KONUŞMA TANIMA TEORİSİ VE TEKNİKLERİ[1]

Nursel YALÇIN
Gazi Ü., Endüstriyel Sanatlar Eğitim Fakültesi, Bilgisayar Eğitimi Bölümü, Ankara.

Özet

Konuşma tanıma, insan sesinin bilgisayar tarafından algılanmasıdır. Bu çalışmada genel bir konuşma tanıma modeli verilmiştir. Konuşma tanıma sistemlerinde kullanılan yöntemler ve teknikler açıklanmıştır. Konuşma tanıma teorisi belirtilmiştir.

Anahtar Kelimeler: *Ayrık Sözcük Tanıma, Sözcük Yakalama Sistemleri, Sürekli Konuşma Tanıma, Yapay Sinir Ağları, Saklı Markov Modelleri, Zaman Eşleştirme,*

SPEECH RECOGNITION THEORY AND TECHNIQUES

Abstract

Speech recognition is perceived human's voice by computer. In this study, it is given a generally speech recognition model. It is explanation techniques and methods using in speech recognition systems. Speech recognition theory is determined.

Key Words: *Isolated Word Recognition, Word Spotting Systems, Continuous Speech Recognition, Neural Networks, Hidden Markov Models, Time Warping*

1. Giriş

Ses tanıma alanı içerisinde bulunan konuşma tanıma disiplini, gelişen teknoloji sürecinde kendine önemli bir yer edinmeye çalışan bir sistemdir ve insan sesinin bir mikrofon vasıtasıyla bilgisayar tarafından algılanarak tanınması işlemidir. Bu işlem ise insan-bilgisayar iletişiminde önemli bir ihtiyaç halini almaktadır. Çünkü artık insanlar klavyeyi kullanmadan bilgisayara bir şeyler yazdırmak veya bir şeyler yaptırmak istemektedirler. Microsoft'un teknolojik bir devrim olarak nitelendirdiği konuşma tanıma sistemi 1950'li yılların sonlarından itibaren üzerinde çalışılan bir alan olmuştur. Ses tanıma başlı başına zor bir çalışma disiplinidir. Ses tanıma problemi birbirinden çok farklı alt problemleri içermektedir. Konuşmacı belirleme, konuşmacı tanıma, konuşmacıdan bağımsız tanıma sistemleri, konuşmacıya bağımlı tanıma sistemleri, ayrık sözcük tanıma, anahtar sözcük yakalama ve sürekli konuşma tanıma sistemleri. Sesin analizinde veya tanınmasında sesli veri girişinin metne dönüştürülmesi üzerinde durulmaktadır. Sayısala dönüştürme işlemleri sırasıyla örnekleme, nicelendirme ve kodlama aşamalarıdır. Örnekleme olarak sayısal işaretten belirlenen anda genlik değerlerinin alınması düşünülmektedir. Nicelendirme, örneklenmiş işareti belirli aralıklara bölme ve basamaklandırma işlemidir. Kuantalama değeri her veriye ayrılacak olan bit sayısını ifade etmektedir. Kodlama ise kuantalanmış işaretin herhangi bir sayı sisteminde gösterilmesidir (1). Konuşma tanıma gerek bu konuda yapılan çalışmaların azlığı gerekse bu alandaki yapılacak çalışmalara gereksinimden dolayı oldukça ilgi çekicidir. Ticari açıdan bakılacak olursa, konuşma tanıma potansiyel olarak çok büyük

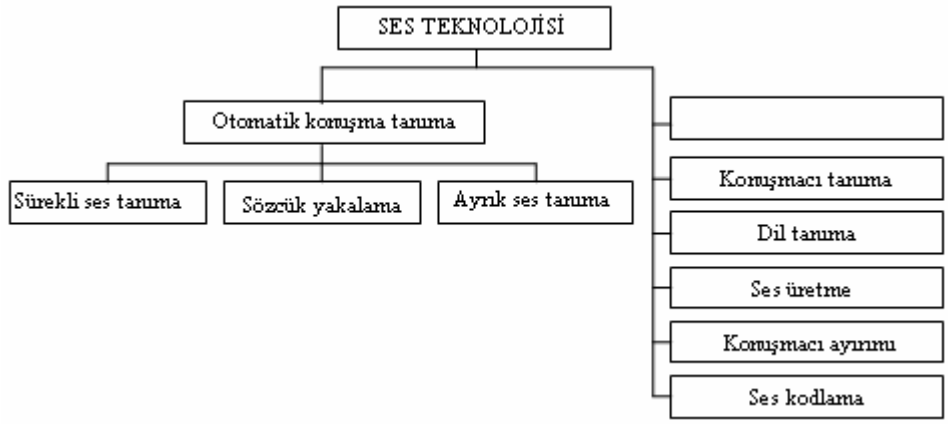
[1] *Bu çalışma Gazi Üniversitesi Rektörlüğü Bilimsel Araştırmaları Projeleri Birimi tarafından desteklenmiştir.*

pazarı olan bir teknolojidir. İnsan sesinin kişiden kişiye göre değişmesi ve üzerinde çalışılan dilin çok geniş bir içeriğe (olası cümle veya sözcük olarak) sahip olması bu çalışma alanının oldukça zor bir disiplin olmasına sebep olmuştur. Ancak insanlığın bu alanda yapılacak çalışmalara ve yeniliklere acil ihtiyacı vardır. Ses tanıma teknolojisindeki gelişmeler sayesinde, ağır işitenler daha iyi işitebilmekte, sağır olanlar ise konuşmanın anında yazıya çevrilmesi ile canlı yayınlardaki konuşmaları anlayabilmektedirler. Ses tanımayla ilgili bu örneklendirmeleri çoğaltmak mümkündür. Sesin bir parmak izi kadar ayırt edilebilir özelliği bulunmasından dolayı güvenlik işlemlerinde mutlaka kullanılması gereken bir teknoloji olarak karşımıza çıkmaktadır. Bu alanda daha ziyade konuşmacı tanıma ve konuşmacı belirleme disiplinleri önem kazanmaktadır. Üzerinde durulması gereken bir diğer ses tanıma alanı ise tıp ve eğitimdir. Eğitim alanında ise (2) "İlköğretim birinci sınıf öğrencilerine konuşma tanıma teknolojisi yardımıyla ilkokuma yazma" konulu çalışmalarıyla ilkokuma yazma öğretimine farklı bir yaklaşım fırsatı sunmuşlardır. Tıp alanında ise konuşma terapisinde kullanılan bir alan olmaktadır.

Bu makalenin 2. bölümünde ses teknolojilerine değinilerek konuşma tanıma teorisi açıklanmıştır. 3. bölümde ise konuşma tanımada kullanılan teknikler üzerinde durulmuştur. 4. Bölümde ise konuşma tanıma üzerine genel bir değerlendirme yapılarak sonuç verilmiştir.

2. Konuşma Tanıma Teorisi

Ses teknolojisi yedi ortak konuşma uygulamalarını içermektedir. Şekil 2.1 de ses teknolojilerinin çeşitleri görülmektedir.

