FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİLİĞİ ANABİLİM DALI

BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

GERÇEK ZAMANLI YÜKSEK KALİTEDE SES TANIMA

YÜKSEK LİSANS TEZİ

Mert Yılmaz ÇAKIR

Danışmanı: Yrd. Doç. Dr. Yahya ŞİRİN

İSTANBUL
Aralık 2017

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİLİĞİ ANA BİLİM DALI

BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

GERÇEK ZAMANLI YÜKSEK KALİTEDE SES TANIMA

YÜKSEK LİSANS TEZİ

Mert Yılmaz ÇAKIR

Danışmanı: Yrd. Doç. Dr. Yahya ŞİRİN

İSTANBUL
Aralık 2017

ONAY SAYFASI

Mert Yılmaz ÇAKIR tarafından hazırlanan "Gerçek Zamanlı Yüksek Kalitede Ses Tanıma" adlı çalışma aşağıdaki jüri üyeleri tarafından BİLGİSAYAR MÜHENDİLİĞİ ANA BİLİM DALI "BİLGİSAYAR MÜHENDİSLİĞİ" Programında "YÜKSEK LİSANS TEZİ" olarak kabul edip onaylanmıştır.

Başkan / Üye

Doç. Dr. Fatih KOÇAN

Danışman

Yrd. Doç. Dr. Yahya ŞİRİN

Üye

Yrd. Doç. Dr. Mehtap YALÇINKAYA

BEYAN

Bu çalışma İstanbul Sabahattin Zaim Üniversitesi Fen bilimleri Enstitüsü BİLGİSAYAR MÜHENDİLİĞİ ANA BİLİM DALI BİLGİSAYAR MÜHENDİSLİĞİ'ndeki öğrenciliğim döneminde hazırlanmış olan YÜKSEK LİSANS TEZİ tarafımdan yapılmış ve kaleme alınmış tamamen özgün bir çalışma olup bu çalışmamın başından sonuna kadar bilimsel ahlak kurallarına uydum. Bu çalışmam süresince elde etmediğim ve tezimde/raporumda kullanmış olduğum bütün bilgiler ve yorumlar için atıf yaptığımı ve kaynak gösterdiğimi, patent ve telif haklarını ihlal edici bir davranışta bulunmadığımı beyan ederim.

İmza

Mert Yılmaz ÇAKIR

TEŞEKKÜR

Bu tez çalışmasında öncelikle beni yetiştiren, hayatım boyunca benden desteklerini esirgemeyen, her zaman yanımda olan anneme ve babama, değerli önerileri ve her türlü yardımlarıyla beni yönlendiren kıymetli danışmanım Yrd. Doç. Dr. Yahya ŞİRİN'e, Bilgisayar Mühendisliği öğretim üyesi değerli hocam Doç. Dr. Fatih KOÇAN'a, değerli vakitlerini ayıran Yrd. Doç. Dr. Mehtap YALÇINKAYA'ya, benden tecrübelerini eksik etmeyen İbrahim GÜMÜŞ'e, bölüm arkadaşım Mehmet Ali KUTLUGÜN'e, moral ve desteklerinden dolayı kardeşlerime, yakınlarıma ve sağladığı burs için İstanbul Sabahattin Zaim Üniversite'sine gönülden teşekkür ederim.

İÇİNDEKİLER

KISALTMALAR

ANN :Yapay Sinir Ağları (Artifical Neural Network)

ASR **:**Otomatik Konuşma Tanıma (Automatic Speech Recognition)

CMS **:**Kepstral Ortalama Çıkarma (Cepstral Mean Subtraction)

DFT :Kesikli Fourier Dönüşümü (Discrete Fourier Transform)

DTW :Dinamik Zaman Bükmesi (Dynamic Time Warping)

FFT **:**Hızlı Fourier Dönüşümü (Fast Fourier Transform)

FNS :Bulanık Sinirsel Sistemler (Fuzzy Neural Systems)

GMM **:**Gauss Karma Modeli (Gaussian Mixture Model)

HMM **:**Saklı Markov Modeli (Hidden Markov Model)

IDFT **:**Kesikli Fourier Dönüşümü Tersi (Inverse Discrete Fourier Transfrom)

LPC **:**Doğrusal Öngörülü Kodlama (Linear Predictive Coding)

MFCC **:**Mel Frekanslı Kepstral Katsayı (Mel Frequency Cepstral Coefficient)

PCM :Darbe Kod Modülasyonu (Pulse-Code Modulation)

PLP :Algısal Doğrusal Tahmin (Perceptual Linear Prediction)

SVM :Destek Vektör Makineleri (Support Vector Machines)

VQ :Vektör Nicemleme (Vector Quantization)

WAV :Dalgaşekli Ses Dosyası Formatı (Waveform Audio File Format)

TABLO LİSTESİ

ŞEKİL LİSTESİ

SEMBOL LİSTESİ

Bu tezde kullanılan semboller açıklamalarıyla birlikte aşağıda ifade edilmiştir.

Semboller	**Açıklama**
α	İleri-Yön değişkeni
β	Geri-Yön değişkeni
δ	Viterbi değişkeni
ξ	Baum-Welch değişkeni

ÖZET

Gelişen teknolojiyle birlikte insan-bilgisayar etkileşiminde birçok arayüz (etkileşim kurma şekilleri) oluşmuştur. Bu arayüzlerden biri de konuşma tanımadır. Konuşma tanıma, insan sesini aracılar olmadan bilgisayar tarafından okunabilecek bir forma çevirir. Böylelikle konuşma ile cihazları yönetme imkânı sağlanır. Sağladığı kolaylıkların kullanılma şekillerine göre değiştiği konuşma tanıma teknolojisi birçok uygulama alanına sahiptir. Bu alanlardan birisi olan konuşmanın yazıya çevrilmesi işlemi, geçmişten günümüze birçok çalışmaya konu olmuştur. Geleneksel çalışmalarda, belirli kişilerin konuşmalarının yazıya çevrilmesi hedeflenmiştir. Bu amaçlı uygulamalar konuşmacı bağımlı sistemlerdir. Fakat konuşmacı bağımlı sistemler, farklı konuşmaları, sisteme tanımlamadan başarılı olamamaktadır. Günümüzde ise akıllı cihazlar başta olmak üzere geliştirilen çoğu sistemler konuşmacı bağımsız olarak tasarlanmaktadır. Bu tezde dil ve konuşmacı bağımsız olarak konuşmaların, söz dizileriyle etiketlenerek gelişmesini hedefleyen sistem önerimi yapılmıştır. Etiketlenen konuşmalar ile bu alandaki araştırmalar için yenilikçi bir bakış açısı sayılabilecek dil bağımsız olarak gelişen metin kütüphanesi (corpus) tabanlı konuşma tanıma sistemi önerilmiştir. İlgilendiği konular kapsamında bu tez, sinyal işleme ve örüntü tanıma gibi farklı bilgisayar bilimlerinin kesişiminde yer almaktadır. Önerilen çalışmada nihai hedef, insanların akıllı cihazlarla etkili iletişim kurmaları için verimli teknikler ile başarısı yüksek gerçek zamanlı bir konuşma tanıma sistemi sunmaktır. Ayrıca bu tez kapsamında, konuşma tanıma alanında kullanılan teknikler karşılaştırılarak önerilen sistemin deneysel çalışması ve değerlendirilmesi yapılmıştır.

Anahtar Kelimeler: Konuşmayı yazıya çevirme, Konuşmacı bağımsız konuşma tanıma, Dil bağımsız konuşma tanıma, Verimli konuşma tanıma, Konuşma metin kütüphanesini geliştirme, Sayısal sinyal işleme, Konuşma tanıma için istatistiksel tabanlı modeller, Çok seviyeli örüntü tanıma.

ABSTRACT

Along with evolving technology, many interfaces (forms of interaction) have occurred in human-computer interaction. One of these interfaces is speech recognition. Speech recognition translates human voice into a form that can be read by the computer without intermediaries. This way, one has the possibility to manage the devices by speaking. The speech recognition technology, which has many application areas, provides facilities that are differentiated according to the ways of use. The process of translating one's speech into one of these areas has been subject to many daily work from past to present. In traditional studies, it was aimed to translate the speeches of certain people into the text. Applications for this purpose are speaker dependent systems. However, speaker-dependent systems are not able to work out, without identifying different speeches to the system first. Nowadays, most of the systems developed, especially smart devices, are designed as speaker independent. In this thesis, a system proposal was made aiming to develop their speech independently from both the speaker and the language by labeling them with their syntax. The tagged speech has been proposed as a corpus-based speech recognition system, which can be considered as an innovative viewpoint for researches in this area. This thesis within the scope of the subjects it is concerned, is in the intersection of different computer sciences such as signal processing and pattern recognition. The ultimate goal in the proposed study is to provide a high level of real-time speech recognition system with efficient techniques for effective communication between humans and smart devices. In addition, in the scope of writing of this thesis, an experimental system is studied and evaluated by comparing the techniques which are used in the field of speech recognition.

Key Words: Speech to text, Speaker independent speech recognition, Language independent speech recognition, Efficient speech recognition, Speech corpus development, Digital signal processing, Statistical based models for speech recognition, Multilevel pattern recognition.

1 GİRİŞ

İlerleyen teknolojinin hayatını birçok alanda kolaylaştırdığı insan, eylemlerini en iyi şekilde konuşma ile anlatmaktadır. Bu kapsamda insan-bilgisayar etkileşimi alanında yapılan çalışmalarda konuşma tanıma sistemlerine ağırlık verilmiştir. Konuşma tanıma sistemi, kullanıcının belirli kurallar ile oluşturulan ve kurallarının bilgisayar tarafından bilindiği, birtakım sesli ifadeleri, bilgisayar tarafından anlaşılabilecek formata dönüştürme işlemidir. Konuşma tanıma; akıllı cihazlarda sesli komut uygulamaları, akıllı ev sistemleri, sesli komutlar ile sağlanan güvenlik sistemleri, eğitim sistemleri, Etkileşimli Ses Yanıtı (Interactive Voice Responce) ve Sesli Yanıt Ünitesi (Voice Response Unit) gibi birçok alanda geliştirilmeye devam etmektedir. Konuşma tanıma sistemleri üzerine yapılan geleneksel çalışmalar, konuşmacı bağımlı olarak eğitilip, eğitilen konuşmacılara göre kişi tanıma üzerine olmuştur. Bu tip konuşmacı bağımlı sistemlerde tanıtılmayan kişiler için sistemin eğitilmesi gerekmektedir. Sınırlı çalışma alanına sahip bu sistemler güvenlik, yetkilendirme gibi alanlarda tercih edilmektedir. Konuşmacı bağımsız sistemler ise konuşmacı bağımlı sistemlere göre karmaşıklığı fazla ve daha zor oluşturulan sistemlerdir. Fakat bu sistemler, konuşmacı bağımlı sistemler gibi bir şablon güncellemesine ihtiyaç duymadan herhangi bir konuşmayı yazıya çevirmeye olanak sağlar. Ayrıca konuşmacı bağımsız sistemler, kaydedilen çok sayıda konuşma örnekleriyle ön öğrenmeden geçirilerek kullanılır. Bu sebeple konuşmacı bağımsız sistemlerde öğrenme kümesi geniş olmalıdır. Akıllı telefonlarda dâhil olmak üzere birçok alanda örneği bulunan konuşma tanıma sistemlerinin önemli özelliklerinden bir tanesi konuşmacı bağımsız olmasıdır. Bu çalışmada da konuşmacı ve dil bağımsız gerçek zamanlı bir konuşma tanıma sistemi için literatür taranmış ve verimli teknikler ile konuşma tanıma önerimi yapılmıştır. Önerimi yapılan sistemin uygulaması yapılmış, sonuçları ile gelecekte

yapılması muhtemel çalışmalar üzerinde durulmuştur. Bu doküman şu şekilde yapılandırılmıştır: Birinci bölümde verimli konuşma tanıma modeline giriş yapılmıştır. Çalışmanın ikinci bölümünde konuşma tanıma ile ilgili genel bilgiler ve literatür araştırmaları verilmiştir. Üçüncü bölümde konuşma tanımada ki süreçlere değinilmiş ve süreçlerdeki teknikler incelenmiştir. Dördüncü bölümde konuşmacı ve dil bağımsız gerçek zamanlı bir verimli konuşma tanıma modeli önerilmiştir. Beşinci bölümde önerilen modelin uygulaması ve kullanımı ile ilgili bilgiler paylaşılmıştır. Altıncı bölümde ise önerilen modelin değerlendirilmesi yapılmıştır, uygulamanın deneysel sorunlarına değinilmiş ve gelecek çalışmalar öngörülmeye çalışılmıştır.

1.1 Çalışmanın Amacı

Bu çalışmanın amacı, konuşmanın yazıya çevrilmesinde geçmişten günümüze kadar yapılan çalışmaların incelenip verimli teknikler ile gerçek zamanlı konuşmacı ve dil bağımsız konuşma tanıma sistemi sunmaktır. Bu amaç doğrultusunda literatür taraması yapılmış, belirlenen teknikler eksikleri ve avantajları yönünden incelenmiş ve sonuçları değerlendirilmiştir. Değerlendirmeler neticesinde sistem önerimi yapılmıştır. Çalışma sonunda önerimi yapılan sistemin deneysel çalışması yapılmıştır.

1.2 Çalışmanın Kapsamı

Literatürde var olan konuşma tanıma tekniklerinin verimlilik yönünden kıyaslanması ve konuşma tanıma alanında tasarlanan sistemlerin yapısal olarak incelenmesi bu çalışma kapsamındadır. Bu inceleme neticesinde önerilen sistemin verimliliği deneysel çalışma ile değerlendirilmiştir.

1.3 Çalışmada Sınırlar

Konuşma tanıma alanında yıllarca süren araştırma ve geliştirmelerden sonra, konuşma tanıma doğruluğu, konuşmacı ve dil değişkenliği, kelime büyüklüğü ve etki alanı, gürültü, konuşma tanıma sisteminin tasarımı, çeşitli konuşma sınıfları, konuşma gösterimi, özellik çıkarma teknikleri, veri tabanı ve performans değerlendirmesi gibi zorluklar konuşma tanımanın en önemli araştırma konuları olmuştur (Saini & Kaur, 2013). Konuşma tanıma sistemlerinde çoğu zaman görülen zorluklar kullanıcının

davranışına ve bilgisine göre değişmektedir. Bunun gibi yetersiz bilgiyi önlemek için sistem iyi hazırlanmış olmalıdır ve güncel teknolojileri bilmelidir (Aydın, 2005).

Konuşma tanıma teknolojisinin ticarileştirilmesi, eğitim ve test koşulları arasındaki çevresel farklılıklardan dolayı sistem performansındaki büyük bozulmayla engellenmektedir (Mammone ve ark., 1996). Böyle bir uyumsuz durumun aksine yapılan ve yapılacak çalışmalar, eğitim sırasındaki koşullar ile işlem sırasındaki koşullara (eşleşen koşullar) benzerse, çağdaş sistemlerin çoğunun iyi bir tanıma performansı sergileyeceği tahmin edilmektedir. Sıklıkla böyle uyumsuz durumlar, hedeflenen uyumlu durumlarla karşılaştırıldığında performansın önemli ölçüde düştüğü görülmektedir. Bu uyumsuzluğa ilişkin yaygın bir örnek, temiz bir konuşmada eğitim yapıldığında ve gürültülü veya kanal bozuk bir konuşmada test yapıldığında geçerlidir. Verimli konuşma teknikleri, bu türden çeşitli çalışma koşulları altında bir konuşma işleme sisteminin performansını korumaya çalışmaktadır.

1.4 Varsayımlar

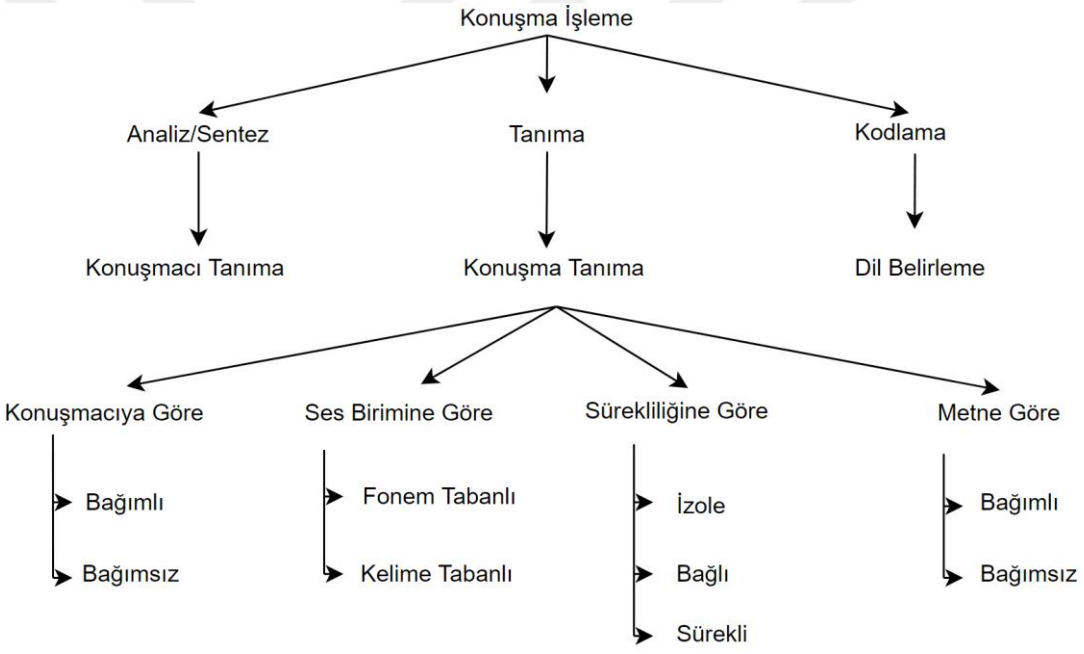
Aşağıda sıralanan varsayımlar kabul görülerek bu çalışma yapılmıştır.

1. Çalışma genelinde sınırlı öğrenme metinleri ile sistem eğitilmiştir. Sistemin eğitim düzeyine göre deneysel çalışma kapsamında testler yapılmıştır. Testlerin yeterli olduğu kabul edilmiştir.

2. Çalışma genelinde literatür incelemesi sonucu tespit edilen teknikler ile tasarlanan sistemin konuşma tanıma için yeterli düzeyde olduğu kabul edilmiştir.

2 KONUŞMA TANIMA TÜRLERİ

Konuşma tanıma, yeteneğine ve kullanımına bağlı olarak farklı türlere ayrılır. Bu türler konuşmacı bağımlılığına göre bağımlı ve bağımsız tanıma, temel alınan ses birimine göre fonem (ses birim) tabanlı ve kelime tabanlı tanıma, sesin sürekliliğine göre izole, bağlı ve sürekli tanıma, metin bağımlılığına göre bağımlı ve bağımsız konuşma tanımadır.



Şekil 1. Konuşma Tanıma Türleri

2.1 Önceki Çalışmalar ve Uygulama Alanları

1952 yılında Davis ve ark. tarafından (Davis ve ark., 1952), Bell Laboratuvarlarında tek bir konuşmacı için, ayrılmış bir rakam tanıma sistemi geliştirilmiştir. 1956 yılında Olson ve ark. tarafından (Olson & Belar, 1956), tek bir konuşmacıya ait 10 ayrı heceyi tanımak için RCA Laboratuvarlarında bir çalışma yapılmıştır. 1959 yılında Fry ve ark. (Fry, 1959), İngiltere'de UCL'de 4 sesli ve 9 sessizi tanıyabilen bir ses birim tanıyıcı geliştirmişlerdir. Çalışmalarında tanıma kararını gerçekleştirmek üzere bir spektrum

çevirici ve bir örüntü eşleştirici kullanılmıştır. 1959 yılında Forgie ve ark. tarafından (Forgie & Forgie, 1959) MIT Lincoln Laboratuvarlarında yapılan çalışmada, konuşmacıdan bağımsız bir konuşma tanıma sistemi ile spektral bilgiyi elde etmek için bir filtre bankası çevirici ve konuşma tanımayı gerçekleştirmek için ses tüpü tınılarının zaman değişimlerini kestiren bir sistem kullanılmıştır.

Şekil 2. IBM Shoebox

1960'lı yılların başında IBM, bugünün ses tanıma sistemlerinin öncüsü olan Shoebox'ı geliştirdi ve gösterdi (IBM, 1960). Şekil 2'deki resimde IBM'in Kaliforniya'daki Gelişmiş Sistem Geliştirme Laboratuvarı'ndaki ileri teknoloji grubunun yöneticisi Dr. E. A. Quade, sesli komutlarla aritmetik gerçekleştiren deneysel bir makine olan Shoebox'u gösteriyor.

1960'lı yıllarda konuşma tanıma ile ilgili birkaç temel fikir ortaya çıkmıştır ve yayınlanmıştır. 1961 yılında Suzuki ve arkadaşları tarafından (Suzuki & Nakata, 1961) Tokyo'da, radyo araştırma laboratuvarlarında gerçekleştirilen sesli bir tanıyıcı donanım çalışması yapılmıştır. 1962 yılında Sakai ve ark. (Sakai & Doshita, 1962), Japonya Kyoto Üniversitesi'nde bir ses birim tanıyıcı donanım gerçekleştirmişlerdir. 1967 yılında Reddy tarafından (Reddy, 1967), makine ile konuşma tanıma çalışması ile sürekli ses alanlarını tanıma için öne sürülen, ses birimlerinin dinamik izlenmesi yöntemi önerilmiştir. 1968'de Vintsyuk (Vintsyuk, 1968), Sovyetler Birliği'nde, bir çift ses ifadesi üzerinde zaman düzenleme (uydurma) için dinamik programlama metotlarını önermiştir.

1970'de Rusya'da Velichko ve ark. (Velichko & Zagoruyko, 1970), konuşma tanıma sistemi içerisinde örüntü tanımanın geliştirilmesine katkıda bulunmuşlardır. Yine 1970'lerin başlarında konuşma tanımada HMM yaklaşımı Princeton Üniversitesi'nde

Lenny Baum tarafından keşfedilmiştir. HMM, karmaşık bir matematiksel örüntü eşleme stratejisi olarak tanımlanabilir ve içinde Dragon Systems, IBM, Philips ve AT&T'nin de bulunduğu birçok konuşma tanıma şirketi tarafından kullanılmıştır (Juang & Rabiner, 2004). 1971 yılında İleri Savunma Araştırma Projeleri Acentesi Topluluğu (Defense Advanced Research Projects Agency, DARPA) tarafından, sürekli konuşmayı anlayabilecek bir bilgisayar sistemi geliştirmek için SUR (Speech Understanding Research) kurulmuştur. Buna ek olarak CMU, SRI, MIT Lincoln Laboratory, Systems Development Corporation (SDC) ve BBN (Bolt, Berenak and Newman)'da kapsamlı SUR projeleri kurulmuştur (Juang & Rabiner, 2004). 1975'de ABD'de Itakura (Itakura, 1975), konuşma tanıma sistemlerinde LPC'nin uygulamasını gerçekleştirmiştir. 1978'de Japonya'da Sakoe ve ark. (Sakoe & Chiba, 1978), konuşma tanıma üzerinde dinamik programlama metotlarının başarılı olarak uygulamasını gerçekleştirmişlerdir.

1984 yılında SpeechWorks şirketi, telefon üzerinden otomatik konuşma tanıma sistemleri üretmiştir (Ford, 2004). 1990'larda ticari olarak başarılı konuşma tanıma sistemlerinin ilk tanıtımları yapılmıştır (Huang ve ark., 2014). 1990 başlarında DARPA sürekli konuşma tanıma sistemlerinin geliştirilmesine destek vermiştir (Pallet ve ark., 1990). İlerleyen yıllarda Vapnik tarafından (Vapnik, 1995), veri sınıflandırılması ile regresyon problemlerini çözüme kavuşturmak amacıyla SVM ortaya atılmıştır. SVM, 2000'lerde konuşma tanıma, konuşmacı tanıma ve doğrulama işlemleri için kullanılmıştır. HMM'yi temel alarak önerilen konuşma tanıma uygulamaları için N tane en iyi aday tabanlı bir eğitim algoritması Chen ve ark. tarafından (Chen & Soong, 1994) 1994 yılında önerilmiştir. 1995 yılında, ilk kez Dragon Systems tarafından üretilen kelime tabanlı dikte yazılımı piyasaya sürülmüştür. Bunun ardından, benzer yazılımlar IBM ve Kurzweil tarafından da üretilmeye başlanmıştır (Koumpis & Pavitt, 1999). 1996'da Charles Schwab ve Nuance tarafından Voice Broker isminde bir konuşma tanıma sistemi geliştirilmiş ve bu sistemle 360 adet müşteri telefon üzerinden aynı anda borsa işlemi yapmıştır. Bu sistem, her gün 50000 adet isteği yerine getirebilmiştir. Sistemin doğrulunun %95 civarında olduğu belirlenmiştir. Yine aynı yıl Dragon Systems "Naturally Speaking" i geliştirmiş ve bu ürün ilk sürekli dikte yazılımı olmuştur. Ayrıca Lernout ve Hauspie'dan Voice Xpress, Dragon Systems'den Naturally Speaking, Philips'den

FreeSpeech, SpeechPro ve IBM’den ViaVoice günümüzdeki dikte paketlerine örnek olarak verilebilir (Öcal, 2005).

2006 yılında, Amerika Birleşik Devletleri'nde, Ulusal Güvenlik Ajansı (National Security Agency) anahtar kelime tespiti için bir konuşma tanıma türünü kullanmıştır (Singh K. , 2016). Bu teknoloji, analistlerin büyük miktarda kaydedilmiş konuşmaları taramasına ve anahtar kelimelerden söz etmelerine izin vermesine olanak tanımıştır. Kayıtlar dizine eklenebilmiş ve analistler ilgi çekici konuşmaları bulmak için veri tabanı üzerinden sorgular çalıştırabilmişlerdir. Bazı devlet araştırma programları, konuşma tanımanın istihbarat uygulamaları üzerine odaklanmıştır; DARPA'nın EARS's programı ve IARPA'nın Babel programı örnek olarak verilebilir (Froomkin, 2015).

2007 yılında Google'ın konuşma tanıma alanında ilk ürünü telefonla çalışan bir dizin hizmeti olan GOOG-411 olmuştur. GOOG-411 kayıtları, Google'ın tanıma sistemlerini geliştirmesine yardımcı olan veriler üretmiştir ve şu anda Google sesli aramada otuzdan fazla dilde desteklenmektedir (Kincaid, 2011). Xuedong Huang, Sphinx-II sistemini geliştirmiştir. Sphinx-II sistemi, konuşmacıdan bağımsız, geniş kelime, sürekli konuşma tanımayı ilk yapan sistemdir. 2012 yılından beri kullanılan Apple'ın Siri teknolojisi, arkasındaki ses tanıma şirketi Nuance tarafından geliştirilmiştir (Wildstrom, 2011).

Şekil 3. iOS 11 Siri ile istek üzerine Seahawks’un fisktürünün dikte edilmesi

Konuşma tanıma, bir metnin dikte edilmesinden gerçek zamanlı olarak bir televizyon yayını için altyazı üretmeye kadar birçok uygulamayı içerir. Konuşma tanıma alanında başlıca uygulama alanlarını sıralayacak olursak;

- Dikte (yazdırım),
- Çeviri,
- Akıllı cihazlar,
- Ev otomasyonu,
- Araba içi sistemler,
- Komut ve kontrol,
- Konuşmacı bağımlı sistemlerde güvenlik kontrolü,
- Telefon üzerinden hizmet (bilgisayar tabanlı telesekreterler gibi),
- Sağlık hizmeti,
- Eğitim alanı,
- Robotların sesle kontrolü,
- Askeri ve istihbari alanlar,
- Gömülü uygulamalar,
- Uzay (örneğin uzay araştırması, uzay aracı, vb.) NASA'nın Mars Polar Lander, Lander'daki Mars Mikrofonunda Sensory, Inc.'den konuşma tanıma teknolojisini kullandı,
- Konuşma tanıma ile otomatik altyazı üretme,
- Mahkeme raporlaması (Gerçek Zamanlı Konuşma Yazma),
- e-Discovery (Yasal keşif),
- Ahizesiz Bilgi İşlem: Konuşma tanıma bilgisayar kullanıcı arabirimi,
- Bilgisayar destekli dil öğreniminde telaffuz değerlendirme,
- Sanal asistan (örneğin, Apple Siri).

