Konuşma Tanıma için İnsan-makine Karşılaştırması

Ayşe Gürel ve Levent M. Arslan
Boğaziçi Üniversitesi

Speech/voice recognition by machines has been a topic of interest since 1950s. Research that initially adopted dynamic programming methodologies now mostly uses the hidden Markov model as the method for speech recognition. Nevertheless, even the most advanced speech recognition system makes, depending on the context, 2-20 times more errors than humans. Although the basic principles behind human speech recognition have not been completely understood, there are some theories that attempt to explain biological mechanisms for speech recognition. This paper aims to provide a review of these theories as well as a brief history of developments in automatic speech recognition technology. Furthermore, the paper discusses some recent studies on Turkish speech recognition. The paper concludes with a comparison between human and machine speech recognition performance.
Key words: *Automatic speech recognition, Human speech recognition, Speech perception, auditory cortex, The Hidden Markov Model, Turkish speech recognition*

Makinelerde ses tanıma, üzerinde 1950'lerden beri çalışılan bir konudur. Bunun için dinamik programlama ile başlayan çalışmalar günümüzde saklı Markov modellerin kullanılmasıyla sürdürülmektedir. Şu an mevcut sistemlerin en gelişmiş olanları, insan konuşma tanıma performansıyla karşılaştırıldığında değişen ortamlara göre 2-20 kat daha fazla hata yapmaktadır. Her ne kadar insan beyninin konuşma tanıma işlemini nasıl gerçekleştirdiği tüm ayrıntılarıyla bilinmiyor olsa da bu konuda birtakım öngörüler vardır. Bu çalışmada bu öngörüler tartışılıp insan ve makinenin konuşma tanımadaki farklılıkları, benzerlikleri, eksik ve üstün yönleri incelenmektedir. Ayrıca, insan ve makinenin konuşma tanıma işlemini nasıl gerçekleştirdiği hakkında şimdiye kadar toplanmış bilgi ve bulgular özetlenmektedir.
Anahtar sözcükler: *Makinelerde konuşma tanıma, insanlarda konuşma tanıma, ses/konuşma algılama, işitme korteksi, Saklı Markov Modeli, Türkçe konuşma tanıma*

1. Giriş

Konuşma tanıma, başka bir deyişle, insanların konuşmalarını anlamak, her birimizin günlük hayatımızda farkında olmadan defalarca zorlanmadan yaptığımız bir işlemdir. İnsan için bu işlemi adımlara ayıracak olursak: önce kulak tarafından karşıdaki kişinin

konuşmasının nereden geldiği tespit edilir. Daha sonra sinyal seviyesi yükseltilerek ve frekans incelemesi yapılarak beyne gönderilir. Beyinde daha önceden öğrenilmiş dil ve kavram ilişkisi ile anlamlı bir hale getirilir. Makinede ise bu işlemler elektronik cihazlar ve bilgisayar algoritmaları tarafından yapılmaktadır. Örneğin kulağın yerini mikrofon almıştır. Akustik sinyal mikrofon tarafından elektrik sinyaline dönüştürülüp analog-dijital dönüştürücü ses kartı vasıtasıyla bilgisayara aktarılır. Bilgisayarda ise istatistiksel yöntemler ve olasılık hesabı kullanılarak o dilde olası söylenebilecek cümleler arasından hangisinin söylendiği tespit edilmeye çalışılır. Aslında insan ve makine karşılaştırıldığında kullanılan bazı yöntemlerin bir kısmı rastlantı sonucu olarak oldukça benzerdir. Bu çalışmada da insan ve makine arasındaki bu farklılıklar ve benzerlikler irdelenecektir.

İnsanlık tarihinde dilin ilk ne zaman ortaya çıktığı hala çözülememiş bir konudur. Bu konuda birtakım savlar vardır. Örneğin Noam Chomsky insanlardaki dil yetisinin doğuştan var olduğunu öne sürmüştür ve bu savı destekleyen gerekçe olarak çevresel verinin çok az olmasına rağmen insanın karmaşık dil yapısını çok kısa sürede öğrenmesini göstermiştir (Chomsky, 1965).

Şu anki insanlığın yaklaşık bir milyon yıl önce Afrika'da yaşadığı düşünülen, anatomik olarak modern yapıya sahip küçük bir insan topluluğunun devamı olduğu kabul edilmektedir. Bu grubun diğer insan türlerinden farklı olarak dil yetisine sahip olduğu düşünülmektedir. Birçok kurama göre dilin evrimi de aynı zaman diliminde başlamıştır. İnsanlarda bulunan FOXP2 adlı genin dilin evrimi ile ilgili olduğu öne sürülmektedir. Bu gendeki hasarların dil yetisinde çeşitli bozukluklara yol açtığı öngörülmektedir (Lai ve diğerleri, 2001).

Makineler için konuşma tanıma çalışmaları 1950'lerde başlamıştır. 1952 yılında Bell Laboratuvarları'nda Davis, Bidduph ve Balashek izole rakam tanıma amaçlı kişiye bağımlı bir sistem geliştirmişlerdir. Bu sistem, ünlü sesbirimlerin spektral tınlaşımlarını sesleri ayırt etmek için temel almaktaydı. 1960'lara kadar değişik araştırmacılar tarafından bu sisteme benzer bir dizi çalışmalar yürütülmüştür (Olson & Belar, 1956; Fry & Denes, 1959). Konuşma tanıma konusunda 1970'lerde önemli ilerlemeler kaydedilmiştir. Örüntü tanıma kavramlarının konuşma tanıma sistemleri için kullanılması ilk olarak Velichko ve Zagoruyko (1970) tarafından önerilmiştir. Ayrıca Sakoe ve Chiba (1978) dinamik programlama yöntemini konuşma tanımada başarıyla uygulamıştır. Doğrusal öngörü incelemesi ile spektrumun modellenmesi ise Itakura (1975) tarafından uygulanmıştır. Aynı tarihlerde IBM araştırma merkezinde de konuşma tanıma üzerine uzun soluklu çalışmalar başlatılmış ve 80'li yıllarda kısa ofis yazışmaları için ilk dikte sistemi çalışması yapılmıştır (Jelinek, 1985). Bu dönemde aynı zamanda şablon karşılaştırma temelli yöntemlerin yerini istatistiksel yöntemler almaya başlamıştır. Bu alanda özellikle Saklı Markov modelleri kullanılmaya başlanmıştır (Rabiner, 1989). Günümüzdeki sistemlerde bu temel teknolojiler temel alınarak akustik inceleme ve modelleme üzerine iyileştirme sağlanmış, aynı zamanda daha fazla veri

kullanılarak sistem performansları önemli ölçüde arttırılmıştır.

