GU J Sci, Part C, 7(4): 834-854 (2019)

Gazi Üniversitesi

Fen Bilimleri Dergisi

PART C: TASARIM VE TEKNOLOJİ

http://dergipark.gov.tr/gujsc

Overview of Automatic Speech Recognition, Approaches and Challenges: Way the Future to Turkish Speech Recognition

Saadin OYUCU[1, *], Hayri SEVER[2], Hüseyin POLAT[3]

[1] *Gazi University Faculty of Technology, Department of Computer Engineering, 06500, Yenimahalle/ANKARA*

[2] *Çankaya University Faculty of Engineering, Department of Computer Engineering, 06790, Etimesgut/ANKARA*

[3] *Gazi University Faculty of Technology, Department of Computer Engineering, 06500, Yenimahalle/ANKARA*

Article Info:

Research article
Received: 08/05/2019
Revision 16/09/2019
Accepted: 06/10/2019

Highlights

- *Speech Recognition.*
- *Feature Extraction.*
- *Turkish Speech Recognition.*

Keywords

Speech Recognition.
Feature Extraction.
Turkish Speech Recognition.

Graphical/Tabular Abstract

Figure A shows the application area of the Applications Automatic Speech Recognition (ASR) system. Adaptation is required for adaptation processes to be carried out according to the Application area. During the Speech Processing phase, which is the first entry point of the ASR system, feature extraction is performed from the audio signal. Individual properties are obtained by different feature extraction techniques. For example, Mel Frequency Cepstral Coefficient (MFCC) is a feature extraction technique commonly used in speech recognition systems. Decoder, one of the other components of ASR, converts the feature vectors obtained by using Acoustic Model (AM) and Language Model (LM) into phoneme sequences. In acoustic modeling, firstly, the posterior probability of the phoneme within a given time signal is calculated. In the artificial neural network-based acoustic model, the posterior probability of phonemes is independent for each window. This independence means that the phonemes in a word are independent of each other.

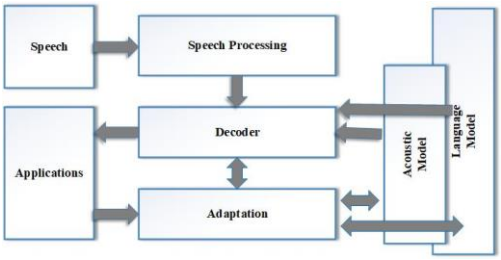


***Figure A**. Basic architecture of speech recognition system.*

Purpose: This study presents a literature review on speech recognition and then discusses the recorded signs of progress made in this area of research for different languages. The data sets used in speech recognition systems, feature extraction approaches, speech recognition methods and performance evaluation criteria are examined and the focus is on the development of speech recognition and the difficulties in this field.

Theory and Methods: In this study, literature review (systematic), which is an important component for a scientific article, was carried out. This process was carried out by the combination of different methods. A combination of review approaches is given.

Results: According to the information obtained as a result of the research; Computational architectures that can be applied to resistance to the acoustic environment, self-learning in ASR, detection of unknown words, the success of the Turkish ASR at a broad and limited repertoire level, insufficient source status and Automatic Speech Recognition ASR were evaluated. In addition, the future of Turkish ASR was discussed and recommendations were made to overcome the current difficulties for Turkish ASR.

Conclusion: The aim of this study is to examine the current speech recognition methods and approaches and to present the developments in this field in detail. For this reason, approaches, datasets and the difficulties faced by the researchers in their studies in this field are discussed in the scope of the study. The effect of deep learning and classical approaches on ASR was investigated. A road map is provided for researchers to incorporate the detailed information necessary for their work in this field to their own work and to overcome the present challenges.

*Corresponding author, e-mail: saadinoyucu@gazi.edu.tr
DOI: 10.29109/gujsc.562111

GU J Sci, Part C, 7(4): 834-854 (2019)

Gazi Üniversitesi

Fen Bilimleri Dergisi

PART C: TASARIM VE TEKNOLOJİ

http://dergipark.gov.tr/gujsc

Otomatik Konuşma Tanımaya Genel Bakış, Yaklaşımlar ve Zorluklar: Türkçe Konuşma Tanımanın Gelecekteki Yolu

Saadin OYUCU[1, *], Hayri SEVER[2], Hüseyin POLAT[3]

[1]*Gazi Üniversitesi, Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü, 06500, Yenimahalle/ANKARA*

[2]*Çankaya Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 06790, Etimesgut/ANKARA*

[3]*Gazi Üniversitesi, Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü, 06500, Yenimahalle/ANKARA*

Makale Bilgisi

Araştırma makalesi
Başvuru: 08/05/2019
Düzeltme: 16/09/2019
Kabul: 06/10/2019

Anahtar Kelimeler

Konuşma Tanıma
Özellik Çıkarımı
Yapay Zekâ
Türkçe Konuşma Tanıma

Keywords

Speech Recognition
Feature Extraction
Artificial Intelligence
Turkish Speech Recognition

Öz

İnsanlar arasındaki en önemli iletişim yöntemi olan konuşmanın, bilgisayarlar tarafından tanınması önemli bir çalışma alanıdır. Bu araştırma alanında farklı diller temel alınarak birçok çalışma gerçekleştirilmiştir. Literatürdeki çalışmalar konuşma tanıma teknolojilerinin başarımının artmasında önemli rol oynamıştır. Bu çalışmada konuşma tanıma ile ilgili bir literatür taraması yapılmış ve detaylı olarak sunulmuştur. Ayrıca farklı dillerde bu araştırma alanında kaydedilen ilerlemeler tartışılmıştır. Konuşma tanıma sistemlerinde kullanılan veri setleri, özellik çıkarma yaklaşımları, konuşma tanıma yöntemleri ve performans değerlendirme ölçütleri incelenerek konuşma tanımanın gelişimi ve bu alandaki zorluklara odaklanılmıştır. Konuşma tanıma alanında son zamanlarda yapılan çalışmaların olumsuz koşullara (çevre gürültüsü, konuşmacıda ve dilde değişkenlik) karşı çok daha güçlü yöntemler geliştirmeye odaklandığı izlenmiştir. Bu nedenle araştırma alanı olarak genişleyen olumsuz koşullardaki konuşma tanıma ile ilgili yakın geçmişteki gelişmelere yönelik genel bir bakış açısı sunulmuştur. Böylelikle olumsuz koşullar altında gerçekleştirilen konuşma tanımadaki tıkanıklık ve zorlukları aşabilmek için kullanılabilecek yöntemleri seçmede yardımcı olunması amaçlanmıştır. Ayrıca Türkçe konuşma tanımada kullanılan ve iyi bilinen yöntemler karşılaştırılmıştır. Türkçe konuşma tanımanın zorluğu ve bu zorlukların üstesinden gelebilmek için kullanılabilecek uygun yöntemler irdelenmiştir. Buna bağlı olarak Türkçe konuşma tanımanın gelecekteki rotasına ilişkin bir değerlendirme ortaya konulmuştur.

Overview of Automatic Speech Recognition, Approaches and Challenges: Way the Future to Turkish Speech Recognition

Abstract

Speech is a paramount means of communication among humans, which makes recognition of the speech by computers is a study area of significance. In this research area, many studies have been carried out based on different languages. Studies in the literature have played an important role in increasing the performance of speech recognition technologies. This study presents a literature review on speech recognition and then discusses the recorded signs of progress made in this area of research for different languages. The data sets used in speech recognition systems, feature extraction approaches, speech recognition methods and performance evaluation criteria are examined and the focus is on the development of speech recognition and the difficulties in this field. Recent studies in the field of speech recognition have focused on developing more powerful methods against adverse conditions (environmental noise, speaker and language variability). Therefore, a general perspective on recent advances in speech recognition in unfavorable conditions is presented. In this way, an attempt has been made to help in selecting the methods that could be used to overcome the congestion and difficulties in speech recognition under adverse conditions. In addition, the well-known methods used in Turkish speech recognition were compared. Difficulty in the recognition of Turkish speech and appropriate methods to overcome these difficulties have been investigated in detail. Accordingly, a future route assessment of Turkish speech recognition has been put forward.

1. GİRİŞ (INTRODUCTION)

Temelde insanların birbirleri ile iletişimini destekleyen teknolojiler zaman içerisinde yapısal olarak değişmiş ve insan-makine etkileşimi kavramı ortaya çıkmıştır [1]. Erken evrelerde basit tuşlu sistemler ile gerçekleştirilen insan makine veya elektronik cihaz etkileşimi günümüzde yerini dokunmatik sistemlere bırakmıştır [2]. Fakat bu durum bile insanların makine veya elektronik cihazlar ile iletişimini sınırlı olarak sağlamaktadır. İnsanların makine veya elektronik cihazlarla ana dillerini konuşarak iletişim kurması, makina veya elektronik cihazların komut ve kontrol mekanizmasının daha hızlı bir şekilde çalıştırabilmesine olanak vermektedir. Bu nedenle Otomatik Konuşma Tanıma (ASR: Automatic Speech Recognition) sistemleri geliştirilmekte ve kullanım alanı her geçen gün artmaktadır.

ASR sistemleri iletişim, eğitim, elektronik eşya, savunma sanayi, adli ve istihbarat gibi alanlarda sıklıkla kullanılmaktadır (Tablo 1) [3]. ASR sistemleri ile metne aktarılan konuşmalar üzerinde yapay zekâ teknolojileri kullanılarak konuşma anlama ve duygu analizi gibi işlemler de gerçekleştirilmektedir [4]. Konuşma anlama ile komuta kontrol işlemleri yapılabildiği gibi aynı zamanda kişisel asistanlar da geliştirilebilmektedir [5]. Duygu analizi ise bir konuşmanın içeriğine göre konuşmacının içerisinde bulunduğu duygu durumunun belirlenmesidir [6]. Konuşma içerisindeki duyguların analizi ile özellikle çağrı merkezlerinde hizmet kalitesinin arttırılması amaçlanmaktadır. Ayrıca konuşma metinleri analiz edilerek müşteri ilişkileri yönetilmekte ve müşterilere özel fırsatlar sunulmaktadır [7].

Tablo 1. *Konuşma tanıma teknolojisinin kullanıldığı alanlar ve uygulamalar*

Kullanım Alanı	***Uygulama***	***Girdi***	***Çıktı***
İletişim alanı	*Telefon rehberinden sesli sorgulama*	*Konuşma sinyali*	*Konuşulan kelimeler*
Eğitim alanı	*Yabancı dil eğitiminde kelimelerin doğru okunuşunu öğretmek*	*Konuşma sinyali*	*Konuşulan kelimeler*
Elektronik	*Fırın, çamaşır makinesi ve buzdolabı gibi ev eşyalarında konuşma tanıma*	*Konuşma sinyali*	*Konuşulan kelimeler*
Savunma	*Savunma sanayide kullanılan araçların kontrol edilmesi*	*Konuşma sinyali*	*Konuşulan kelimeler*
Yapay zekâ	*Robotların kontrolü*	*Konuşma sinyali*	*Konuşulan kelimeler*
Sağlık alanı	*Doktorların muayene raporlarını yazmak*	*Konuşma sinyali*	*Konuşulan kelimeler*
Çağrı merkezleri	*Telefon görüşmelerinin otomatik yazıya dökülmesi ve kalite kontrol*	*Konuşma sinyali*	*Konuşulan kelimeler*

ASR sistemleri sadece sosyal, kültürel ve ekonomik alanda değil aynı zamanda adli ve istihbari alanlarda da yaygın olarak kullanılmaktadır [8]. İstihbari alanda şüpheli şahısların konuşmalarının ASR sistemleri ile metne aktarılması ve bu metinler üzerinde suç analizi yapılması çalışılmaktadır. Yine istihbarat amacıyla terör gruplarına ait karasal, uydu, kablo ve internet üzerinden yapılan radyo ve televizyon yayınlarında geçen konuşmalar ASR sistemleri ile metin haline getirilmektedir. Bu metinler üzerinde kelime yakalama veya büyük veri analizi yapılabilmektedir. Konuşulan ifadenin doğru bir şekilde metne aktarılması ise önem arz etmektedir. Bu nedenle ASR sistemlerinin başarımını arttırmak amacıyla birçok çalışma yapılmıştır. Fakat ASR sistemlerinin başarısı henüz istenilen seviyeye ulaşamamıştır.