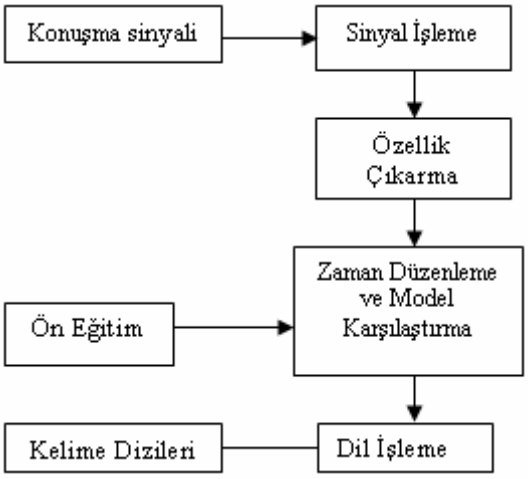


Şekil 2.1. Ses teknolojisi alanları

Ses teknolojisi içerisinde bulunan otomatik konuşma tanıma, sürekli konuşma tanıma, ayrık sözcük tanıma ve kelime yakalama sistemlerini içerisinde bulundurmaktadır. Diğer taraftan ses sentezleme, konuşmacı tanıma, dil tanıma, ses üretme, konuşmacı ayırımı ve ses kodlamayı da içine alan ses teknolojileri içinde en zor olan alan otomatik konuşma tanımayla ilgili olan alandır (3). Otomatik Konuşma tanıma problemi, verilen bir akustik X dizisi için, W kelime dizilerini bulmak için oluşturulmuştur. Konuşma cümleleri, $W = (w_1, w_2, \ldots, w_t)$ şeklinde belirtilen

kelimelerin dizisi olarak gösterilir. w_t, ayrık bir t zamanında söylenmiş belli bir kelimedir. Kelimelerin dizisi söylenen sesli ifade ile bağlantılıdır ve bu sesli ifade X olarak gösterilen akustik sesler dizisidir.Otomatik konuşma tanıma oldukça zor ve karmaşık bir problemdir. Matematiksel olarak bir f fonksiyonu f: X → W şeklinde gösterilir (4). Otomatik konuşma tanıma teorisi oldukça geniş bir konudur. Genel olarak bir örüntü tanıma (pattern recognition) problemi olarak görülebilir. Konuşma sinyalinin stokastik yapısı istatistiksel yöntemlerle konuşma tanıma yapılmasını gerekli hale getirmektedir. Genel bir konuşma tanıma modeli şu modüllerden oluşmaktadır (Şekil 2.2) (3, 5).

- Konuşma tanıma sinyalinin bir gösterimini elde etmek için sinyal işleme modülü
- Bu gösterimin anahtar elemanlarını belirlemek ve fazla bilgiyi çıkarmak için özellik çıkarma modülü
- Kelime tespitini yapmak için zaman düzenleme ve model karşılaştırma algoritmaları
- Bir final kelime dizisi seçmek için dil modeli



Şekil 2.2. Genel Bir Konuşma Tanıma Modeli (3)

Konuşma sinyalinin modellenmesi, sesli ifadelerin, bilgisayar destekli olarak tanıma sürecine sokulabilmesi için bunların öncelikle bu sürece hazırlanmaları gerekmektedir. Bu amaçla sesli ifadelerin bir mikrofon aracılığıyla örneksel sinyallere dönüştürülmesi, sayısallaştırılması, sayısallaştırılan bu sinyallerin gerekirse filtrelenmesi, etiketlenmesi (örneğin sesler, fonemler, sözcükler olarak) ve tanıma işlemlerine taban oluşturacak parametrik yapılar ya da yalın modellerle ifade edilen biçimlere dönüştürülmesi gerekmektedir.

Sinyal işleme modülünün amacı, örneklenen konuşma sinyalini işlemek ve genlik değişimlerinden, konuşmacı stresinden, iletişim araçlarından veya kanalından gelen gürültüden bağımsız bir gösterim elde etmektir. Sayısala dönüştürme işlemleri sırasıyla örnekleme, nicelendirme ve kodlama aşamalarıdır. Örnekleme olarak sayısal işaretten belirlenen anda genlik değerlerinin alınması

düşünülmektedir. Nicelendirme, örneklenmiş işareti belirli aralıklara bölme ve basamaklandırma işlemidir. Kuantalama değeri her veriye ayrılacak olan bit sayısını ifade etmektedir. Kodlama ise kuantalanmış işaretin herhangi bir sayı sisteminde gösterilmesidir (1). Özellik çıkarım modülü, sinyaldeki geçişleri yakalayan parametrelerin kümesini hesaplar ve her fonemi göstermek için yeterince güçlüdür. Bu parametreler genellikle özellikler olarak adlandırılırlar ve genellikle sabit-zaman aralıklarında hesaplanırlar. Sinyaldeki geçişler, konuşma sinyalindeki fonetik bilginin kodlanmasını gösterebilen önemli ipuçlarıdır (3). Giriş ses bilgisi işe yarayan ve yaramayan bir çok bilgi içermektedir. Örneğin, ses bilgisi kelimenin ne olduğu, konuşanın cinsi, duyguları ve fiziksel durumu gibi bize çok sayıda bilgi vermektedir. Ses tanıma siteminde işimize yarayan bilgiler seçilmeli ve diğerleri ise çıkartılmalıdır. Yararlı bilgi özellikleri içermeli ve örüntüleri birbirinden ayırmayı sağlayabilmelidir. Özelliklerin belirlenmesi örüntü tanımanın en önemli aşamasıdır. Segmentasyon işlemi cümledeki kelimeleri ayırma işlemidir. Konuşulan kelimenin başının ve sonunun bulunması çok zor bir problemdir. Bu zorluk sesin bitişi ile diğerinin başlangıcı arasında kesin sınırın belirlenememesinden kaynaklanmaktadır. Zaman düzenleme ve model karşılaştırma modülü, bir kelimeyi, o kelimenin verilen gösterimlerine dayanarak karşılaştırmayı dener. Zaman düzenleme, modellenmiş akustik veya fonetik olayların düzenlenmesi için, konuşma oranındaki değişikliklerden dolayı "sözdeki zaman bozulmalarını" referans alır. Bu zaman bozulması veya geçici değişimler, doğal olarak olur ve özellikle konuşma oranındaki değişimler seslilerin süresini önemli şekilde etkiler. Zaman düzenleme ve model karşılaştırma, bir kelime modelini formüle etmek için kullanılan eğitim yöntemlerine bağlıdır. Kelime modeli, eğitim safhası esnasında çıkarılan parametrelerin kümesinden oluşur. Dil işleme modülü, konuşma tanıma işleminin son aşamasıdır. Hangi dile yönelik konuşma tanıma sistemi hazırlanacaksa o dile ait çeşitli kuralların bilinmesi ve bu kurallara uyarak çalışılması gerekmektedir. Dile ait bu yapılar bilinerek kelime seçimlerinin oluşturulmasıyla ilgili modüldür.

Konuşma tanıma sistemleri kabul ettikleri girdi ses sinyalinin ve içermesi beklenen metin karşılıklarının yapılarına göre çeşitli dallar altında incelenir:

- Ses sinyalinin tek veya çok kişiye ait olabilmesi: Konuşmacıya bağlılık
- Tanınacak metin kümesinin genişliği: Dağarcık
- Tanınacak seste metin elemanlarının yerleşimi: Sürekli, bağlı veya ayrık tanıma

2.1. Konuşma Tanıma Sistemlerinin Sınıflandırılması

Konuşma tanıma programları, alanların ihtiyacına göre tasarlanır. Bu alanlarda çıkan sorunların çözümü için değişik sistemler ve metotlar geliştirilmiştir. Sesli ifade sistemleri ya da sesli ifade tanıyıcılar, artan zorluk sırasına göre aşağıda sıralanmıştır:

- Ayrık sözcük tanıma sistemleri (isolated word recognition systems),
- Sözcük yakalama sistemleri (word spotting systems),
- Sürekli konuşma tanıma sistemleri (continous speech recognition system).

2.1.1. Ayrık sözcük tanıma

Yalıtılmış konuşma tanıma olayı kelime haznesi uyuşmasının en temel formudur ve bu sistem herhangi bir şey kullanıcıya sorulduğunda kullanıcının tek kelimelik bir girdi yapmasını bekler. Sözcükler arası duraklar olmak zorundadır (6). Tanıma sözcük tanımadır. Konuşma tanımada zor olan noktalardan birisi de sözcüklerin başlangıç ve bitiş noktalarının belirlenmesidir. Ayrık sözcük tanımada sözcüklerin birbirinden bağımsız telaffuz edilmesi tanımayı kolaylaştırır. Bu yöntem sadece belirli kelimelerin tanınmasının yeterli olduğu alanlarda kullanılır. Özetle;

- En yalın konuşma tanıma şeklidir.
- Sözcük yakalama ve SKT' ya temel oluşturur.
- Sözcükler arasındaki duraklar belirgindir ve tanıma kolaylaşmaktadır.
- Sözcükler birbirinden bağımsız olarak ele alınır.
- Dolayısıyla birlikte seslendirme sorunu yoktur.