1990 sonrasında insan ve makine arasındaki konuşma tanımadaki farkları inceleyen birtakım çalışmalar yapılmıştır. Örneğin Lippman (1997) ve Moore (2003) makineler ve insanlar için konuşma tanıma performanslarını karşılaştırmışlardır. Moore (2003) ile Lamel ve gurubunun (2002) çalışmalarında insan ve makinenin dil edinimindeki veri gereksinimleri ayrıntılı olarak incelenmiştir.

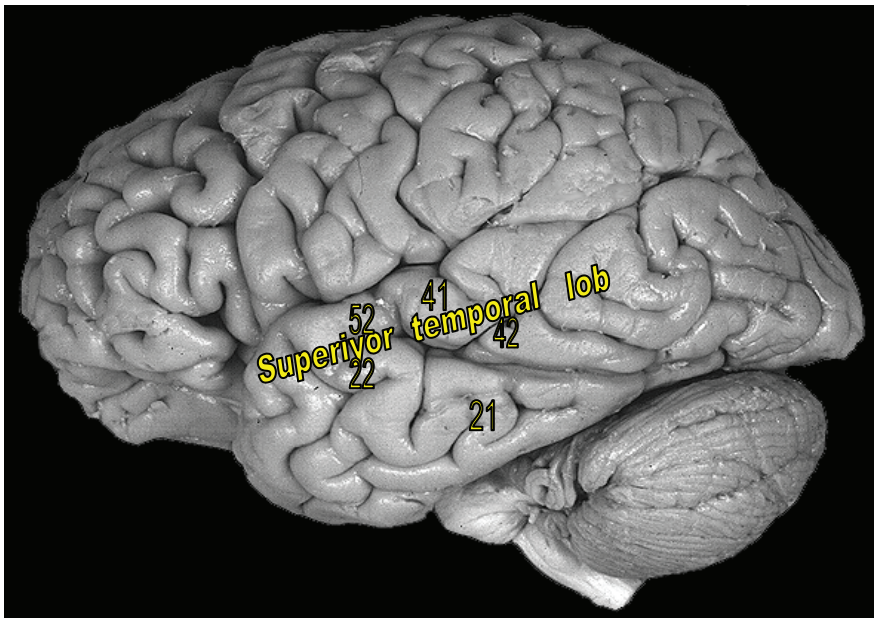
Bu bilgiler ışığında, bu çalışmada daha önceki çalışmalardan farklı olarak insan beyninin algı mekanizması daha ayrıntılı olarak incelenmektedir. İkinci bölümde insanda konuşma tanımanın beyinde nerede ve nasıl gerçekleştiği konusu tartışılmaktadır. Ayrıca, insanda konuşma tanımayı açıklamaya yönelik ortaya atılan modeller incelenmiş, konuşma-dil tanıma yetisinin doğuştan olabileceğini işaret eden çalışmalar özetlenmiştir. Üçüncü bölümde makinede konuşma tanımanın nasıl gerçekleştiği anlatılmıştır. Dördüncü bölümde son yıllarda geliştirilen Türkçe ses/konuşma tanıma uygulamaları ele alınmaktadır. Beşinci bölümde ise insan-makine karşılaştırması yapılmış ve günümüz teknolojisinin eksik yönleri ve insana göre üstünlükleri tartışılmıştır. Son olarak, sonuçlar kısmında geleceğe yönelik öngörülere yer verilmiştir.

2. İnsanda Konuşma Tanıma

2.1 İnsan Beyninde Konuşma Tanımadan Sorumlu Merkezler

İnsan beyninde konuşma tanımadan sorumlu bölgelere ait nöro-anatomik modellerin ilki Alman nörolog Carl Wernicke tarafından 1874'de ortaya atılmış ve daha sonra geliştirilen birçok modele temel oluşturmuştur. Wernicke, konuşulanları anlamakta zorluk çeken afazi hastalarından elde ettiği otopsi bulguları ışığında, dil ve konuşmayı anlamada rol oynayan beyin yapılarının, şu anda nöro-anatomide kendi adıyla bilinen superiyor temporal girusun (Brodmann alanı 22) arka kısmındaki işitsel assosiasyon korteksi ile sınırlanan bölge olduğunu ortaya koymuştur (Şekil 1). Sözcüklerin işitsel gösterimleri ve onlara anlam yükleyen ikinci derece assosiasyonlar arasındaki karşılıklı etkileşimlerin Wernicke alanında koordine edildiği bulgulanmıştır (Mesulam, 2000). Günümüzün modern tarama teknikleri, insan beyninde dil-konuşma tanıma bölgelerini saptamada kullanılan geleneksel lezyon bulgularından edinilen bilgileri doğrular ve tamamlar niteliktedir. Buna göre genel olarak insan beyninin birincil (primary) işitsel korteksi, superiyor temporal planumun posteriyor kısmındaki Heschl girusunda (Brodmann 41) bulunmaktadır. İkincil (secondary) işitsel korteks ise superiyor temporal girusun çevresinde Brodmann alanları 21, 22, 42 ve 52'yi kapsamaktadır (Binder ve diğerleri, 1996; Celesia, 1976; Lauter ve diğerleri, 1985; Mazziotta ve diğerleri, 1982; Mesulam, 2000; Petersen ve diğerleri 1988; Wise ve diğerleri, 1991) (Şekil 1).