Prakoso ve arkadaşlarına göre insanlar tarafından konuşulan kelimeleri bilgisayar tarafından okunabilir metne dönüştürmesini sağlayan bir teknoloji olarak ifade edilen ASR farklı çalışma alanlarının bir arada kullanılmasını gerektirmektedir [9]. Farklı disiplinlerin bir arada kullanılması ve genel yapısının karmaşık olması ASR sistemlerinin başarım oranları üzerinde etkili olmaktadır. Başarım oranlarını olumsuz yönde etkileyen bu karmaşıklık ASR sistemleri üzerine daha birçok çalışma yapılması gerektiğini ortaya koymuştur. ASR sistemlerinde arzu edilen başarı seviyesine ulaşamamanın nedenlerinden birisi de geleneksel yapay sinir ağlarının eğitim zamanın uzun olmasından kaynaklanmaktadır. Fakat bu tıkanıklık

Grafik İşlem Birimi (GPU: Graphics Processing Unit) tabanlı işlemlerin yaygınlaşması ile aşılmış görülmektedir [10]. GPU tabanlı işlemler ile birden fazla katmana sahip derin öğrenme tabanlı yapay sinir ağlarının eğitilmesi çok daha kısa sürmektedir. Bu nedenle daha önceleri klasik yöntemlerle eğitilen ASR modelleri derin öğrenmeye dayalı çok katmanlı yapay sinir ağı mimarileri kullanılarak eğitildiğinde daha yüksek başarım oranlarına sahip ASR sistemleri ortaya çıkmaya başlamıştır.

Son yıllarda Türkçe ASR sistemlerinde de umut vaat eden gelişmeler olmuştur. Türkçe yapısı gereği konuşma tanıma alanında birçok zorluğa sahiptir. Bu zorlukların en başında ise Türkçe'nin sondan eklemeli yapısı ve üretken bir dil olması gelmektedir [11]. Üretken dil yapısı kelime uzayını genişlettiğinden bir kelimeden sonra gelebilecek kelimelerin tahminini zorlaştırmaktadır. Ayrıca Türkçe ASR sistemleri için gerekli olan modellerinin eğitilmesinde kullanılacak olan konuşma ve metin verisinin azlığı Türkçe için gerçekleştirilen çalışmaların önünde büyük bir engel olarak görülmektedir. Türkçe konuşma ve metin verisinin hazırlanması için yapılan çalışmaları yaygınlaştırmak ve geliştirmek için bazı girişimler olmuştur. Interactive Systems Laboratories, her biri 20 dakika olan 100 konuşmacıdan oluşan, Türkçe gazetelerin konuşma bölümlerini içeren Global Phone adlı çok dilli bir konuşma veri tabanı toplamıştır [12]. Orta Doğu Teknik Üniversitesi Bilişim Enstitüsü, Sabancı Üniversitesi ile işbirliği ile Türkçe metin veri kümesi geliştirmek için çalışmaktadır [13]. Bir diğer yakın tarihli proje ise OrienTel projesidir. Bu proje Türkçe telefon konuşma veri tabanının tasarımını içermektedir [14]. Diğer bir veri kümesi ise Boğaziçi Üniversitesi'nin oluşturduğu Türkçe haber yayınlarından elde edilen konuşma veri kümesidir [15].

Türkçe üzerine hazırlanan veri kümeleri Türkçe dili üzerine geliştirilen ASR sistemlerinin geliştirilmesini hızlandırmıştır. Fakat Türkçenin üretken morfolojisi, pek çok eşsiz kelime formunu beraberinde getirerek tahmin olasılığını düşürmektedir [16]. Fince, Estonca ve Çekçe gibi morfolojik açıdan zengin diğer diller de ASR sistemlerinin geliştirilmesini zorlaştırmaktadır [17]. Bu zorluğu aşabilmek için ASR sistemlerinin geliştirilmesinde dil modellemesinin yanında okunuş sözlüğü kullanılmaktır. Bilgi edinimi için gerekli olan istatistiksel dil modelleri, verilen bir kelimeden bir sonraki kelimenin tahmini görevini yerine getirmektedir. N-gram tabanlı dil modelleri ise dilin düzenlerini yakalayabilirken kelime kümelenmesi tabanlı modeller, verilerin ses şiddeti ile ilgili bilgilerini daha iyi tanımlayabilmektedir [18].

Bu çalışma kapsamında literatürde ASR sistemleri üzerine gerçekleştirilmiş çalışmalar detaylı olarak incelenmiş, mevcut ASR sistemlerinin gereklilikleri, bileşenleri ve kullanılan yöntemler araştırılmıştır. Buna bağlı olarak da Türkçe ASR sistemlerinin geliştirilmesindeki zorluklar analiz edilerek, Türkçe ASR sistemlerinin geleceği ile ilgili bir perspektif oluşturulmaya çalışılmıştır. Sonuçta, Türkçe ASR sistemleri üzerine çalışacak araştırmacılara elde edilen bilgi birikiminin aktarılması ve yardımcı olunması amaçlanmıştır. Bu çalışmada ayrıca ASR sistemlerinin başarımlarının arttırılması için gerekli olan akustik model, dil modeli ve okunuş sözlüğünün Türkçe ASR sistemi üzerindeki etkileri de irdelenmiştir. Bu etkiler tartışmaya açılarak Türkçe ASR sistemlerinin başarımının arttırılması ve sondan eklemeli dil yapısının dezavantajının ortadan kaldırılması hedeflenmiştir.

1.1. Literatür Taraması (Literature Review)

Literatürdeki çalışmalar ASR alanındaki mevcut durumu yeni ve birleştirici bir bakış açısıyla tamamlamaktadır. Literatür bu bakış açısıyla incelendiğinde biyolojik öğrenme teorilerinin ortaya çıkması ve yapay sinir ağlarının ilk modellerin üretilmesinin 1940'lı yıllara dayandığı görülmektedir [19]. McCulloh ve Pitts, McCulloch-Pitts nöronu adı verilen nöronların otomata benzer bir modelini geliştirmişlerdir [20]. Model, çıkışın pozitif veya negatif olduğunu test ederek iki farklı girdi kategorisini tanıyan ikili bir cihaz olarak literatüre kazandırılmıştır. Bu modelin başarılı olmamasının sebebi ağırlıkların insan operatörler tarafından ayarlanması gerekliliğidir.

1950'lerde bir psikolog olan Rosenblatt yapay bir nöronun eğitilebilir olduğunu göstermiştir [21]. Nöron eğitiminin ardından bir ya da iki gizli katmanı olan bir sinir ağını eğitmek için geri yayılma ile bağlantılı yaklaşımlar geliştirilmiştir [22]. Rumelhart ve arkadaşları derin sinir ağlarını eğitmek için geri yayılım kullanmış ve geri yayılma algoritmasının popülerleşmesini sağlamıştır. Waibel, bir girişin uzun bir dizi olması zorluğunu yöneten Zaman Gecikmeli Sinir Ağı (TDNN: Time Delay Neural Network) yaklaşımını sunmuştur [23]. Bu yaklaşım özellikle ASR sistemlerinde giriş olarak verilen zamana bağlı konuşma bilgisinin yapay sinir ağları ile işlenebilmesinin önünü açmıştır. Bu yaklaşımda giriş bilgisinin tümüne bir

defa bakmak yerine, her birim bir defa girişin bir alt kümesi olarak sunulmuştur. Böylelikle uzun bağımlılıkların eğitilebilmesi gerçekleştirilebilmiştir.

Yapay sinir ağlarının eğitilebilme ve geri yayılma problemleri giderildikten sonra ASR sistemleri için gerekli olan modelleri geliştirebilmek için farklı yaklaşımlar sunulmuştur. Saklı Markov Modelleri (HMM: Hidden Markov Model) ile ilk modeller üretildikten sonra birden fazla katmana sahip yapay sinir ağı modelleri geliştirilmiştir. Özellikle zaman bağımlı girdileri modelleyebilmek için Zaman Gecikmeli Sinir Ağı (TDNN: Time Delay Neural Network) yaklaşımı geliştirilmiştir. Waibel ve arkadaşları TDNN ile HMM arasında bir karşılaştırma yapmış ve TDNN yaklaşımının özellikle ASR sistemlerinde daha düşük hata oranı verdiği gösterilmiştir [24]. Ancak birçok sinir ağının geri yayılma ve ağırlık güncelleme işlemi ile ilgili sorunları mevcuttur. Özellikle uzun ve zamana bağlı bilgilerin modellemesindeki zorlukları çözebilmek için Uzun Kısa Süreli Bellek (LSTM: Long Short Term Memory) Hochreiter ve Schmidhuber tarafından önerilmiştir [25]. Arısoy ve Saraçlar LSTM yapısını Türkçe yayın haberlerinin transkripsiyonu için kullanmış ve tekrarlayan sinir ağlarını kullanarak bir ASR sistemi geliştirmiştir [26].

Yapay sinir ağlarının çok katmanlı modellerinin geliştirilmesi ve geri yayılma işleminin sorunları nispeten giderildikten sonra derin öğrenme olarak adlandırılan mevcut olarak sıklıkla kullanılan yöntem 2006'da başlamıştır [27], [28]. Bu kapsamda Hinton, Derin İnanç Ağları'nın (DBN: Deep Belief Network) açgözlü yaklaşım ile eğitimini öngören bir strateji kullanarak verimli bir şekilde eğitilebileceğini göstermiştir. Stratejinin amacı ağırlıkları rastgele değil daha akıllı bir şekilde başlatmaktır [28].

Yapay sinir ağındaki ağırlıkların rastgele değil de daha akıllı bir şekilde değiştirilmesi yüksek hesaplama ihtiyacı gerektirmektedir. Ayrıca yapay sinir ağındaki katman sayısının da artması yüksek hesaplama ihtiyacını beraberinde getirmiştir. CPU'ların bu konudaki yetersizliği derin sinir ağlarının eğitimi işleminin GPU'lara aktarılması ile giderilmiştir. Raina ve arkadaşlarının 2009 yılında gerçekleştirdiği bir çalışmada CPU yerine GPU kullanmanın işlem hızının 70 kat daha hızlı olabileceği gösterilmiştir [29]. Böylelikle ASR sistemleri için gerekli olan hesaplama ihtiyacı sağlanmış ve çok katmanlı mimarilerin geliştirilmesine olanak sağlanmıştır.

Derin öğrenme tabanlı yaklaşımlar özellikle görüntü işleme, konuşma tanıma, dil modelleme, ayrıştırma, konuşma çevirisi, araç otomasyonu ve benzeri birçok araştırma alanında yüksek başarım göstermiştir. Derin öğrenme yaklaşımlarının yüksek başarımının nedeni büyük miktarda veri ile eğitildiğinde çok karmaşık ve birleşik bir yapıyı bulma ve farklı öğrenme yetenekleridir [30]. Derin öğrenme yaklaşımları ile birden fazla katmana sahip modeller, ASR için gerekli olan akustik modelleri geliştirmek için kullanılmıştır [31]. Fohr ve arkadaşları derin sinir ağlarını saklı markov modelleri ile birleştirmiş ve hibrid bir mimari sunmuştur [31]. Bu mimari kurgu haber yayınlarından oluşan bir veri kümesi ile eğitilmiştir. Elde edilen başarım oranları tek başına kullanılan HMM veya derin olmayan yapay sinir ağlarına göre çok daha yüksektir.

