayısallaştırılması işlevi gerçekleştirilmektedir. Dijital sinyal işleme tekniklerinin kullanılabilmesi için analog sinyalin, yani konuşmanın, bir seri sayılar şeklinde gösterimi gerekmektedir. Bu işlem, analog sinyalin örneklenmesi ile yapılır [1]. Sesin özelliklerinin korunması için örnekleme, dönüşümü yapılacak ses sinyalinin içerdiği en yüksek sıklıktaki frekansın en az iki katı sıklıkta gerçekleştirilir. Bu; örnekleme teoremi olarak ifade edilir.

Ayrıca kaydedilen sesin içerisinde konuşmanın olmadığı sessiz bölümlerin de ayıklanması ve temizlenmesi ses tanıma sistemleri için gerekli bir işlemdir. İfadenin saptanması sırasında kullanılan teknikler, ses sinyalinin alınan bir çerçevedeki sıfırı geçiş sayısı (1) ile toplam veya ortalama enerjinin (2) hesaplanarak eşik değerlerle karşılaştırılması yöntemidir [2]. Bu iki değer birbirini tamamlamaktadır. Kelimenin olduğu yerlerde enerji seviyesi yüksek, sıfırı geçiş sayısı ise düşüktür.

$$ZCR = \sum_{n=0}^{N-1} \frac{1 - \mathrm{sgn}[x(n+1)]\mathrm{sgn}[x(n)]}{2} \quad (1)$$

$$\hat{E}(n) = \sum_{m=0}^{N-1} |w(m)x(n-m)| \quad (2)$$

Ses sinyali içerisinden ifadenin bulunması işleminde veya sesin öz nitelikleri çıkartılırken ses sinyalinin düzeltilmesi amacıyla filtreler kullanılır [3]. Pencereleme fonksiyonu da diyebileceğimiz bu filtreleme için geliştirilen uygulamada Hamming penceresi kullanılmıştır (3). Şekil 1'de, Hamming penceresi görülmektedir.

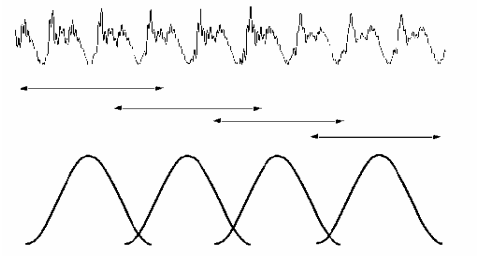


Şekil 1. Hamming penceresi

Hamming penceresi yardımıyla ses sinyalinin merkezi belirginleştirilmektedir.

$$w(m) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}),\ 0 \le m \le N-1 \quad (3)$$

Örneklenen ses bilgisi bir bütün olarak işlenmek yerine daha etkili işlemeye yardımcı olan çerçevelere bölümleme işlemine tabi tutulur. Ve her bölüm yukarıda anlatılan pencereleme işlemine tabi tutulur. Geliştirilen uygulamada kullanılan Hamming penceresi uygulandığı çerçevenin orta noktalarını iyileştiriken uç noktalarında bilgi kaybına neden olmaktadır. Pencereleme yöntemininin bu etkisi pencerelerin üst üste bindirilmesi ile çözülür [3]. Şekil-2'deki sinyal çerçevelenmiş, üst üste bindirilmiş ve pencerelenmiştir.



Şekil 2. Sinyalin üst üste binen çerçeveler ile çerçevelenmesi ve pencere fonksiyonuna tabi tutulması

Sesin sayısallaştırılmasından sonra sesin kodlanması gerçekleştirilir. Geliştirilen uygulamada kodlayıcı olarak, insan gırtlağı ve ağız yapısı özelliklerinin yanı sıra, ses özelliklerini de dikkate alan LPC kodlayıcısından yararlanılmıştır [4].

Ses tanıma sürecinin sonraki aşaması karşılaştırma ve eşleştirmenin yapılmasıdır. Bu aşamada, işlenmiş, kodlanmış ve özellik çıkarımları yapılmış olan ses sinyalinin sistemde kayıtlı şablonlarla karşılaştırma ve eşleştirmesi yapılarak sesli ifadenin kelime karşılığı belirlenmektedir. Bu aşamada iki veri serisinin zamana göre genişletilerek ya da daraltılarak zaman ekseninde örtüştürülerek karşılaştırmasını gerçekleştiren Dynamic Time Warping (DTW) kullanılmıştır.