2.1.2. Sözcük yakalama sistemleri

Sürekli konuşma içinde belirli bir kelimenin ortaya çıkışını belirleme işlemidir. Bu tür uygulamalarda en başarılı sonucu veren dinamik zaman sıkıştırma programlama teknikleri kullanılır. Yakalanacak her kelime şablon tarafından gösterilir. Sadece şablonlar tarafından bilindiğinden dolayı, sesin bitiş noktalarından bağımsız kılmak önemlidir. Potansiyel başlangıç noktası olarak giren ses akıntısının her örneği süreç olarak kabul edilmelidir. Tanıma işlemi, aranan şablonun sesli ifade içinde çakıştığı bir örüntü arama biçiminde gerçekleşmektedir. Konuşma tanıma sistemleri geliştiricileri insanların normal olarak anlaşılmayan kelimeleri konuşmalarda kullanmaları gerçeği üzerinde durmuşlardır ("um" "eee" vs.). Alışıla geldik bir konuşmada bu kelimeleri geçersiz olarak algılar ve bunun yerine mesaj taşıyan gerçek kelimeler üzerine yoğunlaşılır. Bu ek sesler ise konuşma tanımada sorun çıkarabilir. Araştırmacılar anahtar kelimeleri tanımak için bilgisayarlara insanların doğal olarak ne yaptıklarını öğretirler. Sözcük yakalama olarak adlandırılan bu teknik bilgisayara cümle içinde kelimeyi tanımayı mümkün kılar. Özetle;

- Konuşma içinde aranan sözcüklerin yakalanmasını sağlar.
- En çok dinamik zaman eşleştirme tekniğinde kullanılır.
- Her sözcük bir şablon ile ifade edilir.
- Tanıma işlemi aranan şablonun konuşmada çakıştığı bir örüntü tanıma şeklinde olur.

2.1.3. Sürekli konuşma tanıma

Sürekli sesli ifade tanıma sistemleri, bağlı sözcük tanıma sistemleri ve karşılıklı konuşma tanıma sistemleri olmak üzere iki gruba ayrılabilir. Bunlardan ilki tanıma işlemini sözcük bazında yapmayı hedeflerken, ikincisi cümlenin anlamının da anlaşılmasını hedefler. Bu sebeple karşılıklı konuşma tanıma sistemleri sesli ifade anlama sistemleri olarak da adlandırılabilir ve karmaşık dilbilgisi kurallarının da sistemde yer almasını gerektirir. Bu kısım günümüzde ayrı bir alan olarak ele alınmıştır. Bu alan Doğal Dil İşleme ya da Anlama olarak adlandırılmaktadır. Konuşma tanıma mimarisi içindeki yeri önemlidir. Sürekli sesli ifade tanımadaki zorluk üç değişik

nedenden kaynaklanmaktadır. Bunlardan ilki, sürekli ifade içindeki sözcük sınırlarının belirgin olmamasıdır. Sürekli sözcük tanımada, sözcük sınırlarının bulunması zor, hatta kimi zaman olanaksızdır. İkinci sorun, tanınacak ses birimlerinin, ön ve arkasına gelen diğer ses birimleri tarafından etkilenmeleridir. Üçüncü sorun vurgulama, duraklama gibi bürünlerden kaynaklanmaktadır. İsim, sıfat ve fiiller net ve yüksek enerjili seslerle ifade edilirken aradaki bağlaçlar ve kısa süreli sözcükler çoğu kez düşük enerjili olmakta ya da yutulmaktadır. Örneğin, "bir gün" kelime çifti genellikle "bigün" olarak seslendirilir. Özetle;

- Ayrık sözcük tanıma ve sözcük yakalama sistemlerine göre daha zordur.
- Bu zorluk üç nedenden kaynaklanmaktadır.
- Sözcük sınırları belirgin değildir.
- Tanınacak söz birimlerinin ön ve arkasına gelen ses birimleri tarafından etkilenirler. Seslerin birlikte seslendirilmesinden kaynaklanan bir sorundur.
- Sözcüklerin söyleyişte, hızlı konuşma söz konusu olduğu için, kaybedilmesi söz konusudur.

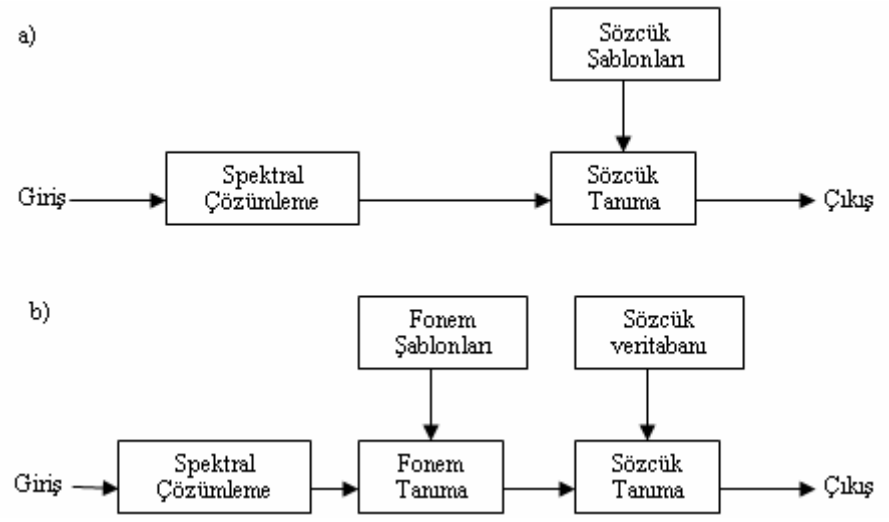
Sürekli ya da ayrık olmalarının dışında sesli ifade tanıma sistemleri konuşmacıya bağımlılığına göre de ikiye ayrılır.

- Konuşmacıya bağımlı sesli ifade tanıma sistemleri (Speaker Dependent).
- Konuşmacıdan bağımsız sesli ifade tanıma sistemleri (Speaker Independent).

Bunlardan ilki tek bir kişi için referans şablonlarının oluşturulmasını öngörür. Yeni kişilerin konuşmalarının tanınabilmesi için referans alınan şablonların günlenmesi gerekir. İkincisinde ise sistem herhangi bir kişi tarafından seslendirilen bir sesli ifadeyi tanıyabilir. Doğal olarak bir sistemin kullanılış alanını artırmak için amaç kişiden bağımsız bir sesli ifade tanıma sistemi olmalıdır. Fakat bunu başarmak kişiye bağımlı bir sistem geliştirmekten daha zordur (7). Bir diğer sınıflandırma da yukarıda söz konusu edildiği gibi tanıma için seçilen birimin büyüklüğüne göre yapılabilir. Bunun için de sesli ifade tanıma sistemlerini ikiye ayırmak mümkündür (7).

- Sözcük tabanlı sesli ifade tanıma sistemleri; tanıma için öngörülen en küçük birim olarak sözcüğün kabul edildiği sistemlerdir (word based speech recognition).
- Fonem tabanlı sesli ifade tanıma sistemleri; tanıma için fonemlerin en küçük birim olarak kabul edildiği sistemlerdir (phoneme based speech recognition).

Bu sistemlerden ilkinin doğruluk derecesi daha yüksek olmaktadır. Çünkü fonemler arası geçişlerin olumsuz etkileri burada gözlenmez. Ancak, bunun yanında sürekli sesli ifade tanıma söz konusu olduğunda sözcükler arası geçişler yine sorun olacaktır. Ayrıca sözcük tabanlı sesli ifade tanıma sistemlerinde referans şablonu olarak sözcüğün tamamı alındığından ve bir dilde çok sayıda sözcük olduğundan sistemin gereksinim duyduğu işleyici zamanı ve bellek gereksinimi çok daha fazla olacaktır. Fonem tabanlı tanımada ise doğruluk yüzdesi bir miktar düşerken, çok az olan fonem sayısı, hızlı sonuç üretme olanağı sayesinde, hata azaltma amaçlı geri dönüşleri mümkün hale getirmektedir. Sözcük ve fonem tabanlı sistemlerden başka bu ikisi arasında sözcük altı birimleri temel alan sesli ifade tanıma sistemlerini de söz konusu etmek mümkündür(7).