Derin öğrenme yaklaşımları çoğunlukla İngilizce, Çince ve İspanyolca gibi kaynak bakımından zengin diller için sıklıkla uygulanmıştır. Türkçe üzerine gerçekleştirilen ASR araştırmaları Türkçe'nin sondan eklemeli bir yapıya sahip olması nedeni ile yetersiz kalmaktadır. ASR sistemlerinde büyük kelime dağarcığının Derin öğrenme tabanlı yaklaşımlar her ne kadar başarım oranlarını arttırsa da hala istenilen seviyeye ulaşılamamıştır. Bu nedenle geniş dağarcık problemini çözmek amacıyla heceleme veya gövde sonlarındaki ekleri ortadan kaldırarak çözümleme gibi yaklaşımlar önerilmiştir [32].

Klamanuka'nın gerçekleştirdiği çalışmada ise akustik modelleme için derin sinir ağlarının kullanılmasının büyük kelime dağarcığına sahip Türkçe ASR için henüz tam olarak kullanılmadığı fark edilmiş ve derin sinir ağlarını kullanarak bir ASR sistemi geliştirilmiştir Daha önceki geleneksel Gauss karışımı modeli ve HMM yöntemiyle derin öğrenme yöntemi karşılaştırılmıştır. Derin öğrenme ve saklı markov modelinin hibrid bir şekilde kullanılması geleneksel olarak eğitilmiş ASR sistemlerine göre daha başarılı olduğu gösterilmiştir. Gerçekleştirilen literatür çalışmasında geleneksel yaklaşımlar ve derin öğrenme tabanlı yaklaşımlar detaylı olarak incelenmiştir.

Tablo 2'de farklı ASR çalışmalarında kullanılan yöntem, uygulama alanı ve bu çalışmalar hakkındaki değerlendirmelerimize yer verilmiştir.

***Tablo 2**. Kullanılan yöntem ve uygulama alanına göre gerçekleştirilen çalışmalar*

Kaynak	***Kullanılan yöntem***	***Uygulama alanı***	***Değerlendirme***
[33]	*Gizli Markov Modeli Araç Seti (HTK: Hidden Markov Model Toolkit) kullanılarak bir ASR sistemi oluşturulmuştur. Veri hazırlama görevleri, dilbilgisi ve okunuş sözlüğü HTK araç seti ile geliştirilmiştir.*	*ASR sisteminin başarı ölçümü, konuşma bozukluğu olan ve Malayca konuşan çocuklar üzerine gerçekleştirilmiştir. Malay-Polinezya dilleri, ayrıca Avustronezya dilleri olarak da bilinen batı alt familyasına ait fonetik bir dildir.*	*Bu çalışmada, konuşma engelli bireylerin konuşma anlaşılırlığını ölçmede ASR uygulamalarının potansiyeli araştırılmıştır.*
[34]	*Bu çalışmada, düşük bant konuşma işlemi ve akustik modelleme yönteminin özelliklerinden faydalanılarak kodlama eğitiminde Mel düşük bant konuşma tanıma işlemi önerilmiştir.*	*Çalışmada konuşma verilerinin yanı sıra, farklı çalışmalar ile karşılaştırma yapabilmek için konuşma tanıma deneyleri sunulmuştur. Gürültülü ortamlarda farklı filtrelerin kullanılması ile elde edilen aurora2 veri kümesi kullanılmıştır.*	*Eğitim sürecinde daha yüksek seviyede gürültü içeren Mel düşük bantlarına daha düşük ağırlıklar tahsis edilerek gürültü davranışı eğitebilir ve gürültü azaltmada daha iyi performans gösterilebilir.*
[35]	*Çocukların konuşmasını tanımak için yetişkinlerden elde edilen veriler kullanılmıştır. İki farklı transfer öğrenme yaklaşımı karşılaştırılmıştır. Önceden eğitilmiş yetişkin modeline akustik model uyarlaması yapılmıştır.*	*Sistemin eğitimi için Mandarin yetişkin konuşma eğitim kümesi "King-ASR-118 mobile speech corpus"dan alınan veriler kullanılmıştır. Dört tip cep telefonuyla 975 kişiden yaklaşık 360 bin farklı konuşmadan oluşmaktadır.*	*Çocukların konuşmalarının işlenmesi, büyük ölçekli çocuk konuşma verilerinin bulunmaması nedeniyle yetişkinler için konuşma tanıma görevinden daha zorlayıcıdır.*
[36]	*Bu çalışmada Kaldi araç seti kullanılarak Hintçe ASR modeli sunulmuştur. Akustik modelleme, Markov modelleri ve Gaussian karışımları kullanılarak yapılmıştır. N-gram dil modeli kullanılarak hem monophone hem de triphone modelinin performansı sunulmuştur.*	*Kaldi araç seti kullanarak Hintçe dilinin ASR performansını değerlendirmek amacıyla MFCC ve Algısal Doğrusal Tahmini (PLP: Perceptual Linear Prediction) özellikleri, AMUAV corpus'un 1000 adet Hintçe cümle verisi kullanılarak elde edilmiştir.*	*Triphone modeli kelime hata oranını azaltmıştır. Fakat Hintçe üzerine eğitim verilerinin çoğaltılması gerekmektedir. Çok katmanlı modeller için 1000 adet cümle verisi yetersiz kalmaktadır.*
[37]	*Bu çalışma Mandarin ve İngilizce üzerine gerçekleştirilmiştir. Kod geçişli konuşma tanıma performansını artırmak için çok görevli öğrenme yaklaşımı önerilmiştir.*	*Bu çalışma kapsamındaki deneyler bir kamu veri kümesi olan LDC2015S04 üzerinde gerçekleştirilmiştir. Konuşmacılar cinsiyete göre dengelenmiştir ve 19 ile 33 yaş arasındadır.*	*Kod geçişli konuşma tanıma görevi için dil modelinin etkisi üzerine araştırmaların yapılması gerekmektedir.*
[38]	*Bu çalışmada ses komutlarına göre kod oluşturma işlemi için Java dili temel alınmıştır. Belli başlı komutların*	*Bu çalışma görme engelliler için programlama kavramlarını öğrenmeyi*	*Sadece belirli komutlar üzerine sistemin geliştirilmesi geniş konuşma alanına sahip*

Kaynak	Kullanılan yöntem	Uygulama alanı	Değerlendirme
	okunuşları ile sistem eğitilmiş ve bir akustik model oluşturulmuştur.	kolaylaştıran bir web uygulaması sunmaktadır.	ASR'nin başarımını net olarak vermemektedir.
[39]	Bu çalışmada, konuşma tanıma için ileri beslemeli yapay sinir ağları ve dikkat tabanlı Seq2Seq modelleri arasında deneye dayalı bir karşılaştırma yapılmıştır.	Tüm modeller, LDC veri tabanında bulunan standart Fisher-Swbd veri kümesi üzerinde eğitilmiştir. Hiper parametre ayarlaması için RT02 corpus'un (2004S11) bir kısmı kullanılmıştır.	Geçici sınıflandırma kaybı ile eğitilmiş uçtan uca modeller, eğitim sürecini basitleştirmektedir. Fakat büyük dil modelleriyle sistemin karşılaştırılması gerekmektedir.
[40]	ASR performansını artırmak için Microsoft uygulama program arayüzü, İngilizce ses modeli ve bir Tayca ses modeli sunan Google kullanarak iki katmanlı konuşma tanıma sistemi elde edilmiştir.	Çalışma kapsamındaki deneyler sırasıyla 46 dB ve 70 dB olan farklı iki gürültü senaryolarında gerçekleştirilmiştir. Önerilen tekniğin performansı, Microsoft ve Google ile karşılaştırılmıştır.	Önerilen tekniğin, %6 hata oranı ile diğer tekniklere göre daha iyi performans gösterdiği görülürken, Microsoft uygulama program arayüzü %14 ve Google uygulama program arayüzü %11 hata oranı ile sonuç vermiştir.
[41]	Bu çalışmada, İngilizce ASR sistemi oluşturmak için Kaldi konuşma tanıma araç setini kullanarak Dil Modeli ve Akustik Modelin oluşturulmasına çalışılmıştır.	Test verileri, kontrolsüz ortamlarda Karnataka çiftçilerinden toplanarak ASR modellerinin geliştirilmesi için kullanılmıştır. Toplanan konuşma verileri hazır bir araç kullanılarak metne aktarılmıştır.	Geliştirilen ASR modellerinin, çiftçilerin tarımsal emtia fiyatlarına ve hava durumu bilgilerine zamanında erişmelerini sağlayan sistemlerde kullanılabileceği belirtilmiştir.
[42]	Bu çalışmada Tayland'ın kuzeydoğusunda konuşulan Isan dili için otomatik bir rakam konuşma tanıma sistemi sunulmuştur. Konuşma özelliklerini çıkarmak için Mel frekanslı katsayılar tekniği ve konuşma tanıma için Saklı Markov modelleri kullanılmıştır.	Bu çalışmada yerli konuşmacılarından elde edilen bir Isan rakam (0-999) okunuş verisi toplanmıştır. Isan rakamlarını söyleyen konuşmacılar için izole edilmiş veriler ve sürekli konuşma tanıma görevi üzerine odaklanılmıştır.	Sadece rakamlar üzerine yoğunlaştığından başarım oranları yüksektir. Geniş kelime haznesine sahip bir ASR ile karşılaştırılması mümkün değildir.
[43]	Bu çalışmada Saklı Markov modelleri kullanıcının sağladığı sesi tanımak için kullanılır. Soru cevaplama sistemi, ilgili belgeleri almak için Lucene arama motorundan vektör uzay modeli kullanmıştır.	Bütün deneyler Endonezce üzerine yapılmıştır. Sistemi eğiten konuşma verileri tek bir kaynaktan alınmıştır. Sisteme giren toplam konuşma verileri 30 kelimelik kompozisyon ve her kelime 30 kez tekrarlanarak oluşturulmuştur.	Konuşma eğitim verilerinin miktarının arttırılması gerektiği belirtilmiştir. Ayrıca belge üzerinden alınacak cevabı görmek için sözdizimsel ve semantik yöntemlerin kullanılması sonuçları iyileştirebilir.

Kaynak	*Kullanılan yöntem*	*Uygulama alanı*	*Değerlendirme*
[44]	*Bu çalışmada açık kaynaklı Sphinx4 araç seti kullanarak sürekli ASR için Endonezce üzerine bir uygulama geliştirilmiştir*	*Çalışmada 407 cümleden oluşan bir metin derlemi ve 4070 konuşma dosyası kullanılmıştır. 10 Endonezyalı konuşmacı tarafından veri oluşturulmuştur.*	*Bu çalışmada kullanılan eğitim verisi sürekli konuşma tanıma işleminde yetersiz kalmaktadır. Verilerin çoğaltılması gerekmektedir.*

Literatürde ASR sistemlerinin çok farklı diller üzerine inşa edildiği görülmüştür. Ayrıca farklı dillerdeki konuşma tanıma görevleri için farklı ortam (gürültü veya diğer akustik parametreler) değerlendirmeleri için yapılan çalışmalar da mevcuttur. Galic ve arkadaşlarının gerçekleştirmiş olduğu kısık sesli konuşmaların tanınması çalışması [45], Shahnawazuddin ve arkadaşlarının gürültülü ortamlarda çocukların konuşmalarının tanınması için yaptıkları çalışma [46] örnek olarak verilebilir. Ayrıca, Google tarafından geliştirilen görsel ve işitsel bilgilerin aynı anda kullanıldığı bir projede gürültülü ortamlardaki konuşmaların daha yüksek başarım oranları ile metne aktarılması çalışılmıştır [47]. ASR işlemi konuşma-metin verileri ile birebir eşleştirilen görsel anlatım ile desteklenmiştir. Bu çalışma da ana sorun gürültülü bir ortamda (üst üste binen karşılıklı konuşmalar, alkış veya yüksek sesle konuşarak diğer konuşmaları bastırmak) geçen konuşmayı tanımlayabilmektir. Bu sorunu çözebilmek için gerçek hayatta olduğu gibi sadece bir konuşmacıya yoğunlaşarak ve konuşmacının yüz hareketleri ele alınarak bir ASR sistemi geliştirilmiştir. Böylelikle gerçek hayatta olduğu gibi gürültülü bir ortamda olunsa dahi hem görsel iletişim hem de duyusal iletişim ile ASR sistemlerinin başarımı arttırılmaya çalışılmıştır.