3. LINEAR PREDICTIVE CODING

Doğrusal önkestirim temel olarak, sesin, periyodik dürtü veya rasgele gürültü ile uyarılan, doğrusal ve zamana göre değişen bir sistemin çıktısı ile modellenebileceği prensibine dayanır. Bu model doğrusal bir filtre olarak (1)'deki transfer fonksiyonu ile ifade edilmektedir. Burada *i*, LPC kodlayıcının seviyesi olarak ifade edilir [5].

$$H(z) = \frac{G}{1 - \sum a_i z^{-i}} \quad (4)$$

4 nolu bağıntıda, ters z-dönüşümü uygulandığında, 5 nolu bağıntı elde edilmektedir.

$$x[n] = \sum_{i=1}^{p} a_i \cdot x[n-i] + e[n] \quad (5)$$

LPC, sıradaki örneğin, önceki bir seri örnekten yaklaşık olarak elde edilebileceği prensibiyle çalışır (6).

$$\hat{x}[n] = \sum_{i-1}^{p} a_i \cdot x[n-i] \qquad (6)$$

Tahmin sonucu elde edilen örneğin asıl örnekle olan farkının; yani hatanın kareleri toplamının minmizasyonu için bir seri parametre hesaplanmaktadır (7).

$$e[n] = x[n] - \hat{x}[n] = x[n] - \sum_{i-1}^{p} a_i \cdot x[n-1] \qquad (7)$$

Hesaplanan bu parametreler LPC parametreleri olarak ifade edilirler. 7 nolu bağıntının çözümü ile *p* sayıda LPC parametresi hesaplanmaktadır.

4. DYNAMIC TIME WARPING

Kayıtlı şablon kelimeler ve tanınacak ses birbirleriyle karşılaştırılacak ve bu eşleşmede en kısa mesafe bize tanıma sonucunu verecektir.

Belirli bir sözcüğün seslendirilmesi, kişiden kişiye hatta aynı kişinin farklı zamanlarda seslendirmesi ile zaman içinde farklılık gösterebilmektedir. Aynı sözcüğün seslendirilmesi, bir seslendirmede uzun, bir seslendirmede ise daha kısa zamanda gerçekleştirilebilir. Aynı zamanda, ses sinyalinde kimi fonemler daha uzun, kimileri ise daha kısa yer almaktadır. Dynamic Time Warping algoritması yardımıyla, bu iki seslendirme, zaman içinde yayılarak ya da daraltılarak birbirine yaklaştırılmaya çalışır [6]. Yani bu iki seslendirmenin, zaman olarak örtüştürülmesi işlevi gerçekleştirilir.

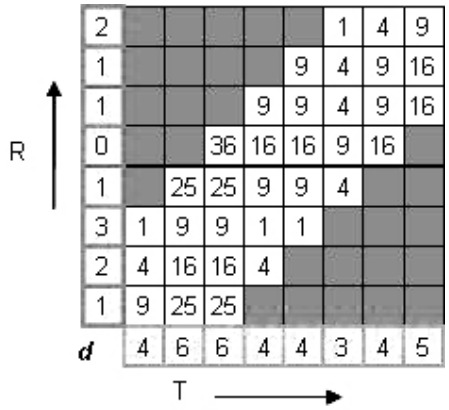
DTW, sözcük tabanlı ses tanıma sistemlerinde etkin ve sıkça kullanılan bir yöntemdir. Bu yaklaşımla, çalışma anında tespit edilen sözcük kesimlemesi, sistemde kayıtlı sözcük şablonları ile seslendirme zamanları örtüştürülerek karşılaştırılması gerçekleştirilebilir.

Geliştirilen uygulamada eşleştirilecek iki veri DTW yöntemi kullanılmak üzere (8), (9) ve (10)'de verilen bağıntılar ile mesafe matriksi oluşturularak Şekil-3'te görüldüğü gibi eşleştirilir.