Şekil 2.3. özcük tanıma sisteminin yapısı: (a) Sözcük tabanlı sesli ifade tanıma; (b) Fonem tabanlı sesli ifade tanıma (8)

Şekil 2.3' deki sözcük tanıma sistemlerinin ikisinde de amaç tüm sözcüğün tanınmasıdır. Birinci yaklaşımda, sesli ifadeden doğrudan sözcük tanımaya geçilirken ikinci yaklaşımda önce fonem tanıma işlemi yapılmakta, daha sonra bir sözcük veritabanı kullanılarak bu fonemler sözcüklere dönüştürülmektedir. Daha önce de söz konusu edildiği gibi (a) şeklindeki sistemin doğruluğu daha fazla fakat tanıyabildiği sözcük sayısı az olacaktır. (b) seklindeki sistemde ise sözcük sayısı için başlangıçta bir sınır yoktur. Bir sözcük veritabanına sözcük eklemek doğal olarak o sözcük için sesli ifade şablonunun sisteme öğretilmesinden daha kolay olduğu için (b) sekli genişlemeye daha elverişlidir (7). Sesli ifade tanımada, sözcük yerine, fonem gibi alt ses birimlerine dayanılabilir. Bir fonem, genellikle, konuşmanın en küçük birimi olarak tanımlanır. Her dilin standart lehçelerinin özgün fonem grupları vardır (9). Ancak bu ses birimlerinin sayıca kısıtlı olması ve sözcükleri bunların kombinasyonu ile ifade edebilme özelliğinin bulunması gerekmektedir (10). Sözü edilen alt birimler hece, fonem ve ses (phon) olabilmektedir. Sözcükler, her biri kendine özgü niteliği bulunan ve çok sayıda öğelerdir. Ses birim sınırlarının belirlenmesi ve birlikte söylenme etkilerinin göz önüne alınması gibi sorunlar ortaya çıkmaktadır. Fonem tabanlı tanıma işlemlerinde önce alt ses birimleri tanımaya çalışılır daha sonra bu birimler birleştirilerek sözcükler oluşturulur. Fonem: bir dilde bir sözcüğün anlamını diğer bir sözcüğün anlamından ayırmaya yarayan en küçük ses birimlerine verilen addır. Eğer bir sesi (phon) değiştirmek, ilgili sözcüğün anlamını değiştiriyorsa bu ses fonem olarak anılmaktadır. Ancak bir sesi değiştirmek, ilgili sözcüğün anlamını değiştirmiyorsa fonemden söz edilmemektedir. Fonemlerin belirlenmesi için en küçük çiftten yararlanılmaktadır. Gel ve kel de olduğu gibi, salt tek bir sesi değişik olan sözcükler en küçük çift olarak anılırlar. En küçük çift içinde değişik gösteren ses, anlam değişikliğine neden oluyorsa sözkonusu ses fonemdir. Gel ve kel örneğinde görüldüğü gibi, /g/ ve /k/ Türkçe'deki iki ayrı fonemi oluşturmaktadır. Ses (phon) taban alınarak bir tanım vermek gerektiğinde fonem (anlam ayırıcı özelliği bulunmayan) benzer seslerden oluşan ses kümesi olarak tanımlanır. Hece ise bir nefeste söylenen sesbirimi (11, 9, 12).

Konuşma tanıma uygulamalarında kullanılan sözcük-altı birimler, hece gibi göreceli olarak büyük birimlerden fon ve fonem gibi küçük birimlere, hatta dilbilimsel karşılığı olmayan ve sadece akustik olarak tanımlanmış özel bazı birimlere uzanan geniş bir yelpaze oluşturmaktadır (13). Sözcük-altı birimlerin seçiminde dikkat edilmesi gereken iki kriter tutarlılık ve eğitilebilirlik olarak sıralanabilir. Tutarlılık sözcük-altı birimin aynı bağlam içinde aynı karakteristiği, yani çok yakın akustik özellikleri göstermesidir (13). Eğitilebilirlik kriteri ise bu birimlerin iyi modellenebilmesi için eğitim süreci içinde yeterli sıklıkta geçmesi ile ilgilidir. Hece, yarım hece gibi büyük sözcük-altı birimler genellikle tutarlıdır ama sayıca çok olduklarından eğitme süreci zordur. Fon ve fonem gibi küçük sözcük-altı birimler ise göreceli olarak daha az sayıda olduklarından eğitme süreci kolay olmakla birlikte tutarlı değildirler (13).

Konuşmacıdan bağımsızlık, ses sinyal niteliği, öğrenme yeteneği, sözlük dağarcığı, dilbilgisi kullanımı gibi özellikler konuşma tanıma sistemine ilişkin önemli özelliklerdir. Seslerin kayıt edildiği ortam ile kayıt koşullarının niteliği konuşma tanıma sürecini etkileyen önemli bir özelliktir. Ses kayıtlarının yapıldığı ortam, ya özel yalıtılmış, yansımasız bir oda ya da her hangi bir ortam olabilir. Öylesine seçilmiş bir ortamda kaydı yapılan seslerin tanınması, ortamdan kaynaklanan gürültüden dolayı daha zordur. Seslerin kayıt ediliş biçimi ve kayıt kalitesi, en az seslerin kayıt ortamı kadar önemlidir. Kayıt esnasında kullanılan araçlar yüksek kalitede elektronik donanımlar (mikrofon ve yükseltici donanımlar), orta kalitede donanımlar ve telefon gibi özel amaçlı ve koşulların zorlandığı donanımlar olarak sınıflandırılmaktadır. Bir konuşma tanıma sisteminde tanınabilen sözcüklerin oluşturduğu küme sistem sözlüğü ya da kısaca sözlük olarak anılır. Sözcük altı ses birimleri tabanında tanıma yapan sistemlerde sözlük, ses kümesi (phone set) olarak adlandırılır. Sözlük büyüklüğü, genelde tanıma başarısını olumsuz yönde etkilemektedir. Sözlük dağarcığının artırılması işlem gücü gereksinimini ve hata oranını artırmaktadır. Sözlük büyüklüğünü şu şekilde gruplamak mümkündür;

- Küçük dağarcıklı: 100 den daha az sözcük içerenler
- Orta dağarcıklı: 100 -1000 arasında sözcük içerenler
- Büyük dağarcıklı: 1000 -10000 sözcük içerenler ve
- Çok büyük dağarcıklı: 10000 sözcükten fazla sözcük içerenler. Yaklaşık 5000–60000 arası sözcükten oluşan dağarcığa geniş dağarcıklı sözlük denilmektedir (6).

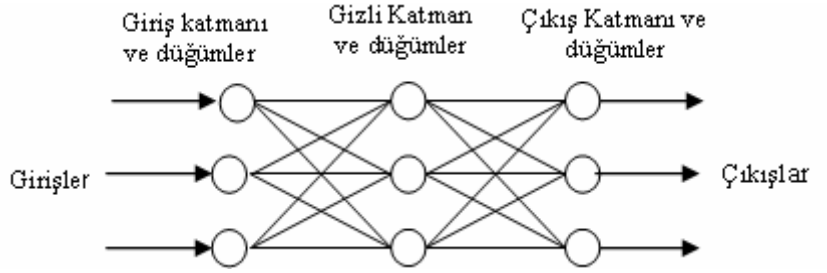
3. Konuşma Tanımada Kullanılan Teknikler

Konuşma tanımada kullanılan belli başlı üç teknik vardır. Bunlar;

- Neural Networks (Yapay Sinir Ağları)
- Time warping - dynamic time warping (Dinamik Zaman Eşleştirme)
- Hidden Markov Models (Saklı Markov Modelleri)