Şekil 1: İşitmeden sorumlu kortikal alanlar



(© 1995. The Digital Anatomist Project, University of Washington)

İnsanlardaki dil-konuşma tanıma işlevinin nöro-fonksiyonel yönü birçok açıdan henüz tam olarak anlaşılamasa da elde edilen bulgulara göre, çevreden gelen ses dalgaları iç kulaktaki kokleanın (kulak salyangozu) içinde bulunan sıvıda titreşim yaratmakta ve bu da kokleada bulunan işitsel alıcıları tetiklemektedir. İşitsel sinirler bu ses sinyallerini kokleadan alıp medulladaki ipsilateral dorsal ve ventral koklea çekirdeğine götürür. İşitsel sinirler (koklear sinirler), kranyal sinirlerden 8. olan vestibüler koklea sinirlerinin parçasıdır. Her bir koklea çekirdeğinden gelen aksonlar medullanın iki tarafındaki medyal ve lateral superiyor olivar çekirdeğe ulaşır. Her bir superiyor olivar çekirdek, bilateral (iki kulaktan da gelen) veriyi aldığı için, işitsel verinin nereden geldiğini sesin iki kulaktaki şiddetine ve iki kulağa ulaşma hızına göre ayırt edebilir. Superiyor olivar çekirdekten, alt beyin sapının ana işitsel sinir yolu olan lateral lemniskiye çıkan sinir lifleri, oradan orta beyin tavanındaki tepeciklere (inferior colliculi) varır. Buradaki aksonlar talamusa ait medyal genikülat çekirdeğe ve oradan primer işitsel kortekse ulaşır (Pinel, 1998; Mesulam, 2000).

2.2 İnsan Beyninde Konuşma Tanıma

Duyulan seslerin akustik özellikleriyle ilgili talamus ve beyin sapında birçok ayarlama (tuning) olsa da buralarda yapılan işlemler tüm çevresel sesleri kapsayan işlemlerdir. Konuşma seslerine ilişkin işlemler ancak sinyaller serebral kortekse

ulaştıktan sonra başlar (Scott & Johnsrude, 2003). Konuşmaya ilişkin seslerle diğer çevresel seslerin insan beynindeki işitsel kortekste farklı statüleri olduğu düşünülmektedir (Binder, 2000). Fonksiyonel beyin tarama tekniklerinde, konuşma seslerinin, diğer seslere oranla, beynin her iki tarafının superiyor temporal bölgesinde daha fazla aktivasyon yarattığı gözlenmiştir (Binder ve diğerleri, 1996; Demonet ve diğerleri, 1992; Zatorre ve diğerleri, 1992). İnsan beyninde yokuş yukarı (ascending) işitsel alanlar (örn. A1 nöronları), sesin frekans ve perde gibi daha temel özelliklerini işlemeye eğilimli iken, yokuş aşağı alanların (descending) (örn. Brodmann 21 alanı), sözcüklerin tanınması, seslerin lokalizasyonu, nesnelere özgü seslerin ve belki de kişisel tarzların sınıflandırılmasına ilişkin daha karmaşık sinir hücreleri içeriyor olduğu düşünülmektedir (Mesulam, 2000).

2.3 İnsan Beyninde Konuşma Tanıma Modelleri

Dile ait seslerin ve konuşma tanımanın insan beynindeki lokalizasyonunun yanı sıra, dilin ve seslerin nasıl algılandığı ve işlendiği sorusu nörodilbilimsel ve ruhdilbilimsel açıdan incelenmiş ve birçok araştırmaya konu olmuştur.

İnsandaki konuşma tanıma mekanizmalarını açıklamak üzere ortaya atılan modellerden bazıları tabandan tepeye (bottom-up) dil işleme üzerine kuruludur. İlk gruba giren Motor Kuramı (Motor Theory) ve Sentezli-Analiz (Analysis-by-Synthesis) 1980'lere kadar yapılan çalışmalara temel oluşturmuştur. Motor Kuramı'na göre konuşma tanıma işlemi sırasında konuşma sinyalleri konuşmanın motor hareketlerine göre çözümlenir ve algılanır. Buna göre konuşma seslerinin nasıl üretildiği onları nasıl algıladığımızı belirler. Liberman ve çalışma arkadaşları tarafından 1970'lerde (Liberman, 1970) ortaya atılan bu model, özellikle ko-artikülasyon gibi birçok nedenle bir sesbirimsel temsilin farklı ortamlarda farklı akustik sinyalle oluşturulması sorununa bir çözüm üretme amacındaydı. Liberman'a göre akustik sinyaller farklı bile olsa bir sesbiriminin oluşturulduğu motor hareketler benzerdi ve bu da sesi tanımayı kolaylaştırıyordu. Bu modelde ortaya atılan diğer bir görüşe göre, konuşma tanıma ve diğer seslerin tanınması beyinde ayrı işlemler gerektirmektedir. Konuşma tanıma insana özel ve doğuştan kazanılmış bir yetidir. Bu yeti, duyulan bir sesi, bu sesin sesletimi için gerekli motor hareketlerle eşlememizi sağlar.

Stevens (1960) tarafından ortaya atılan Sentezli-İnceleme Kuramı'na göre de konuşma tanıma ve konuşma üretme birbirleriyle bağlantılıdır. Buna göre, konuşma incelemesi ancak duyulan konuşmanın örtük biçimde (implicit) senteziyle mümkündür. İnsan duyduğu konuşmanın sentezini yapar ve bunu işitsel uyaranla karşılaştırır. Bu işlem sırasında sesbirimlerinin ayırıcı özellikleri de önemlidir.

Daha sonraki yıllarda ortaya atılan tabandan tepeye modellerden biri de Belirsiz Mantık Modeli'dir (Fuzzy Logical Model). Massaro'nun (1987) bu modeline göre, konuşma tanıma üç işlem üzerine kuruludur: özellik değerlendirme (feature evaluation), özellik bütünleşmesi (feature integration) ve karar verme (decision). Her sesbirimin

ideal bir ilk örnek değeri ve özelliği vardır. Bunlar bir sesbirimin kategorisini belirler. Bu bilgi insan belleğinde mevcuttur. Konuşma tanımada, insan, duyduğu sesbirimin özelliklerinin bellekte tutulan her bir ilk örnek özelliğe uyup uymadığını değerlendirir ve buna göre tanımada karar verir.

Yukarıda sözü edilen üç model de öncelikli olarak sesbirimcik parçaların tanınması ile ilgilenmektedir. Bu modellerde sözcüklerin işitsel düzeyde algılanması öne çıkmakta, sözcüğe anlam yükleme işlemi ele alınmamaktadır.