Gelişen tekniklerle birlikte konuşma tanıma alanında hata oranları sürekli azalmaktadır. Bu da konuşma tanıma teknolojisinin yaygınlaşmasını sağlamaktadır. Konuşma tanıma sistemleri ile yapılabilen işlemlerde sistemler tekrar eden işlemleri hızlı bir şekilde ele alarak maliyetten tasarruf edebilir, anketleri sesli olarak cevaplayabilir, sipariş ve ödemeleri konuşmacı kimliği tanıma ile alabilir, tuşlayarak yapılamayan işlemleri otomatik hale getirebilir. Adresleri ve isimleri toplamada uzun seçenekli listelerden kaçınmayı sağlayabilir.

2.2 Konuşmacıya Göre Konuşma Tanıma

Konuşma tanıma sistemleri, ihtiyaç doğrultusunda konuşmacı bağlılığı temel alınarak iki gruba ayrılır: konuşmacı bağımlı sistemler ve konuşmacı bağımsız sistemler. Bu ayrım ile sistemin uygulanmasında kullanılan teknikler ve sistemin kullanıldığı alanlar değişir.

2.2.1 Konuşmacı Bağımlı Konuşma Tanıma

Konuşmacı bağımlı sistemler, belirli kullanıcı ya da kullanıcılar tarafından önceden sisteme tanıtılmış bir kelime ya da kelime grupları ile tanımlanır. Konuşmacı bağımlı sistemlerde, başka bir konuşmacı sesinin tanınması istenildiğinde, sistem üzerinde kayıtlı olan ve konuşma tanıma için kaynak olarak alınan şablonların güncellenmesi gereklidir (Baygün, 2006). Bu sistemler, yüksek komut sayımına ve kelime tanıma için yüksek oranda doğruluk elde etme yeteneğine sahiptir. Bu yaklaşımın dezavantajı, sistemin yalnızca sistemi eğitmiş olan kişiye doğru bir şekilde tepki vermesidir. Bu tür sistemlerde konuşmacıyı bir veya daha fazla kişi oluşturmaktadır. Tanımlanan her kişinin konuşmasının tanınması için referans şablonları bulunmalıdır (Gelegin & Bolat, 2011).

Furui (Furui, 1991), konuşma dalgalarından konuşmacıya bağımlı özellik çıkarımı, konuşmacının tanımlanması ve doğrulanması, konuşma tanımada konuşmacı uyarlaması ve ses dönüşüm teknikleriyle ilgili araştırmaların son gelişmelerini ve perspektiflerini araştırmıştır. Konuşmacıyla ilgili bireysel bilgilerin, geçici ve dinamik özelliklere ayrılabildiğini göstermiştir.

Bavya ve Steiger (Bayya & Steiger, 2002), tek bir simge eğitimi de dâhil olmak üzere çok sınırlı eğitim verilerini gerektiren, konuşmacı bağımlı konuşma tanıma sistemleri içinde kullanılmak üzere model oluşturulmasını sağlayan bir konuşma tanıma eğitim sistemi geliştirmişlerdir. Çalışmalarında, HMM metodunu kullanarak konuşmacı bağımlı modellerin oluşturulması için basitleştirilmiş bir metot sağlamışlardır.

Murty ve Yegnanarayana (Murty & Yegnanarayana, 2006), konvansiyonel MFCC'yi mevcut bilgilerle karşılaştırıldığında kalıcı fazda bulunan konuşmacıya özgü bilgilerin tamamlayıcı niteliklerini gösteren bir çalışma yapmışlardır. Artık faza dayanan

konuşmacı tanıma sisteminde hata oranı %22, MFCC özelliklerini kullanan sistemde ise %14'lük bir hata oranı vermiştir.

Konuşma tanıma sistemlerinde kişi bağımlılığı temel alındığında kişi yetkilendirmesi üzerine güvenlik alanlarını sıralamak gerekirse;

- Akıllı ev sistemleri,
- Bilgisayarlara veya her türlü kişisel programlara girerken ses kontrolü,
- Üst seviyede önlem gerektiren durumlarda ses ile bilgiye erişme izni,
- Metne bağlı ses tanıma ile yetkilendirme.

2.2.2 Konuşmacı Bağımsız Konuşma Tanıma

Konuşmacı bağımsız sistemlerde konuşmacılar kaydettikleri çok sayıda ses örnekleriyle ön öğrenme ile sistemi kullanmaya başlar. Konuşmacı bağımsız sistemler, konuşmacı bağımlı sistemler gibi bir şablon güncellemesine ihtiyaç duymadan herhangi birinin sesini tanımaya olanak sağlar. Farklı kişilerden alınan sesleri tanımada hazırlanması gereken şablonlar ile sistemin modellenmesi tasarım olarak büyük ölçüde uğraş gerektirir. Böylelikle bir dezavantaj olarak, herhangi bir dil için tüm konuşmacı varyasyonlarını modellemenin olanaksız olduğu gözlemlenebilir. Bu dezavantaj ile konuşmacı bağımsız sistemlerin performansı, konuşmacı bağımlı sistemlere göre daha düşüktür. Fakat kullanım alanı göz önüne alındığında, zorluğuna rağmen konuşmacı bağımsız sistemler, konuşmacı bağımlı sistemlere göre bir adım öne çıkmaktadır. Bu tür sistemlerde özel olarak konuşmacı eğitimi gerekmez (Dede, 2008). Bu nedenle sistem, hedef kelimenin çok çeşitli konuşma kalıplarına ve telaffuzlarına cevap vermelidir.

Seide ve ark. (Seide ve ark., 2011), Bağlam Bağımlı Derin Sinir Ağı HMM'lerinin (CD-DNN-HMM'ler) bir özellik-mühendislik perspektifinden potansiyelini araştırdılar. Son zamanlarda, konuşmacı bağımsız olarak telefon çağrılarının transkripsiyonu için (NIST RT03S Fisher verileri) CD-DNN-HMM'ler ile yalnız HMM'ler ile elde edilen %27,4'lük sözcük hata oranının üçte bir oranında azaltıldığını göstermişlerdir.

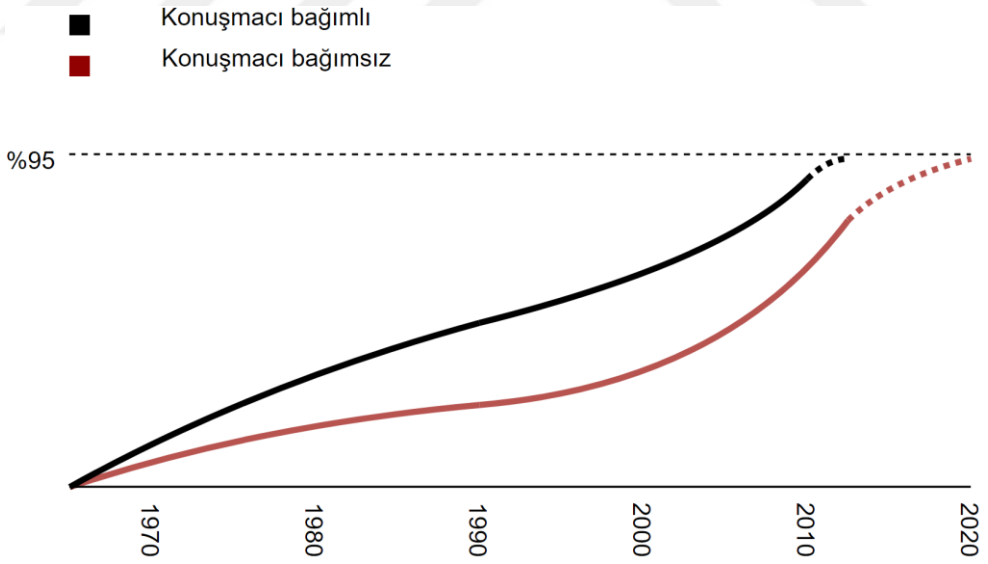
Karthikeyan ve ark. (Karthikeyan & Vijayalakshmi, 2016), ses uygulamaları için konuşma tanımanın performans karşılaştırması üzerine çalışmışlardır. Çalışmalarında

özellikle görme zorluğu çeken kişiler için cihaza dokunmadan tüm cep telefonu uygulamalarını kullanabilmeleri için konuşmacı bağımsız sistem üzerinde durmuşlardır. Önerilen sistem, MFCC özellikleri ile DTW ve HMM / VQ kullanılarak şablon üretimi gibi iki farklı sınıflandırma modellemesi yoluyla değerlendirilmiştir. MFCC özellikleri ile HMM / VQ sınıflandırma modeli ile sesli algılamalar için diğer metotlara göre daha yüksek olarak tanımada %82.77 doğruluk oranı elde etmişlerdir.

Becerra ve ark. tarafından (Becerra ve ark., 2016) akustik modelleme çerçeveleri örneklenerek, kişiselleştirilmiş bir konuşmacı bağımsız metin bağımlı, konuşma tanıma çalışması gerçekleştirmişlerdir. Sonuçlarda, DNN kullanılarak daha iyi bir kelime hata oranı yakalandığını gözlemişlerdir. GMM-HMM oranı %4.20, DNN-HMM modelleri ile %3.33 ile %20.71 arasında göreli iyileşme oranı elde etmişlerdir.

2.2.3 Değerlendirme

Geçmiş çalışmalar incelendiğinde konuşmacı bağımlı konuşma tanıma sistemlerinde başarım oranının konuşmacı bağımsız sistemlere göre daha yüksek seviyelerde olduğu görülmüştür.



Şekil 4. Konuşmacı bağımlılığına göre konuşma tanımanın yıllara göre başarımı

Başarım oranları arasındaki farkın eğitim ve test setlerinden oluştuğu gözlenmiştir. Konuşmacı bağımlı konuşma tanımada sistemi belirli kullanıcı/kullanıcılar tarafından eğitilir. Böylece sistemin testi aşamasında belirli kullanıcılar ile test yapılır. Konuşmacı bağımsız konuşma tanıma sistemi ise herkesin konuşmasını tanımak için

tasarlanır. Bu nedenle eğitim setinin birçok faklı kullanıcılar ile eğitilmesi gereklidir. Bu tür sistemlerin olumsuz yönü, konuşmacı bağımlı olmayan sistemlerin genellikle konuşmacı bağımlı sistemlerden daha az doğru konuşmasıdır. Önerilen modelde ne kadar farklı kullanıcı tarafından eğitilirse, başarımın konuşmacı bağımsız olarak o kadar fazla yükselmesi öngörülmektedir.

Uygulama alanında özellikle konuşmacı bağımsız konuşma tanıma sistemleri daha kullanışlı olacaktır. Buna örnek olarak, emniyet sorgularında, zabıt işlemlerinde, mahkeme duruşmalarında hep karşılıklı konuşmaların gerçek zamanlı bilgisayara yazılması söz konudur. Bu tip alanlarda yapılacak çalışma oldukça kullanışlı olabilir. Günümüz çalışma ortamlarında daha fazla rahatlığa kavuşabilmek amacıyla konuşma tanımayla ilgili uygulanabilecek alanların artırılması gerekmektedir (Yalçın, 2008).

2.3 Temel Alınan Ses Birimine Göre Konuşma Tanıma

Konuşma tanıma sistemleri, ihtiyaç doğrultusunda temel alınan ses birimine göre iki gruba ayrılır: fonem tabanlı sistemler ve kelime tabanlı sistemler. Bu ayrım ile sistemin uygulanmasında kullanılan teknikler ve sistemin kullanıldığı alanlar değişir. Fonem tabanlı konuşma tanıma sistemleri, fonemlerin (harf/hece) en küçük birim olarak kabul edildiği sistemlerdir. Kelime tabanlı konuşma tanıma sistemleri; tanıma için öngörülen en küçük birim olarak kelimelerin kabul edildiği sistemlerdir (Gelegin & Bolat, 2011).

Fonemler arası geçişlerin hata oranı kelimeler arası geçişlere göre daha az olmaktadır. Kelime tabanlı konuşma tanıma sistemlerinde referans şablonu olarak kelimenin tamamı alınır ve bir konuşma dilinde çok sayıda kelime olmasından dolayı sistemin gereksinim duyduğu bellek ihtiyacı daha fazla olacaktır. Fonem tabanlı konuşma tanımada ise doğruluk yüzdesi bir miktar düşerken, az olan fonem sayısı, hızlı sonuç üretme olanağı sayesinde, hataları en aza indirme amaçlı güncellemeleri mümkün hale getirmektedir (Mengüşoğlu, 1999).

2.3.1 Fonem Tabanlı Konuşma Tanıma

Konuşmacı tarafından söylenen fonların simgesel olarak ifadesi fonemdir. Alfabetik harfler fonem olarak tanımlanabilir. Tanıma esnasındaki birimler ikili fonem, üçlü fonem, hece veya kelime olabilir. Fonem tabanlı konuşma tanıma sistemi, fonlardan

fonemlere dönüştürme işlemidir. Bu tür sistemlerde gerekli olan en küçük unsur fonem ve sözcük birimleridir.

Ostendorf ve Roukos (Ostendorf & Roukos, 1989), stokastik (raslantısal, rastsal) segment modeli adı verilen, değişken-süreli fonemlerin modellenmesi için yeni bir yaklaşım modeli üzerinde çalışmışlardır. Bu fonetik modelde HMM ile kelime tanıma sistemine kıyasla sözcük hata oranının üçte bir oranında azaldığı gösterilmiştir.

Mari ve ark. (Mari ve ark., 1996), stokastik yöntemlerle fonem tabanlı sürekli konuşma tanıma alanında, birinci dereceden HMM kullanarak yüksek performans gösterilebileceğini ve metin bağımsız HMM'lerin doğruluğunun %69'undan fazlasını elde edebildiğini göstermişlerdir.

Scheme ve ark. (Scheme ve ark., 2007), akustik konuşma tanıma doğruluğunun gürültülü ortamlarda bozulduğunu göstermişlerdir. Kelimeler, HMM sınıflandırıcısı kullanılarak sınıflandırılmıştır. "Sıfır" ile "dokuz" arasındaki sözcükler toplanmıştır ve %99'luk bir doğrulukla yaklaşık %38'e kadar bozulan 18 biçimlendirme fonemi sınıflandırılmıştır. Simülasyonlarda %94'ün üzerinde doğruluk sağlanırken düşük gürültü seviyelerde %99'luk doğruluk elde edilmiştir. Sonuçlar daha önceki konuşma tanıma doğruluğuna göre yaklaşık %10 oranında iyileşme sağlamıştır.

2.3.2 Kelime Tabanlı Konuşma Tanıma

Konuşma tanıma için gerekli olan en küçük unsurun kelime olarak kabul edildiği sistemdir. Uygulama anlamında yüksek verimlilik derecesiyle birlikte kelime tabanlı konuşma tanıma sistemlerinde gereksinimler fazladır. Bu sistemler üzerinde komuta kontrol uygulamalarının başarılı olabilmesinin sebebi kelime sayısının sınırlı tutulmasıdır. Fakat Türkçenin eklemeli bir dil olduğu düşünüldüğünde kelimelere ekler ekleyerek birçok yeni kelime türetilebilmektedir. Bu durum, kelime tabanlı Türkçe konuşma uygulamaların geliştirilmesinde eğitim setinin büyük tutulmasını gerektirmektedir.

Abdulla ve ark. (Abdulla ve ark., 2003), Türkçe gibi sondan eklemeli diller için kelime tabanlı sürekli konuşma tanımada, teorik olarak sonsuz tam sözlü sözlük boyutu nedeniyle karşılaşılan sorunlara karşı test verisinde sözcük dağarcığının oranını önemli ölçüde azaltmak için alt sözcük sözlük birimleri kullanılabileceğini söylemişlerdir. Bu

sorunları azaltmada, Türkçe için oluşturulan sözlükteki mümkün olan en uzun alt sözcük birimlerini, yani yalnızca yarım sözcükleri ve tam sözcükleri kullanmayı önermişlerdir. Çift gramlı bir modelle yarım kelimeleri kullanmak, iki gramlı tam sözcüklü bir modele kıyasla, kelime-hata oranında belirgin bir düşüş sağlayacağını göstermişlerdir.

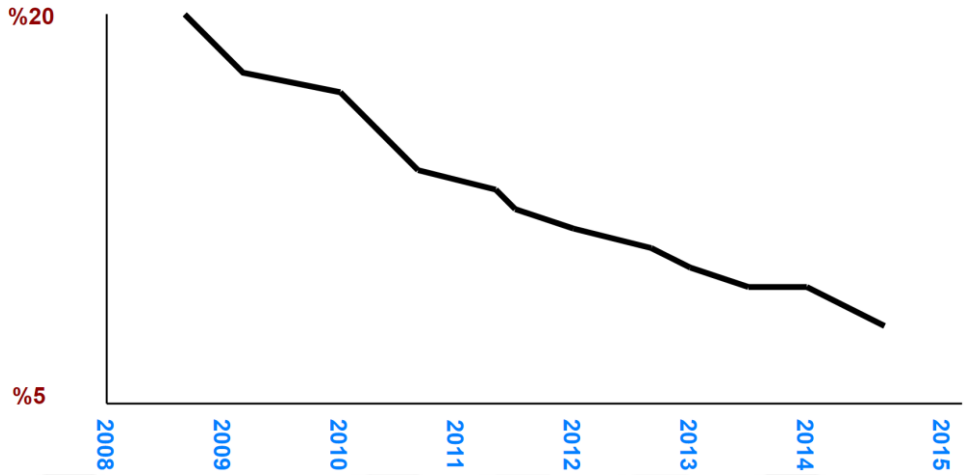
Prakoso ve ark. (Prakoso ve ark., 2016), otomatik konuşma tanıma (Automatic Speech Recognition, ASR) alanında Endonezya dilinde akustik model, dil modeli ve sözlük gerektiren sınırlı veri kümesine sahip, CMUSphinx araç setini (HMM tabanlı ASR aracı) kullanarak sistem tasarlamışlardır. Akustik modelin kelime hata doğruluğunun en iyi başarı ortalamasının %86 olduğunu tespit etmişlerdir.

Tabassum ve ark. (Tabassum ve ark., 2017), geniş bir sözlük kümesinden bazı önemli kelimeleri, konuşmacı bağımsız olarak tanıma sürecini göstermişlerdir. Farklı ünlülerin seslerinden birkaç izole kelimeyi ayırt etmek için, bir dizi rassal erkek ve kadın konuşmacıdan toplanan konuşma sinyallerinden özellikler çıkarmışlardır. Ayıklanan özellikler daha sonra sistemi eğitmede belirli konuşmalar için analiz edilmiştir. Bu çalışmanın özgül (özellikli) hedefleri, etkili bir insan-makine etkileşimi için konuşmayı ve insan ile makine arasındaki bir ses arabirim sistemini tanımanın yanı sıra izole bir otomatik kelime konuşma tanıyıcı uygulamaktır. Sistemi testi neticesinde sonuçların yaklaşık %90'ında tatmin edici olduğunu belirtmişlerdir. Bununla birlikte, bazen benzer sesli seslerle sistemin karışabildiğini gözlemlemişlerdir.

2.3.3 Değerlendirme

Fonem tabanlı konuşma tanıma sistemleri sınırlı sayıda eğitim verisine sahip olsa da kelime tanıma esnasında problem yaşamaktadır. Test aşamasında bir fonem tabanlı sistemin kelime tanımadaki zorluğu, fonemlerin art arda sıralanması esnasında gerçekleşen geçişler ve sesli ifadenin foneme dönüşmesi esnasında fonem sınırlarını belirlemektir. Bu veriyi göz önüne alırsak, fonem tabanlı sistemlerin zorluğu fonemlerin arasındaki geçişlerin, başlangıç ve tespitinin zorluğudur (Yalçın, 2008). Fonem tabanlı konuşma tanıma sisteminde, fonemlerin arasındaki geçişlerin olumsuz etkisi göz önüne alındığında, bu çalışmanın da temelini oluşturan kelime tanıma sistemlerinde kelime tabanlı konuşma tanıma sisteminin verimliliği ve doğruluğu daha

fazladır. Önerilen modelde kullanım alanın genişliği, uygulanabilirliği ve başarım oranı gözlenerek kelime tabanlı olarak planlanmaktadır.



Şekil 5. Dragon Natural Speaking'in yıllara göre kelime hata oranı

Yukarıda da Dragon Natural Speaking tarafından yayınlanan rapora göre son 10 yıl içerisinde ki kelime hata oranlarında (WER) başarılı bir şekilde azalma gösterilmektedir. Ticari alanda gelişen konuşma tanıma teknolojisinde WER'in son yıllarda Google'ın konuşma tanıma teknolojisinde %5'in altına indiği görülmektedir.

2.4 Sesin Sürekliliğine Göre Konuşma Tanıma

Konuşma tanıma sistemleri, ihtiyaç doğrultusunda sesin sürekliliğine göre üç gruba ayrılır: izole konuşma tanıma, bağlı konuşma tanıma ve sürekli konuşma tanıma. Bu ayrım ile sistemin uygulanmasında kullanılan teknikler ve sistemin kullanıldığı alanlar değişir. Tanınacak konuşmada metin elemanlarının yerleşimi; izole, bağlı ya da sürekli olarak değişmektedir.

Sesin sürekliliğine göre konuşma tanıma, izole, bağlı ya da sürekli konuşma tanıma sistemlerini içerisinde bulundurmaktadır. Bu tür sistemlerde verilen bir konuşma (akustik) X dizisi için, W kelime ya da kelime dizisini bulmak için oluşturulmuştur. Konuşma cümleleri, $W = (w_1, w_2, \ldots, w_t)$ şeklinde belirtilen kelimelerin dizisi olarak gösterilir. w_t, ayrık bir t zamanında söylenmiş belli bir kelimedir. Kelimelerin dizisi söylenen sesli ifade ile bağlantılıdır ve bu sesli ifade X olarak gösterilen akustik sesler dizisidir (Becchetti & Ricotti, 1999). Sesin sürekliliğine göre X değişmektedir. Bunlar izole, bağlı ve sürekli konuşma tanımadır. İzole konuşma tanımada kullanıcının tek

kelimelik bir girdi yapması beklenir. Bağlı konuşma tanımada kullanıcı kelimeler arasında mesafe bırakmalıdır. Sürekli konuşma tanımada kullanıcının gerçek zamanlı konuştuklarının tanınması beklenir.

2.4.1 İzole Konuşma Tanıma

İzole yani ayrışık kelime tanıma sistemi, kısa aralıklarla seslendirilen kelimelerin tanınması işlemidir. İzole kelime tanıma sistemlerinde konuşmacı tarafından seslendirilen sözcükler arasında belirli süre ile boşluk olmalıdır. Boşluklar arasında seslendirilen kelimeler birbirinden bağımsız olarak tanıtılmalıdır. Sonrasında bu kelimeler analiz edilerek, sistem üzerindeki daha önceden oluşturulmuş modellerle kıyaslanır.

Choudhary ve ark. (Choudhary ve ark., 2013), izole ve bağlantılı Hintçe dili kelimeleri için Otomatik Konuşma Tanıma uygulaması gerçekleştirmişlerdir. Projelerinde, istatistiksel bir yaklaşım olan HMM temelli HTK (Hidden Markov Model Toolkit)'yı kullanılmışlardır. Başlangıçta sistem, 100 farklı Hintçe sözcük için eğitilmiştir. Sonuç olarak izole kelimeler için %95, bağlı kelimeler için %90 oranında doğruluk gözlemişlerdir.

Cai ve ark. (Cai ve ark., 2016), Çince şarkılardaki izole şarkı sözlerini tanımak için derin öğrenme tiplerinden derin inanç ağları (deep belief network) uygulayan ve bazı ilerlemeler kaydeden bir Çin şarkı sözü veri tabanı oluşturmuşlardır. Deney sonuçlarında, tanıma hassasiyeti yaklaşık %45 olmuştur.

Imtiaz ve Raja (Imtiaz & Raja, 2016), otomatik konuşma tanıma (ASR) sistemi akustik konuşma sinyallerini kelimelerin dizisine dönüştürmek olarak tanımlayarak, MFCC, DTW ve K-En Yakın Komşu (KNN) teknikleri kullanılarak izole sözcük yapısına dayanan ASR sisteminin bir yaklaşımını sunmuşlardır. Konuşma sinyallerinin belirgin özelliklerini yakalamak için kullanılan Mel-Frekans ölçeği ile konuşma özellikleri MFCC kullanılarak çıkartılmıştır. DTW, konuşma özelliği eşlemesi için uygulanmıştır. KNN sınıflandırıcı olarak kullanılmıştır. Deney düzeneğinde, beş konuşmacıdan toplanan İngilizce kelimeler bulunmaktadır. Bu kelimeler, akustik olarak dengeli, gürültülü olmayan bir ortamda söylenmiştir. Önerilen ASR sisteminin deneysel sonuçları, karışıklık matrisi adı verilen matris formunda elde edilmiştir. Bu araştırmada elde edilen tanıma doğruluğu %98.4 olmuştur.

2.4.2 Bağlı Konuşma Tanıma

Bağlı konuşma tanıma sisteminde konuşmacı seslendirdiği sözcükler aralarında kısa boşluklar bırakmalıdır. Bu sistemlerden sonraki evre konuşmacı tarafından seslendirilen sözcüklerin aralarında beklemediği sürekli konuşma tanıma sistemidir (Ghai & Singh, 2012).

Young ve ark. (Young, 1989), bağlı konuşma tanıma sistemlerinde basit bir kavramsal model anlatmışlardır. Çalışmalarında, farklı bağlı kelime algoritmaları, aynı kavramsal çerçeve içerisinde basitçe ağ topolojisini değiştirerek temsil edilebilirliği, dil bilgisel sınırlamaların uygulanmasının basitliği ve tüm yapının asıl alttaki kalıp eşleştirme teknolojisinden bağımsız olması gibi avantajları üzerinde durmuşlardır.

Gorthi ve ark. tasarladıkları sistemde (Gorthi ve ark., 2016), kullanıcılar arasında bağlı bir sesli veya görüntülü arama algılaması ve kısa bir medya örneği kaydetme özelliği bulunmaktadır. Tasarladıkları sistemde konuşma tanıma, aramanın ne zaman bağlandığını belirlemek ve medya örneğinin ses kısmının içeriğini kopyalamak için kullanmışlardır. Kaydedilen medya örneği ve yazılmış içerikler, bir kullanıcının daha sonraki bir noktaya referans verebilmesi için bağlı sesli veya görüntülü arama ile ilişkilendirmişlerdir. Tasarladıkları sistem ayrıca, kopyalanan içeriğe bağlı olarak bağlı sesli veya görüntülü görüşmenin katılımcılarıyla ilişkili iletişim bilgilerini oluşturmayı veya düzenlemeyi önermektedir.

El Maghraby ve ark. (El Maghraby ve ark., 2016), konuşma tanıma alanında yaptıkları çalışmada, tanıma performansını artırmak için hem akustik hem de görsel konuşma bilgisini kullanan İngilizce için bağlı kelimelerle sesli görsel konuşma tanıma sistemi kurmayı amaçlamışlardır. MFCC'yi konuşma dosyalarından ses özelliklerini çıkarmak için kullanılmışlardır. Elde ettikleri özellikleri, kelime düzeyinde akustik modeller kullanarak HMM parametrelerini eğitmek için kullanmışlardır. Önerilen yaklaşımda sürekli İngilizce sesli komutlar içeren görsel-işitsel tanıma sistemi için mevcut en büyük veri tabanlarından bir tanesi olan GRID cümle veri tabanına ilişkin bir ön deneyle göstermişlerdir. Dilbilgisi tabanlı sözcük tanıma sistemi genel konuşmacılar için başarı oranında %3.9 artma gözlemlemişlerdir.

2.4.3 Sürekli Konuşma Tanıma

Sürekli konuşma tanıma sistemi kelimeler arasında ara verilmeden tanımayı amaçlar. Sürekli konuşma tanıma sistemi içerisinde söylenen kelimenin ne zaman söylendiği ya da ne zaman bitirildiği bir sorun teşkil etmez. Kelimeler gerçek zamanlı olarak tanınırlar. Bu sistem içerisinde, konuşma esnasında ki telaffuzlar ve değişkenler başlıca sorunlardır (Aydın, 2005). Sürekli konuşma tanımanın en büyük avantajı, konuşmacı beklemeden doğal bir biçimde konuşur. Bu tip konuşma tanıma, insandan bilgisayara doğru giden ses ile haberleşme arayüzü olabilir.