3.1. Yapay Sinir Ağları

Yapay Sinir Ağları; Yapay Zeka' nın konusudur. Yapay Zeka, 1956 yılında John McCarthy tarafından düzenlenen ve Minsky, Newell, Simon, Shannon başta olmak üzere on bilim adamının, iki ay süre ile Dortmouth College' de yaptıkları çalışmaların sonucunda, John McCarthy' in önerisi ile "Artificial Intelligence" ismi ile ilk kez kullanılmış ve yapay zeka bir araştırma disiplini olarak benimsenmiştir (14). Yapay Sinir Ağı (Artificial Neural Net Model), Connectionist Modeller, Paralel Dağıtılmış İşleme Modelleri veya Neuronorphic System gibi değişik şekillerde isimlendirilmektedir (15). Ancak ismi ne olursa olsun bu modeller, biyolojik sinir sisteminin bilinen yapısını göz önüne alarak, yüksek bir performansın elde edilmesini sağlayacak şekilde basit hesaplama elemanlarının yoğun bağlantılarından meydana gelmiştir. YSA modelleri, paralel ve yüksek hesaplama hızlarının gerekli olduğu ve mevcut en iyi bilgisayar sistemlerinin dahi gerçekleştirmekten oldukça uzak olduğu, özellikle konuşma ve görüntü algılama başta olmak üzere çeşitli sahalarda büyük bir potansiyele sahiptir (16). Son yıllarda yapay sinir ağları kullanılarak yapılan çalışmalar oldukça başarılı sonuçlar vermiştir (17). İnsan beyninin çalışma prensibi üzerine oturtulmuş olan YSA' lar giriş ve çıkış veri kümelerini kullanarak sistem davranışını öğrenebilen yapay sistemlerdir (18). Şekil 3.1 de genel bir YSA modeli görülmektedir. YSA kullanarak sinyal tanıma 1950' li ve 1960'lı yıllarda oldukça pahalı bir araştırmayken, 1969' dan 1982' ye kadarki dönemde bu alandaki pahalılık giderek azalmıştır. Üstelik YSA ile ilgilenen bir çok yeni alan ortaya çıkmıştır ve konuşma tanıma da bu alanlardan birisidir (19).



Şekil 3.1. Genel bir yapay sinir ağı modeli (20)

Konuşma tanıma sistemlerinde YSA yaklaşımına dayalı çeşitli yöntemler geliştirilmiştir (21). Bunlar;

- Perceptron
- Multilayer Networks
- Backward Error Propagation
- Kohonen Self-Organizing Maps
- Hopfield Nets ve Associative Memory

YSA' ları konuşma tanıma alanına uygun bir tekniktir. Çünkü konuşma tanımada kesin sınırlar belli değildir. YSA' larının dezavantajları da bulunmaktadır. Bu teknikteki en büyük problem karmaşık sorunları çözmek için (konuşma tanıma gibi) ya çok büyük ya da çok katmanlı ve çok nöron içeren sinir ağlarına ihtiyaç duyulmaktadır. YSA' lar büyüdükçe çalışmaları üstel bir şekilde yavaşlamaktadır. Bu problemi de paralel işleme teknolojisiyle çözmek mümkündür. YSA' ların eğitme ve kullanma olmak üzere iki

türlü çalışma şekli vardır. Eğitme aşamasında YSA' ndaki düğümlerin birbirine bağlantı yüzdelerini gösteren ağırlık değerleri hesaplanır. Eğitme aşamasında elde edilen bu ağırlık değerleri daha sonra sadece girişlerin verilip çıkışların hesaplanmasının istenildiği kullanma aşamasında işe yararlar. Kullanma aşamasındaki algoritma eğitme aşamasına göre daha kolaydır. Bundan dolayı da YSA' lar bu aşamada daha hızlı çalışırlar. Eğitme aşamasında karşılaşılan sorunlar bu aşamada oluşmaz.

YSA' ları konuşma tanıma sistemlerinde şu şekilde kullanılabilir. Örneğin; belirli bir kelimeden hızlı fourier dönüşüm, doğrusal önkestirim kodu veya kod etkileşimi doğrusal önkestirim ile ya da başka bir yöntemle elde edlien katsayılar, YSA' nın giriş katmanına yüklenir. Kullanılan teknik ve elde edilen katsayıların miktarı, YSA' nın başarısını ve çalışma hızını etkilemesi açısından önem taşımaktadır. Sonra çıkış katmanına bu kelimeyi temsil edecek bir kod yüklenir. Bu kod bu çalışmayı yapacak olan kişinin belirlediği bir teknik olabilir. Ancak genel olarak 0..1 veya -1..1 arasındaki değerler bu tekniğe uygundur. Örneğin 0, 5 ten büyük olan çıkışlar 1 kabul edilip diğerleri 0 kabul edilerek ikili kodlama yapmak mümkündür. Ya da 1' e en yakın çıkışın numarasını kullanmak tercih edilebilir. Sonra YSA' nı eğitmek için hangi algoritma kullanılıyorsa çalıştırılarak işleme devam edilir. Başka bir kelime için yine bu çalışmalar tekrar edilir. Eğitme aşaması yeterli görülene kadar sürmelidir. Yeterli olduğunun anlaşılması için oluşan toplam hatanın belli bir yüzdenin altına inmesi gerekmektedir. Eğitme aşamasında kullanılan verilen sırası rastgele olmalıdır. Eğer benzer karakterdeki veriler öbek halinde eğitme işlemine uygulanırsa, öğrenme en son öbek için daha iyi olabilir. Bu durumda doğru bir çalışma değildir. Çünkü eğitme aşamasındaki döngünün içinde sıra rastgele seçilmemiştir. Ayrıca bazı durumlarda öğrenmenin gerçekleşmeyeceği de düşünülerek döngünün sonlandırma şartına belirli bir tekrar sayısının aşılması koşulu da eklenmelidir. Bu aşamadan sonra kullanma aşamasına geçilir. Bu aşamada konuşulan bir kelimenin hesaplanan katsayıları verilerek çıkışların hesaplanması için YSA' ları çalıştırılır. Programdan elde edilen çıkış kodlarına göre de hangi kelimenin konuşulduğu anlaşılmaya çalışılır.

3.2. Zaman Eşleştirme

Zaman eşleştirme yöntemi konuşma tanıma yöntemlerinde sıklıkla kullanılan bir diğer yöntemdir. Bu yöntem daha çok diğer yöntemlerle birlikte kullanılan ve daha çok tanıma işlemlerinin verimliliğini artırmak amacıyla kullanılan bir yöntemdir. Bu yöntemde, konuşma ifadelerini seslendirme süreleri sıkıştırılarak ya da genişletilerek referanslarla karşılaştırılmaları ilkesi kullanılmaktadır (12). Aynı sözcüğü aynı kullanıcı tekrar seslendirdiğinde bile bir seslendiriliş daha önceki seslendirilişlere benzemeyebilir. Sözcüğün uzunluğu doğrusal olmayan bir biçimde genişleme ve daralma gösterir (7). Zaman eşleştirme yöntemi sözcüğün ya da fonemin sinyalinin, referans şablonu ile aynı zaman aralığında olabilmesi için zaman ekseninde daralma ya da genişleme yapmayı amaçlar. Sözcük tanıma ya da fonem tanıma için genel olarak dinamik zaman eşleştirme yöntemi kullanılmaktadır. Dinamik zaman eşleştirme yönteminde zaman ekseni doğrusal olmayan bir biçimde genişletilip daraltılarak referans şablonu ile tanınacak olan sesli ifade kesiminin başlangıç ve bitiş zamanları çakıştırılmaya çalışılır. Amaç karşılaştırmanın aynı zaman aralıkları için yapılmasını sağlamaktır. Dinamik zaman eşleştirme işlemi, devingen programlama tekniği kullanılarak gerçekleştirilir (7).

Zaman eşleştirme işleminde sorun A ve B örneklerinin çakıştırılmasıdır. Devingen programlamanın uygulanışı için iki örnek zaman dizisi düşünüldüğünde; karşılaştırılacak iki örneğin zaman eksenindeki değerleri,

$A = a_1, a_2, \ldots, a_i, \ldots, a_m$

$B = b_1, b_2, \ldots, b_j, \ldots, b_n$

olarak tanımlanmışsa zaman eşleştirme fonksiyonu:

C = c(1), c(2), …, c(k), …, c(K) olarak yazılabilir.

Burada c, örneklerin kesişen nokta çiftlerini vermektedir. Bunun için devingen programlama yöntemi kullanılarak bir fonksiyon tanımlanır. Bu fonksiyon iteratif bir yaklaşımla sesli ifadeyi daraltarak ya da genişleterek referans şablonu ile aynı zaman aralığına getirir.