İnteraktif modeller olarak bilinen ve konuşma tanımada hem tabandan-tepeye hem de tepeden-tabana işlemleri kullanan modeller içinde Cohort Model (Marslen-Wilson, 1987) ve İz Modelini (Trace Model) (Elman & McClelland, 1984) sayabiliriz. Cohort Modelde, ilk aşamada, tanınacak sözcüğün ilk sesbiriminde olan akustik-fonetik özellikleri paylaşan tüm sözcüklerin aktive olması beklenmektedir. Örneğin işitilen sözcük /bina/ ise /b/ sesiyle başlayan tüm sözcükler aktive olacaktır. Bu ilk aşamada sözcüklerin aktivasyonunda sadece sesbirimcil bilgi kullanılmaktadır. İkinci aşamada ise hedef sözcük dışında aktive olan diğer tüm sözcüklerin değerlendirmeden elenmesi için anlamsal ve sözdizimsel bilgiler kullanılabilir. Burada öngörülen sözcük tanıma modeli aslında makinelerdeki sözcük tanıma sistemine oldukça benzemektedir. Makinelerdeki konuşma tanıma sistemleri de soldan sağa çözümleme yapmakta ve sadece benzer başlangıcı olan sözcükleri olası hedef sözcük olarak değerlendirmektedir. Böylece diğer sözcüklerin olasılık hesabına katılmamasıyla makinenin hızlı bir şekilde yanıta ulaşması sağlanmaktadır (bkz. 3. Bölüm).

Sinir ağı modeli üzerine geliştirilen İz Modelinde ise konuşma tanımada 'node' adı verilen ünitelerin aktivasyonu veya inhibisyonu söz konusudur. Bu üniteler, sesbirimlerinden, sesbirimlerinin ayırıcı özelliklerinden veya sözcüklerden oluşabilir. Her ünitenin durağan (resting) düzeyi, eşik düzeyi ve aktivasyon düzeyi o birimin tanınmasını etkileyebilir. Gelen işitsel veri bir birime (örn. /b/ sesbirimine) işaret ediyorsa, o ünite aktive edilir ve bu birimin aktivasyon eşiğinin düşmesini sağlar. Aktive edilmeyen veya kullanılmayan birimlerin ise aktivasyon eşiği yükselir. Bu modelde, birimler birbirleriyle bağlantılı olduğundan birinin aktivasyonu diğerini de etkiler. Bir sesbirim aktivasyonu sözcük düzeyinde de bir aktivasyona yol açabilir. Örneğin /b/ ve /i/ sesbirimleri *bina*, *bira*, *bir* gibi birçok sözcüğün aktivasyonunu sağlar. Aynı şekilde *bina* sözcüğündeki /b/ sesbirimi, onunla yarış içinde olan diğer sesbirimlerin (örn. /p/) bastırılmasına neden olabilir.

Konuşma tanıma gibi karmaşık bir işlemi çoğu zaman başarıyla ve kolayca yapabilen insanda bu yetinin doğuştan gelen bir dil yetisiyle mi yoksa çevresel veriyle mi açıklanması gerektiği konusu çok tartışılan bir konudur. Bu bağlamda insanda konuşma tanımanın oluştuğu erken evreleri incelemek gerekmektedir.

2.4 İnsanda Konuşma Tanıma Yetisinin Gelişimi

Bebeklerin sesleri tanıma ve ayrıştırmayı çok erken yaşlarda yapabildiği

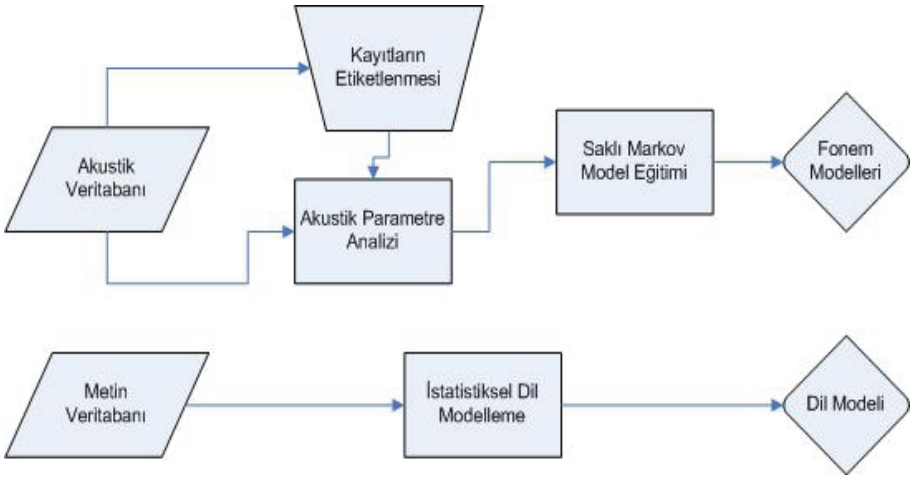
gözlenmiştir. Yeni doğmuş bebeklerle emme hızı (high-amplitude sucking paradigm), kardiyak azalması (cardiac deceleration), koşullanmış baş çevirme (conditioned head-turn) gibi yöntemler kullanılarak yapılan çalışmalar, 3 aydan küçük bebeklerin dile ait sesleri algılayabildiklerini ve ayırt edebildiklerini göstermektedir (Lecanuet & Granier-Deferre, 1993). Bebeklerin doğumdan hemen sonra annelerinin seslerini tanıyabilmeleri (Mehler ve diğerleri, 1988), annelerinin konuştuğu dille başka dilleri ayırt edebilmeleri bebeklerin aslında daha anne rahminde çevresel seslere duyarlı olduklarını göstermektedir. Bu alanda yapılan çalışmalarda, anne karnındaki bebeklerin hamileliğin 20. haftasından sonra duyma yetilerinin gelişmesiyle birlikte ses ve konuşma verilerine tepki gösterdikleri gözlemlenmiştir (Karmiloff & Karmiloff-Smith, 2001). Hatta bebeklerin doğum öncesi sürekli duyup algıladıkları seslerin doğumdan sonraki ses/konuşma tanıma tercihlerini ve yetilerini etkilediği düşünülmektedir (DeCasper ve diğerleri, 1986). Bu da ses/konuşma algılama yetisinin ne kadarının doğuştan ne kadarının çevresel verilerle gerçekleştiğine dair sorular ortaya çıkarmıştır.