Valíček (Valicek, 2017), Lehçe, Slovakça, Rusça ve Macarca dillerinde sürekli konuşma tanıma sistemi için dil modelleme sistemi tasarlamıştır. Serbestçe temin edilebilen kaynakları kullanarak n-gram dil modelleri oluşturulmuş ve yeni kelimeler transkripsiyonuna odaklanılarak telaffuz sözlükleri oluşturmak için bir prosedür tasarlamıştır. Dil modelinin oluşturulması SRILM Toolkit'i kullanılarak yapılmıştır. Bu dillerin her biri için metin kütüphaneleri bulunmuştur. Çalışmanın çıktısı, metin kütüphaneleri işleme yöntemi ve söz konusu yöntemin uygulanmasıdır. Sistemin testinde, dile bağlı olarak elde edilen sonuçlar %13-41 aralığında olmuştur.

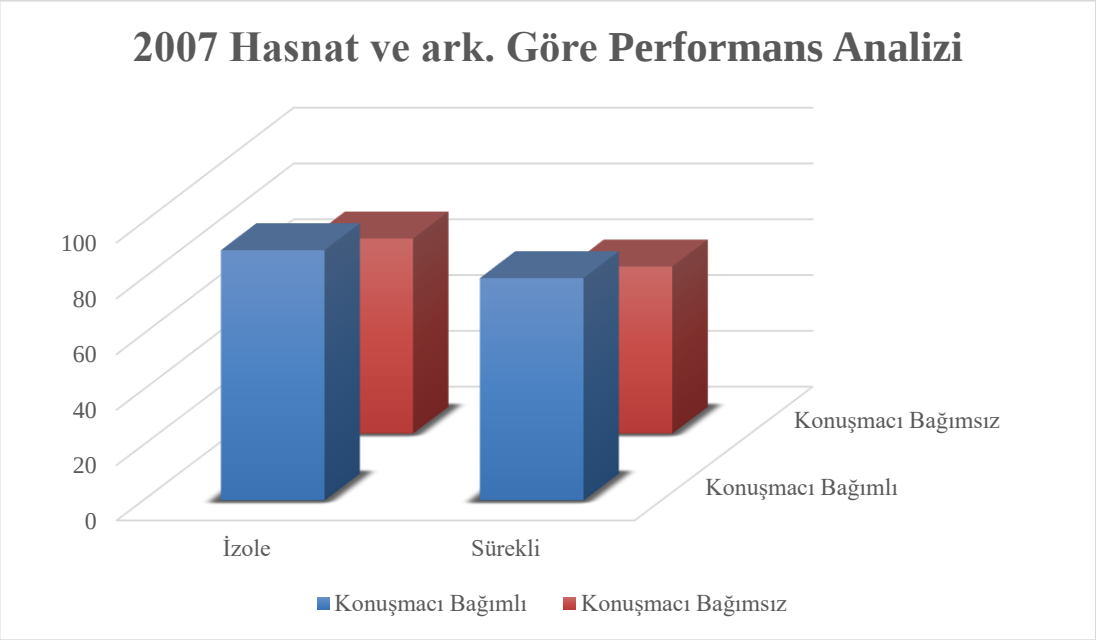
Sallaum ve ark. (Salloum ve ark., 2017), birçok çalışma grubunu içeren çok adımlı bir döngüyü kullanarak büyük ölçekli tıbbi sürekli konuşma tanıma aracının doğruluğunu devamlı olarak geliştirmek için bir yöntem önermişlerdir. Kullandıkları ASR sistemini, MFCC, GMM-HMM ve yüzlerce saat tıbbi dikte sesi konusunda eğitilmiş bir DNN tabanlı akustik modelden oluşturmuşlardır. Test setini ise, yaklaşık 180 doktorun İngilizce konuşan 20 saatlik diktelerinden oluşturulmuşlardır. Başlangıçta 100 milyondan fazla belirteç üzerinde eğitilmiş olan bir tıbbi dikte sisteminin Sürekli İyileştirme Döngüsü'nü kullanarak hata oranının %34.1'den %10.4'e yükseltilebileceğini göstermişlerdir.

Alonso ve ark. (Alonso ve ark., 2017), konuşma duygusu tanımanın, psikoloji, psikiyatri ve duygusal bilgisayar teknolojisi gibi alanlarda insan-bilgisayar etkileşimi uygulamalarında büyük bir potansiyele sahip olduğunu belirterek, konuşma sırasında duygusal değişiklikler yapılan uzun vadeli konuşma örneklerinde sürekli izleme için duygusal sıcaklık stratejisinin kullanılmasını araştırmışlardır. Sırasıyla eylemli ve kendiliğinden konuşma kullanarak, dil ve cinsiyet üzerine bağımlılık ve bağımsızlık analiz edilmiştir. Davranış koşullarında yaklaşım, %67-97 arasında doğruluk ile elde

edilmiştir. Sürekli duygu tanıma konusundaki daha önceki çalışmalarla karşılaştırıldığında, %9 daha yüksek bir oranda ortalamada iyileşme gözlemlemişlerdir.

2.4.4 Değerlendirme

Hasnat ve ark. (Hasnat ve ark., 2007) yaptığı çalışmaya göre hazırlanan aşağıdaki tablo izole ve sürekli konuşma tanımanın, konuşmacı bağımlı ve konuşmacı bağımsız konuşma tanıma türlerine göre performansını göstermektedir. Bu grafiğe göre izole konuşma tanıma, sürekli konuşma tanımaya göre daha yüksek oranda başarı performansına sahiptir.



Şekil 6. Sesin sürekliliğine göre Performans Analizi

Doğal bir konuşma anında bütün kelimeler arasında duraklama olmaz. Sürekli konuşma tanıma sistemi, gerçek zamanlı olduğundan içerisinde söylenen kelimenin ne zaman söylendiği ya da ne zaman bitirildiği bir sorun teşkil etmez. Ancak, izole kelime tanıma sistemlerinde kelimeler arasında duraklamalar vardır. Böylelikle bu sistem kelimeler arasındaki sınırlar ile uğraşmaz. Sürekli konuşma tanıma ile izole kelime tanıma sistemleri arasındaki ara evre olarak görülen bağlı konuşma tanıma sisteminde ise konuşmacı seslendirdiği sözcükler aralarında kısa boşluklar bırakmalıdır. Teknolojik gelişmeler doğrultusunda kullanıcıya hızlı yanıt verebilme ve diğer türlere

göre kullanımının daha kolay olması sebebiyle bu çalışmada sürekli (gerçek zamanlı) konuşma tanıma sistemi önerilmiştir.

2.5 Metne Göre Konuşma Tanıma

Konuşma tanıma sistemleri, ihtiyaç doğrultusunda metne bağlılığı baz alınarak iki gruba ayrılır: metne bağlı sistemler ve metinden bağımsız sistemler. Bu ayrım ile sistemin uygulanmasında kullanılan teknikler ve sistemin kullanıldığı alanlar değişir. Metin bağımlı ve metin bağımsız sistemlerde eğitim seti metne bağlıdır. Bununla birlikte metin bağımlı konuşma tanıma sistemlerinde test aşaması da eğitim verisi gibi aynı metin kütüphanesine bağlıdır. Fakat metin bağımsız konuşma tanıma sistemleri test aşamasında eğitim setinden türetilen söz dizileri kombinasyonlarını da tahmin edebilmektedir.

2.5.1 Metne Bağımlı Konuşma Tanıma

Metne dayalı konuşma tanıma sistemlerinde kullanılan test verisi, eğitim verisi ile sınırlı tutulur. Bu tanıma modelinde, sistem eğitim aşamasında kullanılan kelimelerin farklı seslendirilmeleri ile test edilirler (Ghai & Singh, 2012).

Larcher ve ark. tarafından (Larcher ve ark., 2014), farklı sürelerde ve sözlü kısıtlamalar altında metne bağımlı konuşmacı doğrulama sistemlerini değerlendirmek üzere tasarlanan RSR2015 veri tabanı, Singapur'daki Bilişim Araştırmaları Enstitüsü'nde (Institute for Infocomm Research, I^2R) Human Language Technology (HLT) bölümü tarafından toplanıp piyasaya sürüldü. 151 saatten fazla konuşma verisi, mobil cihazlar kullanılarak kaydedilen bu çalışma iyi performans göstermiştir.

Daoerji ve Guanglai (Daoerji & Guanglai, 2016), otomatik konuşma tanıma (ASR) görevlerinde üstün performans gösteren HMM-derin sinir ağları (Deep Neural Network) hibrit mimarilerini kullanan geniş bir kelime haznesi Moğolca çevrimdışı el yazısı tanıma sistemi önermişlerdir. Önerilen modelin geçerliliğini doğrulamak için, eğitim setinde 100.000 el yazısı, 5.000 test seti ve 14.085 test seti II içeren MHW veri tabanı kullanılarak kapsamlı deneyler gerçekleştirmişlerdir. Ham resim pikselleri üzerinde eğitilmiş olan DNN-HMM, Test seti I üzerinde %97.61 doğrulukla ve Test seti II üzerinde %94.14 hassasiyetle en iyi performansı vermiştir.

Donaj ve Kacic (Donaj & Kačič, 2017), biçim-söz dizili etiketlemeyle elde edilen verileri kullanan dil modelleri oluşturmak için bir yöntem sunmuşlardır. Veriler, tahmin edilen metne dayalı olarak, çalışma zamanında belirlenmiştir. İki geçişli bir tanıma algoritmasında bağlam bağımlı bir model kullanıldığında, genel konuşma tanıma doğruluğunda %1,73 oranında iyileşme sağlamıştır.

2.5.2 Metinden Bağımsız Konuşma Tanıma

Metinden bağımsız konuşma tanıma sistemleri, modelin eğitimi esnasında kullanılan sözcüklerin dışında başka kombinasyonlara da yanıt verebilmektedir. Bu sistemlere örnek vermek gerekirse, sistem, "on" ve "beş" kelimelerini tanıyorsa, o halde "on beş" kelimesini de tanımalıdır (Dede, 2008).

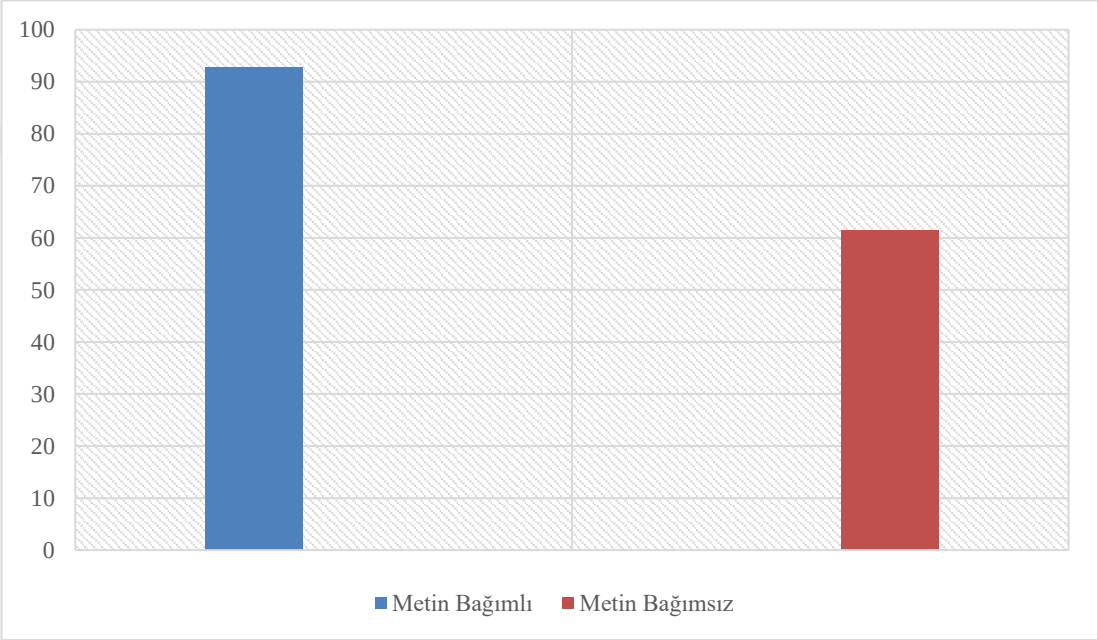
Furui (Furui, 1991), konuşmacı tanıma / doğrulama yöntemlerinin metin bağımlı ve metin bağımsız yöntemlere ayrılabileceğini söylemiştir. Metne bağımlı konuşmacı doğrulama teknikleri pratik uygulama için başarımı yüksek olsa da, metin bağımsız tekniklerin halen başarımda yüksek seviyede olmadığını söylemiştir.

Zhou ve ark. (Zhou ve ark., 2001), stres veya duygu ile ortaya çıkan değişkenliğin, konuşma tanıma doğruluğunu önemli ölçüde azaltabileceğini stres varlığını saptamak veya değerlendirmek için kullanılan tekniklerin konuşma tanıma sistemlerinin sağlamlığını geliştirmeye yardımcı olabileceğini göstermişlerdir. Lineer Olmayan Teager (1980) enerji operatöründen (TEO) elde edilen üç yeni özellik, stres sınıflaması için araştırılmıştır. MFCC'nin esas olarak daha iyi performans sergilediğini göstermişlerdir. TEO tabanlı özelliklerin performansı, metin bağımlı ve metin bağımsız modellerde korunurken, geleneksel özelliklerin performansının metin bağımsız modellerde azaldığı gösterilmiştir.

Shipra ve Chandra (Shipra & Chandra, 2016), elde ettikleri konuşma özelliklerini, gürültülü ortamlarda metin bağımlı ve metin bağımsız vakalar için HMM sınıflandırıcısı ile Hintçe sesli sınıflandırmayla karşılaştırmışlardır. Sınıflandırıcı olarak HMM özellikleri tanıma doğruluğunun, Hintçe ünlüler sınıflandırma görevi için yaklaşık %8'lik bir iyileşme sergilediğini göstermişlerdir.

2.5.3 Değerlendirme

Azim ve arkadaşları (Azim ve ark., 2016), trifon (üç fonem dizisi) HMM'lerini bağlamak için gerekli olan bir Arapça fonetik karar ağacı önermişlerdir. Önerilen karar ağacına dayanan deney sonuçları, aynı eğitim ve test setlerini kullanan geleneksel metin bağımsız modellerle karşılaştırıldığında başarımın daha yüksek olduğunu göstermişlerdir. Önerilen yaklaşımın elde ettiği maksimum metin bağımlı tanıma doğruluğu %92.8 iken metin bağımsız HMM'ler kullanılarak test edildiğinde %61.5 seviyesinde olduğunu gözlemlemişlerdir.

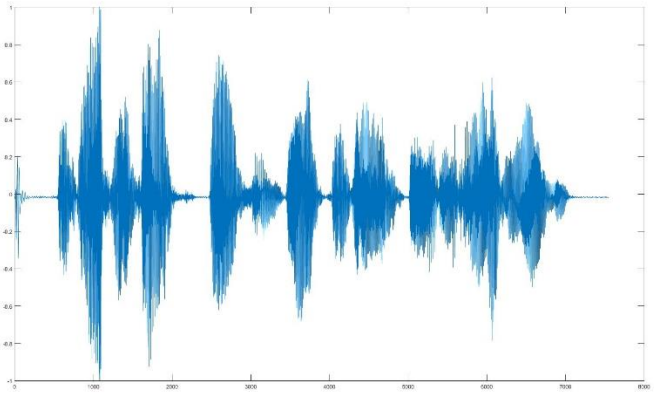


Şekil 7. Metin bağımlılığına göre çalışma performansı (Azim ve ark., 2016)

Yukarıdaki grafikte yapılan çalışmanın (Azim ve ark., 2016) performans sonuçları gösterilmiştir. Yapılan uygulamalara bakıldığında da metin bağımlı sistemlerin başarım oranı daha yüksek olmuştur. Önerilen modelde, kullanıcı konuşmaları, yazısıyla etiketlenmeli ve sistem kendini tanıtılan her yeni konuşmayla metin bağımlı olarak geliştirmelidir.

3 KONUŞMA TANIMA TEKNİKLERİ

Konuşma, insanlar arasında hızlı, etkin ve çok yönlü bir iletişim aracıdır (Baygün, 2006). Konuşma içerisindeki bilgiler, karmaşık bir biçimde kodlanmıştır ve insanlar tarafından şifresi çözülebilmektedir. Bu insan kabiliyeti, araştırmacılara bu yeteneği taklit edecek sistemleri geliştirmeye ilham kaynağı olmuştur. Ses bilgisi uzmanlarından mühendislere kadar birçok araştırmacı, konuşma sinyalindeki bilgileri çözmek için çeşitli alanlarda çalışmaktadırlar. Bu alanlara, konuşulanların sese göre belirlenmesi, konuşulan dilin keşfedilmesi, konuşmanın aktarılması, konuşmanın tercümesi ve konuşmanın tanınması örnek olarak verilebilir.

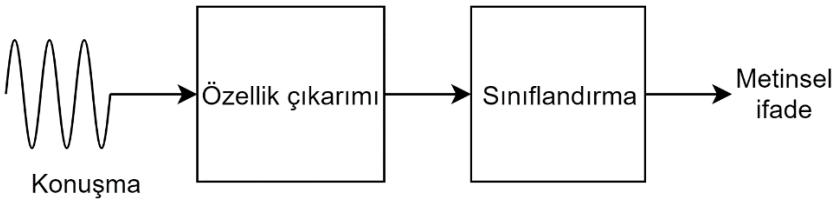


Şekil 8. Konuşma Sinyali Örneği

Konuşma tanıma, kişinin mikrofon ya da benzer bir donanıma ne söylediğini tanımlama ve anlamını metin, resim veya herhangi bir olay gibi gerekli herhangi bir biçimde yansıtma sürecidir (Singh K. , 2016). Konuşma tanıma, birçok araştırmacının uzun yıllardır üzerinde çalıştığı bir alandır. Bu alanda, konuşmacı dil bilgisi mesajı ile ilgilenmektedir. Konuşmacı tarafından dilsel, fizyolojik ve çevresel birtakım faktörlere bağlı olarak konuşmada değişkenlikler gözlenebilir. Böylece araştırmacılar, bir insan yeteneği olmasına karşın konuşmadan bilgi çıkarmanın basit bir süreç olmadığını tecrübe etmişlerdir (Adami, 2010). Bu tecrübeler ile araştırmacılar, konuşma sinyalinden ilgili bilgileri güvenilir bir şekilde çıkartmaya çalışmaktadırlar.

Konuşmanın yazıya çevrilmesi için sesli ifadelerin, bilgisayar tarafından tanıma sürecine dâhil edilmeleri gerekmektedir. Bu amaçla sesli ifadelerin bir mikrofon aracılığıyla sinyallere dönüştürülmesi, sayısal olarak işlenen bu sinyallerin gerekirse filtrelenmesi, etiketlenmesi (örneğin sesler, fonemler, kelime ya da kelime grubu olarak) ve tanıma işlemlerine taban oluşturacak sınıflandırma teknikleriyle parametrik yapılar ya da yalın modellerle ifade edilen biçimlere dönüştürülmesi gerekmektedir (Yalçın, 2008). Böyle etkili konuşma tanıma sistemlerine günümüzde ihtiyaç artmaktadır. Bu sistemler ile uygulama alanına göre kullanım kolaylığı, veri toplama hızı, hareket serbestliği ve uzaktan veri giriş imkânı sağlanabilir.

Sesin dalga şekli incelendiğinde, fiziksel sistem zamana bağlı değiştiğinden, sesin dalga şekli de zamanla değişir. Böylelikle ses, kısa süreler boyunca benzer akustik özellikler gösteren ses parçalarına ayrılabilir (Singh ve ark., 2012). 1998 yılında Kuş'un yaptığı çalışmaya göre (Kuş, 1998), ses sinyallerinin zamana bağlı dalga şekillerine bakılarak, sinyal periyotları, yoğunlukları, süreleri ve her bir ses parçasının sınırları tespit edilebilir. Tespit edilen her bir ses parçası konuşma olarak adlandırılırsa, bu konuşmalardan bilgi elde etme, özellik çıkarımı aşamasıdır. Sonraki aşama ise özellik vektör dizisinin planlanması ve sınıflandırılmasıdır.

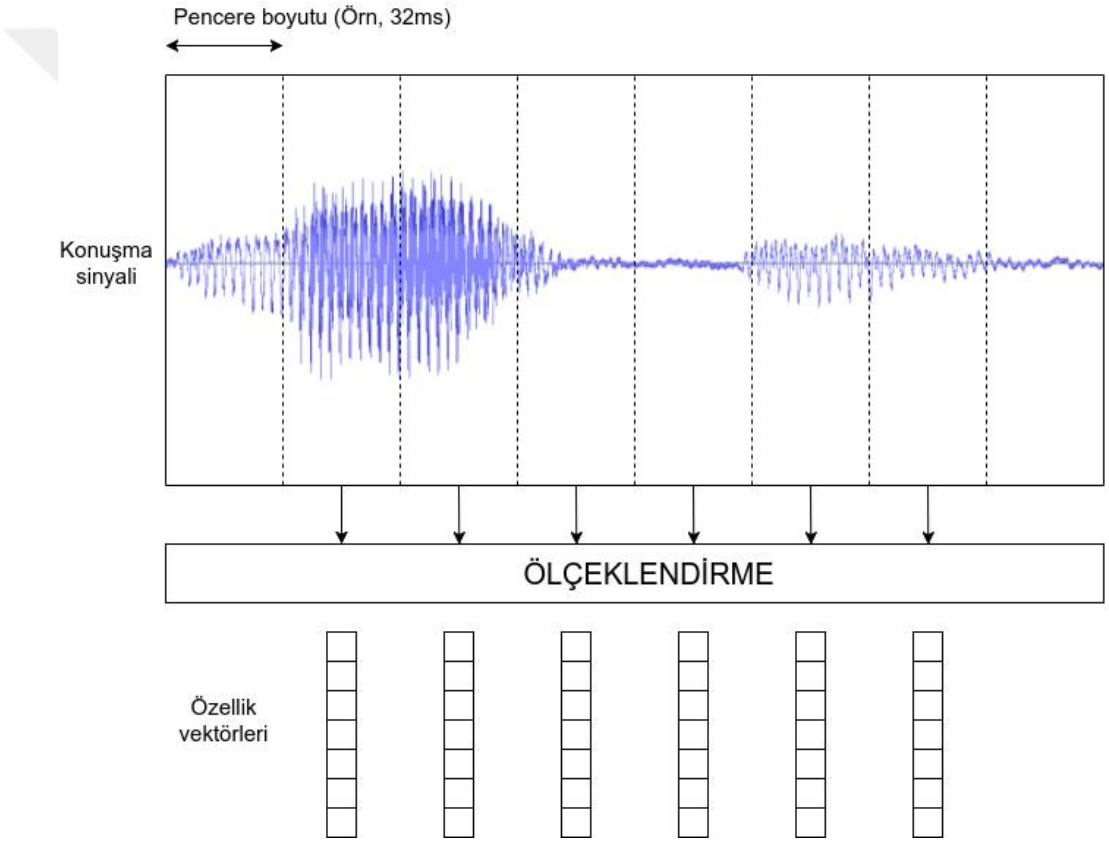


Şekil 9. Teknikler ile konuşmanın yazıya çevrilmesi

3.1 Özellik Çıkarımı

Makine öğrenmede, örüntü tanımada ve görüntü işleme alanlarında kullanılan ve boyutluluğu azaltmakla ilgili olan özellik çıkarımı yöntemi, ölçülen ilk verilerle başlayarak bilgilendirici ve gereksiz olmaması amaçlanan türetilen değerleri (özellikleri) oluşturur (Çakır, 2017). Sonrasında öğrenme ve genelleme aşamalarını kolaylaştırır ve ileriki süreçlere öncü olur. İlk özelliklerin bir alt kümesine rastlanması, özellik seçimi olarak adlandırılır. Seçilen özelliklerin girilen verilerden ilgili bilgileri içermesi beklenir, böylece arzu edilen görev, başlangıç verisi yerine bu azaltılmış gösterimi kullanarak gerçekleştirilebilir.

Bir konuşma sinyalinin, zaman ekseninde dalga formu tüm işitsel bilgileri taşır. Ses bilimsel açıdan, dalga şeklinin kendisi temelinde çok az şey söylenebilir. Bununla birlikte, matematik, akustik ve konuşma teknolojisindeki geçmiş araştırmalar, doğru yorumlanırsa bilgi olarak kabul edilebilecek verileri dönüştürmek için birçok yöntem sağlanmıştır. Gelen verilerden istatistiksel olarak ilgili bazı bilgileri bulmak, ses sinyalindeki her bir bölümün bilgisini nispeten az sayıda parametreye veya özelliklere indirgemek için mekanizmalara sahip olmak önemlidir. Bu özellikler, her bölümleri diğer benzer bölümlerin özelliklerini karşılaştırarak gruplandırılabilecek karakteristik bir şekilde tanımlanmalıdır (Shrawankar & Thakare, 2013). Konuşma sinyalini parametreler açısından tanımlamak için geliştirilen yollar vardır.



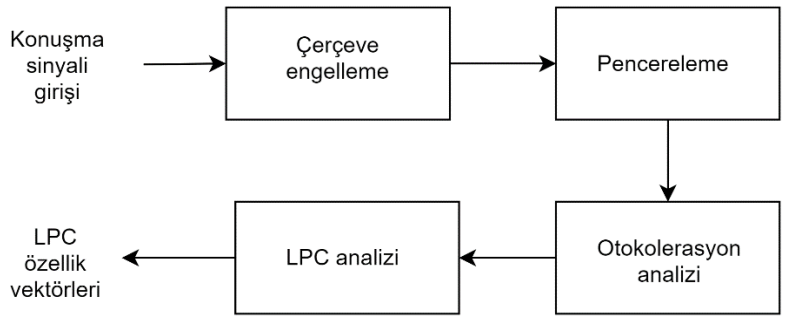
Şekil 10. Özellik çıkarımı örneği

Konuşma tanıma sistemleri ses sinyallerinin analizinden faydalanır. Ses dalgasının içerdiği frekans ve genlik değerleri her bir fonem için farklılık göstermektedir. Ses sinyalinin analizi sonucunda bu frekans ve genlik bilgilerini içeren özellik vektörleri oluşur. Bir özellik vektörleri genelde her bir kısa zaman aralığındaki (10 ms) bir ses sinyali penceresinden (20~30 ms'den) hesaplanır. Söylenen kelime bu özellik vektörlerinin bir dizisi olarak gösterilir. Sonraki aşamada bu özellik vektörleri

konuşma tanıma sistemine giriş olarak verilir. Ses sinyallerinden, konuşmanın özelliklerinin çıkartılması ve ses sinyallerinden elde edilen örneklerin veri sayısı bakımından fazla olması konuşma tanıma sisteminde, daha verimli sonuçlar çıkartılmasını sağlar. Özellik çıkarıma işlemi, konuşma sinyalini insanın duyma sistemine benzeterek duyum olarak anlamlı parametre vektörlerine çevirir. Özellik çıkarma işlemi, ses tanıma sistemlerinin ilk işlem bloğudur.

3.1.1 Doğrusal Öngörülü Kodlama (Linear Predictive Coding (LPC))

Ses özelliklerinin modellendiği sayısal ses işleme tekniği olan doğrusal öngörülü kodlama (LPC), doğrusal öngörülü modelin bilgisini kullanarak, konuşmanın sanal bir sinyalinin spektral zarfını sıkıştırılmış formda göstermek için çoğunlukla ses sinyali işleme ve konuşma işlemlerinde kullanılan bir araçtır (Deng & O'Shaughnessy, 2003). Yaygın olarak kullanılan konuşma analizi tekniklerinden birisidir ve kaliteli konuşmayı düşük bir bit hızında kodlamak için kullanılır. Ses örneğinde, kendinden önceki ses örneklerinin doğrusal kombinasyonu kullanılarak konuşma parametreleri için son derece doğru tahminlerini sağlar. Gerçek ses örnekleri ile tahmin edilen örnekler arasındaki hata minimuma indirilerek öngörü katsayılarından oluşan parametre değerleri elde edilir.