3.3. Saklı Markov Modelleri

Markov zincirleri olarak da bilinir ve olasılık kuramının çok önemli ve iyi çalışılmış bir kavramıdır (22). Saklı Markov Modelleri stokastik prosesleri modelleyebilen sonlu durum ağlarıdır. Her durum, gözlem vektörü uzayına farklı bir bölge ve karakteristiği tanımlar. Durumlar arası geçişler de modellenen öznitelik değişimlerini ele alır. Durumlar içerisinde verilen herhangi bir öznitelik gözlem vektörünün duruma uygunluğunu veren çıktı olasılık dağılım fonksiyonları yer alır. Bu yaklaşım ilk olarak 1965-70 yıllarında kullanılmaya başlanmış ve 1985 -90 yıllarında sesli ifade tanımada çok kullanılan bir yöntem olmuştur. SMM yönteminin popülaritesinin artmasının iki temel nedeni vardır (23). Bu sebeplerden birincisi sahip olduğu zengin matematiksel yapıdır. Bu yüzden çok çeşitli uygulamalarda kullanılabilmesi için bir teorik temeli vardır. Diğer neden ise, uygun bir biçimde uygulandığında, SMM yöntemi, başarılı sonuçlar elde edilmesini sağlamaktadır. Bir araştırmaya göre SMM ile yapılan konuşma tanıma sistemleri diğer tekniklerle de başarı sağlamasına rağmen en başarılı sonuçlar veren teknik olmuştur (24). SMM bir sinyalin stokastik olarak modellenmesidir. Sesli ifade tanımada kullanılması kısaca, ardışık kısa süreli sesli ifade kesimlerinin birlikte ele alınması ile ardı ardına gelebilecek bu kesimler için bir model oluşturmak ve bu modelden yararlanarak uzun süreli sesli ifade kesimlerinin tanımasını sağlamak şeklinde özetlenebilir (23). Markov işlemi yada zinciri kesikli bir *t* zamanındaki N durumdan biridir. Durum değişkeni *qt* ile belirtilir. Genelde sistemin tam bir tanımı önceki durumların sırasının bilinmesine bağlıdır. Birinci dereceden bir Markov işleminde sistemin şu andaki durumu sadece bir önceki duruma bağlıdır. Durum geçiş olasılıkları zamandan bağımsız olduğundan dolayı, bir Markov işlemi durum geçişleri ile tanımlanabilir (7). Durum geçis matrisi (Eş. 3.1)' deki gibidir.

$$A = \left[a_{ij}\right],$$

$$a_{ij} = P(q_t = j / q_{t-1} = i), i, j = 1,\ldots, N \qquad 3.1$$

Buradaki olasılıklar (Eş. 3.2)' deki sınırlamayı sağlamak zorundadır.

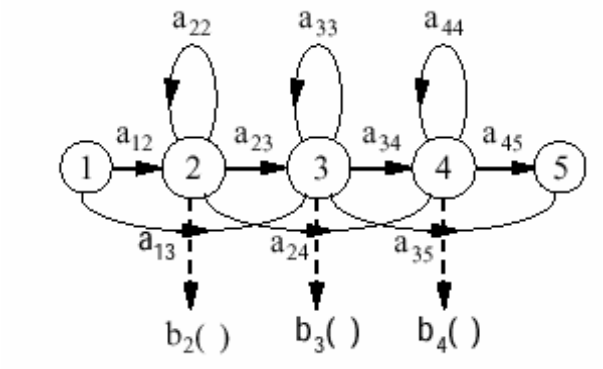
$$\sum_{j=1}^{N} a_{ij} = 1,$$

$$a_{ij} \geq 0. \qquad 3.2$$

Sistemin ilk durumu q_0 olarak tanımlanır ve bu durum ilk değer olasılıklar vektörü $\pi = [\pi_1, \pi_2, ..., \pi_N]^T$ ile belirlenir (7). Bundan sonra gelecek olan herhangi bir durum dizisi $q = (q_0, q_1, ..., q_r)$, için bir Markov işlemi tarafından üretilebilme olasılığı (Eş. 3.3)' deki gibi tanımlanır.

$$P(q \mid A) = \pi_{q_0} a_{q_a q_1}, a_{q_1 q_2}, ..., a_{q_{T-1} q_T} \qquad 3.3$$

Bir Markov işlemi için eğer durum dizisi q doğrudan gözlenemiyorsa "gizli" olarak adlandırılır. Bu durumda sadece durumların dolaylı olarak gözlemlenmesi söz konusudur. Bir gözlem ve bir durum arasında her ne kadar birebir bir gereklilik yoksa da her bir durumun belirli bir olasılıkla gözlenmesi gerekmektedir (Şekil 3.2).



Şekil 3.2. Basit bir soldan sağa SMM (25)

Sesli ifade tanıma bağlamında gözlemler özellik vektörleridir. SMM durumları ise temel alınan sesli ifade birimlerine denk gelir. Bu durumda amaç saklı olan durum dizisini gözlemlerden yararlanarak bulmaktır (7). Fonem tabanlı bir sesli ifade tanıma sisteminde bulunan durum dizisi fonem dizisine denk gelir. Her fonem için ayrı bir model tanımı yapılır. Her bir fonem için böyle bir model oluşturulduğu düşünülürse ardarda gelen fonem zincirleri bu modellerin ardarda sıralanması ile modellenebilir. Bu durumda her bir fonemin son durumundan bir sonraki fonemin ilk durumuna bir geçiş söz konusudur (7). Böylece tüm sesli ifade tanıma sistemi bir Markov işlemi olarak tanımlanabilir. Böyle bir model farklı uzunluktaki fonem dizilerini tanımak için kullanılabilir. Saklı Markov Model yaklaşımı daha biçimsel tanımlanacak olursa; herhangi bir *t* anında yapılan bir gözlem O_t ile belirtilsin (7). Herhangi bir *i* durumu tarafından O_t' nin üretilmesi olasılığı (Eş. 3.4)' de belirtilmiştir.

$$b_i(O_t) = P(O_t \mid q_t = i) \qquad 3.4$$

$b_i(O_t)$ şeklinde tanımlanmış olasılıkların kümesi $B = \{b_i(O_t)\}_{i=1}^{N}$ olarak tanımlanır. Durumlar tarafından üretilen gözlemlerin birbirinden bağımsız olduğu varsayılır ve bir durum dizgisinin ürettiği gözlemler $O = (O_1, O_2, ..., O_T)$ şeklinde ifade edilecek olursa bu gözlemlerin olasılığı (Eş. 3.5) deki gibidir.

$$P(O \mid \mathrm{q}, \mathrm{B}) = b_{q_1}(O_1) b_{q_2}(O_2), .., b_{q_T}(O_T) \qquad 3.5$$

son olarak gözlem dizisi O ve durum dizisi q dizilerinin birleşik olasılıkları da bu ikisinin olasılıklarının çarpımıyla bulunur.

$$P(O, q \mid \pi, A, B) = \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}} b_{q_t}(O_t) \qquad 3.6$$

Bu durumda bir HMM, $\lambda = (\pi, A, B)$ şeklinde ifade edilebilir.

SMM geliştirme aşaması üç alt probleme ayrılabilir (11, 7). Bunlar;

- Hesaplama problemi; verilen bir model için belirli bir gözlem dizisinin olasılığının hesaplanması
- Çözümleme problemi; verilen bir model için belirli bir gözlem dizisini oluşturan durum dizisinin bulunması.
- Öğrenme problemi; verilen bir model için bir dizi gözlemi oluşturan bir modelin olasılığının yüksek olması için model parametrelerinin kararlaştırılması.

(11) SMM' ye basitce bir örneği şu şekilde vermiş; "örneğin, bir öğrencinin genellikle ya kafede, ya okulda olabildiği durumunu göz önüne alalım. Eğer öğrencinin birkaç hafta aktivitesini gözlemlersek şöyle bir modelle karşılaşırız.

Kafe → okul → kafe → uyku → kafe → uyku → okul → kafe → uyku → kafe

Böylece belli bir süre boyunca öğrenciyi on kez gözlemlemiş oluruz. Öğrencinin okuldan kafeye gitme olasılığını bulabiliriz (bu yüksek bir olasılık çünkü her seferinde öyle yapmış). Aynı zamanında öğrencinin okuldan uykuya gitme olasılığını hesaplayabiliriz (pek olası değil çünkü hiç yapmamış). Bunu temsil etmenin diğer yolu da Çizelge 2.3 deki gibi 0 (imkansız) dan 1 (mutlaka) kadar bir ölçek ile temsil edilmiştir.