Şimdiye dek yapılan çalışmalar, bebeklerde erken gelişen bu yetinin beyinde bazı bölgelerin bu işlev için doğuştan kodlandığını, ancak ilerleyen yaşlarda kayba uğradığını göstermiştir. Bu bulgu, dil edinimi için kritik veya hassas bir dönem olduğu varsayımıyla doğrudan ilişkilidir (Lenneberg, 1967). Bebeklerde evrensel sesbirimleri ayırt etme yetisi, kendi anadillerini öğrendikçe sadece kendi dilinde bulunan seslerdeki farkları algılayabilecek şekilde azalmaktadır. Örneğin, İngiliz bebekleri 6 aylıkken Hintçe sesleri ayırt edebilirken, 12 ay civarında bu yetilerinin büyük ölçüde azaldığı gözlemlenmiştir (Werker ve diğerleri, 1984). Bu yetinin kaybolması yetişkin olarak öğrenilen yabancı dillerdeki sesbirimlerin algılanmasında ve üretilmesinde önemli problemler yaratmaktadır.

3. Makinede Konuşma Tanıma

İnsanlar gibi davranan, konuşan, hareket eden robotların yapımı bilim adamlarının uzun zamandır yoğun çaba harcadığı bir araştırma konusudur. Bu bölümde makinelerin konuşma anlama yetisinin modellenmesi konusu incelenecektir. Özellikle entegre devre tasarımının 1960'larda hızlı işlem yapma ve veri depolama imkanını sağlaması üzerine bu alanda hızlı gelişmeler kaydedilmiş ve artık otomatik konuşma tanıma sistemleri günlük hayatımızda kullanılmaya başlanmıştır. Otomatik konuşma tanıma sistemlerinin temel olarak iki aşaması bulunmaktadır. İlk aşama bilgisayarın daha önceden kaydedilmiş seslerle ve metinlerle eğitilerek dile özgü akustiğin ve dilbilgisinin modellenmesi, ikinci aşama ise bu modeller kullanılarak kaydedilen herhangi bir konuşmanın içeriğinin metne dönüştürülmesidir.

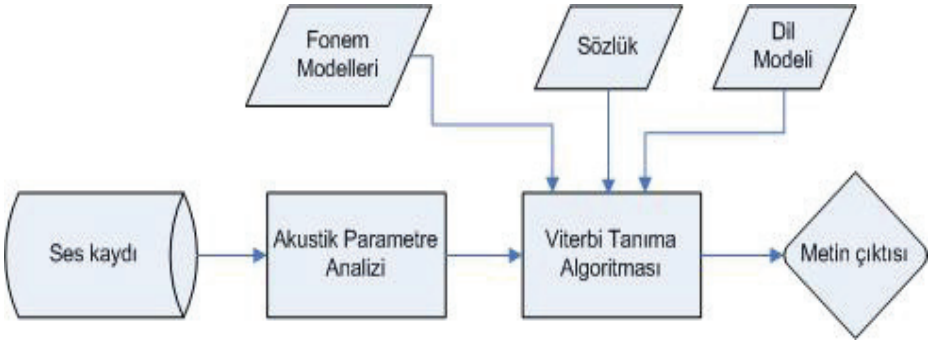
Şekil 2: Makine için akustik ve dil modellerinin eğitimi



Şekil 2'de model eğitimi için yapılan işlemler özetlenmiştir. Model eğitimi de temel olarak iki amaca yöneliktir. İlki sesbirimlerin her birinin akustik incelemesinin yapılıp ortalama istatistiklerinin çıkarılmasıdır. Bu aşamada amaç farklı seslerin spektrumlarındaki farkların matematiksel olarak modellenmesidir. Örneğin /a/ sesini /e/ sesinden ayırt edebilmemiz için iki sesin birçok kişiden alınan örneklerinin frekans incelemesi yapılır ve her bir ses için ortalama spektrum elde edilir. Seslerin birbirinden ayırt edilmesi için sadece akustik olasılıkların hesap edilmesi yeterince yüksek konuşma tanıma performansı vermemektedir. Çünkü herhangi bir dil için söylenebilecek cümlelerin sayısı çok fazladır ve akustik ayrımlama tek başına söylenenlerin makine tarafından yazıya çevrilmesi için yeterli olmamaktadır. Dille ilgili birtakım istatistikler bu aşamada devreye sokulmaktadır. Örneğin sözcüklerin sıklık oranları veya birbirlerini takip etme sıklıkları hesaplanıp buna göre herhangi bir cümlenin söylenme olasılığı bulunabilir. Bu olasılığın da akustik olasılığa eklenmesi ile daha yüksek performansta konuşma tanıma yapılabilmektedir.

Şekil 3'te model eğitiminden sonraki aşama olan konuşma tanıma işleminin makine tarafından nasıl yapıldığı özetlenmiştir. İlk aşamada mikrofon vasıtasıyla akustik sinyal elektrik sinyaline dönüştürülür. Analog-dijital dönüştürücü sinyalin dijital ortama aktarılmasını sağlar. Bilgisayarda Viterbi algoritması (bkz. Huang ve diğerleri, 2001) kullanılarak söylenen cümlenin olası sözcük dizelerinden hangisinin olma olasılığının daha yüksek olduğu hesaplanır. Olasılık bulunurken dil modeline ve akustik modele uygunluk bir arada değerlendirilir. Örneğin bir sözcük dizesi akustik olarak daha yüksek bir olasılık verse de o dilin genel dilbilgisine uygun söylenmediyse makine tarafından farklı şekilde tanınabilir. Bu bakımdan insanlar da çoğu zaman benzer şekilde davranırlar. Aslında fark etmeden sıkça duyduğumuz sözcüklere veya sözcük dizelerine öncelik veririz.