Şekil 11. LPC adımları

Moonasar ve Venayagamoorthy'nin yaptığı çalışmada (Moonasar & Venayagamoorthy, 2001), konuşmacı doğrulama sistemlerinin sonuçlarının nasıl iyileştirilebileceğini açıklamışlardır. Örüntü sınıflandırıcısı olarak denetlenmiş Öğrenme Vektörü Nicemleme (LVQ) sinir ağı kullanımını göstermişlerdir. Linear Predictive Coding (LPC) tekniği ile tanınacak olan konuşmacıların sayısını etkilemede kullanmışlardır. LPC ile ANN tekniğinde %70 doğruluk elde etmişlerdir.

Sunny ve ark. (Sunny ve ark., 2012), Doğrusal Tahmin Edici (Öngörülü) Kodlama (LPC) özellik çıkarımı tekniğiyle konuşmacıdan bağımsız olarak konuşulan izole kelimeleri tanımak için bir çalışma yapmışlardır. Çalışmalarında ses sinyallerini doğrudan mikrofon aracılığıyla almışlar ve daha sonra LPC tekniğini kullanarak özellik vektörlerini çıkarmışlardır. Bir Hindistan dili olan Malayalam dili sözcüklerini tanımak için seçmişlerdir. Önerilen yöntemde, her biri 20 izole kelime söyleyen 50 konuşmacı için uygulanmıştır. LPC tabanlı yöntem ile %81.2 doğruluk elde etmişlerdir.

Cahavan ve Sable (Chavan & Sable, 2013), konuşma tanıma için özellik çıkarma tekniklerini tartışmışlar ve konuşma tanıma için genel bir bakış açısı sunmuşlardır. Çalışmalarını, özellik çıkarma ve özellik tanıma olmak üzere iki farklı kısma ayırmışlardır. Özellik çıkarmada LPC tekniği ile özellik tanımada DTW tekniğinde %69, HMM tekniğinde %77 doğruluk oranı elde etmişlerdir.

3.1.2 Mel Frekanslı Kepstral Katsayılar (Mel Frequency Cepstral Coefficients (MFCC))

En güçlü konuşma analizi tekniklerinden biri olan MFCC kullanım oranı yüksek olan ses işleme tekniklerinden biridir. İnsan ses algılamasını taklit eden ve FFT tabanlı olarak hesap edilen bir sayısal analiz tekniğidir. Bu teknik analiz ile elde edilen sayılar MFC katsayıları (MFCC) olarak adlandırılır. Ses işlemede, Mel frekans kepstrum (MFC), bir frekansın doğrusal olmayan Mel skalasında kısa süreli güç spektrumunun bir gösterimidir. Mel frekanslı kepstral katsayıları (MFCC), bir MFC'yi topluca oluşturan katsayılardır. Sesin bir tür kepstral gösteriminden türemiştir. MFC'de frekans bantları, insan ses sistemi tepkisini yaklaşık olarak Mel ölçeğinde eşit aralıklarla yerleştirilmiştir. Böylece ses sıkıştırmasında daha iyi bir ses temsili sağlanabilir.

Cahavan ve Sable (Chavan & Sable, 2013), konuşma tanıma için özellik çıkarma tekniklerini tartışmışlar ve konuşma tanıma için genel bir bakış açısı sunmuşlardır. Çalışmalarını, özellik çıkarma ve özellik tanıma olmak üzere iki farklı kısma ayırmışlardır. Özellik çıkarmada MFCC tekniği ile özellik tanımada DTW tekniğinde yaklaşık olarak %90, HMM tekniğinde %90-96 arası doğruluk elde etmişlerdir.

Joshi ve Cheeran (Joshi & Cheeran, 2014), konuşma tanıma için özellik çıkarma tekniklerini tartışmışlar ve konuşma tanıma için genel bir bakış açısı sunmuşlardır.

Çalışmalarını, özellik çıkarma ve özellik tanıma olmak üzere iki farklı kısma ayırmışlardır. Özellik çıkarmada MFCC tekniği ile özellik tanımada ANN tekniğinde yaklaşık olarak %80 doğruluk elde etmişlerdir.

Ananthi ve Dhanalakshmi (Ananthi & Dhanalakshmi, 2014) konuşma tanıma yaklaşımını işitme engelli insanlar için daha yararlı olabilecek konuşma sözlüğünden gelen metni tanımayı amaçlamada kullanmışlardır. Sınıflandırmada, konuşma tanıma sistemi için yaygın olarak kullanılan teknikler olan Destek Vektör Makinesi (SVM) ve Saklı Markov Modeli (HMM)'ni kullanmışlardır. Özellik çıkarmada Mel Frekans Kepstral Katsayıları (MFCC) ile akustik özellikleri çıkarmışlardır. Sistem, HMM için %98.92 SVM için %91.46'lık bir doğruluk sergilemiştir. Ayrıntılı analiz sonucu, MFCC ile HMM'nin SVM gibi diğer modelleme tekniklerinden daha iyi performans gösterdiğini belirtmişlerdir.

3.1.3 Değerlendirme

Tablo 1. LPC ve MFCC tekniklerinin sınıflandırma teknikleri ile uygulandığındaki başarım oranı

Teknik	***Başarım Oranı***	***Referans***
LPC ve ANN	%70	(Moonasar ve Venayagamoorthy, 2001)
LPC ve VQ, HMM	%96,5	(Rabiner ve ark., 1993)
LPC ve DTW	%69	(Chavan ve Sable, 2013)
MFCC ve DTW	%91,46	(Ananthi ve Dhanalakshmi, 2014)
MFCC ve HMM	%98,92	(Ananthi ve Dhanalakshmi, 2014)
MFCC ve ANN	%80	(Joshi ve Cheeran, 2014)

Tablo 1'de görüleceği üzere sık kullanım oranı da göz önüne alındığında MFCC tekniği LPC tekniğine göre kıyaslandığında başarım oranı daha yüksektir (Bu sonuçlar neticesinde deneysel çalışmada MFCC özellik çıkarımı tekniği önerilmiştir). MFCC'ler yaygın olarak, bir telefonla konuşulan numaraları otomatik olarak tanıyabilen sistemler gibi konuşma tanıma sistemlerinde (Ganchev ve ark., 2005), tür sınıflaması, ses benzerlik ölçüleri gibi uygulamalarda da yaygın olarak kullanılır (Müller, 2007).

Konuşma tanımada, mel-frekans kepstrum (MFC) bir konuşmanın kısa vadeli güç spektrumunun (tayfının) bir gösterimi, frekansın doğrusal olmayan bir mel skalasında bir log güç spektrumunun doğrusal bir kosinüs transformasyonuna dayanır. MFCC'ler genellikle aşağıdaki adımlardan (ön işlemeden geçirildiği farzedilerek) türetilmektedir (Sahidullah & Saha, 2012).

1. Pencerelenmiş bir sinyalin Fourier dönüşümü gerçekleştirilir.
2. Birinci adımdan elde edilen spektrum güçleri üçgen örtüşen pencereler kullanılarak mel ölçeğine eşlenir.
3. Her bir mel frekansındaki güçlerin log'ları alınır.
4. Mel log güçleri listesinin ayrık kosinüs dönüşümü bir sinyal olduğu varsayılarak alınır.
5. MFCC'ler ortaya çıkan spektrumun genlikleridir.

Bu adımlar üzerinde; ölçeği eşleştirmek için kullanılan pencereleme tekniğinde veya aralığında farklılıklar olabilir, birinci (delta) ve ikinci (delta-delta) dereceden çerçeveleme katsayıları farkı olabilir. Bu işlemler ile elde edilen MFCC değerleri, konuşma haricindeki gürültünün etkisini azaltmada çok başarılı değildir. Bu sebeple konuşma tanıma sistemlerindeki değerleri normalleştirmek gereklidir.

3.2 Sınıflandırma

İnsanlar olarak beyinlerimiz yüzleri tanıma, sesleri tanıma gibi her gün hayatımızın her anında sınıflandırma yapmaktadır. Sınıflandırma, farklı bilgi parçalarını tanıma mekanizmasıdır. Örüntü tanımada toplanan bilgilere dayanarak kararlar veya tahminler yapmaya yarar. Toplanan bilgileri kullanım amacı için kalıpları farklı kategorilere ayırarak modeller ve sistemler tanımlamaya yarar. Bu sistemler daha sonra, konuşma tanıma, parmak izi tanıma, DNA dizilimi doğrulamaya kadar çeşitli farklı uygulamalara sahip kalıp tanıma mekanizmaları tasarlamamızı sağlamaktadır (Therrien, 1989).

Sınıflandırma, kategori üyeliği olarak bilinen gözlemleri veya örnekleri içeren bir eğitim seti temelinde yeni bir gözlemin ait olduğu bir grup kategorinin tanımlanmasıdır. Bunun bir örneği, belirli bir e-postayı spam veya spam olmayan sınıflara atamak veya belirli bir hastaya gözlenen özelliklerle (cinsiyet, kan basıncı,

belirli belirtilerin varlığı veya yokluğu vb.) açıklandığı gibi bir tanı tayin etmek olacaktır. Sınıflandırma, denetlenen öğrenmenin bir örneğidir. İlgili denetimsiz işlem, kümeleme olarak bilinir ve veriyi, doğal bir benzerlik veya uzaklık ölçüsüne dayalı olarak kategorilere ayırmayı içerir.

Yaşaroğlu, konuşma tanıma işleminin bir kelimenin söylenmesine karşılık gelen özellik vektör dizisinin planlanması ve sınıflandırılması olarak düşünülebileceğini söylemiştir (Yaşaroğlu, 2003). Sınıflandırma, konuşmadaki ses özellik parametreleri bulunduktan sonra istatistiki bir model bulunmasıdır. Bu sırada algoritmalar ile konuşmadan bilgi çıkarılır ve sınıflandırılır. Sınıflandırmada kurallar, sistemin kendini konuşma verisi ile güncellemesi ve modellerin gelişerek sonuç verisini bulması ile oluşur. Bu bağlamda sınıflandırma evresinde öğrenme kümesi ne kadar fazla olursa sistemin tanıma başarısı da o kadar artmaktadır. Önerilen konuşma tanımada sınıflandırma aşamasında, özellik vektörlerinin bulunmasından sonra, bu kelimelere karşılık gelen istatistiksel model ile veri tabanı oluşturulur. Söylenen kelimeye karşılık, tüm veri tabanı içerisinden arama yapılır ve verilen sinyale en uygun eşleşme seçilir.

3.2.1 Dinamik Zaman Bükmesi (Dynamic Time Warping (DTW))

DTW, zaman serisi analizinde hız açısından değişiklik gösterebilen iki zaman aralığı arasındaki benzerliği ölçmek için kullanılan algoritmalardan biridir (Silva & Batista, 2016). Örneğin, yürüme benzerlikleri göz önüne alındığında, bir kişi diğerinden daha hızlı yürürse veya bir gözlem sırasında hızlanma ve yavaşlama DTW kullanılarak tespit edilebilir. Bilimsel çalışmalarda DTW, video, ses ve grafik verilerinin zamansal dizilerine uygulanmaktadır. Doğrusal bir sıraya dönüştürülebilen herhangi bir veri, DTW ile analiz edilebilir. Konuşma tanıma alanında DTW, farklı konuşma hızlarıyla baş etmek için otomatik konuşma tanıma uygulaması için kullanılmaktadır. DTW konuşma tanıma alanında, konuşmacı tanıma ve çevrimiçi imza tanıma gibi uygulama alanlarında da kullanılır.

Belli kelimenin seslendirilmesinde, konuşmacıdan konuşmacıya, farklı zamanlarda, kelimenin uzun ya da kısa seslendirilmesinde farklılık oluşabilir. Fakat DTW algoritması ile bu seslendirmeler belli zaman aralığında yayılarak veya daraltılarak birbirlerine yaklaştırılır. Bu yöntem ile çalışma anında belirlenen kelimenin

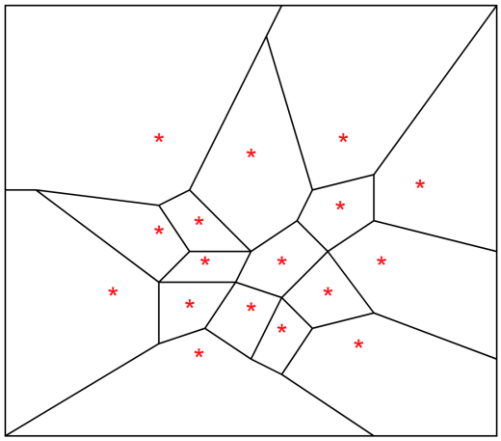
bölümlenmesi, sistem üzerinde kayıtlı bulunan kelime şablonuyla seslendirilme zamanının örtüşerek karşılaştırılması yapılabilir (Yaldır & Baygün, 2006).

Abdulla ve ark. (Abdulla ve ark., 2003), Dinamik Zaman Bükmesi (DTW) tabanlı konuşma tanıma sistemlerinde İngilizce dili üzerinde test etmişler ve konuşmacıya bağlı tanıma doğruluk oranını %85.3 olarak elde etmişlerdir.

Gawali ve ark. (Gawali ve ark., 2011), MFCC ve DTW teknikleri ile Marathi dili üzerinde veri tabanı ve izole edilmiş kelime tanıma sistemi sunmuşlardır. Veri tabanı, Marathi ünlüleri, her sesli harfle başlayan izole kelimeler ve basit Marathi cümlelerinden oluşturulmuştur. Her kelime 35 konuşmacı tarafından üç kez tekrarlanmıştır. DTW ile doğruluk oranını %73,25 olarak elde etmişlerdir.

Cahavan ve Sable (Chavan & Sable, 2013), konuşma tanıma için özellik çıkarma tekniklerini tartışmışlar ve konuşma tanıma için genel bir bakış açısı sunmuşlardır. Çalışmalarını, özellik çıkarma ve özellik tanıma olmak üzere iki farklı kısma ayırmışlardır. Özellik çıkarmada LPC tekniği ile özellik tanımada DTW tekniğinde %69 doğruluk oranı elde etmişlerdir.

3.2.2 Vektör Nicemleme (Vector Quantization (VQ))



Şekil 12. İki boyutlu vektörel sınıflama

Vektör nicemleme, vektörel uzaklık ölçümüdür. Şekil 6'da iki boyutlu bir vektör ölçümü bulunmaktadır. Kod vektörü * sembolüyle ifade edilmektedir. Kodlama bölgesi, kod vektörüne yaklaştırılan bölgeleri ifade etmektedir. Kod kitabı ise kod vektörlerin bulunduğu gruba denilmektedir (Uzunçarşılı, 2005). Bu yönteme dayalı konuşma tanıma sisteminde eğitim örüntüleri ile tüm örüntülere ilişkin özellik vektörü

çıkartılabilir. Özellik vektörleri, sayısal ses işleme yöntemlerinden elde edilebilir. Kod kitabı için her örüntünün özellik vektörleri çıkartılır. Kümeleme algoritması ile eğitilen vektörler ile her örüntünün optimum referans modeli tasarlanır. VQ sisteminin test aşamasında, test örüntüleri ile referans modeli hazırlanan kod kitabı vektörleri ile en yakın uzaklık veren kod vektörleri elde edilir. Böylece tanınmayan örüntü, belirlenen karar ölçütüne göre örüntülerden en yakın olana tayin edilir. Bu yöntem, veri miktarının indirgenmesini sağlayan kümeleme metodudur. Aslında günlük yaşantımızda yiyecek satın alırken aynı gruplardaki yiyecekler aynı bölümlere yerleştirilir. Bu da kümelemeye örnek olarak verilebilir (Çelebi & Buldu, 2014).

Rabiner ve ark. (Rabiner ve ark., 1983), sınıflandırmada VQ ve HMM teknikleriyle özellik çıkarımında LPC tekniğini, konuşmacıdan bağımsız, izole kelime tanıma uygulaması için geliştirmişlerdir. VQ ve HMM'yi tanınan sözcük dağarcığı için eğitmişlerdir. Uygulamayı (eğitim ve test) 10 kelimeli bir basamakta değerlendirmişlerdir. Eğitim için, 100 konuşmacından oluşan bir grup, her bir basamağı bir kez konuşmuştur. Tanıma doğruluğunu 100 konuşmacı test seti için %96,5 olarak elde etmişlerdir.

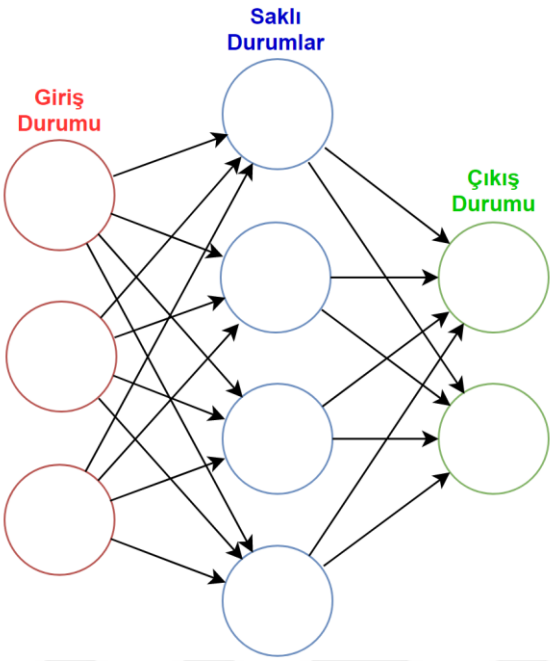
Debyeche ve ark. (Debyeche ve ark., 2007) ayrık HMM ses tanıma sistemleri için yeni bir VQ yaklaşımı önermişlerdir. Bu yeni yaklaşım HMM durumları üzerinde VQ kod kitabı bileşenlerinin optimal olarak dağıtılmasını gerçekleştirmiştir. Bu yeni yöntem Dağıtılmış VQ olarak isimlendirilmiştir.

Çelebi ve ark. (Çelebi & Buldu, 2014), MFCC ve VQ kullanarak, gürültüsüz ortamda, sesli komutlar ile gezgin araç kontrolünde %95'in üzerinde başarı elde etmişlerdir.

3.2.3 Yapay Sinir Ağları (Artifical Neural Networks (ANN))

ANN bilgiyi sınıflandırmak ve bilgiyi yorumlamanın gerektirdiği problemlerin çözümlerinde kullanılır. Konuşma tanıma esnasında özellik vektörleri çıkarılan ses sinyallerini tanıma işlemi, ANN konusuna dâhildir. ANN biyolojik olan sinirsel ağdan etkilenerek üretilmiş bir bilgi sistemidir. ANN, bilgi depolamada temel ünitelerden oluşturulan paralel olarak dağıtılan işlemcidir. Bu tip sinirsel ağlar bilgi depolamada iki nöron arasındaki bağlantı kuvvetini kullanır. ANN çokça işlemcilerden oluşmaktadır ve birbirlerine veriler ya da ağırlık taşıyan bağlantılarla bağlıdır. ANN üzerinde bilgiler, nöronların birbirleri arasındaki bağları üzerinde ağırlık değerlerinde

tutulmaktadır. ANN'nin eğitim ve testi bu ağırlık değerlerinin değişmesini sağlar. ANN sistemlerinde tüm modeller ağ üzerinde gizlidir. Ağ üzerinde konuşma tanımanın ne şekilde gerçekleştiğini bulmamız bu yüzden zorlaşır.



Şekil 13. Birbirine bağlı düğümler grubu olan Yapay Sinir Ağı

Nicholson ve ark. (Nicholson ve ark., 2000), konuşmada duygu tanımanın bugüne kadar çok az araştırmada işlendiğini, konuşmada duygu tanımanın neden önemli, uygulanabilir bir araştırma konusu olduğunu belirtmişler ve ANN kullanarak duygu tanıma için bir sistem üzerinde tartışmışlardır. Geniş bir veri tabanı kullanarak, sistemlerini konuşmacı ve metin bağımsız olarak oluşturmuşlardır. Sekiz duyguyu test ederken yaklaşık %50 oranında bir doğruluk oranı elde etmişlerdir.

Polur ve ark. (Polur ve ark., 2001), sınıflandırmada ANN tekniği ile konuşma tanıma için proje geliştirmişlerdir. MFCC ile özellik vektörleri çıkarılan bu proje neticesinde %75'lerde başarı elde etmişlerdir.

El-Ramly ve ark. (El-Ramy ve ark., 2002) ANN ile bir ses tanıma çalışması gerçekleştirmişlerdir. Arapça ses birimler üzerine geliştirilen sistem üzerinde konuşulan ifadeleri tanıma başarısı %80'lerde gerçekleşmiştir.

3.2.4 Destek Vektör Makineleri (Support Vector Machines (SVM))

SVM, istatistiki olarak öğrenmek, yapısal riskleri en aza indirmek, regresyon ve sınıflandırma yapabilmek için önerilmiştir (Vapnik, 1995). Günümüzde SVM, üç boyutlu nesne tanımada, metin sınıflandırılmasında, bilinmeyen el yazılarını tanımada, yüz tanımada, konuşmacı tanımada ve ses tanımada kullanılmaktadır (Eray, 2008).

Lin ve Wei (Lin & Wei, 2005), konuşma esnasında otomatik duygu tanımanın insan-makine etkileşiminde geniş bir uygulama yelpazesine sahip olduğunu belirtmişlerdir. Çalışmalarında beş duygusal durumu sınıflandırmak için HMM ve SVM olmak üzere iki sınıflandırma yöntemi kullanmışlardır: öfke, mutluluk, keder, şaşkınlık ve tarafsızlık. Özellik çıkarımı tekniği olarak MFCC kullanılmışlardır. HMM ile tanıma başarım oranları, kadınlar için %98.9, erkeklerde %100 ve cinsiyete bağlı olmayan vakalarda %99.5 olarak elde etmişlerdir. SVM ile sırasıyla erkek, kadın ve cinsiyete bağımsız durumlarda sırasıyla %89.4, %93.6 ve %88.9 başarım oranları elde etmişlerdir.

Pao ve ark. (Pao ve ark., 2006), insanoğlunun dünyaya nasıl tepki gösterdiğini ve birbiriyle nasıl etkileşime girdiğini araştırmada bir kişinin duygusal durumları için duygusal konuşma tanıma sistemi geliştirmişlerdir. Mandarin dili üzerinde duygusal konuşma tanımada beş duyguyu sınıflandırmak için SVM ve ANN sınıflandırma tekniklerini karşılaştırmışlardır. SVM için %84.2, ANN için %80.8 doğruluk oranı elde etmişlerdir.

Al-Haddad ve ark. (Al-Haddad ve ark., 2008) DTW ve HMM sınıflandırma teknikleriyle 'Malay' ses tanıma problemi üzerinde konuşma tanıma çalışması yapmışlardır. Son-nokta tarama, çerçeveleme, normalizasyon, MFCC ve VQ tekniklerini tanıma işleminde kullanılacak ses örneklerini işlemek için kullanmışlardır. Ses örüntülerinin tanınması aşamasında da bu iki yöntemin ayrı ayrı kullanımı durumunda; DTW ile ses örüntüleri %80.5 doğruluk oranıyla, HMM ile ses örüntüleri %90.7 doğruluk oranı ile tanınmıştır.

3.2.5 Saklı Markov Modelleri (Hidden Markov Models (HMM))

Saklı Markov modeli, gözlem dizilerinin bilinmesine karşın temel durum dizileri bilinmediğinden saklı ibaresi ile "Saklı Markov Model" olarak adlandırılır. HMM ile istatistiki süreç verileri çok iyi biçimde tanımlanabilir. HMM ile geçmişten günümüze

birçok alanda çalışma yapılmıştır. HMM, konuşma tanımada, yüz tanımada, vücut hareketleri tanımada, el yazısı tanımada, biyoinformatikte (biyolojinin bilgisayar ile incelenmesi ve işlenmesi), gen tahmininde, kripto analizinde, protein yapısı & DNA dizilimlerinde, örüntü tanımada, gizli olasılıkların hesabında ve en iyi olasılığı bulmasından dolayı yaygın olarak kullanılmıştır. Ayrıca Markov modeller, sistem olasılıksal dağılıma bağlı olarak kendi durumundan başka bir duruma geçebilir veya aynı durumda kalabilir. Markov modellerde, bulunulan durumdan meydana gelen olasılıklar, geçiş olasılıkları olarak isimlendirilir. Bununla birlikte HMM'de duruma bağlı olan geçişler görülebilir.

HMM konuşma tanımada, ses sinyallerinin en iyi şekilde parametrik olarak rastgele karakterize edilmesini sağlar. HMM iki stokastik aşamadan oluşur. İlki Markov sürecidir ve zamanla alakalı değişikliklerde kullanılır, durumların içerildiği Markov zincirini üretir. Öteki süreçler gözlemlenebilir ve özellik parametreleriyle rastgele değişkenleri içerir. HMM, dinamik zaman serilerinin modellenmesi için çok uygundur ve özellikle bol bilgi, tekrar hesap edilebilirlik özelliklerine sahip bir sinyal için güçlü bir örüntü sınıflandırma kabiliyetine sahiptir. HMM, rasgele uzun dizileri işleyebilir. Aynı zamanda HMM, çok çeşitli zaman serisi verilerini modellemek için popüler istatistiksel bir araçtır. Doğal dil işleme uygulamalarında, HMM'ler konuşmanın parçası olarak etiketleme (part of speech tagging) ve isim öbekleri bölme (noun-phrase chunking) gibi çalışmalara da büyük başarı ile uygulanmıştır (Blunsom, 2004).

Abushariah ve ark. (Abushariah ve ark., 2010), yüksek performanslı doğal konuşmacı bağımsız bir Arapça sürekli konuşma tanıma sistemi geliştirmek için bir araştırma çalışması tasarlamışlardır. Konuşma sinyallerinin özelliklerini çıkarmak için bir dizi özellik vektörü üretmek üzere MFCC tekniğini uygulamışlardır. Ardından, sistem üzerinde üçlü telefon akustik modellemesi için üç ayrı yayın durumu bulunan beş durumlu HMM kullanmışlardır. Sistem konuşma korpusunda 7 saat boyunca eğitilmiştir ve bir saatte test edilmiştir ve %93.88'lik bir başarı sağlanmıştır.

Utane ve Nalbawar (Utane & Nalbalwar, 2013) HMM model sınıflandırıcılarını kullanarak, konuşmacıların beş temel duygusal durumunu; öfke, mutluluk, üzgün, sürpriz ve tarafsız olarak tanımlayan bir çalışma yapmışlardır. Ayrıca çalışmalarında, vurgu, perde, tonlama, duraksama gibi (prosodik) dilin özelliklerini MFCC ile çıkartmışlardır ve bu özelliklere bağlı HMM sınıflandırmalarının performansını

tartışmışlardır. HMM test sonuçlarında %80'lerin üzerinde başarım sağlanmıştır. Doğru bir duygusal konuşma veri tabanı ile sistemin verimli çalışacağı belirtilmiştir.

Bhaskar ve Rao (Bhaskar & Mohana Rao, 2014), Telugu dili üzerinde HMM ve MFCC özelliklerine dayanan izole kelime konuşma tanıma sistemini geliştirmişlerdir. Eğitim için 25 konuşmacı tarafından 5 kez kaydedilen 250 Telugu (Hindistan bölgesinin bir dili) sözcüğü kullanmışlardır. Performans aşamasında, 10 farklı konuşmacı, performans değerlendirmesi için bazı sözcükleri söylemişlerdir. Ortalama %91 doğruluk oranı elde etmişlerdir.

3.2.6 Değerlendirme

DTW, ANN, SVM, HMM gibi tekniklerin dışında Bulanık Sinir Sistemleri (FNS) ve Gauss Karma Modelleri (GMM) de konuşma tanımanın sınıflandırma aşamasında kullanılmaktadır. Üyelik fonksiyon tekniği olan FNS gibi sistemlerde ağ yapısı tasarımı, üyelik fonksiyonlar üretilmesi şeklindedir. Zadeh tarafından ortaya atılan FNS (Zadeh, 1965), pek iyi seçilemeyen, vasfı iyi anlaşılamayan, saf olmayan, açık şekilde görünmeyen biçimde tanımlanır. FNS, ANN ile Bulanık Mantığın karışımıyla oluşmuştur. Literatürde uygulamaları az olsa da FNS, sistem modellenmesi, tıbbi tanı koyma ve konuşma tanıma sistemi gibi sahalarda kullanılabilir. İstenilen sonuç değerine sahip olunduğunda, üretilen üyelik fonksiyonlar optimal sayıda olup, ANN devre dışında bırakılır. Gauss karma yoğunluk fonksiyonu ise, n bileşenli yoğunluk fonksiyonunun ağırlıklandırılmış toplamlarıdır. Ayrıca GMM genelde konuşmacı bağımlı konuşma tanıma sistemlerinde kullanılır (Uzunçarşılı, 2005).