Çizelge 3.1. Saklı Markov Model Örneği

	Uyku	okul	Kafe
Uyku	0	0.3	0.2
Okul	0.1	0	0.8
Kafe	0.9	0.7	0

Öğrencinin davranışını tahmin etmek için bu diyagramı kullanabiliriz. Eğer öğrenci kafedeyse gitmesi en muhtemel yer okuldur. Çünkü kafede en yüksek sayı 0.8 dir. Eğer öğrenci okuldaysa (yine üst sırada) o zaman gitmesi en muhtemel yer kafedir. Çünkü en yüksek değere sahip (0.7). Bu yöntem Saklı Markov Modeli olarak bilinir.

Bir diğer tarafından takip edilen sesin frekansına bakarak sesler için bir SMM model oluşturabiliriz. Ve bunu bir sonraki en muhtemel sesi tahmin etmekte kullanabiliriz. Bu bizim olası olmayan adayları elememize ve konuşulan cümle için olası ses kümesini ortaya çıkarmamıza yardımcı olur. Şekil 3.3 de basit bir şekilde oluşturulmuş bir SMM tanımı görülmektedir.

```
~h "hmm1"
<BeginHMM>
    <VecSize> 4 <MFCC>
    <NumStates> 5
    <State> 2
       <Mean> 4
          0.2 0.1 0.1 0.9
       <Variance> 4
          1.0 1.0 1.0 1.0
    <State> 3
       <Mean> 4
          0.4 0.9 0.2 0.1
       <Variance> 4
          1.0 2.0 2.0 0.5
    <State> 4
       <Mean> 4
          1.2 3.1 0.5 0.9
       <Variance> 4
          5.0 5.0 5.0 5.0
    <TransP> 5
       0.0 0.5 0.5 0.0 0.0
       0.0 0.4 0.4 0.2 0.0
       0.0 0.0 0.6 0.4 0.0
       0.0 0.0 0.0 0.7 0.3
       0.0 0.0 0.0 0.0 0.0
<EndHMM>
```

Şekil 3.3. Basit bir soldan sağa SMM tanımı (25)

Viterbi algoritması

Viterbi algoritması genel anlamıyla bir dinamik eşleştirme (dynamic warping) yöntemidir. SMM için kullanımı, tanınacak sese ait öznitelik gözlem vektör sekansının modele ait durumlara dağıtılarak en uygun durum-gözlem vektörü eşleşmesini bulmak ve buna ait olurluk değerini hesaplamak olarak gerçekleşir (4).

Sürekli yoğunluklu Gauss karışımları

İstatistikte en yoğun kullanılan olasılık dağılım fonksiyonu Gauss dağılımıdır. İstatistiksel konuşma tanımada SMM durumları, Gauss dağılımları ile ifade edilebilir. Eğitim verisinin durumuna göre birden çok sayıda Gauss dağılımının ağırlıklı toplamı ile sürekli yoğunluklu Gauss karışımları (Continuous Density Gaussian Mixtures) elde edilir (25).

Model parametre tahmini

Parametre tahmini işlemi SMM' lerin eğitilmesidir. Bir SMM' nin eğitilecek parametreleri şunlardır:

- Her durum için olasılık dağılım fonksiyonunun parametreleri. Gauss karışımları için bu parametreler: karışım sayısı, karışım ağırlıkları, karışım ortalamaları ve kovaryansları.
- Durumlar arası geçiş tablosu olasılık değerleri.

Baum-Welch algoritması SMM parametre tahmini işlemini maksimum olurluk yaklaşımı (maximum likelihood sense) ile çözen bir en iyileştirme tekniğidir. Ayrıca HTK tarafından kullanılan gömülü yeniden tahmin yöntemi (embedded re-estimation) geniş veritabanları üzerinden Baum-Welch uygulanmasına olanak tanır (25).

Sınırlı dağarcıklı ayrık sözcük tanıma

SMM uygulamalarının ilk örneğidir. Sistem dağarcığı sınırlı sayıda sözcükten oluşur. Tanınacak konuşma da bu sözcüklerden herhangi birinin bir kez söylenmesinden ibarettir. Bu tür uygulamalarda her sözcük için bir SMM modellenir. Tanıma işlemi ses sinyalinin bu modellerden geçirilerek olurluk değerlerinin hesaplanması ve en yüksek olurluk değerini veren modelin tanınan sözcük olarak seçilmesi şeklinde gerçekleştirilir.

Geniş dağarcıklı tanıma

Geniş dağarcıklar söz konusu olduğunda her sözcük için bir SMM üretilmesi pratik olmaktan çıkmaktadır, çünkü:

- Her sözcük için ayrı ayrı eğitim verisi gerekmektedir.
- Sistem dağarcığına yeni bir sözcük eklenmesi için o sözcüğe ait eğitim verisi bulunması ve yeni bir model üretilmesi gerekecektir.
- Sürekli konuşma tanıma için sözcük modelleri uygun değildir çünkü bir sözcüğün ayrık ve sürekli konuşma içindeki söylenişleri farklıdır.

Bu nedenle SMM' ler sözcükler yerine sözcük-altı birimler, örneğin fonem, için oluşturulur. En yaygın kullanılan seçim bağlam bağımlı modeller kullanmaktır. Burada da üçlü-ses modelleri (tri-phone) karşımıza çıkmaktadır.

4. Sonuç

Şimdiye kadar konuşma tanımayla ilgili olarak Dragon Naturally Speaking, IBM Via Voice Pro, Nuance gibi yazılımlar hazırlanmıştır. Ancak bu yazılımlar daha çok İngilizce dil özelliklerine göre düzenlenmiştir. Bu yazılımların Fransızca, Almanca versiyonları da bulunmaktadır. Microsoftun Office XP programlarında sesli komutlarla çalışan bir araç çubuğu oluşturma imkanı bulunmaktadır. Ancak bu imkan sadece İngilizce(ABD), Japonca ve basit olarak kısmen Çince dillerinde insanların hizmetine sunulmaktadır. Konuşma tanıma alanında günümüzde yaşanan en büyük zorluklardan biri hiç kuşkusuz, konuşmacıdan bağımsız konuşma tanıma sistemleri için hala yüksek