$$M = \frac{A'*B}{(EA)'*(EB)} \qquad (8)$$

$$EA = \sqrt{(\sum A^2)} \text{ ve } EB = \sqrt{(\sum B^2)} \qquad (9)$$

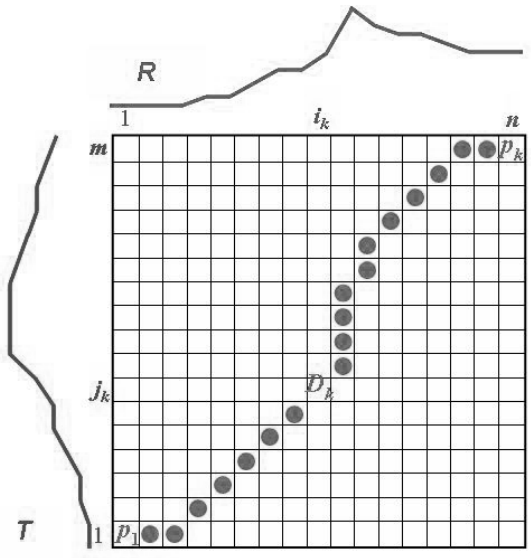
$$N = 1 - M \qquad (10)$$



Şekil 3. Mesafe matriksine DTW uygulaması

DTW algoritması oluşturulan fark matrisi içerisinde (11)'deki bağıntı kullanarak Şekil-4'te görüldüğü gibi minimum mesafeyi arar. Şablonların tamamına uygulanan bu işlem neticesinde en kısa sonuca sahip kelime, tanıma neticesidir.

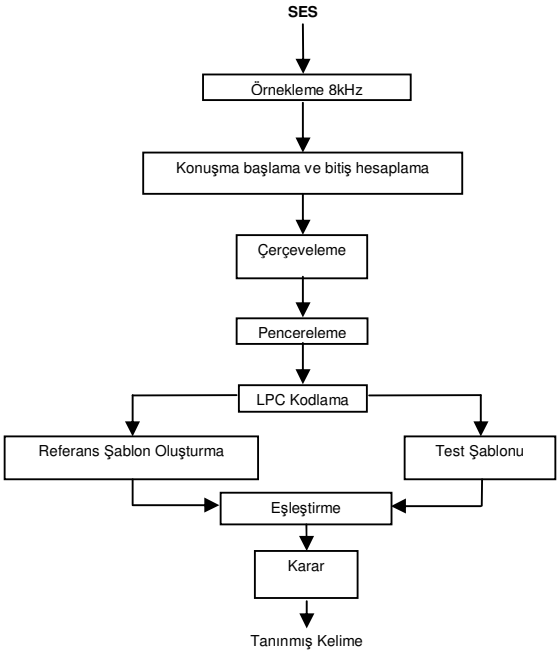
$$D(i_k, j_k) = \min\left[D(i_{k-1}, j_{k-1})\right] + d_f(i_k, j_k) \qquad (11)$$



Şekil 4. DTW algoritmasının en kısa mesafeyi bulması

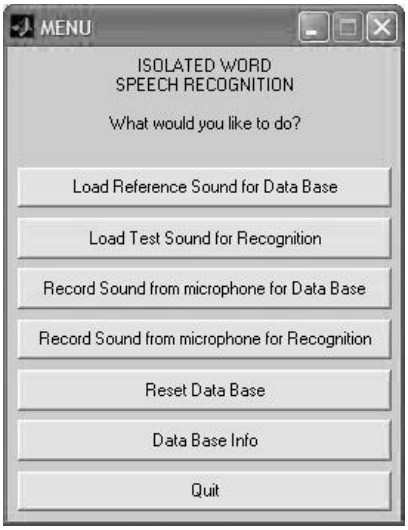
5. GELİŞTİRİLEN SİMULASYON UYGULAMASI

Geliştirilen simulasyonun modeli, Şekil 5'de veri akış diyagramı ile verilmektedir.



Şekil 5. Geliştirilen uygulama modeli.

Şekil 6'da kullanıcı grafik arayüzü verilmektedir.



Şekil 6. Geliştirilen simulasyonun kullanıcı grafik arayüzü.

Ana İş Parçacığı, kullanıcı grafik arabirimi olup, sesin Mikrofon aracılığı ile yüklenmesi veya daha önce bilgisayara kaydedilmiş seslerin programa yüklenmesi ve alt programlar yardımı ile LPC kodlama yapılarak veri bankasına kaydedilmesi, test kaydının yapılarak tanıma işlevinin DTW algoritması kullanılarak gerçekleştirilmesi, veri bankasının temizlenmesi ve programdan çıkış aşamalarını içerir.

Kaydetme işlemi 8Khz'lik örnekleme ve 8 bit çözünürlük ile yapılmıştır. Ses sinyali içerisinden başı ve sonu bulunan kelimenin verisi çerçevelenmiş ve bu çerçeveler üst üste bindirilmiştir. Ardından Hamming pencereleme fonksiyonundan geçirilerek ifadenin orta kesimleri belirginleştirilmiş ve ses sinyali düzeltilmiştir.