ASR sistemlerinin başarımını attırmaya dayalı çalışmalar yapılırken sadece yetersiz veri kaynağı değil yetersiz donanım kaynağı da bir sorun olarak karşımıza çıkmaktadır. ASR sistemlerinin gerçek hayattaki uygulamalarda çevrimiçi olarak kullanılması ve hesaplama yoğunluğunun GPU üzerine aktarılması ile yetersiz donanım sorunu nispeten giderilmiştir. Çevrimiçi kullanımda ASR için gerekli olan donanım kaynağı uzak ortamlarda bulunan güçlü sunucular ile sağlanmıştır. Ancak ASR için gerekli olan yüksek hesaplama ihtiyacı çevrimdışı sistemlerde hala bir sorun olarak karşımıza çıkmaktadır. Bu nedenle çevrimiçi ASR sistemlerinin sunduğu yüksek hesaplama alt yapısı sayesinde bulut tabanlı ASR uygulamaları çoğalmış ve yaygınlaşmıştır.

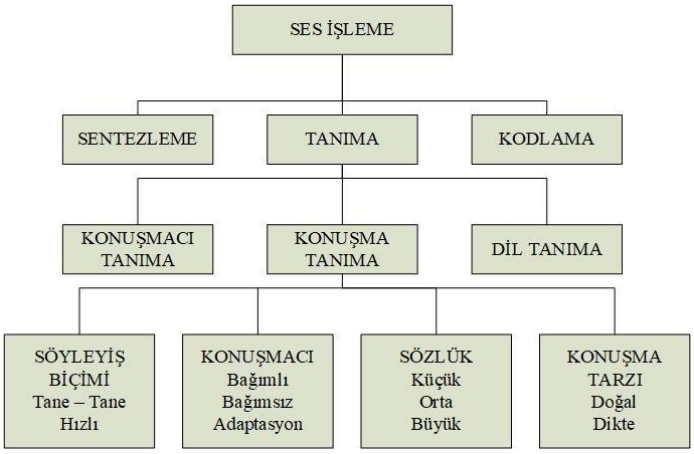
ASR sistemlerinin yaygınlaşması ve ekonomik kazanımların ortaya çıkması ile bu alana yapılan yatırımlar artmıştır. Amazon firması "Amazon Triscribe", Microsoft firması "Azure Bing Speech", Google firması "Cloud Speech-to-Text", IBM ise "Watson Speech-to-Text" projesini başlatmıştır [48]. Bu kapsamda bulut platformu üzerinden sunulan Uygulama Programlama Arayüzleri (API: Application Programming Interface) sayesinde farklı uygulamalar geliştirilmiştir. Özellikle Google'un sunduğu "Ok Google", Microsoft'un sunduğu "Cortana", IBM'in sunduğu "Watson" ve Apple'ın sunduğu "Siri" kişisel asistan uygulamaları en iyi örneklerdendir. Ayrıca bulut tabanlı ASR servis sağlayıcıları sundukları API'ler yardımı ile ASR sistemlerinin kullanımı yaygınlaştırmıştır. Ancak kullanıcılar bulut tabanlı ASR sistemlerinin gizliliğinden ve güvenliğinden tam anlamı ile emin değildirler. Bu nedenle çevrimdışı ASR sistemlerin yetersiz donanım kaynağına rağmen başarımlarını arttıracak yaklaşımlar gerçekleştirilmelidir [49].

ASR sistemleri üzerinden sadece konuşma bilgisinin metne dönüştürülmesi değil farklı çalışmalarda gerçekleştirilmiştir. Örneğin, konuşma duygusunun tanımlanması [50], [51], negatif etki ve saldırganlığın otomatik olarak tanımlanmasını sağlayan konuşma analizinin yapılması [52] ve cinsiyet tanınması [53] gibi çalışmalar da mevcuttur. Ayrıca aksan tanıma çalışmaları da ASR sistemlerinin başarımını artırmada önemli rol oynayacağı gibi aynı zamanda konuşmacı hakkında detaylı bilgiler vermektedir [54]. Aksan tanıma çalışmalarında dil bilimsel bir yaklaşım izlenmektedir. Bir dile ait bir kelimenin okunuş biçiminde sergilenen morfolojik yaklaşım o konuşmacının konuştuğu aksan hakkında bilgiyi içermektedir [55]. Sagha ve arkadaşlarının yaptığı bir çalışmada ise konuşma bilgisi üzerinden yaş bilgisi, cinsiyet bilgisi ve konuşma bilgisinin ASR sistemleri üzerindeki performansı araştırılmıştır [56].

Güncel ASR sistemlerinin performanslarının gerçek kişilerin konuşma tanımasına göre çok daha düşük olduğu bilinmektedir. Yapay sinir ağları, konuşma tanıma çalışmalarının erken evrelerinde başarımı arttırmak için kullanılmıştır. Ancak bazı çalışmalarda konuşma verisinin elde edilmesinde özel koşullar oluşturulmuş ve bu konuşma verileri kullanılarak yapay sinir ağı ile konuşma tanıma gerçekleştirilmeye çalışılmıştır. Bu çalışmalardan Lin ve arkadaşlarının boğaz mikrofonu kullanan kişiler üzerine uyguladıkları ASR sistemi dikkat çekicidir [57]. Geleneksel bir akustik mikrofon sinyaline göre boğaz mikrofonunun işlenmesi zorlu bir görevdir. Bunun gibi zorlu görevlerin üstesinden gelebilmek için gerçekleştirilen ilk çalışmalar, zaman gecikmeli çok katmanlı yapay sinir ağlarının küçük harf grupları arasında ayrımcılık sağlayabileceğini göstermiştir.

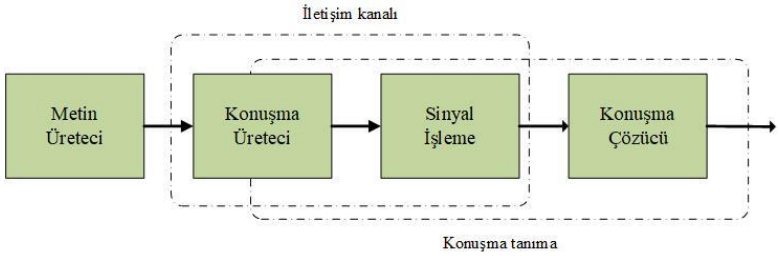
2. OTOMATİK KONUŞMA TANIMA SİSTEMLERİ (AUTOMATIC SPEECH RECOGNITION SYSTEMS)

ASR sistemleri, konuşmacıya bağımlılık, sözlük boyutu, konuşma biçimi ve konuşma türü açısından çeşitli sınıflarda incelenir [58]. Şekil 1'de genel olarak ses işlemedeki süreç ve ASR sistemlerinin sınıflandırılması verilmiştir.



Şekil 1. *Ses işlemindeki süreç ve ASR sistemlerinin sınıflandırılması*

ASR sisteminin görevi konuşma sinyalini birtakım algoritma ve modelleri kullanarak işleyip metine dönüştürmektir. Şekil 2'de gösterildiği gibi öncelikle konuşma tanıma işlemi için metin üreteci aracılığıyla kaynak kelime dizisi hazırlanır. Kaynak kelime dizisi sisteme giriş olarak verilen konuşma bilgisinden üretilmektedir. Konuşma sinyalinin çıkarılması için kaynak dizisi hazırlanır. Farklı format veya kod yapısındaki konuşma bilgisi ilk giriş aşamasında istenilen diziye dönüştürülmektedir. Kaynak kelime dizisi konuşmanın ses dalga formunu ve konuşma sinyali bileşenini üretmek için bir iletişim kanalından geçirilir. Son olarak konuşma kod çözücü sayesinde akustik sinyalin, orijinal kelime dizisine yakın olan en ideal kelime dizisi haline getirilmesi sağlanmaktadır [59].

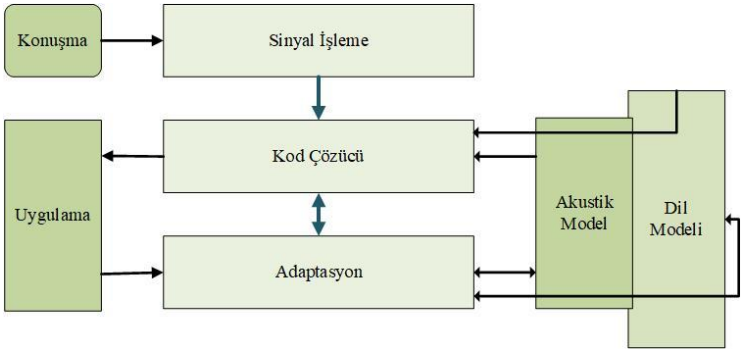


Şekil 2. *Konuşma tanıma sistemi için kaynak-kanal modeli*

Klasik bir konuşma tanıma sisteminin mimarisinde üç önemli bileşen bulunmaktadır (Şekil 3). Sistemin giriş noktası olan ilk aşamada, örüntü tanıma (pattern recognition) açısından önemli olan öznitelik vektörleri ön-uç bileşeni tarafından elde edilmektedir [60]. Daha sonra, deşifre (decoder) modülü akustik

model, dil modeli ve okunuş sözlüğü kullanarak elde edilen öznitelik vektörlerini fonem dizilerine dönüştürmektedir. Bu bileşenlerin her biri farklı model ve eğitim verisi ile eğitilip bütün bunların birlikte çalışması sonucunda girdideki ses sinyali harf veya kelime dizisine aktarmaktadır.

Akustik modeller arasında akustik ortam bilgisi, fonetik, mikrofon ve çevre değişkenliği, konuşmacılar arasındaki cinsiyet ve lehçe farklılıkları hakkında detaylı bilgiler yer almaktadır [61]. Dil modelleri bir sistem dizisinde hangi kelimelerin olası bir kelime dizisi oluşturduğunu, hangi kelimelerin birlikte ortaya çıkacağını vermektedir. Her bir konuşmacının farklı aksanlarının ve konuşma biçimlerinin akustik belirsizliği, dil modelinde temsil edilen sözel ve dil bilgisel karmaşıklık ile birleştirilmektedir.



Şekil 3. *Konuşma tanıma sisteminin temel mimarisi*

ASR sistemlerinde kullanılan akustik modelleme ve dil modelleme işleminin bölünmesi, istatistiksel konuşma tanıma sistemlerinde Denklem 1'deki gibi tanımlanabilir [62].

$$W = \arg\max P(W \mid X) = \arg\max \frac{P(W)P(X|W)}{P(X)} \quad (1)$$

Denklem 1'de, akustik fonem veya özellik vektör dizisi X = X1, X2, … , Xn için konuşma tanıma sisteminde karşılık gelen kelime dizisini vermektedir. Denklem 1'de ifade edilen maksimum arka (posterior) olasılık P(W|X) değerine sahiptir. Denklemin en üst düzeye çıkarılması sabit X gözlemiyle gerçekleştirildiğinden payın maksimize edilmesine doğrudan bağlıdır. P(W) ve P(X|W) bileşenleri dil modellemesi ve akustik modelleme tarafından hesaplanan olasılıksal büyüklükleri oluşturmaktadır. Büyük kelime dağarcığına sahip ASR sistemlerinde çok sayıda kelime olduğu için bir kelimeyi bir kelime dizisine parçalamak gerekir. Bu işleme kısaca okunuş sözlüğü modellemesi denmektedir [63]. P(X|W) fonetik modelleme ile yakından ilişkilidir. P(X|W), konuşmacı varyasyonlarını, telaffuz varyasyonlarını, çevresel varyasyonlar ve içeriğe bağlı fonetik varyasyonları dikkate almaktadır. ASR sistemlerinde X, giriş konuşma sinyalini kelime ile eşleştirmek için en iyi eşleşen kelime dizisini "W" bulmaktadır. Bu noktada gerçekleştirilen kod çözme işlemi, basit bir kalıp tanıma probleminden daha fazlası olarak göze çarpmaktadır. Buradaki zorluk geniş kelime haznesi olan bir dilde ASR işlemini gerçekleştirmek için neredeyse sonsuz sayıda kelime kalıbına bakılmasının gerekliliğidir.