Veri tabanlı istatistiksel yaklaşım, son dönemde konuşma tanıma alanında yüksek kullanım oranı ve yüksek verimliliğe ulaşmıştır. Ses sinyalinin sınıflandırılması, bu yaklaşımda ses verisi ile bilgi çıkaran algoritmadır. DTW, ANN ve HMM gibi istatistiksel yöntemler konuşma tanıma sistemlerinin yanında duygu tanıma gibi karmaşıklığı ve zorluğu artan sistemler üzerinde de etkili olarak kullanılmaktadır. Bu çalışmalardan biride 2005 yılında Lin ve arkadaşları tarafından (Lin & Wei, 2005) otomatik duygu tanıma sistemi için geliştirilmiştir. Çalışmalarında, duygu tanıma işlevini ses sinyaliyle gerçekleştirmişlerdir. Çalışmada HMM sınıflandırma tekniği kullanmışlardır. Özellik alt kümesinin sınıflandırma performansı MFCC ile karşılaştırmışlardır. Duygusal konuşma veri tabanı üzerinde cinsiyete bağımlı ve

bağımsız uygulamalar yapmışlardır. Sistem ile test aşamasında %93.6 başarım sağlamışlardır.

Tablo 2. Sınıflandırma tekniklerinin MFCC özellik çıkarımı ile uygulandığındaki başarım oranı

Teknik	***Başarım Oranı***	***Referans***
DTW ve MFCC	%69	(Chavan ve Sable, 2013)
ANN ve MFCC	%75	(Polur ve ark., 2001)
SVM ve MFCC	%88.9	(Lin ve Wei, 2005)
HMM ve MFCC	%93.88	(Abushariah ve ark., 2010)
GMM ve MFCC DTW ve MFCC HMM ve MFCC	%84.26 %87.26 %98.34	(Bakır, 2016)

Bakır (Bakır, 2016), Almanca ses biçim ve özelliklerine bakılarak konuşmacının cinsiyetinin otomatik olarak tanınması için bir sistem tasarlamıştır. 50 erkek ve 50 kadından Almanca farklı uzunlukta kelime ve cümle ile yaklaşık 3000' e yakın ses örneği alınmıştır. Ses örnekleri üzerinde özellik vektörleri, MFCC kullanılarak elde edilmiştir. Elde edilen ses örnekleri HMM, DTW ve GMM yöntemleri ile eğitilmiştir. Test aşamasında ise ses örneklerine bakılarak verilen ses örneğinin cinsiyeti belirlenmeye çalışılmıştır. GMM ile %84.26 oranında, DTW ile %87.37 oranında başarım gerçekleşirken, HMM ile %98.34 oranında başarım sağlanmıştır.

Tablo 2'nin sonuçlarına da bakıldığında, HMM yönteminin diğer sınıflandırma yöntemlerine göre daha başarılı sonuçlar verdiği görülmektedir. Ayrıca son zamanlarda birden çok özellik çıkarımının hibrit modeli ile HMM'nin kullanılması ile geliştirilen uygulamalarda da doğruluk oranında iyileşme görülmektedir. Buna bir örnek olarak 2015 yılında, Kepuska ve Elharati (Këpuska & Elharati, 2015) yeni özellikler elde etmek için, MFCC, LPC, PLP ve RASTA-PLP gibi özellik çıkarma yöntemlerinin bir kombinasyonu ile geliştirilen hibrit algoritmaları, değişkenli HMM kullanarak her birinin performansını incelemişler ve her bir özellik çıkarımı için ortalama %95'in üzerinde başarım elde etmişlerdir.

4 ÖNERİLEN ÇALIŞMA

Bu bölümde konuşmacı ve dil bağımsız gerçek zamanlı verimli bir konuşma tanıma sistemi için önerilen çalışma yer almaktadır. İlk olarak sistemin özelliklerinden bahsedilmektedir. Sonrasında sistemin aşamaları ve bu aşamalar için uygulanan tekniklerden söz edilmektedir. Daha sonrasında ise sistem tasarımı detaylı bir şekilde yapılacaktır.

4.1 Genel Yapı

Literatür araştırması neticesinde konuşma tanıma üzerine önerilen çalışmada, konuşmacı bağımsızlık, dil bağımsızlık ve gerçek zamanlı verimlilik hedeflenmiştir. Kullanıcıların verimli sonuç alması adına sistem eğitilmelidir. Her bir izole konuşma, sisteme metni ile kaydedilir. Sonrasında sistemin güncellenmesi gereklidir. Güncelleme sonrasında sisteme yüklenen konuşmanın metin karşılığı gerçek zamanlı olarak ekranda gösterilir. Önerilen sistem, eğitim şeklinden dolayı dil bağımsızdır. Konuşmacı tanımak yerine konuşmaların tanınması ise konuşmacı bağımsız bir özelliktir.

4.2 Konuşma Tanıma Evreleri

Sistem, algılanan konuşmanın yazıya çevrilmesinde, eğitim ve tanıma evresi olmak üzere iki aşamada önerilmektedir.

4.2.1 Eğitim Evresi

Sistemin eğitilmesi, konuşmaların metinleri ile etiketlenerek sisteme tanıtılması ile olmaktadır. Aynı konuşmanın farklı konuşmacılar tarafından seslendirilmesi ile sistemin başarım oranının artması hedeflenmektedir. Konuşmalar sisteme izole bir şekilde metinleri ile etiketlenerek kaydedilmektedir. Kaydedilen konuşmaların özellik vektörleri çıkartılmaktadır. Sonrasında bulunan özellik vektörleri veri kümesi giriş

parametresi olarak kabul edilerek k-ortalama algoritması ile kümelenir ve Kod Kitabı oluşturulur. Bu özellik vektörleri HMM için gözlenen durumları oluşturulur. Gözlem durumları ile geçiş durumları arasındaki gözlem olasılıkları Baum-Welch algoritması ile hesaplanır. Böylece test aşaması için sistem hazır olur.

4.2.2 Tanıma Evresi

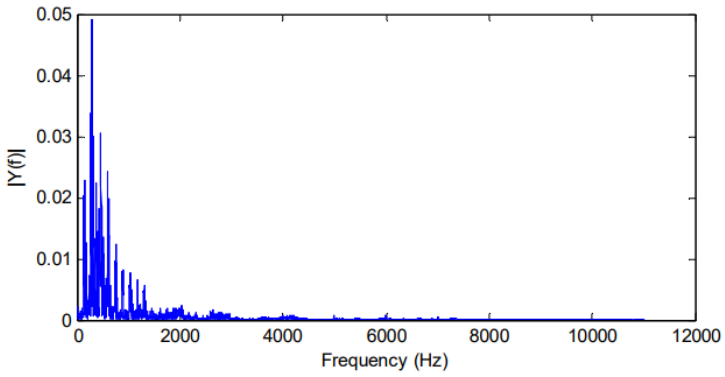
Kaydedilen konuşmalar, tanıma evresinde yazıya çevrilmektedir. Tanıma evresi, konuşmanın algılanması ile başlar. Konuşma sayısal olarak anlamlı özellik vektörlerine çevrilir. Sınıflandırma aşamasında Viterbi algoritması ile en iyi olasılıklı konuşma metni ekrana yazılır.

4.3 Özellik Çıkarım Yöntemi

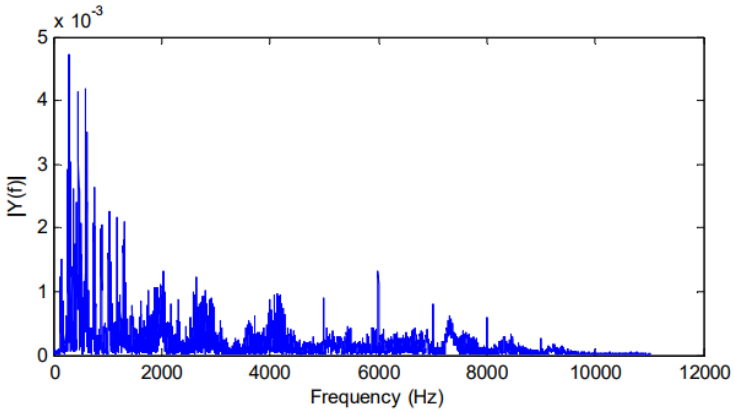
Önerilen çalışmada konuşma tanıma sürecine konuşmaların kaydı sonrası özellik vektör dizilerinin çıkartılması ile başlanır. Özellik vektörleri aşaması öncesi ön işleme olarak adlandırılır. Önerilen çalışmada ön işlemede ön vurgulama, çerçeveleme ve pencereleme adımları yer almaktadır. Sonrasında ise özellik çıkarımında MFCC tekniği önerilmektedir.

4.3.1 Ön Vurgulama Tekniği

Konuşma tanımada en önemli sorunlardan birisi gürültü (kirlilik) dür. Tespit edilmek istenilen sinyal ile sinyal üzerindeki gürültü arasındaki orana sinyal gürültü oranı (Speech/Noisy (S/N)) denilmektedir. Ön vurgulama konuşma tanımada sinyal gürültü oranını arttırmak için kullanılmakta olan bir tekniktir. Önerilen çalışmada konuşmadan yeterli bilgiyi çıkarmak için S/N oranı yüksek tutulmalıdır.



Şekil 14. Ön vurgulama öncesi konuşma sinyali

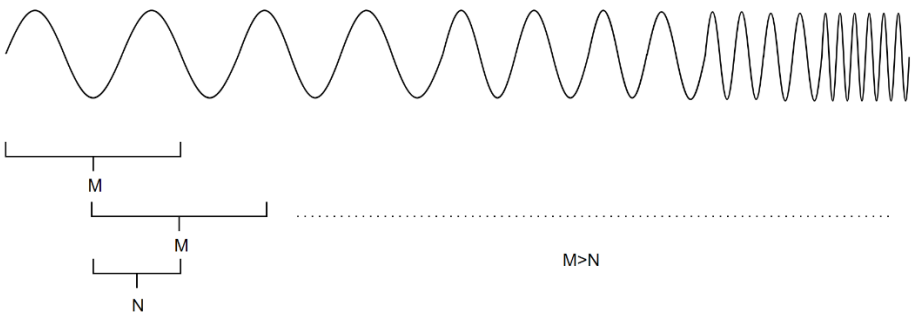


Şekil 15. Ön vurgulama sonrası konuşma sinyali

Yüksek hızlı dijital iletimde, bir veri iletiminin çıkışında sinyal kalitesini arttırmak için ön vurgulama kullanılır. Sinyalleri yüksek veri hızlarında iletirken, iletim ortamı bozulmalara neden olabilir, bu yüzden bu bozulmayı düzeltmede iletilen sinyali bozmak için ön vurgulama kullanılır. Düzgün bir şekilde yapıldığı zaman, daha yüksek frekansların kullanımına izin veren veya daha az bit hatası üreten, orijinal veya arzu edilen sinyali daha yakından izleyen bir sinyal üretir. Ön vurgulama aşaması, örneklendirilmiş ses sinyallerinin filtre yardımıyla spektral olarak düzlenmesinde de kullanılmaktadır.

4.3.2 Çerçeveleme

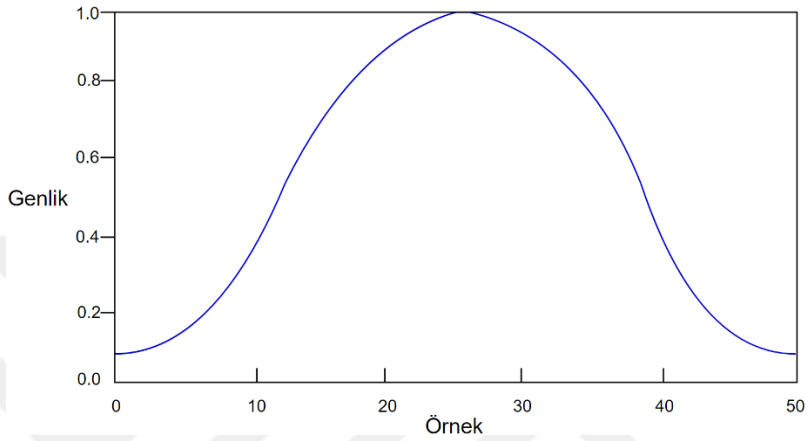
Çerçeveleme ile konuşma sinyali M kadar örnekle çerçevelere ayrılır. İkinci çerçeve, ilkinden N kadar sonra başlar. Böylece ikinci çerçeve M-N sayısınca üstüne eklenir. Örneğin, 5000 H ile örneklenen sinyalden 80 ms'lik kesim kullanıldığında, 400 örneklik ses sinyali çerçeve olarak kullanılır. Burada n ve (n-1). çerçeveler arası 40 ms olursa, bu seri ikili 40 ms boyunca üst üste biner.



Şekil 16. Çerçeveleme

4.3.3 Pencereleme

Pencereleme aşamasında, alınan sinyal içerisindeki devamsız kısımların dikkate alınmaması ses tanıma için kritik bir eşiktir. Bu işlem pencereleme ile gerçekleşmektedir. Pencereleme, elde edilen ses sinyalinde spektral analiz yapabilmemizi sağlar. Pencereleme fonksiyonları arasında konuşma tanımada en yaygın olarak kullanılan Hamming Pencereleme'dir.



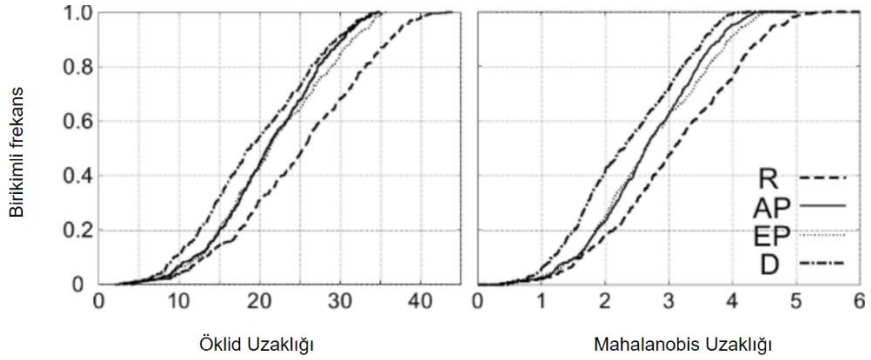
Şekil 17. Hamming Pencereleme

Önerilen sistemde yaygın olarak kullanılan, sesin frekans dönüşümü esnasındaki bozulmalarının azaltılabilmesinde ve konuşmanın spektral analizinde, Hamming Pencereleme ile farklı teknikler kıyaslanmıştır. Hamming Pencereleme w(n), her bir pencere içindeki örnek sayısı N olacak şekilde formülü şu şekilde gösterilir;

$$w(n) = \begin{cases} 0.54 - 0.46\left(\frac{2\pi n}{N-1}\right), & 0 \leq N-1 \\ 0, & n < 0, n > N-1 \end{cases} \quad (3)$$

4.3.4 Mahalanobis Uzaklığı

Çerçeveleme ve pencereleme ile elde edilen ilkten sona kadar tüm örneklerin bir ses numunesi olarak kabul edilip edilmemesinde kullanılması önerilmektedir. Ortamdaki gürültünün anlık değişiminde eşikteki sınırları belirler.



Şekil 18. Mahalanobis uzaklığı (Këpuska & Elharati, 2015)

Şekilde 2015 yılında yapılan çalışmada (Këpuska & Elharati, 2015) Mahalanobis uzaklığı ile ses numunesi olarak mesafelerin tespitinin daha iyi olduğu saptanmıştır.

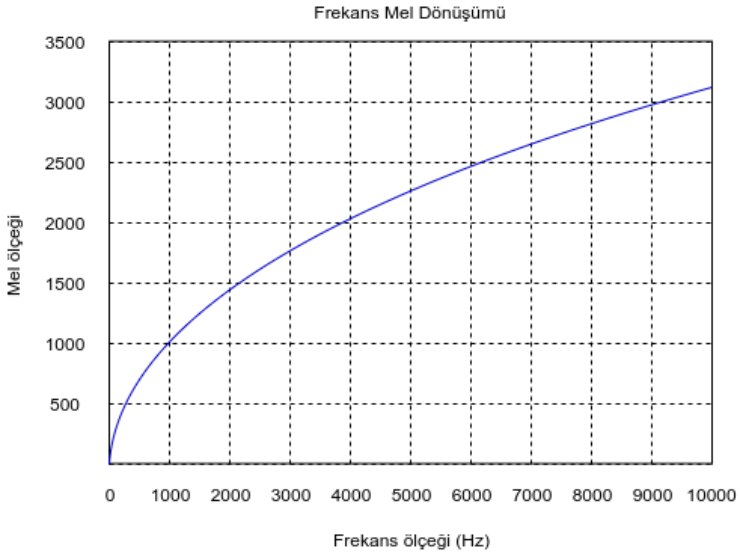
$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2/\sigma_i^2} \quad (1)$$

Yukarıdaki formül, i ve j fonemleri arasındaki Mahalanobis uzaklığıdır ve σ, varyanstır.

4.3.5 MFCC ile Özellik Vektörlerinin Elde Edilmesi

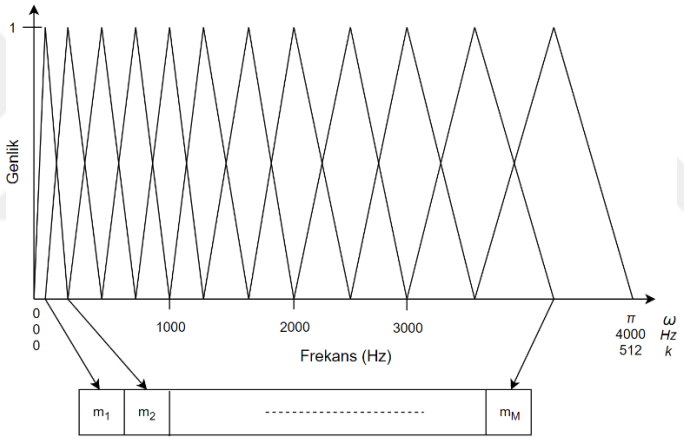
Stevens ve arkadaşları (Stevens ve ark., 1937) tarafından ortaya atılan Mel ölçümü, bir tondaki frekansı ifade etmektedir ve insan tarafından algılanan ses sinyali frekansının ölçümüdür. 1000 Hz'lik bir tonun referans olarak seçilmesi halinde bu 1000 Mels'e karşılık gelmektedir. Buna göre Mel ile frekans arasındaki ilişkisel formül şu şekildedir (Uzunçarşılı, 2005).

$$F_{mel}(f_{Hertz}) = 1127In(1 + \frac{f_{Hertz}}{700}) \quad (2)$$



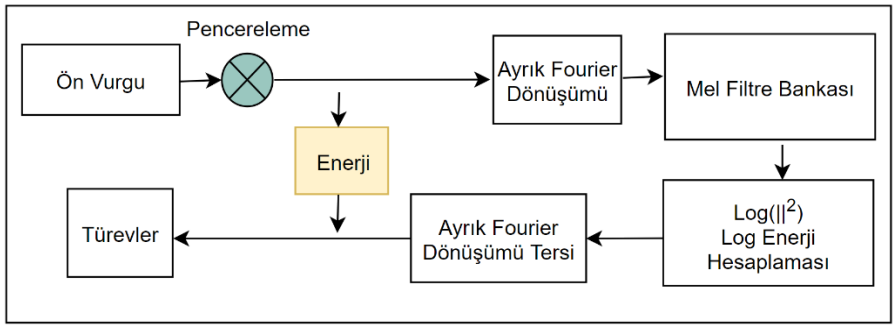
Şekil 19. Frekans ve Mel arasındaki ilişki

Şekle göre Mel filtre bankasını açıklamak gerekirse;



Şekil 20. Mel Filtre Bankası

Frekans – genlik grafiğinde 1000 Hz eşik değerdir. Grafiğin altında ise her bir frekans anında Mel katsayıları gösterilmektedir (Jurafsky & Martin, 2006).



Şekil 21. MFCC Adımları

4.4 Sınıflandırma Yöntemi

Konuşmaların özellik vektörlerinin bulunmasından sonra, bu konuşmalara karşılık gelen istatistiksel model ile veri tabanı oluşturulur. Sonrasında konuşmalara karşılık, tüm veri tabanı içerisinden arama yapılır ve en uygun eşleşme (metin) sonuç olarak seçilir.

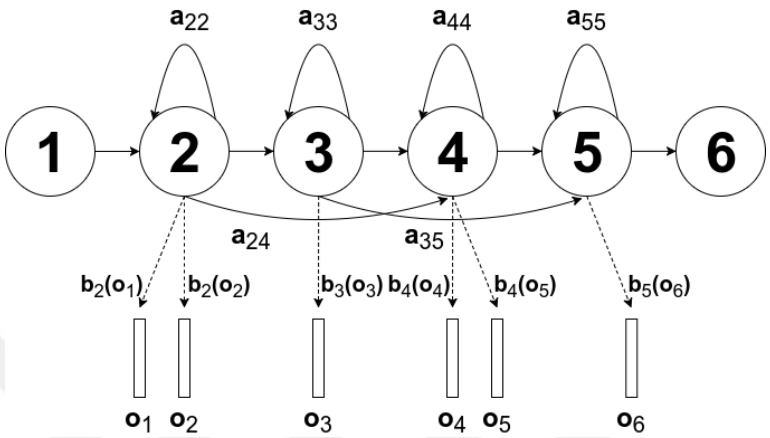
4.4.1 VQ ile Kod Kitabı

Konuşma tanıma HMM için oldukça uygundur. Çünkü konuşma sinyali kısa süreli durağan sinyal olarak görülebilir. Kısa sürede konuşma durağan bir süreç olarak hesaplanabilir. Konuşma sinyali içerisindeki her akustik özellik vektörü, belirli bir noktadaki farklı frekans bantlarındaki enerji miktarı gibi bilgileri temsil eder. Konuşma tanıma için gözlem sırası, bir dizi akustik özellik vektörü (MFCC vektörleri) olup, fonemler gizli durumlardır. MFCC vektörlerini, simüle edebilecek yöntemlerden birisi her girdi vektörünü az sayıda simgeden birine eşleyen bir Kod Kitabı oluşturmaktır. Durumsal olasılıklar için istatistiksel bir çerçeve sunmasından dolayı (Debyeche ve ark., 2007) giriş vektörlerini ayrık sembollere eşlemek için Vektör Nicemleme veya VQ kullanılmaktadır. Bununla birlikte VQ, Kod Kitabı sözcüklerinin sayısını arttırmaya çalışsak bile konuşma sinyalinden bazı bilgileri kaybedebilir.

4.4.2 HMM ile Sistemin Eğitilmesi ve Testi

Önerilen konuşma tanıma sisteminde sınıflandırma aşamasında, özellik vektörlerinin, özünü bilmeden sinyalin kaynağıyla ilgili modelleme yapabilmesinden dolayı HMM tekniği önerilmiştir. HMM'in geri görünümünde çalışan saklı bir Markov işlemi bulunur. Bu modeller gözlem vektörleri üretir ve HMM için gözlem dizileri oluşur. HMM'nin buradaki amacı gözlemlenen durumlara karşı olan durumları tahmin etmektir. HMM'de durumlar, gözlemler ve durumlar arası geçişler vardır. Konuşma tanımada gözlemler, konuşma sinyalinden elde edilen özellik vektörlerinden oluşur. Durumlar ise temel olarak alınan konuşmaların karşılığı olan söz dizisine denk gelir. Bu durumda konuşma tanımada ki amaç saklı olan durum dizisini gözlemlerden yararlanılarak çıkarılan olasılıklardan bulmaya çalışmaktır. Her konuşma ifadesi için ayrı bir model tanımı yapılır. Her konuşma ifadesi bir model olarak düşünülürse gerçek zamanlı olarak gelen konuşmalar bu modellerin art arda sıralanması ile modellenir. Bu durumda her bir konuşmaların son durumda bir sonraki konuşmaların ilk durumuna

bir geçişi söz konusudur. Böylece konuşma tanıma HMM ile gerçekleşmiş olur. Önerilen sistemde her bir konuşma olasılıksal olarak bir HMM'ye sahip olmalıdır. Konuşmaların sonuç için metin ile karşılaştırılmasında HMM, veri tabanında bulunan bütün kelimeler için en yüksek olasılıklı olanı bulmaktadır.



Şekil 22. HMM'nin standart gösterimi

HMM belirlemede N ve M gibi iki model parametresi, gözlem sembolleri ile olasılıksal ölçüme yarayan A, B, π bilinmelidir.

S = {1, 2, ..., N}: Oluşan konuşma sinyallerinin mevcut durumu

A: Durum geçiş olasılığı

B: Gözlem sembol olasılığı

$\pi = \{ \pi_i \}$: i. durumda olma olasılığı belirtilen başlangıç durumu

N: 1'den N'ye kadar olan HMM durum sayısıdır. t anında durum q_t'dir.

M: Tüm durumlardaki diğer gözlem sayılarıdır. Gözlem sembolleri şu şekilde gösterilir:

$$O = \{o_1, o_2, \dots, o_T\} \quad (4)$$

M parametresi sürekli gözlem dizilerinde bulunmaktadır.

$$A = \{a_{ij}\} \quad (5)$$

i'den j'ye durum geçiş olasılığıdır.

$$a_{ij} = P(q_{t+1} = j \mid q_t = i);\ 1 \leq i, j \leq N \quad (6)$$

i'den j'ye geçiş olmadığı durumlardır.

$$\{a_{ij}\} = 0, \quad (7)$$

Gözlem sembolleri ile

$$B = \{b_j(o_t)\}, \quad (8)$$

gözlem olasılık dağılımı olsun.

$$b_j(o_t) = P(o_t \mid q_t = j);\ 1 \leq t \leq T \quad (9)$$

Bunların akabinde HMM'de tam parametre seti göstermek için HMM yoğunluk gösterimini

$$\lambda = (A, B, \pi) \quad (10)$$

şeklinde belirtebiliriz. Bu sistem üzerinde HMM, gözlem dizisi olasılığı hesabını bulur. Durum dizisi,

$$q = (q_1, q_2, \dots, q_t) \quad (11)$$

ile T uzunluğunda bulunan durum dizileri hesaplanır. Böylelikle gözlem dizisi "O" şöyledir,

$$P(O \mid q, \lambda) = \prod_{t=1}^{T} P(o_t \mid q_t, \lambda) \quad (12)$$

Formülü sadeleştirilirse,

$$P(O|q, \lambda) = b_{q1}(o_1).b_{q2}(o_2)\dots b_{qT}(o_T) \quad (13)$$

q durum dizi olasılığı ise

$$P(q|\lambda) = \pi_{q1} a_{q1q2} a_{q2q3} \dots a_{qT-2qT} \quad (14)$$

şeklindedir. Olası durumların hepsi q toplamı ile O olasılığını verir.

$$P(O, q|\lambda) = P(O|q, \lambda).P(q|\lambda) \quad (15)$$

$$P(O|\lambda) = \sum_{q1,q2,\ldots,qt} P(O|q,\lambda).P(q|\lambda) \quad (16)$$

Bu formülden çarpma işleminde pay ve paydadakiler sadeleşir ve aşağıdaki denklem elde edilir.

$$= \sum_{q1,q2,\ldots,qt} \pi_{q1} b_{q1}(o_1) a_{q1q2} b_{q2}(o_2) a_{q2q3} \ldots a_{qT2qT} b_{qT} \quad (17)$$

şeklindeki hesabı o_1 ve $b_{q1}(o_1)$ olasılığı ilk durumundan t zamanında q_{t-1}'den q_t durumuna giden ve o t olasılığını elde eden özyinelemeli ve herhangi bir zamanda durum olasılığını bulmamızı sağlar (Manning & Schütze, 1999).