Sonrasında bu örnek serisi LPC kodlayıcıya gönderilmiştir. LPC kodlayıcısı, her bir çerçeveye karşılık, RMS değeri, perde değeri ve 10 LPC parametresinden oluşan toplam 12 değer göndermektedir. Bu değerlerin bir serisi oluşturulmakta ve tamamlandığında LPC kuyruğuna eklenmektedir. Sonrasında bulunan bu değerler ile sistemde şablon kayıt işlemi gerçekleştirilmektedir. Bu şekilde sistemin eğitimi gerçekleştirilmektedir.

Şablonlar kaydedilirken daha sonar tanınmak amacı ile etiketlenmektedir. Tüm kelimeler için bu şekilde şablon kaydı gerçekleştirilmesi gerekmektedir.

Karşılaştırma için test şablonu tüm kayıtlı şablonlarla DTW algoritması ile birebir karşılaştırılmakta ve en yakın şablon aranmaktadır.

Tüm şablonlarla karşılaştırma tamamlandıktan sonra en yakın şablonla eşleştirme yapılmakta ve bu şablonun etiketi arayüze yansıtılmaktadır.

6. SONUÇ

Geliştirilen uygulamanın başarısını test etmek için tek bir kişiden 10 ayrı kelime kaydedilerek referans şablonu oluşturulmuş ve sonradan 10 ayrı kişiye teker teker bu 10 kelime test ettirilmiştir. Yapılan testin tanıma sonuçları Tablo-1'de görülmektedir.

Yapılan araştırmalar ve çalışmalar doğrultusunda başarının ortamın gürültüsü, sesin bir seslendirilişi ile bir diğer seslendirilişi arasında çokça farklılık göstermesi, kullanılan mikrofon ve ses kartının

özelliği, seslendirme sırasında mikrofona olan yakınlık veya uzaklık ile etkilendiği görülmüştür. Ayrıca system kişiye bağlı olarak kullanıldığında tanıma oranı artmaktadır.

		TEST ŞABLONU									
		PENCERE	KİTAPLIK	BİLGİSAYAR	İLAÇ	DENEME	ÇİÇEK	ARABA	GÖZLÜK	RESİM	KEDİ
REFERANS ŞABLONU	PENCERE	9					1				1
	KİTAPLIK		7					1			
	BİLGİSAYAR			6							
	İLAÇ			2	8						
	DENEME	1		1	1	9			2		
	ÇİÇEK				1		6			1	
	ARABA		3			1	1	9			
	GÖZLÜK								7		
	RESİM			1					1	7	
	KEDİ						2			2	9

Tablo 1. Geliştirilen uygulamanın tanıma sonuçları

Önerilebilecek yaklaşımlar, gürültü ayıklama programlarının simulasyona ilavesi, farklı zamanlarda kayıtlar sonrası ortaya çıkan şablonlardan, ortalama bir şablon hesaplanması ve tanıma için bu şablonun kullanılması ile başarının arttırılması, C++ veya makine dili gibi farklı proglamlama dilleri kullanımı olabilir..

Sadece LPC parametreleri üzerinde DTW uygulanması yerine, kaydedilmiş sesin farklı tekniklerle başka özellikleri de çıkarılarak LPC'ye ek parametreler olarak saklanabilir ve bu özelliklerin tümünün üzerinde DTW algoritmasının uygulanması ile başarım arttırılabilir.

KAYNAKLAR

[1] Smith, S.W., 'The Scientist's and Engineer's Guide to Digital Signal Processing'(2nd Ed.) California Technical Publishing, ISBN 0-96-601764-1, 1999.

[2] Rabiner, L. R., M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", *The Bell System Technical Journal*, Vol. 54, No. 2, pp. 297-315, 1975.

[3] L.R Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, N.J., 1978.

[4] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 67-72, February 1975

[5] Huang, X., Acero, A. and Hon, H.W., 'Spoken Language Processing: A Guide to Theory, Algorithm and System Development'(1st Ed.) Prentice Hall PTR, ISBN 0-13-022616-5, 2001.

[6] H.F. Silverman and D.P. Morgan, "The Application of Dynamic Programming to Connected Speech Recognition," IEEE ASSP Magazine, pp. 7-25, July 1990.