2.1. Konuşma Özelliklerinin Çıkarılması (Feature Extraction of Speech)

Özellik çıkarımı, konuşma tanımanın en önemli adımlarındandır. Çünkü bir konuşmayı farklı diğer konuşmalardan ayırmak için önemli rol oynamaktadır. Konuşma eyleminin doğası gereği her konuşmanın, konuşma bilgisi içerisine entegre edilmiş farklı bireysel özellikler bulunmaktadır [64]. Bu özellikler, ASR sistemleri için önerilen ve başarılı bir şekilde kullanılan farklı özellik çıkarım teknikleriyle elde edilmektedir. Çalışma kapsamında güncel olarak sıklıkla kullanılan özellik çıkarım teknikleri açıklanmıştır.

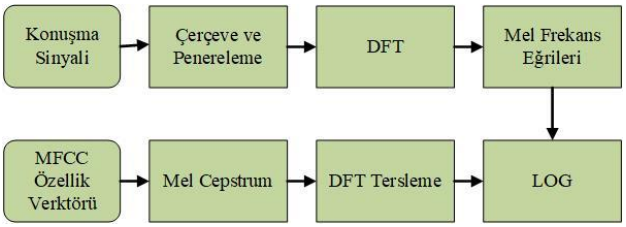
2.1.1. Mel frekanslı cepstral katsayıları (Mel frequency cepstral coefficients)

MFCC, konuşma tanıma sistemlerinde yaygın olarak kullanılan bir özellik çıkarım tekniğidir. Frekans bantları MFCC'de logaritmik olarak konumlandırılmıştır. MFCC hesaplamanın tekniği kısa vadeli analize

dayanmaktadır. Bu nedenle her çerçeveden bir MFCC vektörü hesaplanmaktadır. MFCC, Denklem 2 kullanılarak hesaplanmaktadır [62].

$$\text{Mel(f)} = 2595 * \log10(1 + \frac{\text{f}}{700}) \quad (2)$$

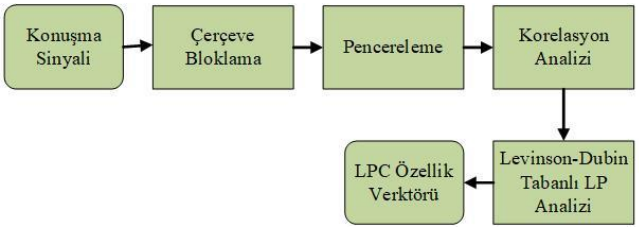
Cepstral katsayıları çıkarmak amacıyla konuşma örneği giriş olarak alınmakta ve sinyalin süreksizliğini en aza indirmek için Hamming penceresi uygulanmaktadır. Bu pencereler daha sonra Mel filtre bankasını oluşturmak için Ayrık Fourier Dönüşümü (DFT: Discrete Fourier Transform) ile birlikte kullanılacaktır. Mel frekans eğrisine göre filtrelerin genişliği değişmekte ve böylece merkez frekansı etrafındaki kritik bant üzerindeki özellikler hesaplanmaktadır. Son olarak ise elde edilen katsayılar ters yönlü Fourier Dönüşüm katsayıların hesaplaması için kullanılmaktadır [65]. MFCC özellik çıkarım işleminde yer alan adımlar Şekil 4'te gösterilmiştir.



Şekil 4. *MFCC özellik çıkarım işlem adımları* [62]

2.1.2. Doğrusal öngörülü kodlama (Linear predictive coding)

En güçlü sinyal analiz tekniklerinden biri olan Doğrusal Öngörülü Kodlama (LPC: Linear Predictive Coding,) konuşmanın temel parametrelerini tahmin etmek için kullanılmaktadır [66]. LPC'nin arkasındaki temel fikir bir konuşma özelliğinin geçmiş konuşma örneklerinin doğrusal bir kombinasyonu olarak tahmin edilebilmesidir. Gerçek konuşma örnekleri ile tahmin edilen değerler arasındaki kareler arası farkların en aza indirilmesiyle, benzersiz bir parametre ya da bağımsız katsayılar elde edilmektedir. Bu katsayılar konuşmanın LPC'si için temel oluşturmaktadır. Analiz işlemi zaman içindeki konuşmanın doğrusal tahmin modelini hesaplama yeteneğini sağlamaktadır [67]. LPC konuşma özellik çıkarma işleminde yer alan adımlar Şekil 5'te gösterilmektedir.



Şekil 5*. LPC özellik çıkarım işlem adımları*

2.2. Akustik Modelleme (Acoustic Modeling)

ASR sistemlerinin doğruluğunu arttırmak için, konuşmacı varyasyonları ve ortam varyasyonları önemlidir. Bir konuşma işleminin akustik modellenmesi, tipik olarak konuşma dalga formundan hesaplanan özellik vektör dizileri için istatistiksel bilgilerin oluşturulması işlemidir. Akustik modelleme, ASR'de gürültü sağlamlığının sağlanmasında, konuşmanın özellik vektörlerini yeniden şekillendirmek için tanıyıcıdan gelen geri bildirim bilgisinin kullanımını da içermektedir [68].

Akustik modellemede ilk olarak belirli zaman sinyali çerçevesindeki foneme ait sonsal olasılık hesaplanmaktadır. Tekrarlanan sinir ağı akustik model olarak kullanılır ise belirli bir zaman aralığında bir fonemin benzerlik olasılığını hesaplamak için sonsal olasılığı basit bir şekilde ön olasılığa bölerek elde edilebilir. Fonem sırası HMM Triphone yapısına birebir eşleştirilebilir. Genel olarak fonemlerin hizalanması normal Gauss dağılımını kullanarak ile elde edilmektedir [69]. HMM'de, durumlar arasındaki

geçişlerde Markov varsayımları yapılıp bir durumdan başka duruma geçişte sadece bir önceki durum göz önünde bulundurulmaktadır. Bu kısıtlama uzun kelime bağımlılığı olan dizilerin modellenmesini zorlaştırmaktadır. Ayrıca, HMM'deki durumların yayım olasılığı da birbirinden bağımsızdır. HMM temelde bir zaman diliminin durumunu bir kez değiştiren sonlu durum makinesi olarak düşünülebilir. Çok sayıda HMM değerlendirme yöntemi, durum geçiş olasılıklarının ve HMM'nin her bir durumundaki yayılma olasılık yoğunluklarının parametrelerinin değerlerini tahmin etmek için geliştirilmiştir.

Akustik modelleme Derin Sinir Ağları (DNN: Deep Neural Network) ile de oluşturulabilmektedir [70]. DNN tabanlı akustik modelde fonemlerin sonsal olasılığı her bir pencere için bağımsızdır. Bu bağımsızlık kelimede bulunan fonemlerin birbirinden bağımsız olması anlamına gelmektedir. DNN tabanlı akustik modelde, ses niteliklerini alarak bu özelliklere bir fonem etiketi tahsis edilmektedir. Temelde her HMM durumunun gözlem olasılığı Gauss karışımları kullanılarak hesaplanmaktadır. Ancak DNN tabanlı akustik model durumunda, her HMM triphone durumunun bu gözlem olasılığı, DNN kullanılarak hesaplanmaktadır.

2.3. Dil Modelleme (Language Modeling)

Dil modelleme, dildeki kelimelerin dizilişini akustik özelliklerinden tamamen bağımsız olarak modellemektedir. Genel olarak n-gram tabanlı modelleme yöntemi kullanmaktadır. Bu modeller de Markov varsayımı bulunup genelde 2 ile 4 arası geçmişteki kelime sırası göz önünde bulundurarak olasılık hesaplamaları gerçekleştirilmektedir [71]. Dolayısıyla, uzun bir cümledeki kelimelerin dizilişini modellemek n-gram'lar ile mümkün olmayıp sadece kısıtlı kelime geçmişi modellenebilmektedir. İleri beslemeli sinir ağını kullanan dil modellerinde Markov varsayımı bulunmadığı için bu modeller ile kelimelerdeki uzun bağımlılıklar modellenebilmektedir [72].

Dil Modeli, bir dildeki kelimelerin ve cümlelerin yapısı ve sırasını modelleyerek o dile ait bir istatistiksel model üretmektedir [73]. En basit ifade ile dil modeli bir kelime diziden sonra hangi kelimelerin gelebileceğini modelleyip deşifre zamanında olası dizilişleri üretmektedir. Gerçekçi bir istatistik elde etmek için olasılık hesaplamaları yeterince büyük metin verileri kullanarak yapılmalıdır. Bu veriler genelde çevrimiçi gazetelerden, elektronik kitaplardan ve dijital metin içeren farklı kaynaklardan elde edilip bazı ön işlemelerden sonra eğitim verisi olarak kullanılmaktadır.

Dil modeli eğitilirken metin verisi içerisinde çok sık geçen kelime dizilerinin olasılık değeri seyrek geçen kelime dizilerine göre daha yüksek çıkıp sistemin çıktısını da bu yönde etkilemektedir. Bu tür sorunları çözmek için dil modeli üzerinde düzleme yöntemleri uygulanmaktadır [74]. Düzlemedeki amaç olasılığı çok yüksek olan n-gram'ların olasılık değerini biraz düşürürken olasılığı düşük olan n-gram'ların ise olasılık değerini artırmaktır. Ayrıca, metin derlemi içerisinde hiç geçmeyen n-gram'lar için mevcut olasılıkları kullanarak bir olasılık hesabı yapılabilmektedir [75].

2.4. Okunuş Sözlüğü (Lexicon)

Sözlük içerisinde konuşma tanıma sisteminin tanıması gereken sözcükler yer alıp sistemin kullanım alanına göre sözlükteki kelime sayısı değiştirilebilmektedir. Büyük sözlüklü ASR sistemlerinde yüzbinlerce kelimenin farklı okunuşları yer almaktadır [76]. Sözlük hazırlama aşaması, HMM tabanlı konuşma tanıma sistemlerinin tasarlanması sürecindeki önemli adımlardan birisidir. Sözlük dışı kelimelerin azaltılması ve gereksiz kelimelerin sözlükte bulundurulmaması arasında bir denge kurulması gerekmektedir.

Okunuş sözlüğündeki tüm kelimelerin okunuşu fonetik semboller ile belirlenip her bir HMM bu sembollerden birisini ifade ederek modellemektedir. Türkçe, yazıldığı gibi okunan bir dil olduğu için sözlük hazırlama süreci daha kolay olup kelimedeki harfler fonetik sembol olarak da kullanılabilmektedir. Aynı kelime farklı şekillerde okunabildiği için sözlükte bulunan her kelimenin çeşitli okunuş şekilleri fonetik olarak belirtilmelidir. Ayrıca, yabancı kelimeler ve kısaltmalar gibi kelimelerin de yazılışı ile okunuşları farklı olabileceği için bütün bu okunuşlar sözlükte bulunmalıdır. Genel bağlam için hazırlanan sözlükler genelde büyük metin derleminden elde edilen tekil kelime listesinden oluşmaktadır. Kelimeler dil bilim uzmanları tarafından incelenip farklı okunuş ve yazılışları ile sözlüğe dâhil edilmektedir. Dâhil edilen

kelimelerde imla kurallarının doğru olması gerekir. Güncel konuları içermesi gerekir. Bu nedenle geniş konuşma haznesinde olması gereken kelimeler sisteme eklenmelidir.