Şekil 3: Makine ile konuşma tanıma işlemi



4. Türkçe'de Konuşma Tanıma

Konuşma tanıma teknolojisindeki gelişmelere koşut olarak son yıllarda, içinde Türkçe seslerin süre özelliklerinin de incelendiği birçok Türkçe ses/konuşma tanıma çalışması yapıldığı görülmektedir (Arısoy ve diğerleri, 2004; Şaylı, 2002; Şaylı ve diğerleri, 2002a, b; Türk ve diğerleri, 2004). Ancak Türkçe için geliştirilmeye çalışılan ses tanıma yazılımlarında, dilin yapısından kaynaklanan bazı önemli sorunlarla karşılaşılmaktadır. Türkçe'nin eklemeli bir dil olması, üretilebilecek sözcük sayısını önemli ölçüde arttırmaktadır. Bu da Türkçe için hazırlanabilecek geniş dağarcıklı ses tanıma uygulamalarında (örn. değişik gazete haber ve yazılarını kapsayan veriler) sorun olmaktadır (Büyük ve diğerleri, 2007). Bu tür uygulamalarda, tanıma birimi olarak sözcükler kullanıldığında sınama verisindeki sözcüklerin %15-20'sinin tanıma sözlüğünün dışında kaldığı ve dolayısıyla kod çözücü tarafından tanınamadığı bildirilmiştir (Arısoy, 2004; Büyük, 2005). Bu durum, Türkçe ses tanıma işlemlerinde diğer dillere oranla daha fazla hata oranı çıkmasına neden olmaktadır (Büyük ve diğerleri, 2007). Bu kapsama sorununa çözüm olarak tanıma birimi olarak sözcük altı birimlerin (seslem, ek veya kök) kullanımı denenmiş ve tanıma oranında daha yüksek bir yüzde elde edilmiştir (Arısoy & Arslan, 2005; Çarkı ve diğerleri, 2000; Erdoğan ve diğerleri, 2005).

Türkçe için hazırlanan geniş dağarcıklı ses tanıma uygulamalarında kapsama sorunu olmasına rağmen, halen kullanılmakta olan yazılımlar daha sınırlı sayıda sözcük ve dilbilgisi yapılarının kullanıldığı alanlarda daha başarılı sonuçlar verebilmektedir. Örneğin, radyoloji raporları gibi sınırlı dil yapıları içeren verilerin metne çevrilmesinde ses tanıma başarısı %85'lerin üzerine çıkmakta ve tanıma hızı artmaktadır (Arısoy & Arslan, 2004; Büyük ve diğerleri, 2007).

Türkçe ses tanıma uygulamaları kapsamında geliştirilen dikte yazılımlarında önemli bir gelişme, kullanıcının, önceden eğitilmiş dil modellerine istediği sözcük veya cümle

yapılarını sonradan ekleyerek modeli ihtiyaca göre genişletebilmesi ve diğer alanlara uyarlayabilmesidir. Bu da kapsama sorununa sınırlı da olsa bir çözüm olabilmektedir (Büyük ve diğerleri, 2007).

5. İnsan-makine Karşılaştırması

5.1 Konuşma Tanımada Hata İncelemesi

Konuşma tanıma için günümüzün en son teknolojisi dahi insan konuşma tanıma performansıyla karşılaştırıldığında oldukça yetersiz kalmaktadır. Tablo 1'de farklı sözlük boyutuna sahip değişik konuşma tanıma uygulamalarında insan ve makine arasındaki performans karşılaştırması verilmiştir (Huang ve diğerleri, 2001). Tablo incelendiğinde makinedeki hata oranının insandaki hata oranından 5-80 kat daha fazla olduğu görülmektedir. Şu anki makine tanıma performansı her ne kadar tabloda listelenen değerlerden daha yüksek olsa da hala insan tanıma performansı ile arasında önemli bir fark bulunmaktadır.

Tablo 1: Değişik ortamlarda insan-makine konuşma tanıma performansı karşılaştırması. WSJ, Wall Street Journal ses veritabanını ifade etmektedir (Huang ve diğerleri, 2001).

Ortam	Sözlük boyutu	İnsan	Makine
Rakamlar	10	0.009%	0.72%
Alfabedeki Harfler	26	1%	5%
Telefon Konuşması	2000	3.8%	36.7%
WSJ temiz ortam	5000	0.9%	4.5%
WSJ gürültülü ortam	5000	1.1%	8.6%

5.2 Makinenin Eksik Yönleri

İnsanla makineyi konuşma tanıma performansı açısından karşılaştırdığımızda günümüzde makinelerin eksik olduğu yönlerin daha fazla olduğunu görmekteyiz. Örneğin, şu aşamada makine bir ortama bırakıldığında, bir insan gibi kendi kendine dili öğrenememektedir. Makinenin eğitimi için verinin çok iyi bir süzgeçten geçirilmesi ve etiketlenmesi gerekmektedir. Oysa insan için durum farklıdır. Bulunduğu ortamdaki tüm sesleri herhangi bir ek girdi gerekmeden yorumlayıp anlamlı veri olarak kullanabilmektedir. Makineler için bir başka eksiklik adaptasyon algoritmalarının yetersiz düzeyde olmasıdır. Eğitildiği ses kayıtlarından farklı ortamlardan ve mikrofonlardan gelen ses kayıtlarını incelerken makinelerin hata oranı oldukça yüksek olmaktadır. Bunun yanı sıra makineler aksanlı konuşmayı algılamada oldukça yetersiz kalmaktadır (Arslan, 1996). Geriplan gürültüsünün ve farklı kaynaklardan gelen seslerin filtrelenmesi gerekmektedir, ancak bu işlem insandaki kadar iyi yapılamadığından tanıma performansı önemli oranda düşük kalmaktadır (Gong, 1995). Makinelerin eksikliği olan bir diğer konu da mevcut otomatik sistemlerin çok fazla veriye ihtiyacı olmasıdır. İstatistiksel dil modelleri için yüz milyonlarca sözcük, ses kaydı olarak ise binlerce saat