boyutlu sözcüklerde yüksek doğrulukta tanıma oranlarının elde edilememesidir. Özellikle yurt dışında İngilizce veya Japonca dilleri için bu alanda yapılan çalışmaların sayısı tatminkar sonuçlar elde edilemese de oldukça fazladır. Ancak ülkemizde Türkçe için yapılan çalışmalar diğerleri ile karşılaştırıldığında oldukça düşük sayıdadır. Türkçe dil yapısına yönelik yapılmış çalışmalara henüz sıklıkla rastlanamamaktadır. Konuşma tanıma teknolojisiyle Türkiye' de ticari manada ciddi olarak ilgilenen pek fazla firma yoktur. Bu konuda daha çok çeşitli üniversitelerde doktora tezleri şeklinde çalışmalar yapılmaktadır. Bu konuyla ilgili olarak GVZ ve SYS yazılım ses teknolojileri firmalarının yapmış olduğu çalışmalar vardır. Bu grubların yapmış olduğu çalışmalar sadece metni konuşmaya dönüştürme şeklinde olup Türkçe yazılan metni tam bir insan sesiyle seslendirmektedir ve ancak konuşmayı metne dönüştürme teknolojisiyle ilgili çalışmalarına da devam etmektedirler. Diğer taraftan TÜBİTAK ODTÜ BİLTEN' de bulunan proje grupları arasında Konuşma İşlemeye yönelik çalışmalar söz konusudur. Ayrıca Bilkent Üniversitesi, Türkçe Dil ve Konuşma İşleme Merkezi kurarak bu konuyla ilgili araştırmalarını sürdürmektedirler. Konuşma Tanıma sistemleri dilden bağımsız olarak hazırlanabilmelidir. Bu amaç doğrultusunda yeni algoritmalar geliştirilmelidir. Ülkemizde konuşma tanıma çalışmalarında yapay zeka tekniklerinden yapay sinir ağlarına ait çeşitli algoritmalar üzerinde çalışılmıştır. SMM kadar çok sıklıkla kullanılmasa da yapay sinir ağlarıyla ilgili yapılan çalışmalarda, ses eğitme aşamasının oldukça uzun sürdüğü yapılan araştırmalarda görülmüştür. Kelime veya cümle sayısının artmasıyla bu başarı oranı daha da düşmektedir. Diğer ülkelere bakıldığında konuşma tanıma sitemlerinde genel olarak SMM yapısı kullanılmıştır. Bunun yanı sıra diğer yapay zeka teknikleriyle de çalışılmıştır (Genetik algoritma, fuzzy logic, uzman sistemler). Öyle ki hybrid sistemler denilen karma veya melez sistemlerle bu çalışmalara ağırlık verilebilir. Hem yapay sinir ağı ve hem SMM sistemleri birlikte kullanılabilir. Bu örnek diğer yapay zeka teknikleriyle de çoğaltılabilir. Ülkemizde konuşma tanıma teknolojilerinde yapay sinir ağları haricinde diğer yapay zeka teknikleriyle çalışılmadığı gözlenmiştir. Ayrıca konuşma tanıma çalışmaları için hazır ses veri tabanı kütüphanesi oluşturulmalıdır. Tubitak bu konuyla ilgili olarak çalışmalarına devam etmektedir. Türkçe konuşma tanıma yapabilen sistemlerin geliştirilmesi üzerine daha çok araştırma yapılmalıdır. Özellikle konuşmacıdan bağımsız bağlı konuşma tanımayı sağlayan çalışmalar daha kullanışlı olacaktır. Örneğin mahkeme duruşmalarında, emniyet sorgularında, zabıt işlemlerinde hep karşılıklı konuşmaların anında bilgisayara yazılması söz konudur. Böyle bir alana yönelik yapılan çalışma oldukça kullanışlı olacaktır. İnsanlığı daha rahat bir çalışma ortamına kavuşturabilmek için konuşma tanımayla ilgili çalışılabilecek alanların çeşitliliğinin artırılması gerekmektedir (26). Yazar, "Konuşma teknolojisi yardımıyla ilköğretim birinci sınıf öğrencilerine ilkokuma yazma öğretimi için bir yazılım geliştirme" adlı doktora teziyle konuşma tanıma teknolojisini ilköğretim alanına uygulamıştır.

Kaynaklar

1. Nabiyev, V. "Yapay Zeka", ISBN 975 347 985 9, ***Seçkin Yayıncılık San. Ve Tic. A. Ş***., s. 704-714, Ankara, 2005.
2. Yalçın, N. ve Bay, Ö.F. "İlköğretim birinci sınıf öğrencilerine konuşma tanıma teknolojisi yardımıyla ilkokuma yazma öğretimi***", 6th International Educational Technılogy Conference***, Eastern Mediterranean University, , vol:3, , 19-21 April, pp:1659, Famagusta, North Cyprus, 2006.
3. Morgan, D. and Scofield, L. C., "Neural Networks and Speech Processing", ***Kluwer Academic Publishers***, pp. 102-108, USA, 1991.
4. Becchetti, C., and Ricotti L. P., "Speech Recognition Theory and C++ Implementation", ISBN 0-471-97730-6, ***John Wiley & Sons Ltd,*** 167-188, 310-311, England, 1999.
5. Gökhan, A., "Yapay Sinir Ağları İle Ayrık Türkçe Sözcüklerin Tanınması", Yüksek Lisans Tezi, ***Fırat Üniversitesi Fen Bilimleri Enstitüsü***, , s. 1-17, Elazığ, 1997.
6. Jurafsky, D. & Martin, J. H., "Speech and Language Processing An Inroduction to Natural Language Processing, Computational Linguistics, and Speech Recognition", ISBN 0-13-122798-X, ***Prentice Hall Upper Saddle River***, pp. 235-249, New Jersey, USA, 2000.
7. Mengüşoğlu, E., "Bir Türkçe Sesli İfade Tanıma Sisteminin Kural Tabanlı Tasarımı ve Gerçekleştirimi", Yüksek Mühendislik Tezi, ***Hacettepe Üniversitesi Fen Bilimleri Enstitüsü***, s. 14-16, 22-26, Ankara, 1999.
8. Furui, S., "Digital Speech Processing", Synthesis and Recognition", ***Marcel Dekker Inc***., New York, 1989.
9. Akçay, B., "Yapay Sinir Ağları İle Türkçe Konuşma Tanıma", Yüksek Mühendislik Tezi, ***Hacettepe Üniversitesi Fen Bilimleri Enstitüsü***, s. 1-10, Ankara, 2004.
10. Durmuş, E., ve Gül E., "Yazıyı sese çevirim", ***SİU' 98 6. Sinyal İşleme Uygulamaları Bildirimler Kitabı***, , s. 1:233, Kızılcahamam, Ankara, 1998.
11. Doğan, S., "PC Ortamında Sesli Komutları Tanıma", Yüksek Lisans Tezi, ***Marmara Üniversitesi Fen Bilimleri Enstitüsü***, , s. 4, 15-27, İstanbul, 1999.
12. Artuner, H., "Bir Türkçe Fonem Kümeleme Sistemi Tasarımı ve Gerçekleştirimi", Doktora Tezi, ***Hacettepe Üniversitesi Fen Bilimleri Enstitüsü***, s.47-55, Ankara, 1994.
13. Bayri A. ve Bingöl S., "Sözcük-altı Modeller Kullanarak Ayrık Sözcük Tanıma", ***SİU' 98 6. Sinyal İşleme Uygulamaları Bildirimler Kitabı***, 2:s.502-503, Kızılcahamam, Ankara, 1998.
14. Charniak, E., and Mcdermott, D., "Introduction to Artificial Intelligence", ***Addison Wesley Longman Publishing Company***, ISBN 0-201-11946-3, Boston, USA, 1985.

15. Lippmann, R.P., "An Introduction to Computing with Neural Nets." ***IEEE Acoustics, Speech and Signal Processing Magazine***, 4:pp.4-22, 1987.

16. 16. Akpınar, H., "Yapay Sinir Ağları ve Kredi Taleplerinin Değerlendirilmesinde Bir Uygulama Önerisi", ***İstanbul Üniversitesi İşletme Fakültes*i**, İstanbul, s. 4-7, 1993.

17. Maren, A. et al, "Handbook of Neural Computing Applications", ***Academic Press Inc***., ISBN 0-12-4711260-6, Orlando, USA, 1990.

18. Rumelhart, D.E. ve McClelland, J.L, "Parallel Distributed Processing", ***MIT Press***, July, pp.1:10-16, Cambridge, 1986.

19. Chassaing, R., "Digital Signal Processing with C and the TMS320C30", ***A Wiley-Interscience Publication***, ISBN 0-471-55780-3, pp.324-325, USA, 1992.

20. Akçayol, M.A., "Bir Anahtarlamalı Relüktans Motorun Sinirsel-Bulanık Denetimi", Doktora Tezi, ***Gazi Üniversitesi Fen Bilimleri Enstitüsü***, Ankara, 2001.

21. Schroeder, M. R., "Computer Speech Recognition, Compression, Synthesis", ISBN 3-540-21267-1, ***Printed in Germany***, s.55-59, 166**,** Springer, 2004.

22. Jelinek, F., "Statistical Methods for Speech Recognition", ISBN 0-262-1006-5, ***The MIT Preess Cambridge***, Massachusetts London, s.14-18, pp.15-27, England, 1997.

23. Rabiner, L. R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", ***Proceedins of IEEE***, 77(2):pp.257-286, 1989.

24. Niesler, T., "Category-based Statistical Language Models", P.h.D. Thesis, ***Cambridge University***, June, s.11-12, Cambridge, 1997.

25. Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., and etc., "The HTK Book (for HTK Version 3.1)", ***Copyright (1995-1999) Microsoft Corparation, Copyright (2001-2002), Cambridge University Engineering Department***, , s.3-19, Cambridge, 2002.

26. Yalçın, N. "Konuşma Tanıma Teknolojisi Yardımıyla İlköğretim Birinci Sınıf Öğrencilerine İlkokuma Yazma Öğretimi İçin Bir Yazılım Geliştirme", Doktora Tezi, ***Gazi Üniversitesi Fen Bilimleri Enstitüsü***, s.115, Ankara, 2006.