2.5. Deşifre (Transcription)

Bir ASR çalışmasında akustik ve dil modelleme giriş özellik vektörleri sırasına göre en iyi konumlanan bir kelime dizisi sunmaktadır. Bu nedenle eğitilmiş akustik ve dil modelleri ile birlikte bir deşifre işleminin süreci genellikle bir arama işlemi olarak adlandırılabilir [77]. ASR sistemlerinde kullanacak arama algoritmasının karmaşıklığı, dil modellemenin getirdiği kısıtlamalarla belirlenen arama alanıyla yüksek oranda ilişkilidir. Sonlu durum dilbilgileri ve n-gramlar dâhil olmak üzere farklı dil modellerinin etkisi kod çözme verimliliğinde kritik öneme sahiptir. Kod çözücüsü konuşmayı soldan sağa bir süreç olarak işaret eden bilgileri ve zamana bağlı bir süreç tarafından sağlanan verimliliği içermektedir.

Deşifre akustik ve dil modellemeden çıkan ve en uygun arama işlemini gerçekleştiren aynı zamanda büyük arama alanlarını ele alma yeteneğini barındıran bir bileşendir [78]. Deşifre sırasındaki amaç elde edilen X sinyaline ait en olası W kelime sırasını bulmaktır. Ancak gerçek uygulamalarda akustik modelin tek başına kullanılması kelime hata oranını oldukça yükseltmektedir. HMM tabanlı konuşma tanıma sistemlerinin başarısı daha çok dil modeline bağımlı kalmaktadır. Dolayısıyla, oldukça güçlü bir dil modelinin eğitilmesi doğruluk oranı yüksek bir konuşma tanıma sistemi elde etmek için zorunlu hale gelmektedir.

Deşifre işlemlerinde kullanılan Zamansal Bağlantılı sınıflandırma (CTC: Connectionist Temporal Classification) algoritması, hizalama çıktısındaki bir alt sembolü ile x dizisi arasında bağımsızlık varsayımını temel almaktadır [79]. Girdi ile çıktı arasında derin bir yapı olsa bile girdi ile çıktı arasında güçlü bir Markov varsayımı bulunmaktadır. Bir çerçeve üzerindeki tahmin ile komşusundaki çıktı arasında koşullu bağımsızlık mevcuttur. Bu nedenle CTC algoritması çıktısındaki sembollerin öğrenemeyip deşifre aşamasında güçlü bir dil modeline ihtiyaç duymaktadır.

3. KONUŞMA TANIMA ZORLUKLARI VE GELECEK ARAŞTIRMALAR (SPEECH RECOGNITION CHALLENGES AND FUTURE RESEARCHES)

Konuşma tanıma, teknik zorluklarla dolu bir çalışma alanını içermektedir. Yapılan araştırmalarda iki temel zorluk üzerinde durulmaktadır. Bunlardan ilki gürültülü ortamlardaki konuşmaların tanınması diğeri ise doğal serbest stil konuşma tanımadır. Bu zorluklar üzerine çalışmalar sürekli olarak devam etmektedir.

3.1. Akustik Ortamlara Karşı Dayanıklılık (Resistance to the Acoustic Environment)

Güncel konuşma tanıma sistemleri sağlam bir istatistiksel çerçeve üzerine inşa edilmiştir. İstatistiksel çerçeveler doğal konuşma verilerinde meydana gelen değişkenliği temsil etmektedir [80]. Sisteme daha önce verilen örnek konuşma verisinden tahmin edilen parametreler yardımıyla olasılık modelleri hazırlanmaktadır. ASR sistemlerinin altında yatan temel sorun, doğal konuşma sinyalinde var olan değişkenliğin (ortam, gürültü, aksan, konuşmacı, mikrofon vb.) karmaşıklığıdır [81].

Konuşma sinyalinde, dil bilgisine yabancı olan yaygın çeşitlilik, akustik ortamdan kaynaklanmaktadır. Konuşmanın gerçekleştiği akustik ortam ve konuşma sinyalinin ön işlemden önce iletildiği iletişim kanalı, sistem performansının önemli ölçüde bozulmasından sorumludur. Mevcut teknikler, ham gürültü veya doğrusal çarpıklıkların neden olduğu değişkenliği azaltabilmekte ve yavaşça değişen doğrusal kanalları telafi edebilmektedir [82]. Bununla birlikte yankılanma ya da hızlı değişen gürültü gibi daha karmaşık kanal çarpıklıkları ve ayrıca Lombard etkisi gelecekteki araştırmalarda üstesinden gelinmesi gereken önemli bir zorluk olarak görülmektedir.

Konuşmacı özelliklerinin konuşmacı fizyolojisi, konuşmacı tarzı ve aksanları içeren birçok faktör nedeniyle konuşmacılar arasında büyük ölçüde değiştiği bilinmektedir [83]. Konuşma tanıma sistemlerini mikrofon karakteristiğindeki değişikliklere karşı daha güçlü kılmak için kullanılan birincil yöntem model eğitiminde kullanılan verilerin çok sayıda farklı konuşmacılar içermesidir. Ayrıca mevcut ASR sistemleri, bir dilin ana konuşmacılarını ve dilin çeşitli ana dil konuşanlarından gelen çok sayıda konuşma verisini

modelleyen bir okunuş sözlüğü varsaymaktadır. Okunuş sözlüğü ile desteklenmeyen sadece uçtan uca akustik modelleme ile yapılan ASR sistemlerinin başarımı okunuş sözlüğü kullanan sistemlere göre düşük olacaktır.

Konuşma tanıma sistemlerindeki teknik zorluk, akustik ortamlardaki değişiklikler, yankı, dış gürültü kaynakları ve iletişim kanalları da dâhil olmak üzere her türlü değişkenliğe karşı daha güçlü olan ASR sistemlerinin oluşturulmasını ve geliştirilmesini gerektirmektedir. Bu önemli konular için yeni teknikler ve mimariler geliştirilmelidir. Günümüzün ASR sistemlerinde kullanılan akustik modeller, çalışma kapsamında açıklanan akustik değişkenliğin altında yatan nedenlerin çoğunu barındırmaktadır.

Sonuç olarak bir konuşma tanıma sistemine sunulan konuşma, parametre tahmininde kullanılan eksenlerden biri boyunca saptığında, modellerin tahminleri oldukça şüpheli hale gelmektedir. Bu değişkenlik faktörlerine karşı ASR sistemlerinin sağlamlığı bu alanda önemli bir teknik zorluk teşkil etmektedir. Bu zorluğun giderilmesi yalnızca konuşma değişkenliğinin gerçek doğası için açık mekanizmaları akıllıca temsil edebilen yenilikçi mimariler ve teknikler kullanmak ile mümkün değildir. Daha da önemlisi ASR modellerini geçmişte mümkün olmayan çok katmanlı şekillerde eğitmek ve uyarlamak için mevcut olan ve sürekli artan verileri ASR sistemlerinin kullanacağı şekle getirilmesi ile teknik zorlukları aşabilmek mümkün olacaktır.

3.2. ASR'de Kendi Kendine Öğrenme (Self-Learning in ASR)

ASR için konuşma-metin eşleştirmesi yapılmış eğitim verileri ve telaffuz edilen kelimeler, gerçek kullanıcılar tarafından sağlanmaktadır [84]. Bu verilerden elde edilen bilgilerden istatistiksel modeller üretilmektedir. Bu yaklaşım ile geliştirilen ASR sistemlerinin başarımı giderek düşmektedir. Başarımı arttırmak için ise tekrar gerçek kullanıcıların müdahalesi gerekmektedir. Konuşmanın doğal dengesi göz önüne alındığında bu sürecin sonsuza kadar devam etmesi muhtemeldir. Buradaki zorluk, konuşma tanımayı insanın kendi kendine öğrenme kabiliyetinin en azından temel bir biçimiyle donatacak olan kendi kendini uyarlama veya kendi kendine öğrenme teknikleri oluşturmaktır.

Değişen ortamlar, konuşulmayan ifadeler, farklı konuşmacılar, farklı telaffuzlar, lehçeler, aksanlar, kelimeler, anlamlar ve konularla başa çıkabilmek için konuşma ve dil işlemenin her seviyesinde öğrenmeye ihtiyaç duyulmaktadır. Bu alandaki araştırmalar hem yeni modellerin öğrenilmesini hem de bu modellerin önceden var olan bilgi kaynaklarına entegrasyonunu ele almaktadır [85]. Örneğin, bir ASR sisteminin giriş konuşmasında yeni bir özel isimle karşılaşabilir ve özel ismin yazımını bulabilmek için eş zamanlı metni uygun içerikle incelemeye ihtiyaç duyulabilir. Etiketlenmemiş veya kısmen etiketlenmiş verilerin kullanılması bu tür bir öğrenme için gerekli olacaktır. Tanınamayan kelimelerin işaretlenmesi otomatik veya el yordamı ile yapılabilir. Burada kendine özgü öğrenme yaklaşımları geliştirilebilir. Ayrıca genelleme yapılabilir. Genelleme yaklaşımı konuşma tanıma sistemlerinde hızı arttıran farklı yaklaşımlardan biri olarak görülmektedir.

3.3. Bilinmeyen Kelimelerin Tespiti ve Yetersiz Kaynak Durumu (Detection of Unknown Words and Insufficient Resource Status)

ASR sistemleri zengin kelime haznesi ve sondan eklemeli yapıya sahip dillerde üretilen kelimeleri tahmin etmekte zorluk çekmektedir. Bu durum kelime haznesi, yabancı veya eğitim kümesinde yer almayan kelimeleri içeren konuşmalarda başarımı düşürmektedir. Bilinmeyen kelimelerin tespiti ASR sisteminin kelime haznesini ve okunuş sözlüğünü oluşturmak için yetersiz kaynak bulunan dillerde büyük bir sorundur. Bu konudaki temel sorun yüksek olasılık değerine sahip kelime terimlerinin diğer ortak ve benzer kelimeleri yanlış tanımlamasıdır [86]. Çözümden istenilen amaç ise bir kelime ASR sistemine daha önce eğitim verisi olarak verilmediyse bu kelimeyi güvenilir bir şekilde tahmin eden sistemler oluşturmaktır. Bu nedenle dil modelleme ve okunuş sözlüğü büyük önem arz etmektedir. Ayrıca bu tür durumların tespiti için sistemin kelime hipotezinin güvenilmez olduğu varsayılmalı ve hata düzeltme planlarının tasarlandığından emin olunması gerekmektedir.

3.4. ASR için Hesaplamalı Mimariler (Computational Architectures for ASR)

Bilgisayar donanım alt yapıları gelişmekte ve veri gereksinimi olan sistemler için hesaplama veya depolama ihtiyaçları giderilmektedir [87]. ASR sistemleri üzerinde yapılan çalışmalar daha büyük eğitim veri setlerinin kullanılması gerektiğini belirtmektedir. Büyük veri setlerinin işlenmesi hesaplama yoğunluğu arttırmaktadır. Mevcut mikroişlemcilerdeki güç ve hesaplama seviyelerinin düşüklüğü nedeniyle büyük veri setleri mikroişlemciler üzerinde işletilememektedir. Mikroişlemciler üzerinde veri setlerini işletebilmek için hesaplama kümelerinin azaltılması gibi farklı yaklaşımlar sergilense de bu işlemler uzun zaman aldığından tercih edilmemektedir. Bu durumda paralel işletim yeteneğine sahip sistemler üzerinde çalışmak performansı arttıracaktır. Çoğu zaman konuşma sistemleri için algoritma tasarımcıları bu paralellik araştırmasını göz ardı etmektedirler. Odaklanılan tek sorunu başarımı arttırmak olarak nitelendirmişlerdir. Gelecekteki araştırma önerileri ve bu çalışmada tartışılan yaklaşımlar daha fazla hesaplama birimi gerektirecektir. Sonuç olarak ASR ile ilgili araştırmacılar, tasarımlarında paralellik yaklaşımını düşünmek zorundadır. Ayrıca mikroişlemci tabanlı yaklaşımlar yerine grafik işlemciler üzerinde hesaplama işlemlerini yapmalıdırlar.