konuşma kaydı kullanılmaktadır. Televizyon kayıtlarının transkripsiyonu için 10 dakika ile 135 saat arasında değişen miktarlarda ses kaydıyla eğitildiğinde hata oranı %65.3'ten %37.4'e düşmektedir (Lamel ve diğerleri, 2002). Bu veriye dayalı olarak öngörü analizi yapıldığında insandaki hata oranlarına erişmek için makinelerin şu anki algoritmalarla yaklaşık 1 milyon saat akustik veriye ihtiyacı vardır (Moore, 2003). İnsanlardaki veri gereksinimi ise bu rakamla karşılaştırıldığında oldukça azdır. ABD'de yapılan bir çalışmada (Hart ve Risley, 1995) 2-3 yaşında bir çocuğun yıllık ortalama 800 saatlik konuşma duyduğu saptanmıştır. Bu rakam çocuğun yetiştiği sosyal ortama bağlı olarak 600 ile 2100 arasında değişmektedir. Buna göre bir bebeğin dili anlayıp konuşabilmesi için yaklaşık 2-3 bin saat yeterli olmaktadır. 10 yaşındaki bir çocuk ise ortalama 10 bin saat konuşma duymaktadır. Makine ile karşılaştırıldığında bu rakam çok düşüktür. Bunun temel sebebi insanlardaki verinin yorumlanmasını sağlayan "algoritma" ya da "model"in makinelerde kullanılan yöntemlere göre çok daha gelişmiş olmasıdır.

5.3 Makinenin Üstünlükleri

Konuşma tanıma sistemlerinin temel amacı şimdiye kadar insanı modellemek, bir anlamda taklit etmek olmuştur. Bu çalışmada temel olarak makinenin modellemedeki eksiklikleri özetlenmiştir. Ancak makinenin insana göre hali hazırda önemli üstünlükleri de bulunmaktadır. Örneğin makine yeryüzündeki tüm dilleri tanıyabilir. Öte yandan insanlar en çok 5-10 dil konuşabilmektedir. Makine insanla karşılaştırıldığında birçok işlem için daha düşük maliyetlidir. Makine insandan daha hızlı öğrenebilir. İnsanın yeni bir dili öğrenmesi 1-2 sene alırken makine veri uygun şekilde sunulduğu taktirde 1-2 günde yeni bir dili öğrenebilir. Makine için anadil ile ikinci dil ayrımı yoktur. Tüm dilleri benzer performansla tanıyabilir. İnsanın kritik yaş döneminden sonra anadilde olmayan sesleri algılama ve ayırt etme yetisi önemli ölçüde azalır. Makine için böyle bir problem bulunmamaktadır.

6. Sonuç

Makineler için konuşma tanıma teknolojisi gün geçtikçe gelişmekte ve hızla insanın konuşma tanıma performansına yaklaşmaktadır. Ancak yine de günümüz teknolojisi insan beyniyle kıyaslandığında birçok konuda oldukça zayıf kalmaktadır. Örneğin, insan beyni geri plan seslerini çok iyi bir şekilde filtre edebilmekte ve esas konuşmaya odaklanabilmektedir. Makinenin gelişime açık yönlerinden birisi insanla karşılaştırıldığında çok fazla miktarda veri ihtiyacının olmasıdır. Bilgisayarların hızları ve bellek kapasiteleri arttıkça ve teknolojideki gelişmeler bu hızla devam ettikçe önümüzdeki 50-100 yıl içerisinde insan gibi anlayan ve konuşan bilgisayarların üretilmesi oldukça yüksek bir olasılık olarak görünmektedir.

Kaynakça

Arısoy, E. (2004). *Turkish dictation system for radiology and broadcast news applications*. Yüksek Lisans Tezi, Boğaziçi Üniversitesi.

Arısoy, E., & Arslan, L. M. (2004). *Turkish radiology dictation system*. The 9th International Conference "Speech and Computer" (SPECOM 2004), St-Petersburg, Rusya.

Arısoy, E., & Arslan, L. M. (2005). *Turkish dictation system for broadcast news applications*. The 13th European Signal Processing Conference (EUSIPCO 2005), Antalya, Türkiye.

Arısoy, E. , Arslan, L. M., Demiralp-Nakipoğlu, M. , Ekenel, H. K. , Kelepir, M., Meral, H. M., Özsoy, A.S., Şaylı, Ö., Türk, O., & Yolcu, B. (2004). *Acoustic analysis of Turkish sounds*. The 12th International Conference on Turkish Linguistics, İzmir, Türkiye.

Arslan L. M. (1996). *Foreign Accent Classification in American English*. Doktora Tezi, Duke Üniversitesi.

Binder, J.R. (2000). The new neuroanatomy of speech perception. *Brain, 123 (12),* 2371-2372.

Binder, J. R., Frost, J. A., Hammeke, T. A., Rao, S.M., Cox, R. W. (1996). Function of the left planum temporale in auditory and linguistic processing. *Brain, 119*, 1239-1247.

Büyük, O. (2005). *Sub-word language modeling for Turkish speech recognition*. Yüksek Lisans Tezi, Sabancı Üniversitesi.

Büyük, O., Haznedaroğlu, A., & Arslan, L. M. (2007). *Dil modeli uyarlanabilir Türkçe ses tanıma yazılımı*. Sinyal İşleme Uygulamaları 2007 Konferansı, Eskişehir, Türkiye.

Çarkı, K. , Geutner, P., Schultz, T. (2000). *Turkish LVCSR: Towards better speech recongition for agglutinative languages.* The Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (*ICASSP 2000),* Vol. 3, 1563-1566, İstanbul, Türkiye.

Celesia, G. G. (1976). Organization of auditory cortical areas in man. *Brain, 99*, 403-414.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Davis, K.H. , Biddulph, R. & Ballashek, S. (1952). Automatic recognition of spoken digits, *Journal of Acoustical Society of America, 24* (6), 637-642.

DeCasper, A. J., Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior & Development*, 9(2), 133-150.

Demonet, J-F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J-L, Wise, R. (1992). The anatomy of phonological and semantic processing in normal subjects. *Brain, 155*, 1753-1768.

Elman, J.L., & McClelland, J. L. (1984). Interactive activation model of speech perception. N. Lass (haz.), *Language and speech* içinde, 337-374. New York: Academic Press.