HMM üç temel probleme sahiptir. Bunlar, ilerleme ve gerileme algoritması, Viterbi algoritması, Baum-Welch algoritması. İlerleme algoritması, ayrık-kelime tanıma işleminde yararlıdır. $\alpha_t(i)$, ilerleme algoritması için değişken olarak tanımlarsak,

$$\alpha_t(i) = P(o_1 o_2 \ldots o_t, q_t = i|\lambda) \quad (18)$$

Herhangi t zamanında i durumundan λ modeliyle gözlem dizisi olan o_1, o_2, ..., o_t olasılıklarını verir. İlerleme ve gerileme algoritmaları uygun durumu bulma dizi hesabında kullanılmaktadır.

$$\beta_t(i), \quad (19)$$

gerileme algoritması için değişken olarak tanımlarsak,

$$\beta_t(i) = P(o_{t+1} o_{t+2} \ldots o_t \mid q_t = i, \lambda) \quad (20)$$

Gerileme algoritması da ilerleme algoritması gibi HMM'nin eğitim aşamasında yararlıdır. Herhangi bir t zamanından t+1 zamanına i durumundan λ modeliyle gözlem dizi olasılıklarını verir. Viterbi algoritması ise konuşma tanıma sistemlerinde yaygın olarak kullanılmaktadır.

$$\delta_t(i), \quad (21)$$

Viterbi algoritması, özellikle Markov bilgi kaynakları ve HMM bağlamında, gözlemlenen olayların bir dizilimiyle sonuçlanan, gizli durumların en olası sırasını bulmak için kullanılır. Uygun dinamik programlama algoritması ile bulunan en uygun

şablon model yazı olarak ekrana yazılır. Viterbi algoritması için değişkendir ve bir yol ile en yüksek olasılık hesabıdır.

$$\delta_t(i) = \max_{q_1 q_2, \ldots, q_{t-1}} P(q_1 q_2 \ldots q_{t-1}, q_t = i, o_1 o_2 \ldots o_t | \lambda) \quad (22)$$

Gözlem dizisi,

$$q = (q_1 q_2 \ldots q_t) \quad (23)$$

En iyi durum dizisi,

$$O = \{o_1, o_2, \ldots, o_T\} \quad (24)$$

t zamanında, i durum hesabı için Viterbi değişkeni özyinelemeli olarak,

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{ij}\right] b_j(o_{t+1}) \quad (25)$$

algoritması en iyi olasılığı bulmak için konuşma tanımada başarılıdır.

Burada sistemin her bir söz dizisi için eğitilmesi, HMM Baum-Welch algoritması ile durumlar arasında olasılıkların hesabı ile yapılmaktadır. Baum-Welch, HMM'de uygun durum hesaplamasıdır. Viterbi algoritması ile bulunan

$$P(q|O, \lambda) \quad (26)$$

olasılığının azami durumunu bulur. A, B ve π gibi durum olasılıkları ve sabit değer ile hesaplanır. Baum-Welch aslında beklenti maksimizasyonu olarak da bilinir.

$$\xi(i, j) = argmax_\lambda P(q|O, \lambda) \quad (27)$$

4.5 HMM Üç Temel Problemi

HMM modelinin konuşma tanımada kullanılabilmesi için 3 temel problemin çözülmesi gereklidir. Bunlar; gözlem olasılığı, saklı durum dizisi tahmini ve modelin yeniden yapılandırılmasıdır.

4.5.1 1. Problemin Çözümü İleri-Yön ve Geri-Yön Algoritması

$\lambda = (A, B, \pi)$ ile parametreleri verilen $O = o_1 o_2 \ldots o_T$ gözlem dizisi olsun. Herhangi bir andaki durum olasılığı $P(O|\lambda)$ nasıl hesaplanabilir.

İleri-Yön değişkeni $\alpha_t(i) = P(o_1 o_2 \dots o_t = S_i|\lambda)$ (28)

Geri-Yön değişkeni $\beta_t(i)$ = 1 olacak şekilde $1 \leq i \leq N$ ise

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j (O_{t+1}) \beta_{t+1}(j), \qquad t = T-1, T-2, \dots, 1 \quad (29)$$

4.5.2 2. Problemin Çözümü ve Viterbi Algoritması

$\lambda = (A, B, \pi)$ ile parametreleri verilen $O = o_1 o_2 \dots o_T$ gözlem dizisi olsun. Herhangi bir durum $Q = q_1 q_2 \dots q_T$ nasıl seçilir. Buda saklı kısımda yapılan doğru durum dizisinin bulunmasıdır. (25) numaralı formülün diğer bir gösterim şekli şu şekildedir.

$$\delta_t(i) = \max_{q_1, q_2, \dots, q_{t-1}} P[q_1 q_2 \dots q_t = i, o_1 o_2 \dots o_t|\lambda] \quad (30)$$

$\delta_t(i)$, t anına kadar tek yol boyunca en yüksek olasılığı gösterir ve t anında S_i durumuna ulaşır ve formül (25) elde edilir.

4.5.3 3. Problemin Çözümü ve Baum-Welch Algoritması

HMM modelinin eğitimi için İleri-Yön ve Geri-Yön değişkenlerine ait olasılıklar üzerine kurulu olan benzerlik maksimizasyonudur. $\lambda = (A, B, \pi)$ ile parametreleri verilen $O = o_1 o_2 \dots o_T$ gözlem dizisi olsun. Herhangi bir durum $\xi(i,j)$ değişkeni, t anında S_i durumunda ve $t+1$ anında S_j durumunda olma olasılığını gösterir.

$$\xi(i,j) = P(q_t = S_i, q_{t+1} = S_j|O, \lambda) \quad (31)$$

İleri-Yön, Geri-Yön Algoritma ifadeleri yerlerini yazılırsa aşağıdaki formül elde edilir.

$$\xi(i,j) = \frac{\alpha_t a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)} = \frac{\alpha_t a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) \alpha_{ij} b_j(O_{t+1}) \beta_{t+1}(j)} \quad (32)$$

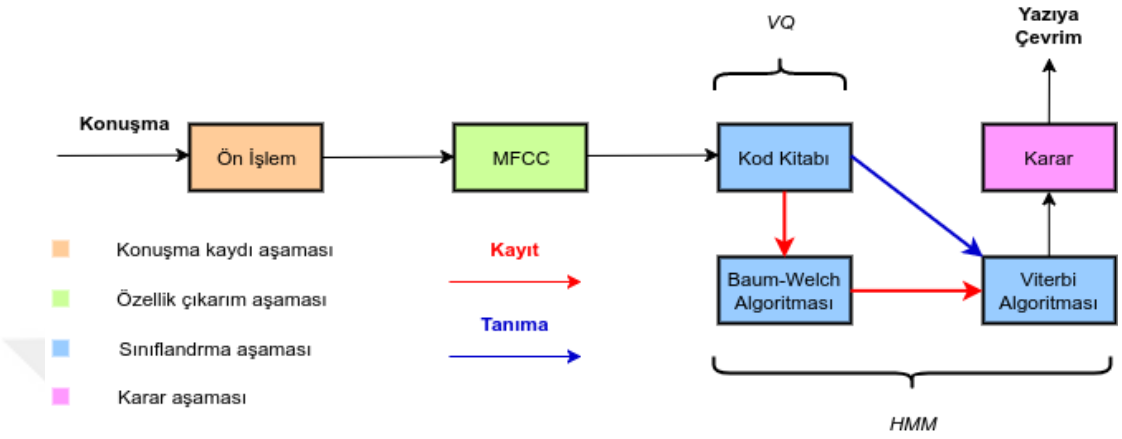
İleri-Yön, Geri-Yön Algoritma ifadelerini basit bir şekilde ifade edersek Baum-Welch formülü şu şekilde olmaktadır.

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \quad (33)$$

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \quad (34)$$

4.6 Önerilen Çalışmanın Mimarisi

Önerilen çalışmada konuşma konuşmanın kaydı ile özellik vektörleri çıkartılır. Bu özellikler sınıflandırma yöntemleri ile tanınır. Son olarak yüksek olasılıklı eşleşme ekrana yazdırılır.



Şekil 23. Önerilen çalışmanın mimarisi

Ön işlem aşamasında konuşma wav (Waveform Audio File Format) formatında kaydedilerek özellik çıkarımı için hazır hale gelir. Konuşma özellik çıkarımı aşamasında sınıflandırma aşamasına parametre olacak şekilde işlenir ve Kod kitabı oluşturulur. Oluşturulan Kod kitabı ile HMM, Baum-Welch algoritması kullanılarak test aşamasına hazır hale gelir. VQ ile ifade edilen gözlem dizileriyle test aşamasında Viterbi algoritmasıyla en uygun kelime seçilir. Karar aşamasında kelime ekranda gösterilir.

5 DENEYSEL ÇALIŞMA

Bu bölümde, gerçek zamanlı ve kaliteli konuşma tanıma için önerilen sistemin uygulaması anlatılmaktadır. Bu bağlamda bölüm içerisinde deney, deney sonuçlarına ait değerlendirmeler ve bu değerlendirmelere ilişkin yöntemlerden bahsedilmektedir.

5.1 Kullanılan Teknolojiler

1950'li yıllardan itibaren konuşma tanıma üzerine çalışmalar devam etmektedir. Çalışmaların önemli araştırma konuları ise yüksek performanslı konuşma tanımadır. Bu hedefler gözetilerek yapılan literatür taramasıyla sistem önerimi yapılmıştır. Bu bölümde de önerilen sistemde bahsedilen konuşma tanıma teknikleriyle uygulanan çalışmanın detayları ve değerlendirilmesi yapılmaktadır.

Tez çalışması kapsamında uygulama yazılımında kullanılan teknoloji Java'dır. Deneysel çalışma esnasında tekniklerin uygulanmasında üçüncü bir kütüphane kullanılmamıştır. Uygulama aşamasında kullanılan kütüphaneler;

java.awt : Kullanıcı arabirimi oluşturmaya ve resim boyama için tüm sınıfları içerir (Oracle, 2017). Çalışmada konuşma sinyali görselini çizmeye yarar.

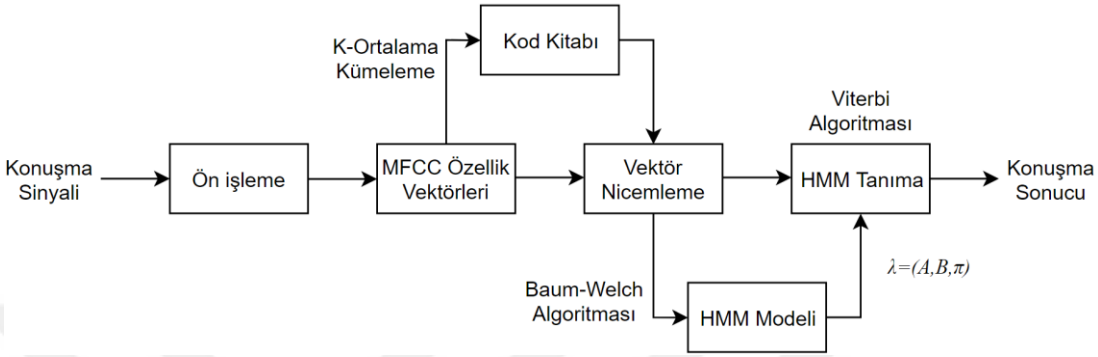
javax.swing : Uygulamalar için grafik kullanıcı arabirimi (GUI) oluşturulmasını sağlar (Oracle, 2017). Çalışmada arayüz sağlamaya yarar.

java.util : Koleksiyon sınıfları, olay modeli, tarih ve saat özellikleri ve çeşitli yardımcı program sınıfları (rasgele sayı üreteci ve bir bit dizisi) içerir (Oracle, 2017). Çalışmada koleksiyon sınıfları oluşturmaya yarar.

java.io : Veri akışları, serileştirme ve dosya sistemi aracılığıyla sistem girişi ve çıkışı sağlar (Oracle, 2017). Çalışmada konuşmaların dosyaya kaydedilmesini sağlar.

javax.sound : Java Sound API, genişletilebilirlik ve esnekliği artıran bir çerçevede normalde ses girişi ve çıkışı için gerekli olan yetenekler üzerinde açık kontrol sağlar (Oracle, 2017). Ses işlemlerine yarar.

5.2 Uygulama Mimarisi

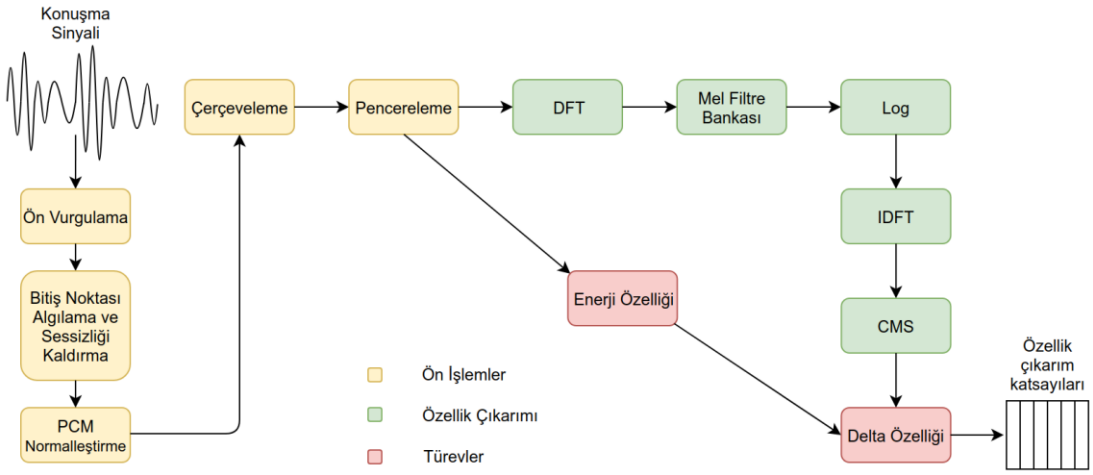


Şekil 24. Uygulama Mimarisi

4.Bölümde önerilen tekniklerle özellik çıkarımında MFCC ve sınıflandırma aşamasında VQ ve HMM kullanılmıştır. Gerçek zamanlılık için her 3 saniyede bir sistem ortamdaki konuşmayı algılamaya çalışmaktadır.

5.3 Ön İşleme

Bu çalışmada konuşma tanıma sürecine konuşmaların ön işlenmesi ile başlanmaktadır. Ön işleme ile özellik çıkarımı aşamasına uygun örneklenmiş ses numuneleri verilmektedir. Ön işleme aşamasında konuşmanın kaydı, bitiş noktası algılama ve sessizliği bozma, PCM normalleştirme, ön vurgulama, çerçeveleme ve pencereleme adımları bulunmaktadır.



Şekil 25. Özellik Çıkarımı katsayılarının elde edilmesi

5.3.1 Konuşmanın Kaydı

Konuşmanın işlenmesindeki ilk adım, mikrofondaki analog elektrik sinyallerini dijital bir sinyal (x[n]) haline dönüştürmektir (x konuşma örnekleri, n zaman endeksidir). Konuşma analizinde enerjinin 4 kHz'e kadar bulunduğu varsayılarak kullanılan ses formatı; 22050 Hz, 16 bitlik imzalı, tek kanallı ve WAV uzantılıdır.

5.3.2 Bitiş Noktası Algılama Algoritması ve Sessizliği Bozma

Algılanan konuşma sinyali, sinyalin başı veya sonu gibi farklı konumlarda sessizlik içerebilir (konuşma olmayan alan). Böyle sessiz alanlar dâhil edilirse, modelleme kaynakları sinyali tanımlamaya katkıda bulunmayan kısımlar için harcanır. Bu sebepten sessiz alanlar, bir sonraki aşamaya geçmeden önce kaldırılmalıdır.

Konuşmacının kayda başladığı anda konuşması biraz zaman alacağından bir konuşma kaydının genellikle ilk 100 ms'si arka plan gürültüsü ile ilgilidir.

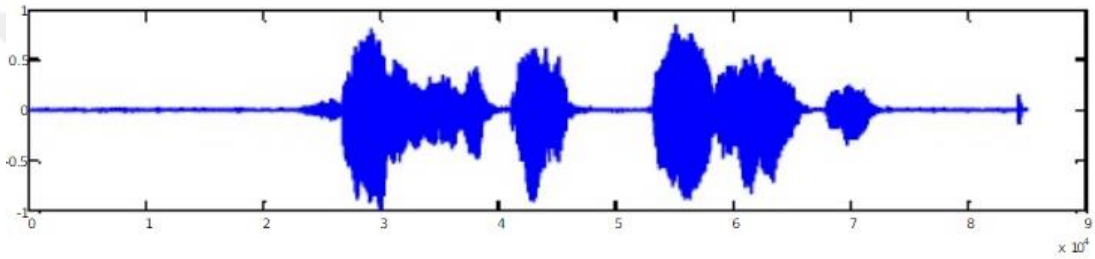
1.Adım : Kaydedilen konuşmanın ilk 100 ms'lik örneklerinin ortalaması (μ) ve standart sapması (σ) hesap edilir. Arka plan gürültüsü μ ve σ ile karakterize edilir.

2. Adım : İlk örnekten son örneğe kadar her bir örnekte, tek boyutlu Mahalanobis uzaklık fonksiyonlarının, yani | x-μ | / σ değeri 3'ten büyükse (Keerio ve ark., 2009) bu bir ses numunesi kabul edilir. Ses numuneleri sesli örnek olarak kabul edilir.

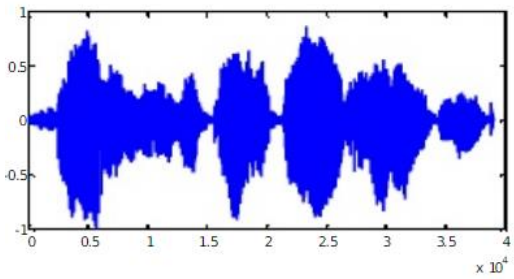
3. Adım : Sesli örnekler 1 olarak işaretlenir. Sessiz örnekler 0 olarak işaretlenir. Sonrasında tüm konuşma işaretleri 10 ms örtüşmeyen pencerelere bölünür ve konuşma yalnızca sıfırlarla ve birlerle temsil edilir.

4. Adım : Bir pencerede sıfırların sayısı M, birlerin sayısı N olsun. M ≥ N ise, birler sıfırlara çevrilir ve tersi yapılır. Bu yöntem konuşmanın 10 ms'lik kısa bir süre penceresinde aniden değişemeyeceğinden dolayı yapılır.

5. Adım : Sesli bölüm için pencereli dizideki sadece '1' etiketli örnekler toplanır ve yeni bir dizi oluşturulur. Böylece '1' etiketli örnekle konuşma sinyalinin sesli bölümü alınır.



Şekil 26. Kelime sonu tespiti öncesi giriş verisi



Şekil 27. Kelime sonu tespiti sonrası sinyal verisi

5.3.3 PCM Normalleştirme

Darbe kod modülasyonu (Pulse-code modulation (PCM)), dijital olarak örneklenmiş analog sinyalleri temsil etmek için kullanılan bir yöntemdir. Bilgisayarlarda dijital ses, kompakt diskler, dijital telefon ve diğer dijital ses uygulamalarında standart şekildir. Audio CD, WAV gibi formatları içerir. Çıkartılan PCM modüle edilmiş genlik değerleri normalize eder ve yakalama sırasındaki genlik değişimi önlenir.

5.3.4 Ön Vurgulama

Konuşma sinyali genellikle özellik çıkarımından önce ön-işleme aşamasında vurgulanır. Ünlüler gibi sesli bölümlere bakılırsa, daha düşük frekanslarda daha yüksek frekanslardan daha fazla enerji bulunduğu gözlenir. Frekanslardaki enerji düşmesi, gırtlak darbesinin doğasından kaynaklanır. Yüksek frekanslı enerjiyi artırmak, bu bilgilerin akustik modele daha fazla erişilebilir olmasını sağlar ve algılama doğruluğunu geliştirir.

Ön-vurgu filtresi, birinci derece yüksek geçiş filtresidir.

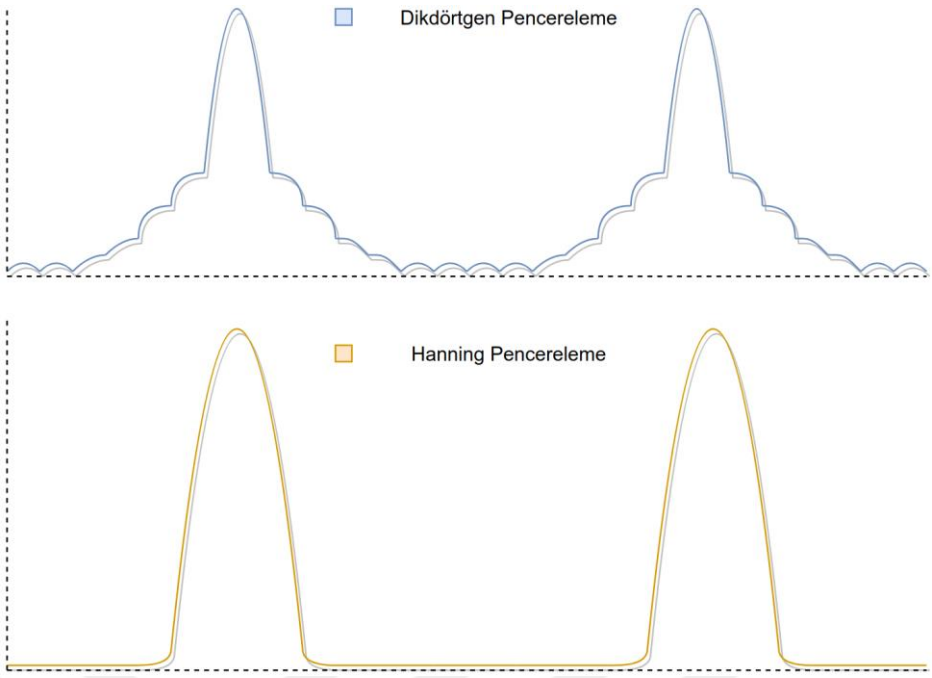
Zaman alanı içerisinde x[n] ve $0.9 \leq \alpha \leq 1$ girişi ile filtre denklemi şu şekildedir:

$$y[n] = x[n] - \alpha x[n-1] \tag{35}$$

$\alpha = 0.97$ olarak alındı (Zeghidour ve ark., 2017).

5.3.5 Çerçeveleme ve Pencereleme

Konuşma, durağan olmayan bir sinyaldir. Bu durum konuşma içerisinde istatistiksel özelliklerinin zaman içinde sabit olmadığı anlamına gelir. Bu çalışmada belirli bir konuşmadaki alt çerçevelerin karakterize edildiği ve sinyalin durağan olduğu (yani istatistiksel özellikleri bu bölge içinde sabit olduğu) varsayımını yaparak bir konuşma penceresinden spektral özellikleri çıkarmak istiyoruz. Bu nedenle, özellik çıkarımından önce ön işlemede bir konuşma sinyali pencerelenir. Buna göre bu uygulamada her bir çerçeve bloğunun %50 çakışması sağlanacak şekilde, çerçeve başına 512 örnek kullanılmıştır.



Şekil 28. Dikdörtgen ve Hanning Pencereleme

Pencereleme aşamasında konuşma tanıma alanında Hamming, Hanning, Blackman, Dikdörtgen pencereleme kullanılmaktadır. Fakat Dikdörtgen Pencereleme, Fourier analizi yapılırken problemlere sebep olmaktadır. Dikdörtgen Pencereleme ile ilgili bir problem, analiz edilecek sinyalde bozulmaya neden olabilecek kenardaki ani değişimdir (aslında pencere tarafından sinyal değiştirildiğinden herhangi bir pencere işlemi bozulmaya neden olur). Bu çarpıklığı azaltmak için genellikle Dikdörtgen Pencereleme yerine daha pürüzsüz bir pencereleme tekniği kullanılır. Bu pencereleme tekniklerine örnek olarak Hamming ve Hanning Pencereleme, kenarlarında sıfırdır ve kademeli olarak ortada 1 olacak şekilde yükselir. Bu pencerelemeler kullanıldığında, sinyal kenarları daha az vurgulanır ve kenar efektleri azaltılır.

5.4 Özellik Çıkarımı

Bu çalışmada konuşma tanıma sürecinde konuşmaların ön işlenmesinden sonra pencerelenmiş örnekler üzerinden özellik vektör dizilerinin çıkartılması ile devam edilir. Bu aşamada MFCC ile 12 katsayılı özellik vektörleri elde edilir.

5.4.1 Kesikli Fourier Transformu

Geçerli çerçevenin frekans içeriğini (spektrumu) ayıklamak için pencereli sinyalin Kesikli Fourier Dönüşümü (DFT) kullanılır. Spektral bilginin çıkartılması için araca

örnek olarak sinyalin, ayrık zaman (örneklenmiş) sinyal için ayrı frekans bantlarında ne kadar enerjiyi içerdiğini DFT ile bulabiliriz. DFT'ye giriş, pencereli bir sinyal x [n]…x [m] olup, N ayrı frekans bandının her biri için çıktı, frekans bileşeninin büyüklüğünü ve fazını temsil eden bir karmaşık sayı orijinal sinyaldeki X [k]'dır. DFT, Fourier dönüşümünün eşit aralıklı frekanslardaki örneklerine özdeştir. Sonuç olarak N -noktalı bir DFT'nin hesaplanması Fourier dönüşümünün N örneğinin, N eşit aralıklı frekanslarla ($w_k = 2\pi k/N$), z-düzlemindeki birim çember üzerinde N nokta ile hesaplanmasına karşılık gelir. Buradaki temel amaç N -noktalı DFT'nin hesaplanması için verimli algoritmaların kullanılmasıdır. (36) Numaralı formül DFT'nin hesaplanmasıdır.

$$bin_k = \left|\sum_{n=1}^{N} S_w(n) e^{-i(n-1)\frac{2\pi k}{N}}\right|, k = 0,1,2,\dots,N-1 \quad (36)$$

5.4.2 Mel Filtresi

MFCC'nin hesaplanması için (2) numaralı formül uygulanır. (2) numaralı formülün bir diğer gösterim şekli x doğrusal frekans olacak şekilde aşağıdaki gibidir:

$$Mel(x) = 2595\log[1 + \frac{x}{700}] \quad (37)$$

Ardından, mel-skala spektrumunun genliğine bir filtre bankası uygulanır. Mel frekans çarpması, Mel frekanslarına göre merkezlenmiş filtreler içeren bir filtre bankası kullanılarak yapılır. Üçgen filtrelerin genişliği, Mel ölçeğine göre değişir, böylece merkez frekans etrafındaki kritik bir banttaki günlük toplam enerjisi dahil edilir. Filtrelerin merkezleri Mel ölçeğinde eşit aralıklarla yerleştirilir. 30 Filtre kullandığımız filtre bankasında, Mel filtreleme sonucunda her bir Mel ölçek bandı enerji dağılımı hakkında bilgi verir. Her bir filtreden bir çıktı vektörü elde edilir. (37) numaralı formülden x değeri aşağıdaki şekilde elde edilir:

$$x = 700(10^{\frac{mel}{2595}} - 1) \quad (38)$$

Mel filtreleme ile elde edilen katsayılara IDFT'nin Kepstrumundan önce Log enerji hesaplaması şu formüle göre yapılır:

$$f_i = \ln(f_{bank_i}) \quad (39)$$

5.4.3 IDFT'nin Kepstrumu

Her çerçevenin MFCC özelliğini elde etmek için, kepstrum dönüşümü filtre çıkışlarına uygulanır. Üçgen filtre çıktıları $\mathrm{Y(i)}, \mathrm{i} = 0, 1, 2 \ldots, \mathrm{M}$ logaritma kullanılarak sıkıştırılır ve ayrık kosinüs dönüşümü (DCT) uygulanır. Burada M, filtre bankasındaki filtre sayısına, yani 30'a eşittir. c[n], her çerçeve için MFCC vektörü olacak şekilde aşağıdaki gibidir:

$$c[n] = \sum_{i=1}^{M} logY(i)\cos\left[\frac{\pi n}{M}(i - \frac{1}{2})\right] \quad (40)$$

Her konuşma çerçevesinden çıkarılan özellik vektörü Mel frekans kepstrum (MFC) olarak adlandırılır ve tek tek bileşenler Mel frekanslı kepstral katsayılarıdır (MFCC).

5.4.4 Son İşlemler

Konuşmanın kaydından sonra ön işlemlerin oluşturduğu, bitiş noktası algılama, sessizliği bozma, ön vurgulama, çerçeveleme ve pencereleme aşamalarından sonra MFCC özellik çıkarımı yapılır. Tüm bunlardan sonra son işlemler aşamasında Kepstral Ortalama Çıkarma uygulanır.