3.5. Türkçe ASR'nin Geniş ve Sınırlı Dağarcık Düzeyindeki Başarısı (The Success of Turkish ASR at Wide and Limited Vocabulary)

ASR sistemlerinde dağarcık sistemin tanıyabileceği kelimeleri ifade etmektedir. ASR sisteminin tanıyamadığı kelimeler dağarcık dışı olarak kabul edilir. Her bir dağarcık dışı kelime konuşma tanımada ortalama 1.5 hata oranına yol açmaktadır [88]. Literatürdeki çalışmalar sınırlı dağarcığa sahip ASR sistemlerinin daha doğru sonuçlar verdiğini göstermektedir [89]. Yer isimleri veya rakamlardan oluşan sınırlı dağarcığa sahip konuşmalarda kelime hata oranları %1, yer isimlerinden oluşan cümlelerde ise kelime hata oranı yaklaşık %4 civarındadır. Geniş dağarcığa sahip konuşmalar ile yapılan deneylerde ise kelime hata oranları çok daha yüksek çıkmaktadır. Aksoylar ve arkadaşlarının gerçekleştirdiği çalışmada geniş dağarcığa sahip konuşma tanıma sistemi spor haberleri ile test edilmiş ve %46'lık bir kelime hata oranının tespit edildiği gösterilmiştir [89]. Parlak ve arkadaşlarının gerçekleştirdiği çalışmada ise eğitim ve test işlemlerinde aynı veri kümesinin parçaları kullanılmıştır [90]. Bu kullanım şeklinde kelime hata oranı %26.9 olarak belirlenmiştir. Eğitim işleminde 184 saatlik test işleminde ise 3 saatlik bir veri kullanılmıştır. Bu çalışma Türkçe için gerçek anlamda bir geniş kelime dağarcığına sahip değildir.

Akın ve arkadaşlarının gerçekleştirdiği çalışmada önceki çalışmalardan farklı olarak kelime altı birimler bulunurken kelimeler öncelikle bir biçimbirimsel çözümleyiciden geçirilmiştir [91]. Türüne göre sınıflara ayrılıp kelimeler daha sonra biçimbirimsel tabanlı yâda istatistiksel yöntem kullanılarak kelime altı birimler elde edilmiştir. Yaklaşık 60 saatlik bir konuşma verisi kullanılarak akustik model ve yaklaşık 614 milyon kelimeden oluşan bir dil modeli hazırlanarak bir ASR sistemi geliştirilmiştir. Toplam %6,2 dağarcık dışı kelimenin olduğu bir test verisinde %24,16'lık bir kelime hata oranı veren bir test işlemi gerçekleştirilmiştir. Farklı deneylerde dağarcık dışı kelimelerin azalması ile sistemin başarımının arttığı gözlemlenmiştir. Ancak bir ASR sisteminden istenilen hangi dil üzerine geliştiriliyor ise o dildeki bütün kelimeleri rahatlıkla tanıyor olmasıdır. Ancak bu durum sondan eklemeli bir yapıya sahip diller için oldukça zorlu bir görevdir. Bu nedenle geniş kelime dağarcığına sahip veri setlerinin hazırlanması gerekmektedir.

3.6. Türkçe ASR'nin Geleceği (The Future of Turkish ASR)

Yapılan araştırmalar sonucu Markov sürecindeki geçiş oranlarının Türkçe konuşma tanıma sistemlerinin performansını ve davranışını belirlemek için kolaylıkla kullanılabileceği sonucuna varılmıştır. Ani ortam değişikliklerinin meydana getireceği güç değişkenlikleri, ağ kontrol sistemleri ve üretim sistemlerinde bulunan bazı pratik yöntemleri modellemek için en uygun yaklaşımlar Markov modellerinde mevcuttur [92]. Markov zincirleri, gürültü modellemesi ve konuşma sinyalinin tahmini için etkili bir şekilde uygulanmaktadır.

Bu çalışmada ele alınan ASR sistemin bileşenleri (akustik modelleme, dil modelleme ve deşifre) modern konuşma tanıma sistemlerinde yer almaktadır. Türkçe ASR için bu bileşenlerin her birinin ve okunuş sözlüğünün önemli katkısı olacaktır. Çok sayıda yeni kelimenin tek bir kökten meydana gelmesi Türkçenin morfolojik karakteristiğidir. Türkçe dilinin sondan eklemeli yapısı, kelime dağarcığını genişletmektedir.

Kelime sırasının düzensizliği, eğitim için gerekli olan veri kümesinin azlığı Türkçe ASR sistemlerinin geliştirilmesinin önünde önemli bir sorundur. Veri kümesinin azlığı ve kelime dağarcığının geniş olması dil modelinin olasılık tahminlerinde başarısız olmasına sebep olmaktadır. Bu durumda Türkçe için öncelikle geniş kelime haznesine sahip eğitim veri setlerinin hazırlanması ve ardından istatistiksel modeller üzerine çalışmaların gerçekleştirilmesi gerekmektedir. Böylelikle Türkçe dil yapısı modellenebilecek ve başarımı daha yüksek ASR sistemleri geliştirilebilecektir.

Gelecek çalışmalarda özellikle mobil iletişim ve çok kullanıcı arabirime sahip yayın organları üzerinden elde edilen verilerin Türkçe ASR sistemlerinde kullanılması ile daha güçlü Türkçe ASR sistemlerinin geliştirileceği tahmin edilmektedir. Dolayısıyla gelecek yıllarda Türkçe ASR sistemlerinin başarımını arttırmaya ve hesaplama zamanının azaltılmasına yönelik birçok çalışma yapılacağı ön görülmektedir.

4. SONUÇ VE ÖNERİLER (CONCLUSION AND SUGGESTIONS)

Bu çalışmanın amacı, mevcut konuşma tanıma yöntemlerini ve yaklaşımlarını inceleyerek bu alandaki gelişmeleri detaylı olarak sunmaktır. Bu nedenle çalışma kapsamında araştırmacıların bu alanda yaptıkları çalışmalarda kullandıkları ölçütler, yaklaşımlar, veri setleri ve bu alanda karşılaştıkları zorluklar ele alınmıştır. ASR çalışmaları incelenirken ASR'nin uygulama alanı, kullanılan materyal ve metotlar dikkate alınmıştır. Ayrıca başarım ölçütlerinde kullanılan teknikler açıklanmıştır. Araştırmacıların hangi dil üzerine çalıştığı ve bu dil üzerindeki zorluklar belirtilmiştir. Türkçe üzerine geliştirilen çalışmalarda karşılaşılan zorluklar verilmiştir. Araştırmalar sonucu elde edilen bilgiler doğrultusunda; akustik ortamlara karşı dayanıklılık, ASR'de kendi kendine öğrenme, bilinmeyen kelimelerin tespiti, Türkçe ASR'nin geniş ve sınırlı dağarcık düzeyindeki başarısı, yetersiz kaynak durumu ve ASR üzerine uygulanabilecek hesaplamalı mimariler üzerine değerlendirmelere yer verilmiştir. Ayrıca Türkçe ASR'nin geleceği tartışılmış ve Türkçe ASR için mevcut zorlukların üstesinden gelebilmek amacıyla önerilerde bulunulmuştur. Derin öğrenme ve klasik yaklaşımların ASR üzerine etkisi araştırılmıştır. Araştırmacıların bu alandaki çalışmaları için gerekli olan detaylı bilgiyi kendi çalışmalarına dâhil edebilmesi ve mevcut zorlukların üstesinden gelebilmesi için bir yol haritası sunulmuştur.

TEŞEKKÜR (ACKNOWLEDGMENTS)

Bu çalışma, EMFA Yazılım Danışmanlık A.Ş. tarafından desteklenmiştir. Desteklerinden dolayı EMFA Yazılım Danışmanlık A.Ş. yönetim kurulu başkanı Emre EVREN'e teşekkürlerimizi sunarız.

KAYNAKLAR (REFERENCES)

[1] Moore, R. K. (2007). Presence: A human-inspired architecture for speech-based human-machine interaction. IEEE Transaction Computer, 56(9), 1176-1188.

[2] Ghorbel, M., Haariz, M., Grandjean, B., & Mokhtari, M. (2005). Toward a generic human machine interface for assistive robots: The amor project. International Conference Rehabilitation Robotics, 68-172.

[3] Abushariah, M., Gunawan, T. S., Khalifa, O. O., & Abushariah, M.A.M. (2010). English digits speech recognition system based on hidden markov models. International Conference Computer Communication Engineering, 1-5.

[4] Kurian, C., & Balakrishnan, K., (2009). Speech recognition of Malayalam numbers. World Congress National Biology Inspired Computer NABIC, 1475-1479.

[5] Paraiso, E. C., & Barthès, J. P. A. (2006). An Intelligent speech interface for personal assistants in R&D projects. Expert System Application, 31(4), 673-683.

[6] Busso, C. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal

information. International Conference Multimodal Interfaces, 205-212.

[7] Myakotnykh, E. S., & Thompson, R. A. (2009). Adaptive Speech quality management in voice-over-ip communications. Advanced International Conference Telecommunication, 64-71.

[8] Xiao, Q. (2007). Biometrics-technology, application, challenge, and computational intelligence solutions. IEEE Computer Intelligence Magazine, 2(2), 5-10.

[9] Prakoso, H., Ferdiana, R., & Hartanto, R. (2017). Indonesian automatic speech recognition system using CMUSphinx toolkit and limited dataset. International Symposium Electronic Smart Devices, 283-286.

[10] Miao, Y. (2014). Kaldi+PDNN: Building DNN-based ASR Systems with Kaldi and PDNN. IEEE Computer Intelligence Magazine, 14(6), 1-4.

[11] Greibach, S. (2010). A mixed trigrams approach for context-sensitive spell checking. Lecture Notes in Computer Science, 939-953.

[12] Salor, O., Pellom, B., Ciloglu, T., Hacioglu, K., & Demirekler, M. (2002). On developing new text and audio corpora and speech recognition tools for the Turkish language, International Conference Spoken Language Processing, 349-352.

[13] Salor, O., Ciloglu, T., Hacioglu, K., & Demirekler, M. (2002). On developing new text and audio corpora and speech recognition tools for the Turkish language. International Conference Spoken Language Processing, 367-372.

[14] Salor, Ö., Pellom, B., Ciloglu, T., & Demirekler, M. (2007). Turkish speech corpora and recognition tools developed by porting SONIC: Towards multilingual speech recognition. Computer Speech Language, 21(4), 580-593.

[15] Arisoy, E., Can, D., Parlak, S., Sak, H., & Saraclar., M. (2009). Turkish broadcast news speech and transcripts. IEEE Transactions on Audio Speech and Language Processing 17(5), 874 - 883.

[16] Coltekin, C. (2010). A freely available morphological analyzer for Turkish. International Conference Language Resource Evalution, 820-827.

[17] Jeanmonod, D., Rebecca, J., & Suzuki, K. (2018). Control of a proportional hydraulic system, Intech Open, 2(1), 64-72.

[18] Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., & Mercer, R. L. (1990). {Class-Based} {N-Gram} Models of natural language. Computer Linguistic, 18(1950), 14-18.