Erdoğan, H. , Büyük, O. , & Oflazer, K. (2005). *Incorporating language constraints in sub-word based speech recognition*. Automatic Speech Recognition and Understanding Workshop (IEEE-ASRU 2005), Cancun, Meksika.

Fry, D.B., & Denes, P. (1959). Theoretical aspects of mechanical speech recognition: The design and operation of the mechanical speech recognizer. *Journal of British Inst. Radio Engr, 19 (4)*, 211-229.

Gong, Y. (1995). Speech recognition in noisy environments: a survey. *Speech Communication, 16*, 261-291.

Hart, B., & Risley, T.R. (1995). *Meaningful differences in the everyday experiences of young American children*. Baltimore: Paul H. Brookes Publishing Company.

Huang, X., Acero, A., & Hon, H.W. (2001). *Spoken language processing*. New Jersey: Prentice Hall.

Itakura, F. (1975). Minimum prediction residual applied to speech recognition, *IEEE Transactions on Acoustics, Speech, Signal Processing, 23* (1), 67-72.

Jelinek, F. (1985). The development of an experimental discrete dictation recognizer. *Proceedings of IEEE, 73* (11), 1616-1624.

Karmiloff, K. & Karmiloff-Smith, A. (2001). *Pathways to language: From fetus to adolescent.* Cambridge, M.A: Harvard University Press.

Lai, C.S.L,. Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., & Monaco, A. P. (2001). A novel forkhead-domain gene is mutated in a severe speech and language disorder. *Nature, 413*, 519-523.

Lamel, L., Gauvain, J-L. ve Adda, G. (2002). Unsupervised acoustic model training. *Proceedings of IEEE International. Conference on Acoustics, Speech and Signal Processing, I*, 877-880.

Lauter, J. L., Herskovitch, P., Formby, C., Raichle, M.E. (1985). Tonotopic organization in human auditory cortex revealed by positron emission tomography. *Hearing Research 20*: 199-205.

Lecanuet, J.P., & Granier-Deferre, C. (1993). Speech stimuli in the fetal environment. B. De Boysson-Bardies, S. De Schonen, P. Jusczyk, P. MacNeilage, & J. Morton (haz.) *Developmental neurocognition: Speech and voice processing in the first year of life* içinde. Dordrecht: Kluwer.

Lenneberg, E. (1967). *The biological foundations of the language*, New York: Wiley.

Liberman, A. M. (1970). The grammars of speech and language. *Cognitive Psychology, 1*: 301-323.

Lippmann R. (1997). Speech recognition by machines and humans. *Speech Communication, 22*, 1-16.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition, 25*, 71-102, 1987.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for pscyhological inquiry.* Hillsdale, NJ: Lawrence Erlbaum.

Mazziotta, J. C., Phelps, M. E., Carson, R. E., Kuhl, D. E. (1982). Tomographic mapping of human cerebral metabolism: auditory stimulation. *Neurology, 32*, 921-937.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants, *Cognition, 29*, 143-178.

Mesulam, M. (2000). Principles of behavioral and cognitive neurology.Oxford: Oxford University Press.

Moore R K. (2003). A comparison of the data requirements of automatic speech recognition systems and human listeners, *Proceedings of. Eurospeech*, Geneva, 2582-2584.

Olson, H. F., & Belar, H. (1956). Phonetic typewriter. *Journal of Acoustic Society of America, 28* (6), 1072-1081.

Petersen, S. E., Fox, P. T., Posner, M. I. Mintun, M. Raichle, M. E. (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature, 331*: 585-589.

Pinel, J. P. J. (1998). A colorful introduction to the anatomy of the human brain Boston: Allyn and Bacon.

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE,77*, 257-286.

Sakoe, H. & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition, *IEEE Transactions on Acoustics, Speech, Signal Processing, 26* (1), 43-49.

Scott, S. K. ,& Johnsrude, I. S. (2003). Neuroanatomical and functional organization of speech perception. *Trends in Neurosciences, 26* (2), 100-107.

Stevens, K. N. (1960). Toward a model for speech recognition. *Journal of the Acoustical Society of America, 32*, 47-55.

Şaylı, Ö. (2002). *Duration analysis and modelling for Turkish text-to-speech synthesis*. Yüksek Lisans Tezi, Boğaziçi Üniversitesi.

Şaylı, Ö., Arslan, L. M. , & Özsoy, A. S. (2002a). *Türkçe ses sentezi için süre modellemesi*. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU-2002), Denizli-Türkiye, 841-845.

Şaylı, Ö., Arslan, L. M. , & Özsoy, A. S. (2002b). *Duration properties of the Turkish phonemes*. 11. Uluslar arası Türk Dilbilim Konferansı bildirisi, Gazimağusa, KKTC.

Türk, O., Şaylı, Ö., Özsoy, A.S. , Arslan, L. M. (2004). *Türkçede ünlülerin formant frekans incelemesi*. 18. Ulusal Dilbilim Kurultayı bildirisi, Ankara, Türkiye.

Velichko, V.M. & Zagoruyko, N.G. (1970). Automatic recognition of 200 words. *International Journal of Man-Machine Studies, 2*: 223.

Werker, J.F., & Tees, R.C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavioral Development, 7*, 49-63.

Wernicke, C. (1974). The aphasic symptom complex: a psychological study on a neurological basis, Kohn & Weigert, Breslau. Tekrar basım: R.S. Cohen, & M. W. Wartofsky (haz). *Boston studies in the philosophy of science* içinde. Cilt 4. Reidel, Boston, Mass.

Wise, R., Chollet, F. Hadar, U., Friston, K., Hoffner, E., Frackowiak, R. (1991). Distribution of cortical neural networks involved in word comprehension and word retrieval. *Brain, 114*: 1803-1817.

Zatorre, R. J., Evans, A.C., Meyer, E., Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science, 256*, 846-849.

Yard. Doç. Dr. Ayşe Gürel
Boğaziçi Üniversitesi
Yabancı Diller Eğitimi Bölümü
Bebek, 34342, İstanbul
agurel@boun.edu.tr

Prof. Dr. Levent M. Arslan
Boğaziçi Üniversitesi
Elektrik Elektronik Mühendisliği Bölümü
Bebek, 34342, İstanbul
arslanle@boun.edu.tr