5.4.5 Kepstral Ortalama Çıkarma (Cepstral Mean Subtraction (CMS))

Bir konuşma sinyali, kaydedildiğinde bazı kanal gürültüsüne maruz kalabilir, buna kanal etkisi (channel effect) denir. Belirli bir kişi için eğitim verilerini kaydederken kanal etkisi, kişi sistemi kullandığında sonraki kayıtlarda kanal etkisinden farklıysa, bir sorun ortaya çıkar. Sorun, eğitim verileri ile yeni kaydedilen veriler arasındaki yanlış uzaklığın, farklı kanal etkilerinden dolayıdır. Kanal efekti, Mel kepstrum katsayılarının ortalama Mel kepstrum katsayılarıyla çıkarılmasıyla ortadan kalkar:

$$mc_j(q) = c_j(1) - \frac{1}{M}\sum_{i=1}^{M} c_i(q),\ q = 1,2,\ldots,12 \quad (41)$$

5.5 Sınıflandırma ve Tahmin

Sınıflandırma aşamasında sırasıyla aşağıdaki adımlar takip edilmektedir:

1. Kod Kitabı, eğitim verilerinden özellik vektörü kullanılarak üretilir ve VQ, özellik vektörünü ayrı gözlem simgesine eşlemek için Kod Kitabını kullanır.

2. Kelimeler içerisindeki her v kelimesi için, bir HMM λ_v oluşturulmuştur, yani, v kelimesi için eğitim seti gözlem vektörlerinin olasılığını optimize eden model parametrelerini $\lambda = (A, B, \pi)$ tahmin etmeliyiz. Tüm model parametrelerinin güvenilir tahminlerini yapmak için birden fazla gözlem dizisi kullanılmalıdır. Baum-Welch algoritması, HMM parametrelerinin tahmini için kullanılır.
3. Tanımlanacak olan her bilinmeyen kelime için, sözcüğe karşılık gelen konuşmanın özellik analizi yoluyla $O = \{o_1, o_2, \dots, o_T\}$ gözlem dizisinin ölçülmesi ve bazı adımların işlenmesi gerekir; Viterbi Algoritmasıyla olası tüm modeller için model olasılıkları hesaplanır. $P(O|\lambda^v), 1 \leq v \leq V;$ ardından model olasılığı en yüksek olan kelime seçilir.

   $$v = arg_{1 \leq v \leq V}^{max}[P(O|\lambda^v)] \quad (42)$$

5.5.1 K-Ortalama Kümeleme

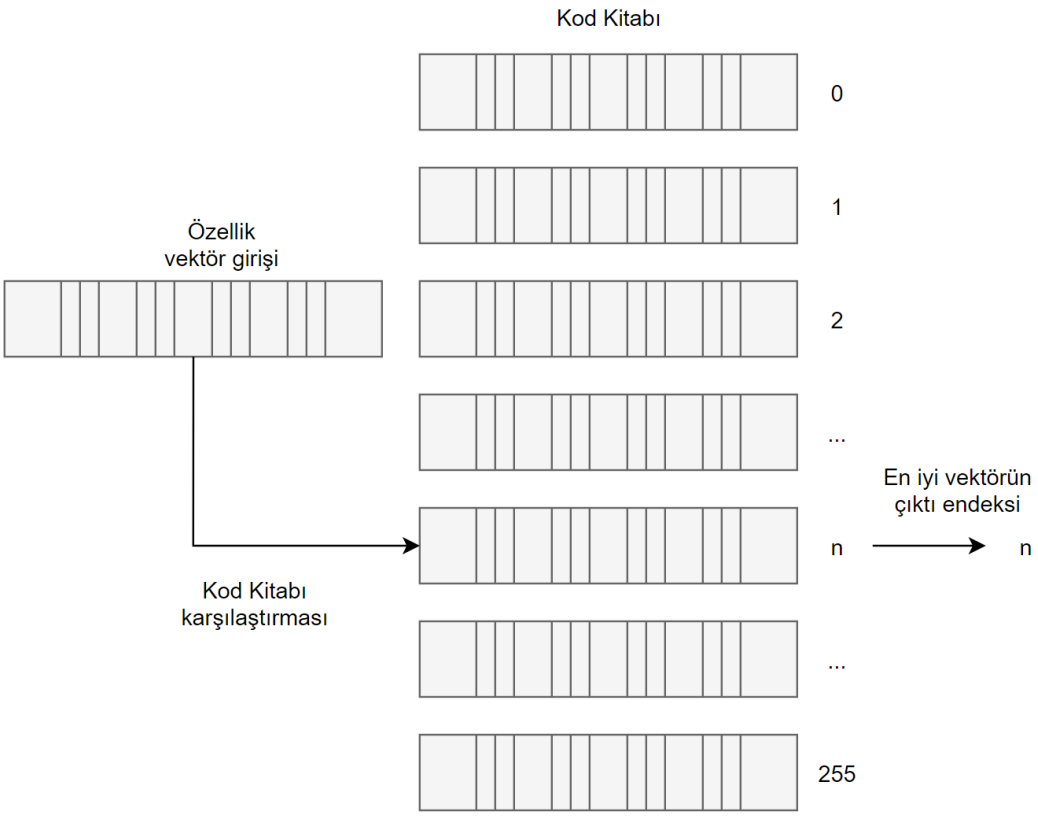
Kod Kitabı için kümeleme probleminin çözümünde kullanılmaktadır. Özellik vektörlerini k adet merkezle ifade etmeye yarar. K-Ortalama Kümeleme algoritması algoritması şu şekildedir:

1. İlk küme merkezleri belirlenir.
2. Giriş kümesindeki tüm örneklerin merkezlere olan uzaklıkları hesaplanır ve en yakın kümeye yerleştirilir.
3. Oluşturulan kümeler, içerisindeki örneklerin ortalaması ile güncellenir.
4. Ağırlık merkezleri değişmeyene kadar 2 ve 3. Adımlar tekrarlanır.

5.5.2 Kod Kitabı Oluşturulması

Bir konuşma sinyali bir parça halinde durağan sinyal veya kısa süreli durağan bir sinyal olarak görülebildiğinden dolayı konuşma tanımada yaygın olarak HMM kullanılır. Kısa sürede konuşma durağan bir süreç olarak hesaplanabilir. Her akustik özellik vektörü, belirli bir noktadaki farklı frekans bantlarındaki enerji miktarı gibi bilgileri temsil eder. Konuşma tanıma için gözlem sırası, bir dizi akustik özellik vektörü (MFCC vektörleri) olup, fonemler gizli hallerdir. MFCC vektörlerini, güvenebileceğimiz sembollere benzetmenin bir yolu, her girdi vektörünü az sayıdaki simgeden birine eşleyen bir haritalama fonksiyonu oluşturmaktır. Giriş vektörlerini ayrı nicelenmiş sembollere haritalama fikrine vektör nicemleme veya VQ denir.

VQ ile Kod Kitabını artırmaya çalışsak bile VQ, konuşma sinyalinden bazı bilgileri kaybetmekten sorumludur. Bu kayıp nicemleme hatasından (bozulma) kaynaklanmaktadır. Bu bozulma Kod Kitabındaki kod sözcüklerinin sayısını değiştirebilir ancak ortadan kaldırılamaz. VQ, her konuşma çerçevesini tanımlamak için gereken bit sayısını en aza indirgeyen artıklık kaldırma işlemidir. VQ işleminde, her eğitim özellik vektörünü az sayıda sınıflara eşleyerek küçük sembol seti oluşturulmaktadır ve her sınıf ayrı bir sembolle temsil edilmektedir. Bir VQ sistemi, bir Kod Kitabı, bir kümeleme algoritması ve bir mesafe metriği ile karakterize edilir.



Şekil 29. Vektör Nicemleme ile Kod Kitabı

Bir Kod Kitabı olası sınıfların bir listesi, $F = \{f_1, f_2, ..., f_n\}$ özellik vektörlerini oluşturan bir sembol kümesidir. Konuşma verilerini eğitmekten gelen tüm özellik vektörleri, 256'lı sınıflara kümelenerek, K-Ortalama kümeleme tekniği yardımıyla 256 ağırlık merkezli bir Kod Kitabı üretir. VQ, mesafe metriğini Kod Kitabına uygulayarak girdi özellik vektöründen ayrı gözlem dizisi elde etmek için kullanılır. Şekil 29'e göre özellik vektörlerini ayrık yapmak için, gelen her özellik vektörü, kod kitabındaki 256 prototip vektörün her biriyle karşılaştırılır. Sonrasında yakın olan vektörü (Öklid uzaklığı) seçilir ve giriş vektörü Kod Kitabındaki ilgili ağırlık merkez indeksi ile

değiştirilir. Böylece, tüm sürekli girdi özellik vektörleri, ayrık semboller kümesine VQ ile nicelenir.

5.6 Deney Setleri

Uygulamanın testi aşamasında 9 erkek, 6 kadın toplam 15 kişinin 15 kelimeyi üçer defa seslendirmeleri istenmiştir. Bu test seti için farklı test yolları denenmiştir.

A deney seti : 1 kişinin üçer defa seslendirdiği 6 kelime için her kelime 15 kişi tarafından üçer defa test edilmiştir.

B deney seti : 6 kelimenin her biri 15 kişi tarafından üçer defa seslendirilmiş ve her biri 3 kişi tarafından üçer defa test edilmiştir.

C deney seti : 10 farklı kelime için her bir kelime 15 kişi tarafından üçer defa seslendirilmiştir ve her bir kelime 3 farklı kişi tarafından üçer defa test edilmiştir.

D deney seti : 15 farklı kelimenin her biri 15 kişi tarafından üçer defa seslendirilerek her bir kelime 4 farklı kişi tarafından dörder defa test edilmiştir.

5.7 Deney Çalışmaları ve Değerlendirme Yöntemleri

Geçmişte yapılan çalışmalarda özellik çıkarımı aşamasında önerilen MFCC ile 12 katsayılı özellik vektörleri (Rabiner & Juang, 1993) elde edilmiştir. Bu çalışmada genelde konuşmacı tanıma uygulamalarında kullanılan ses bantlarında saklı konuşmacı kimliklerini ölçmeye de yarayan enerji ve delta türevleri ile 39 katsayılı özellik vektörleri (Tiwari ve ark., 2011) elde edilerek 12 katsayılı özellik vektörleri ile yapılan çalışma kıyaslanmış ve performans değerlendirmesi yapılmıştır. Yine geçmişte yapılan çalışmalar incelendiğinde ön işlemede pencereleme aşamasında Hamming Pencereleme sıkça önerilmiştir. Bu çalışmada diğer pencereleme teknikleri ile Hamming Pencereleme tekniği uygulanarak performans değerlendirilmesi yapılmıştır.

5.7.1 Enerji Özelliği

Bir çerçevedeki enerji, çerçevedeki örneklerin gücü toplamıdır; böylece zaman örneği t_1'den zaman örneği t_2'ye kadar bir penceredeki bir sinyal x için enerji şu şekildedir:

$$Energy = \sum_{t=t_1}^{t_2} x^2[t] \quad (43)$$

5.7.2 Delta Özelliği

Konuşma sinyali, kareden kareye sabittir. Hız (delta) ve ivmelenme (delta delta) katsayıları genellikle statik pencere tabanlı bilgi ile elde edilir. Bu delta ve delta delta katsayıları, bitişik pencereler arasındaki kepstral özellik vektörlerinin değişim hızını ve ivmesini modeller (Kinnunen, 2003). Deltalar, çerçeveler arasındaki farkın hesaplanması ile elde edilir; böylece belli bir kepstral c(t) zamanı t için delta değeri d(t) şu şekilde tahmin edilebilir:

$$d(t) = \Delta f_k[i] = f_{k+M}[i] - f_{k-M}[i] \quad (44)$$

Farklılaştırma yöntemi basittir fakat parametre alanında yüksek geçiren bir filtreleme işlemi görevi görmesi nedeniyle, gürültüyü yükseltme eğilimi gösterir. Bu soruna çözüm olarak doğrusal regresyon, yani birinci derece polinom alınır ve regresyon penceresi boyutu M=4 olacak şekilde en küçük kareler çözümü aşağıdaki formda gösterilir:

$$\Delta f_k[i] = \frac{\sum_{m=-M}^{M} m f_{k+m}[i]}{\sum_{m=-M}^{M} m^2} \quad (45)$$

Bu katsayılarla birlikte 12 MFCC, 12 Delta, 12 Delta Delta, 1 Enerji, 1 Delta Enerji, 1 Delta Delta Enerji toplam 39 özellik elde edilmiştir. Bu özellikler sınıflandırma aşamasına verilerek tahmin yapması beklenmektedir. C deney seti için performans değerlendirmesinde Hamming Pencereleme kullanılarak, 12 katsayılı özellik vektörleri ile %84,3, 39 katsayılı özellik vektörleri kullanılarak %86,6 başarı elde edilmiştir. Sonrasında 39 katsayılı özellik vektörleri ile Pencereleme yöntemleri kıyaslaması yapılmıştır.

5.7.3 Pencereleme Yöntemlerinin Kıyaslaması

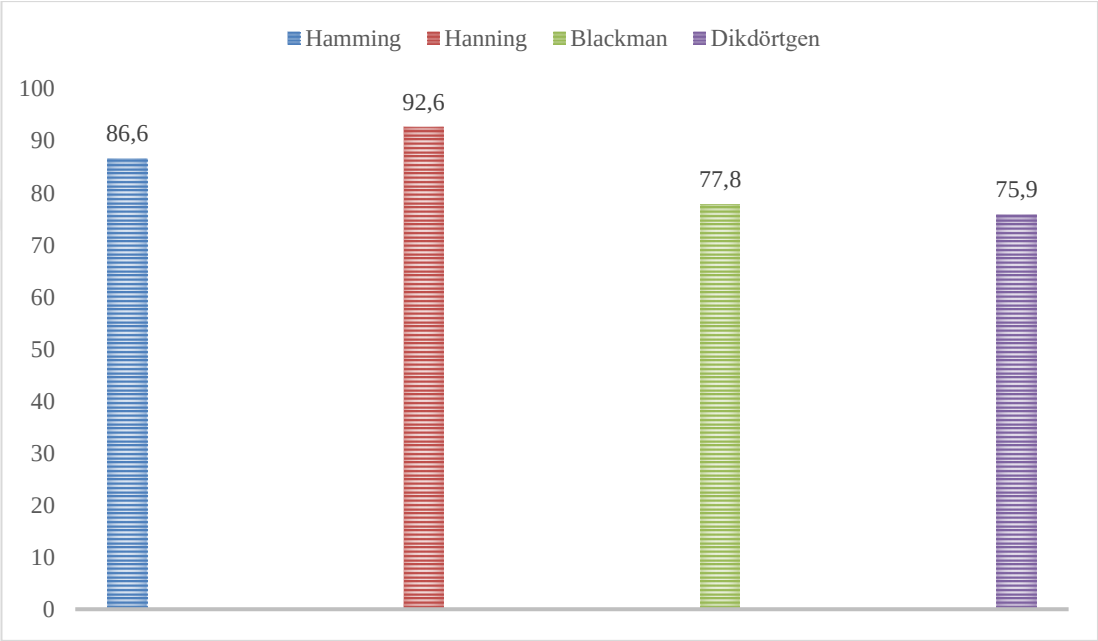
Pencereleme aşamasında Hamming, Hanning, Blackman ve Dikdörtgen Pencereleme yöntemleri B deney seti ile deneysel çalışmada kıyaslanarak değerlendirilmiştir. Pencereleme tekniği w(n), her bir pencere içindeki örnek sayısı N olacak şekilde formüller şu şekilde gösterilir;

Hamming Pencereleme: $w(n) = 0.54 - 0.46\left(\frac{2\pi n}{N-1}\right),\ 0 \leq N-1$ (46)

Hanning Pencereleme: $w(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right),\ 0 \leq N-1$ (47)

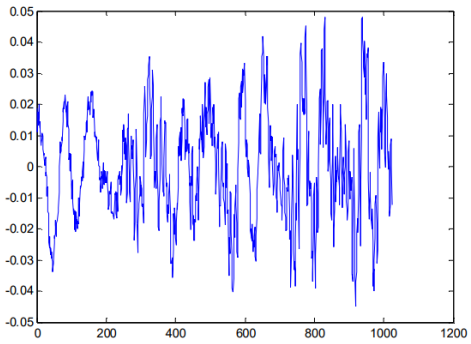
Blackman Pen.: $w(n) = 0.42 - 0.5\cos\left(\frac{2\pi n}{N-1}\right) + 0.08\cos\left(\frac{4\pi n}{N-1}\right),\ 0 \leq N-1$ (48)

Dikdörtgen Pencereleme: $w(n) = 1,\ 0 \leq N-1$ (49)

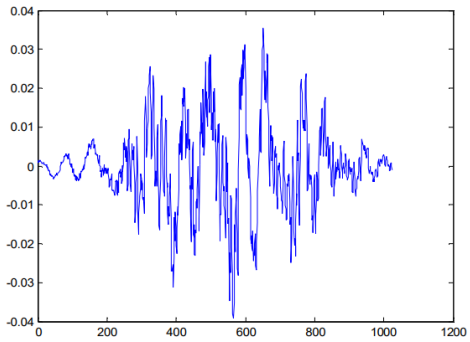


Şekil 30. Pencereleme tekniklerinin kıyaslanması

B deney seti ile Hanning Pencereleme'de %92,6 ile en yüksek başarım elde edilmiş ve önerilen çalışma için bu pencereleme türü kullanılmıştır.



Şekil 31. Hanning Pencereleme Öncesi

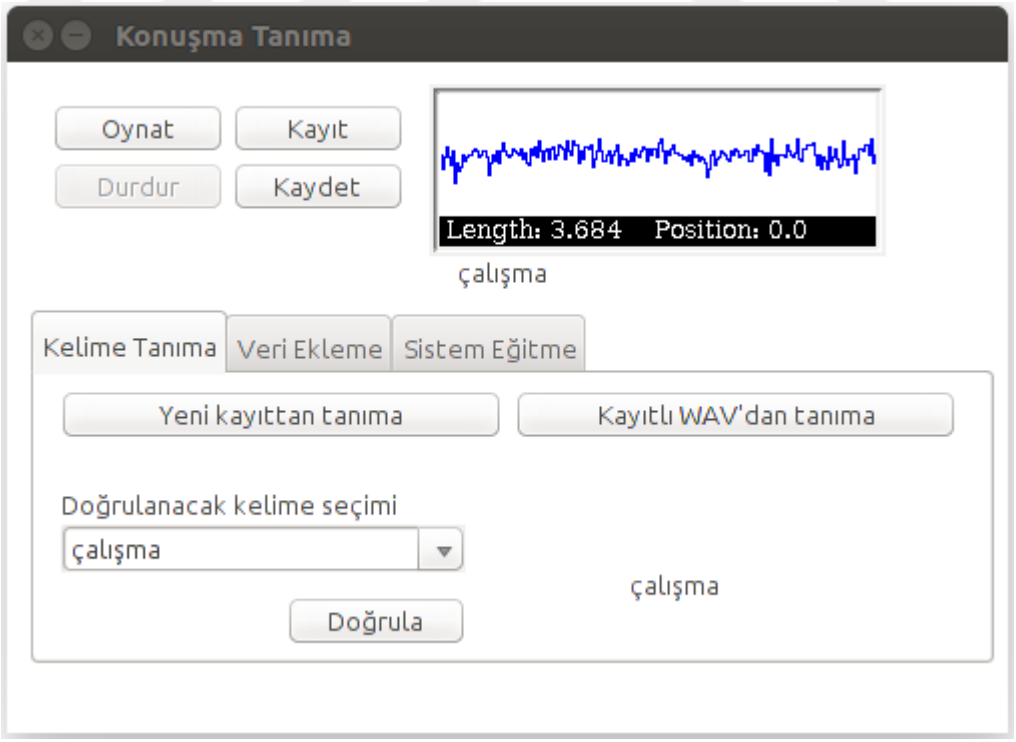


Şekil 32. Hanning Pencereleme Sonrası

Bu çalışmada ön işleme aşamasında çerçevelemeden sonra Hanning Pencereleme yapılmış, DFT alınmadan önce Mel filtre bankasında 30 filtre kullanılmış ve Delta-Delta enerji çıkarımı ile MFCC tekniği algoritmasında değişiklik yapılarak düşük enerjili bileşenlerin ve arka plandaki gürültünün etkisinde azalma sağlanmıştır.

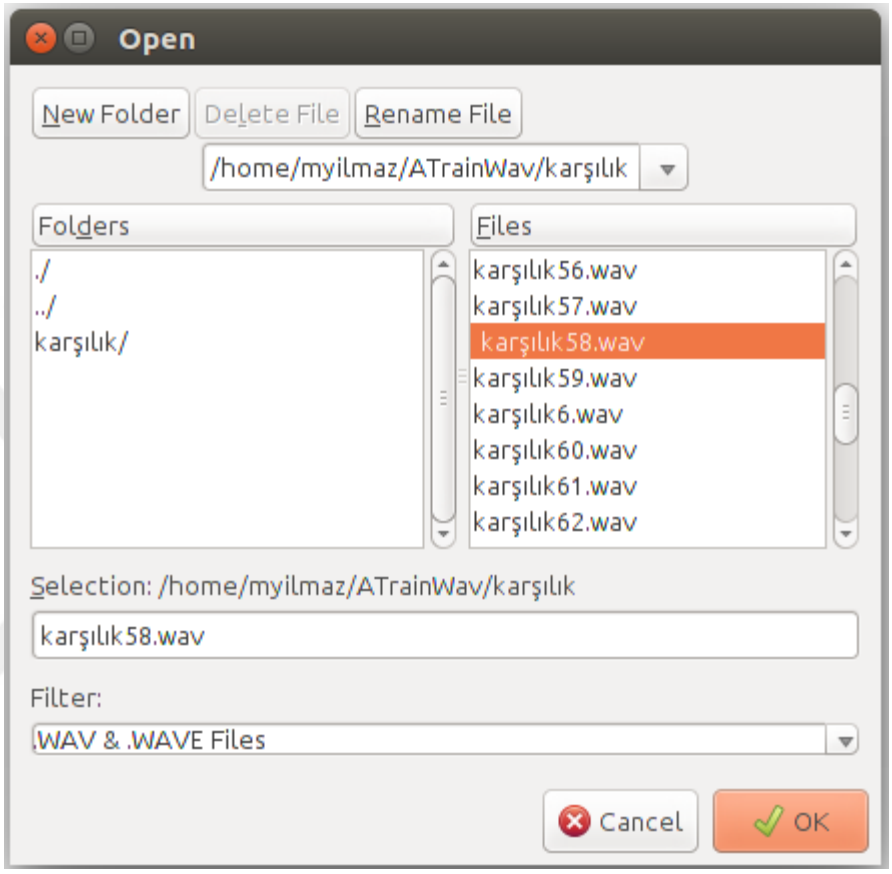
5.8 Uygulama Kılavuzu

Bu bölümde geliştirilen uygulamanın kullanıcı arayüzü tanıtılmaktadır. Uygulama ekranı Şekil 33'deki gibidir.



Şekil 33. Uygulama ekran görüntüsü

Uygulama 3 ana bölümden oluşmaktadır. Her bölüm programın alt yarısında gösterilmektedir. Üstte gösterilen bölüm konuşmanın kaydı, durdurulması, oynatılması ve kaydedilmesini sağlamaktadır. İlk bölüm "Kelime Tanıma" kısmıdır. Bu bölümde "Kayıt" ile konuşma kaydı yapılırken gerçek zamanlı (eş zamanlı) olarak ses dalgalarını gösteren bölümün altında olacak şekilde ekranda yazısı gözükmektedir.



Şekil 34. Önceden kayıtlı konuşmanın seçilmesi

Ayrıca kısa süreli konuşmalar için "Kayıt" ve "Bitir" (Kayıt'a basılınca gelmektedir) butonlarına basılarak alınan konuşma "Yeni kayıttan tanıma" butonuyla ya da önceden kayıtlı olan dosyalar için Şekil 34'de gösterildiği üzere "Kayıtlı WAV'dan tanıma" butonuna basılarak ses dosyası dizini seçilir ve tanınması istenir. Bu şekilde izole tanınan konuşmaların metni "Doğrula" butonun yanında bulunmaktadır.

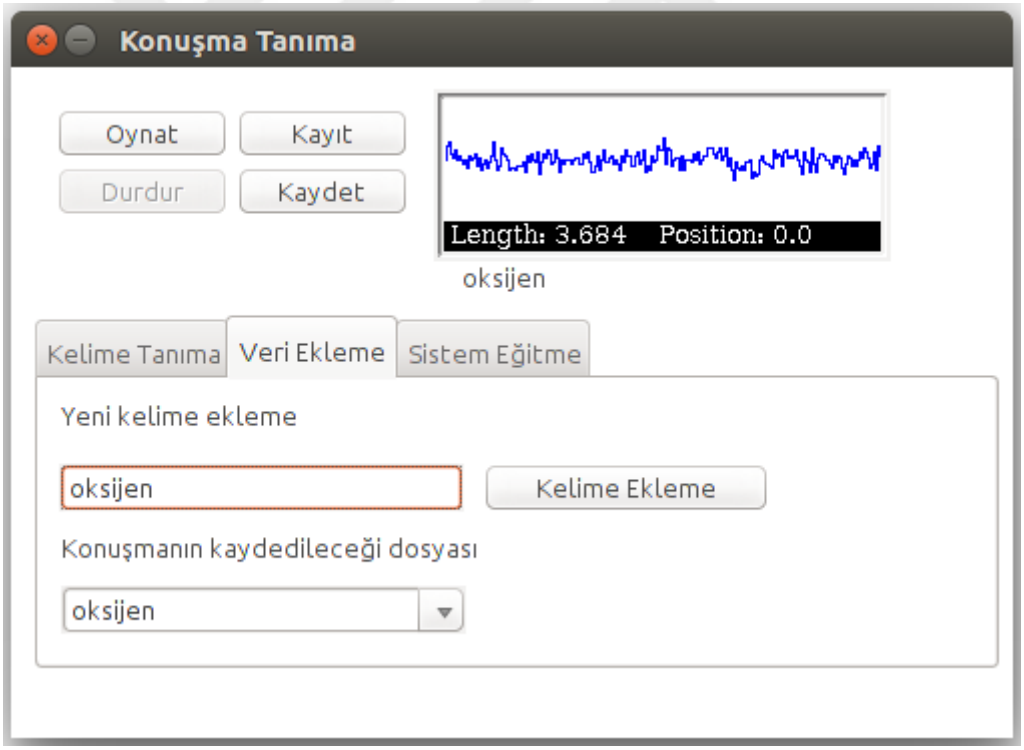
```
Likelihood with sigorta is -2521.438922909982
Likelihood with yönetici is -2403.405686255526
Likelihood with oksijen is -2333.6351221814357
Likelihood with cehalet is -2341.816959865233
Likelihood with karşılık is -2287.770440477055
Likelihood with mahalle is -2459.6984357779456
Best matched word karşılık
```

Şekil 35. Kelime tanıma konsol sonucu

Şekil 35'de gözükeceği üzere konuşma kaydına en çok benzeyen kelime ekranda yazdırılır. Tanınan kelime hatalı ise "Doğrulanacak kelime seçimi" listesinden seçilen kelime ile "Doğrula" butonuna basılarak sisteme geri bildirim yapılarak eğitilir.

5.8.1 Uygulamanın Eğitilmesi

Sistemin eğitilmesi için belirlenen konuşmaların metinleriyle etiketlenerek kaydedilmesi gerekmektedir. Bu çalışmada da ilk aşama olarak konuşma kaydı yapılır, dosya dizini oluşturulur ve konuşma dizin altına kaydedilir.