[19] Burrell, A. T., & Papantoni-Kazakos, P. (2012). Stochastic binary neural networks for qualitatively robust predictive model mapping. Communication Network System Science, 5(9), 603-608.

[20] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 5(4), 115-133.

[21] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization. Psychology Revulation, 65(6), 386-408.

[22] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986) Learning represenations by back propogation errors. Natural Product Letters, 5(3), 533-536.

[23] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. IEEE Transaction Acoustic, 37(3), 328-339.

[24] Hinton, G., & Shikano, I. C. (1988). Phoneme recognition: neural networks vs hidden markov models. ATR Interpreting Telephony Research Laboratories, 07-110.

[25] Hochreiter, S., & Schmidhuber, U. J. (1997). Lstm. Neural Computing, 9(8), 1735-1780.

[26] Arisoy, E., & Saraclar, M. (2016). Compositional neural network language models for agglutinative languages. Annual Conference International Speech Communication Association, 3494-3498.

[27] Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. Advances in Neural Information Processing Systems, 19(1), 153-162.

[28] Morandi, A. (2012). A fast learning algorithm for deep belief nets geoffrey. Monthly Notices of the Royal Astronomical Society, 425(3), 2069-2082.

[29] Raina, R., Madhavan, A., & Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. International Conference on Machine Learning, 873-880.

[30] Kımanuka, U. A., & Buyuk, O. (2018). Turkish speech recognition based on deep neural networks .Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 22(Özel), 310-319.

[31] Fohr, D. (2017). New paradigm in speech recognition. Deep Neural Networks, Hal Id: hal-01484447.

[32] Arisoy, E., & Arslan, L. M. (2005). Turkish dictation system for broadcast news applications. European Signal Processing Conference, 1351-1354.

[33] Rosdi, F. (2017). Assessing automatic speech recognition in measuring speech intelligibility : a study of malay speakers with speech impairments. International Conference on Electrical Engineering and Informatics, 1-6.

[34] Baniardalan, F., & Akbari, A. (2017). A weighted denoising auto-encoder applied to mel sub-bands for robust speech recognition. Iranian Conference on Intelligent Systems and Signal Processing, 38-42.

[35] Tong, R., Wang, L., & Ma, B. (2017) Transfer learning for children's speech recognition. International Conference on Asian Language Processing, 36-39.

[36] Guglani, A. N., & Mishra, J. (2018). Continuous Punjabi Speech recognition model based on kaldi ASR toolkit. International Journal of Speech Technology, 5:(6), 1-6.

[37] Song, X., Zonu, Y., & Chen, S.(2017). Investigating multi-task learning for automatic speech recognition with code-switching between Mandarin and English. International Conference on Asian Language Processing, 27-30.

[38] Lunuwilage, K., Abeysekara, S., Witharama, L., & Mendis, S. (2017). Web based programming tool with speech recognition for visually impaired users. International Conference on Software, Knowledge, Information Management and Applications, 1-6.

[39] Battenberg, E. (2017). Exploring neural transducers for end-to-end speech recognition. Computation and Language, 206-213.

[40] Sirikongtham, P. (2017). Improving speech recognition using dynamic multi - pipeline API. International Conference on ICT and Knowledge Engineering, 234-240.

[41] Thimmaraja, Y. G., & Jayanna, H. S. (2017). Creating language and acoustic models using kaldi to build an automatic speech recognition system for Kannada language. IEEE International Conference Recent Trends Electronic Information Communication Technology, 161-165.

[42] Sasithon, P. S. (2017). Isarn digit speech recognition using HMM. International Conference Informatic Technology, 1-5.

[43] Ho, H., Mawardi, V. C., & Dharmawan, A. B. (2017). Question answering system with hidden markov model speech recognition. International Conference on Science in Information Technology, 257-262.

[44] Syadida, A. Q., Ignatius, D. R., Setiadi, M., & Setyono, A. (2017). Sphinx4 for indonesian

continuous speech recognition system. International Seminar on Application for Technology of Information and Communication. 264-267.

[45] Gali, J., Šumarac, D., Jovi, S. T., & Markovi, B. (2017). Prepoznavanje bimodalnog govora bazirano na metodi potpornih vektora. Telecommunications Forum, 73-76.

[46] Shahnawazuddin, S., Deepak, K. T., Pradhan, G., & Sinha, R. (2017). Enhancing noise and pitch robustness of children's ASR. India Department of Electronics and Communication Engineering, 5225-5229.

[47] Ephrat, A. (2018). Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM Transaction Graphics, 37(4), 1-11.

[48] Cloud Comparison RightScale, https://resources.flexera.com/web/www/cloud-comparison-tool/index.html.

[49] Bumbalek, Z., Zelenka, J., & Kencl, L. (2012). Cloud-based assistive speech-transcription services, Springer Berlin, 113-116.

[50] Yu, B., Li, H., & Fang, C. (2012). Speech emotion recognition based on optimized support vector machine, Journal of Software, 7(12), 2726-2733.

[51] Sivanagaraja, T. Ho., M. K. M. K., Khong, M. K., & Wang, Y. (2017). End-to-end speech emotion recognition using multi-scale convolution networks. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 1-4.

[52] Lefter, I., & Jonker, C. M. (2017). Aggression recognition using overlapping speech. Seventh International Conference on Affective Computing and Intelligent Interaction, 299-304.

[53] Engineering, E., Mara, U. T., & Pauh, P. (2017). Automatic gender recognition using linear prediction coefficients and artificial neural network on speech signal, IEEE International Conference on Control System, Computing and Engineering, 24-26.

[54] Wai K. L., & Fung, P. (1999). Fast accent identification and accented speech recognition. IEEE International Conference on Acoustics, Speech, and Signal Processing, 221-224.

[55] Fung, P.(1999). Fast accent recognition. IEEE International Conference on Acoustics, Speech, and Signal Processing, 225-229.

[56] Sagha, H., & Deng, J. (2017). The effect of personality trait, age, and gender on the performance of automatic speech valence recognition. Seventh International Conference on Affective Computing and Intelligent Interaction, 86-91.

[57] Lin, S., Tsunakawa, T., Nishida, M., & Nishimura, M. (2017). DNN-based feature transformation for speech recognition using throat microphone. Asia-Pacific Signal Information Processing Association, 1-4.

[58] Shafran, I., Rose, R., Park, F. (2003). Robust speech detectıon and segmentatıon for real-time ASR applications. Izhak Shafran & Richard Rose Labs Research, 432-435.

[59] Ochiai, T., Watanabe, S., & Katagiri, S. (2017). Does speech enhancement work with end-to-end ASR objectives?: experimental analysis of multichannel end-to-end ASR. International Workshop on Machine Learning for Signal Processing, 1-5.

[60] Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transaction Acoustic, 28(4), 357-366.

[61] Hinton, G. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82-97.

[62] Huang, X., & Deng, L. (2010). An overview of modern speech recognition, Handbook Natural

Language Processing, 339-367.

[63] Huang, C., Chang, E., Zhou, J., & Lee, K. (2000). Accent modeling based on pronunciation dictionary adaptation for large vocabulary mandarin speech recognition, Interspeech, 818-821.

[64] Narang, S., & Gupta, M. (2015). International journal of computer science and mobile computing speech feature extraction techniques: a review, Journal Computer Science Mobil Computing, 4(3), 107-114.

[65] Winursito, A., Hidayat, R., & Bejo, A. (2018). Improvement of MFCC feature extraction accuracy using pca in Indonesian speech recognition, International Conference on Information and Communication Technology, 379-383.

[66] Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC, International Journal for Advance Research in Engineering and Technology, 1(6), 1-5.

[67] Gupta, H., & Gupta, D. (2016). LPC and LPCC method of feature extraction in speech recognition system. International Conference Cloud System Big Data Engineering, 498-502.

[68] Haridas, A. V., Marimuthu, R., & Sivakumar, V. G. (2018). A Critical Review and Analysis On Techniques Of Speech Recognition: The Road Ahead, International Journal of Knowledge-Based and Intelligent Engineering Systems, 39-57.

[69] Hinton, G. (2012). Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine, 29(6), 82-97.

[70] Li, L. (2013). Hybrid deep neural network - hidden markov model (DNN-HMM) based speech emotion recognition. International Conference on Affective Computing & Intelligent Interaction, 312-317.

[71] Saul, L., & Pereira, F. (1997). Aggregate and mixed-order markov models for statistical language processing. Second Conference on Empirical Methods in Natural Language Processing, 81-89.

[72] Karafi, M., & Cernock, J. H. (2010). Recurrent neural network based language model. Interspeech, 1045-1048.

[73] Elmaghraby, A.S. (1989). Voice recognition applications for programming environments, IEEE Energy and Information Technologies in the Southeast, 655-659.

[74] Chen, S. F., & Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. Annual Meeting of the Association for Computational Linguistics, 310-318.

[75] Stolcke, A. (2000). Entropy-based pruning of backoff language models. Proceedings DARPA Broadcast News Transcription and Understanding Workshop, 270-274.

[76] Bocchieri, E., & Caseiro, D. (2010). Use of geographical meta-data in asr language and acoustic models, International Conference on Acoustics, Speech, and Signal Processing, 5118-5121.

[77] Aubert, X. L. (2012). An overview of decoding techniques for large vocabulary continuous speech recognition. Computer Speech Language, 16(1), 89-114.

[78] Hoffmeister, B., Heigold, G., Rybach, D., Schluter, R., & Ney, H. (2012). WFST enabled solutions to ASR problems: beyond HMM decoding. IEEE Transactions on Audio, Speech, and Language Processing, 20(2), 551-564.

[79] Kang, S. S. (2008). Regulation of early steps of chondrogenesis in the developing limb. Animal Cells System, 12(1), 1-9.

[80] Frederick, J. Google Kitaplar, 2019. "Statistical methods for speech recognition" https://books.google.com.tr/books.

[81] Gemmeke, J. F., Hurmalainen, A., Virtanen, T. & Sun, Y. (2011). Toward a practical

implementation of exemplar-based noise robust ASR. European Association for Signal Processing, 1490-1494.

[82] Macho, D. (2007). Narrowband to Wideband feature expansion for robust multilingual ASR. Interspeech, 1118-1121.

[83] Zheng, Y. (2005). Accent detection and speech recognition for shanghai-accented mandarin. Interspeech, 7-10.

[84] Das, B., Mandal, S., & Mitra, P. (2011). Bengali speech corpus for continuous automatic speech recognition system. International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques, 51-55.

[85] Herbig, T., Gerl, F., & Minker, W., (2012). Self-learning speaker ıdentification for enhanced speech recognition. Computer Speech Language, 26(3), 210-227.

[86] Ogawa, A., Hori, T. (2017). Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks. Speech Communication, 89(1), 70-83.

[87] Yu, D., & Deng, L. (2015). Automatic Speech Recognition, Springer London, 124-129.

[88] Hetherington, I. L. (1995). A characterization of the problem of new, out-of-vocabulary words in continuous speech recognition and understanding. Doctoral Dissertation, Massachusetts Institute of Technology Cambridge.

[89] Aksoylar, C., Mutluergil, S. O., & Erdogan, H. (2009). The anatomy of a Turkish speech recognition system. Signal Processing and Communications Applications Conference, 512-515.

[90] Parlak, S., & Saraçlar, M. (2012). Performance analysis and Improvement of Turkish broadcast news retrieval. Transaction Audio, Speech Language Processing, 20(3), 731-741.

[91] Akin, A. A., Demir, C., & Dogan, M. U. (2012). Improving sub-word language modeling for Turkish speech recognition. Signal Processing and communications Applications Conference, 1-4.

[92] Tombaloğlu, B., & Erdem, H. (2016). Development of a MFCC-SVM based Turkish speech recognition system, Signal Processing and communications Applications Conference, 1-4.