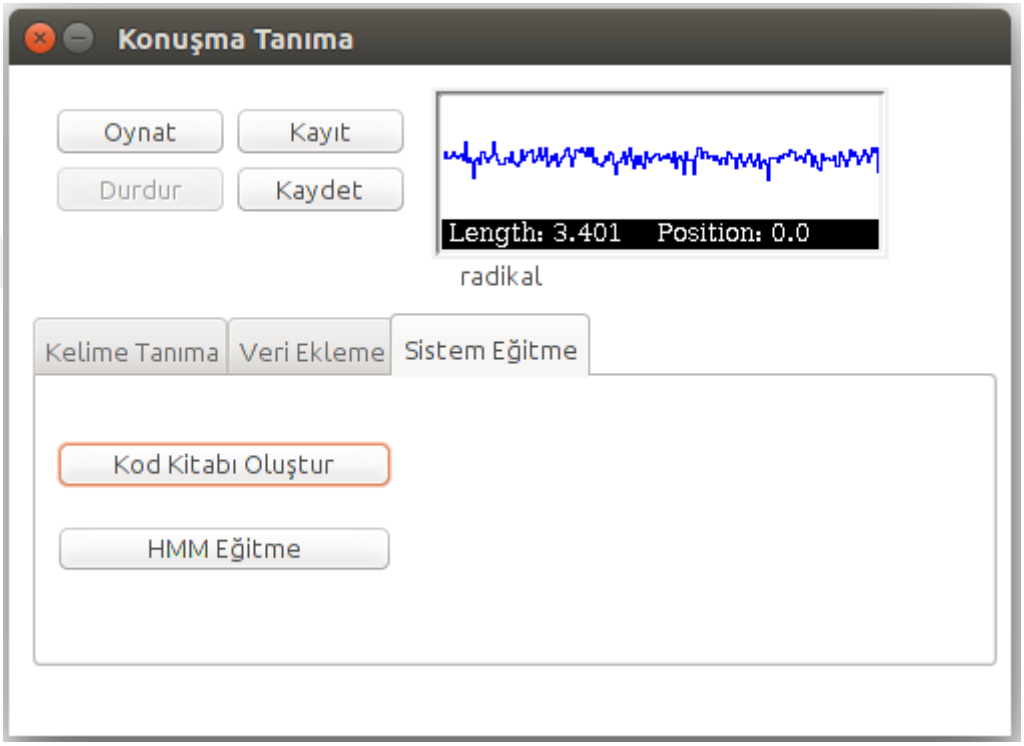


Şekil 36. Örnek konuşma ekleme

Uygulamanın konuşmaların kaydedildiği aşamasında Şekil 36'de gösterildiği üzere "Veri Ekleme" başlığı altındaki bölüme gidilir. "Yeni kelime ekleme" kısmına girilen metin adında klasörü "Kelime Ekleme" butonuna basarak oluştururuz. Sonrasında

"Kayıt" ve "Bitir" butonları ile kaydedilen konuşma için "Konuşmanın kaydedileceği dosyası" listesinden seçilen dosya dizininin altına konuşmayı kaydederiz.

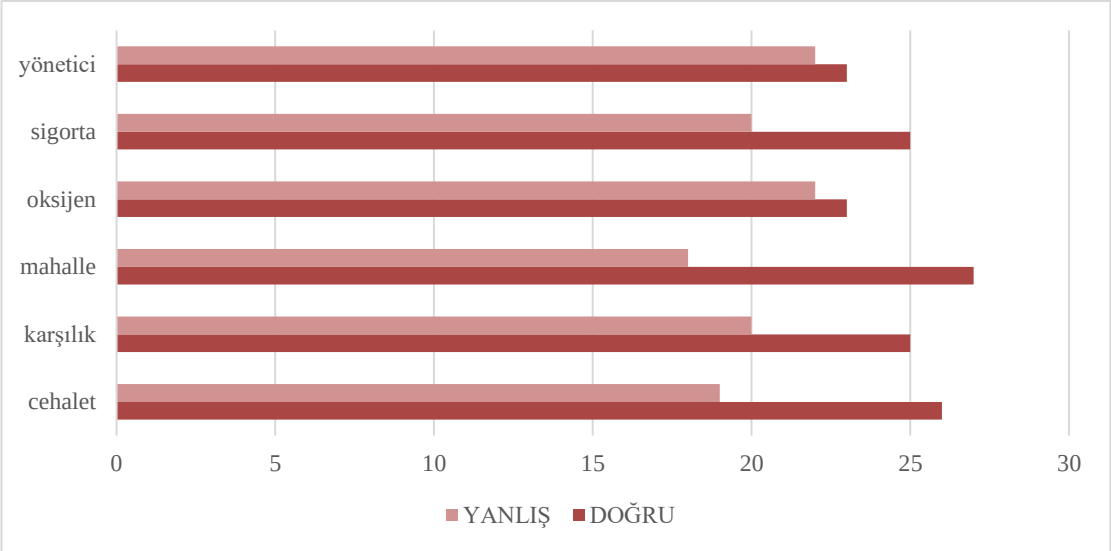
Uygulamanın eğitilmesi aşamasında Şekil 37'daki gibi "Sistem Eğitme" başlığı altındaki bölüme gidilir. Dizinler altına oluşturulan konuşmalar için ilk aşamada "Kod Kitabı Oluştur" butonu ile Kod Kitabı oluşturulur. Sonrasında "HMM Eğitme" butonu ile Kod Kitabında bulunan her bir kelimenin özellik vektörleri sınıflandırılır.



Şekil 37. Uygulamanın Eğitilmesi

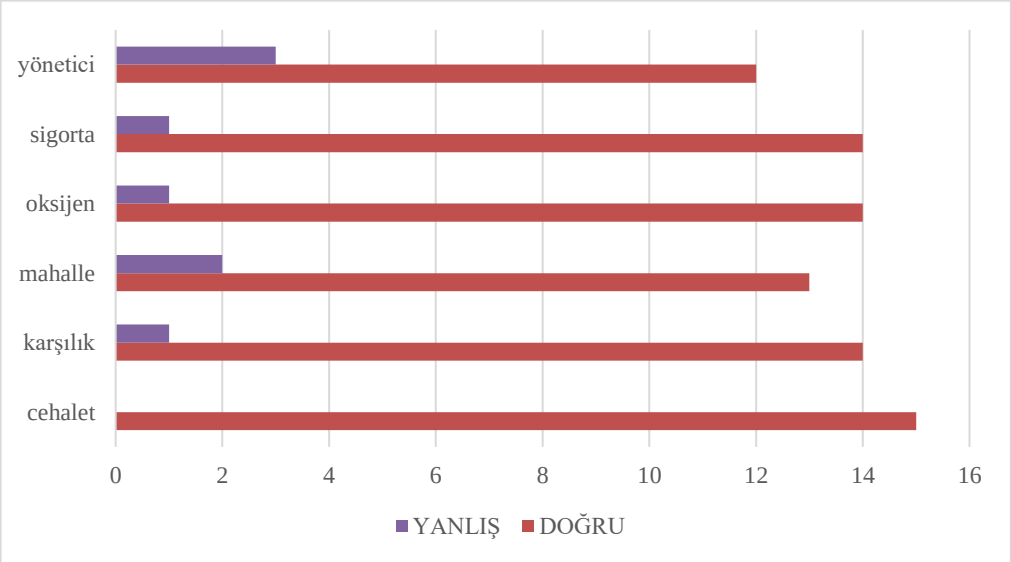
5.9 Uygulamanın Testi

Uygulamanın testi aşamasında farklı test yolları denenmiştir. İlk olarak A deney seti test edilmiştir.



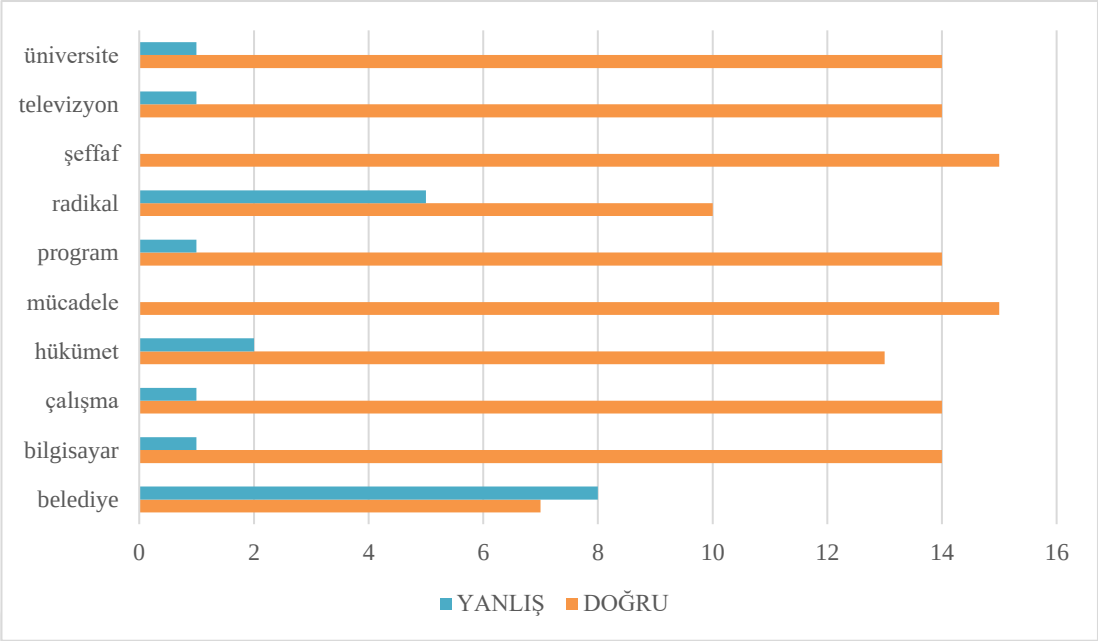
Şekil 38. A deney seti doğru-yanlış cevap sayısı

A deney test sonuçları Şekil 37'de gösterilmektedir. A deney testi sonucunda ortalama %55,2 başarı oranı sağlanmıştır. İkinci olarak B deney seti test edilmiştir.



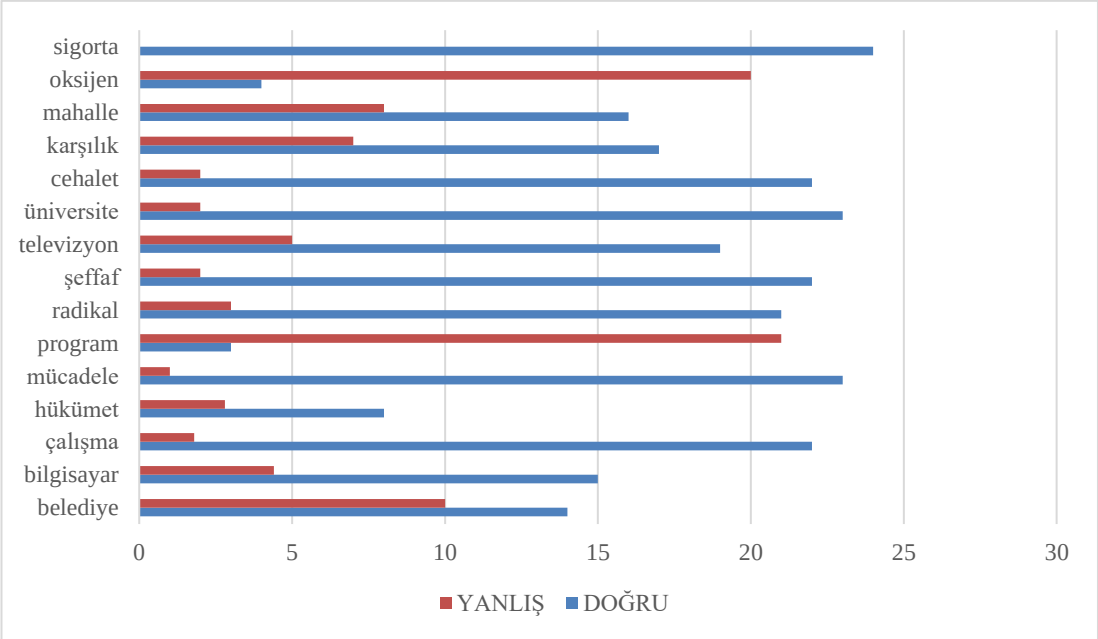
Şekil 39. B deney seti doğru-yanlış cevap sayısı

B deney testi sonrasında elde edilen doğru-yanlış sayıları Şekil 38'de gösterilmektedir. B deney testi aşamasında ortalama %92,6 doğruluk oranı elde edilmiştir. Üçüncü olarak C deney seti test edilmiştir.



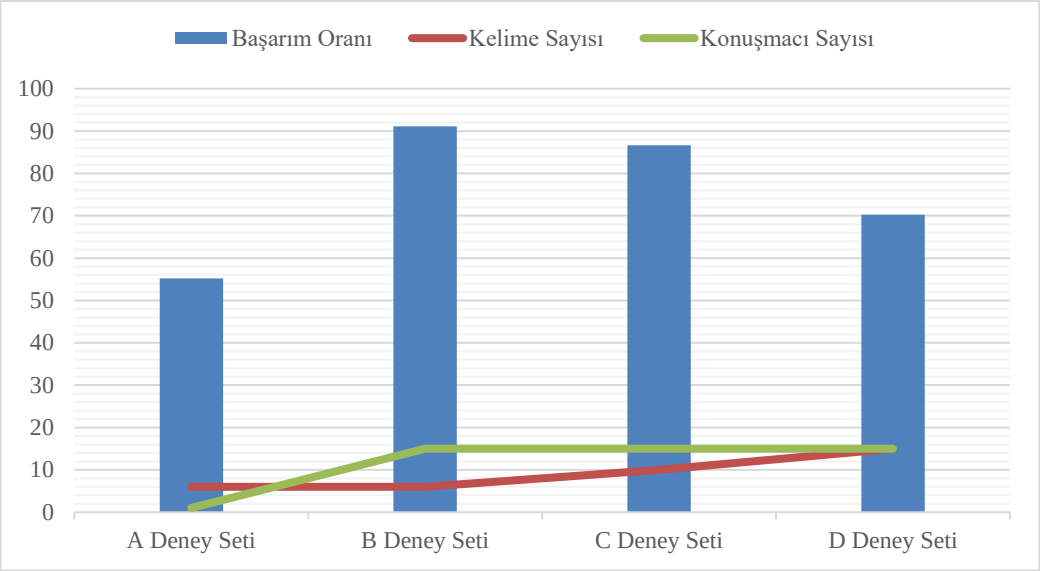
Şekil 40. C deney seti doğru-yanlış cevap sayısı

C deney testi sonrasında elde edilen doğru-yanlış sayıları Şekil 39'da gösterilmektedir. C deney testi aşamasında ortalama %86,67 doğruluk oranı elde edilmiştir. Dördüncü olarak D deney seti test edilmiştir.



Şekil 41. D deney seti doğru-yanlış cevap sayısı

D deney testi sonrasında elde edilen doğru-yanlış sayıları Şekil 40'da gösterilmektedir. D deney testi aşamasında ortalama %70,28 doğruluk oranı elde edilmiştir.



Şekil 42. Testlerin başarım oranları

5.10 Uygulama Değerlendirmesi

Konuşma tanıma teknolojisi, akademik araştırmalarda ve ticari uygulamalarda her geçen gün büyüyen bir alandır. Konuşma tanıma alanındaki zorluklar, tanıma oranı, arka plan gürültüsü, konuşmacı değişkenliği, konuşma oranı, aksan vb. gibi durumlardır. Bu zorlukları dışında bir konuşma tanıma sistemindeki uygulama performansı esas olarak özelliklerin çıkarımı ve sınıflandırma yöntemlerine bağlıdır (Kaur ve ark., 2017).

Konuşmacı bağımsız ve dil bağımsız olarak konuşmanın yazıya çevrilmesi konusunda her aşamanın detaylı olarak anlatıldığı bu çalışma alanına yeni bir bakış açısı katmaktadır. Tasarlanan konuşma tanıma sisteminde her bir kelime için telaffuz modelleri önem arz etmektedir. Buna bağlı olarak durumlar ve tanınmak istenen konuşma kaydında geçiş olasılıkları önemlidir. Aynı kelimenin telaffuzu kişiden kişiye farklılık gösterebilir. Bu nedenle her bir kelimeyi olabildiğince geniş bir grup için eğitmeliyiz, böylece oluşabilecek her bir kelimenin telaffuzundaki olası değişimleri modelleyebiliriz.

Uygulamanın testi aşamasında bahsedilen 4 test ile şu sonuçlar görülmüştür.

- Bu çalışma ile sisteme tanıtılan kelimelerde, konuşmacıların sayısı arttıkça sistemin doğruluk oranının da artacağı gözlenmiştir.

- Konuşmacılar, sistem eğitiminde her kelime için ne kadar fazla konuşmayı kaydederse başarımın da artacağı gözlenmiştir.
- Gerçek zamanlı seslendirme esnasında ortamda bulunan gürültü sistemin başarım oranını olumsuz olarak etkilemiştir.
- Aynı ortam içerisinde kaydedilen konuşmalar ile daha yüksek bir başarım oranı elde edilmektedir.
- Konuşmacı sayısı sabit tutulup kelime sayısı artan testlerde başarım oranının düştüğü gözlenmiştir.
- İzole kelime tanıma ile elde edilecek başarı oranının gerçek zamanlı kelime tanıma ile elde edilecek başarı oranından yüksek olduğu gözlenmiştir.
- Seslendirilen kelimelerde dil olarak farklı ağızların sisteme eğitim için verilmesi başarımı olumlu olarak etkilemektedir.
- Konuşmalar eğitim için sisteme izole olarak kaydedilirken konuşma anı dışında sessiz ortam dikkate alınmamaktadır. 0-100 ms gürültü olarak ele alınmaktadır. Bu sebeple konuşma yapılırken kaydetme işlemine başlandığında programın yanılma payı yükselmektedir.

6 SONUÇLAR VE ÖNERİLER

Bu tezde konuşmacı ve dil bağımsız gerçek zamanlı ve kaliteli bir konuşma tanıma sistemi önerilmiştir. Çalışma, her bir konuşmanın sisteme metni ile etiketlenmesiyle geliştirilir. Her bir konuşma özellik çıkarımı ve sınıflandırma aşamalarından geçerek tanınır. Bu aşamaların seçimi esnasında geçmişte yapılan çalışmalardan yola çıkarak en verimli teknikler belirlenerek sistem önerimi yapılmıştır. Özellik çıkarımı aşamasında tekniğin daha iyi sonuçlar vermesi için bazı karşılaştırmalar yapılmıştır. Sınıflandırma aşamasında teknik içerisinden bazı algoritmalar ile sistemin eğitilmesi ve testi gerçekleştirilmiştir. Yapılan bu çalışmayla konuşma tanıma sisteminin zenginleşmesi ve yüksek verime ulaşması için her kullanıcının sisteme katkı sağlaması planlanmaktır.

Tez kapsamında konuşmacı sayısı ve kelime sayısı değişen 4 farklı test yapılarak deneysel çalışma değerlendirilmiştir. Yapılan değerlendirme ile bu çalışmanın artan konuşmacı sayısıyla başarımının da artacağı gözlenmiştir. Gelecekte önerilen çalışma için daha yüksek başarısı olabilecek çalışmalar yapılması planlanmaktadır. Özellik çıkarımı yönteminde uygulanan adımlardan sonra sınıflandırma alanında farklı tekniklerin kullanılmasıyla çalışmalar yapılması planlanmaktadır. Bunlara ilave olarak gerçek zamanlı konuşma anında algılanamayan kelimeler için muhtemel kelime tavsiyesinin yapılması planlanmaktadır.

Bu tez ile elde edilen sonuçlara göre çalışmanın katkıları aşağıda sunulmuştur.

- Literatür tarama ile konuşma tanıma için gerekli adımlar oluşturularak bu adımlar için kullanılabilecek teknikler ifade edilmiştir.
- Kullanılabilecek olan teknikler için önceden yapılan çalışmalardan elde edilen başarı performansları çizelge olarak gösterilmiştir.
- Önerilen çalışma için özellik çıkarımı aşamasında farklı yöntemlerin karşılaştırılması yapılmıştır.

- Önerilen çalışma için sınıflandırma aşamasında verimli teknikler belirlenerek olasılık hesaplanması ile Kod Kitabı oluşturulmuş ve kullanıcıya en yüksek olasılıklı kelime gösterilmiştir.
- Önerilen çalışma için farklı veri setleri oluşturularak kelime, konuşmacı ve kelimelerin kaç defa seslendirildiği göz önüne alınarak başarı performansları hesaplanmıştır.
- Testler sonucunda hesaplanan performanslara göre başarıyı etkileyen faktörler belirtilmiştir.
- Her bir kelime için farklı konuşmacılardan alınan sesler ile sistemin eğitilmesi sonucunda başarı performansının artacağı ifade edilmiştir.

KAYNAKÇA

Abdulla, W. H., Chow, D., & Sin, G. (2003). Cross-words reference template for DTW-based speech recognition systems. *TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region, 15-17 Oct.* Bangalore, India, India : IEEE.

Abushariah, M. A., Ainon, R. N., Zainuddin, R., Elshafei, M., & Khalifa, O. O. (2010, May). Natural speaker-independent Arabic speech recognition system based on Hidden Markov Models using Sphinx tools. *IEEE Computer and Communication Engineering (ICCCE).*

Adami, A. G. (2010). Automatic Speech Recognition: From the Beginning to the Porteguese Language. *Universidade de Caxias do Sul, Centro de Computação e Tecnologia da Informação Rua Francisco Getúlio Vargas.* Brasil.

Al-Haddad, S. R., Samad, S. A., Hussain, A., & Ishak, K. A. (2008). Isolated Malay Digit Recognition Using Pattern Recognition Fusion of Dynamic Time Warping and Hidden Markov Models. *American Journal of Applied Sciences 5 (6): 714-720* (s. 714-720). 2008 Science Publications.

Alonso, J. B., Cabrera, J., Travieso, C. M., López-de-Ipiña, K., & Sánchez-Medina, A. (2017). Continuous tracking of the emotion temperature. *Neurocomputing, Volume 255, 13 September* (s. 17-25). Available online at www.sciencedirect.com.

Ananthi, S., & Dhanalakshmi, P. (2014). SVM and HMM Modeling Techniques for Speech Recognition Using LPCC and MFCC Features. *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)* (s. 519-526). Part of the Advances in Intelligent Systems and Computing book series (AISC, volume 327).

Aydın, Ö. (2005). *Yapay sinir ağlarını kullanarak bir ses tanıma sistemi geliştirilmesi Yüksek Lisans Tezi.* Edirne: Trakya Üniversitesi Fen Bilimleri Enstitüsü.

Azim, M. A., Hamid, A. A., Badr, N. L., & Tolba, M. F. (2016). Tree-Based HMM State Tying for Arabic Continuous Speech Recognition. *International Conference on Advanced Intelligent Systems and Informatics, AISI 2016: Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, 18 October* (s. 96-103). Springer.

Bakır, Ç. (2016). Alman Dili Üzerinde Konuşmacı Cinsiyetinin Otomatik Olarak Belirlenmesi. *Akademik Platform* (s. 52-58). 10.21541/apjes.24088.

Baygün, M. K. (2006). *Türkçe Komutları Tanıyan Ses Tanıma Sistemi Geliştirilmesi Yüksek Lisans Tezi.* Denizli: Pamukkale Üniversitesi.

Bayya, A., & Steiger, D. L. (2002). Speaker dependent speech recognition training using simplified hidden markov modeling and robust end-point detection. *US6405168 B1*. Conexant Systems, Inc.

Becchetti, C., & Ricotti, L. P. (1999). *Speech Recognition Theory and C++ Implementation.* England: John Wiley & Sons Ltd, 167-188, 310-311.

Becerra, A., Rosa , J. I., & González, E. (2016). A case study of speech recognition in Spanish: From conventional to deep approach. *ANDESCON.* Arequipa, Peru: IEEE.

Bhaskar, P. V., & Mohana Rao, S. R. (2014, February). Telugu Speech Recognition System development using MFCC based Hidden Markov Model technique with Sphinx-4. *IJECEAR, 2*(2), 141-147.

Blunsom, P. (2004, August 19). Hidden Markov Models. Y. Changhui (Dü.), *CS7960 Machine Learning Methods for Bioinformatics.* içinde Utah.

Cai, J., Wang, N., Wang, H., & Zhu, B. (2016). Research on the recognition of isolated Chinese lyrics in songs with accompaniment based on deep belief networks. *Signal Processing (ICSP), 2016 IEEE 13th International Conference, 6-10 Nov.* Chengdu, China: IEEE.

Chavan, R. S., & Sable, G. S. (2013). An Implementation of Text Dependent Speaker Independent Isolated Word Speech. *International Journal of Engineering Sciences & Research, Chavan, 2(9): September* (s. 2311-2318). India: www.ijesrt.com.

Chen, J. K., & Soong, F. K. (1994). An N-Best Candidates-Based Discriminative Training for Speech Recognition Application. *IEEE Transactions on Speech and Audio Processing, 2*(1), 206-216.

Choudhary, A., Chauha, R. S., & Gupta, G. (2013). Automatic Speech Recognition System for Isolated & Connected Words of Hindi Language By Using Hidden Markov Model Toolkit (HTK). *Association of Computer Electronics and Electrical Engineers*, 847-853.

Çakır, M. Y. (2017). Derin Sinir Ağları ile Konuşma Tespiti ve Cinsiyet Tahmini. *22.Türkiye'de İnternet Konferansı.* Beşiktaş, İstanbul: Bahçeşehir Üniversitesi.

Çelebi, M., & Buldu, A. (2014). Ses Komut Tanıma İle Gezgin Araç Kontrolü. *Akademik Platform* (s. 395-404). Karabük: ISITES2014.

Daoerji, F., & Guanglai, G. (2016). DNN-HMM for Large Vocabulary Mongolian Offline Handwriting Recognition. *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on, 23-26 Oct.* Shenzhen, China: IEEE.

Davis, K. H., Biddulph, R., & Balashek, S. (1952). Automatic Recognition of Spoken Digits. *Journal of the Acoustic Society of America, 24*(6), 637-642.

Debyeche, M., Haton, J. P., & Houacine, A. (2007). A New Vector Quantization front-end Process for Discrete HMM Speech Recognition System. *International Journal of Computer, Electrical, Automation, Control and Information Engineering, 1*(6), 1627-1632.

Dede, G. (2008). *Yapay Sinir Ağları ile Konuşma Tanıma Yüksek Lisans Tezi.*

Ankara: Ankara Üniversitesi Fen Bilimleri Enstitüsü.

Deng, L., & O'Shaughnessy, D. (2003). *A Dynamic and Optimization-Oriented Approach.* Inc. New York, NY, U.S.A.: Published by Marcel Dekker.

Donaj, G., & Kačič, Z. (2017). Context-dependent factored language models. *EURASIP Journal on Audio, Speech, and Music Processing, December.* Springer.

El Maghraby, E. E., Gody, A. M., & Farouk, M. H. (2016). Enhancing quality and accuracy of speech recognition system by using multimodal audio-visual speech signal. *Computer Engineering Conference (ICENCO), 2016 12th International, 28-29 Dec.* Cairo, Egypt: IEEE.

El-Ramy, S. H., Abdel-Kader, N. S., & El-Adawi, R. (2002). Neural Networks Used for Speech Recognition. *Nineteenth National Radio Science Conference*, (s. 200-207). Alexandria.

Eray, O. (2008). *Destek Vektör Makineleri İle Ses Tanıma Uygulaması Yüksek Lisans Tezi.* Denzili: Pamukkale Üniversitesi.

Ford, T. L. (2004). Speech Idiosncrasies are the Nemesis of Speech Recognition Software. *University of Maryland University College.*

Forgie, J. W., & Forgie, C. D. (1959). Results Obtained From a Vowel Recognition Computer Program. *Journal of the Acoustic Society of America, 31*(11), 1480-1489.

Froomkin, D. (2015, 5 5). *The Computers Are Listening*. The Intercept: https://theintercept.com/2015/05/05/nsa-speech-recognition-snowden-searchable-text/ adresinden alındı

Fry, D. B. (1959). Theoretical Aspects of Mechanical Speech Recognition. *Journal of the British Institution Radio Engineers, 19*(4), 211-229.

Furui, S. (1991). Speaker-dependent-feature extraction, recognition and processing techniques. *Speech Communication, Volume 10, Issues 5–6, December*, (s. 505-520).

Ganchev, T., Fakotakis, N., & Kokkinakis, G. (2005). Comparative evaluation of various MFCC implementations on the speaker verification task. *10th International Conference on Speech and Computer (SPECOM 2005)*, s. 191-194.

Gawali, B. W., Gaikwad, S., Yannawar, P., & Mehrotra, S. C. (2011). Marathi Isolated Word Recognition System using MFCC and DTW Features. *Int. J. on Information Technology, Vol. 01, No. 01, Mar* (s. 21-24). ACEEE.

Gelegin, İ., & Bolat, B. (2011). *Ayrık Kelime Tabanlı Bir Konuşma Tanıma Sistemiyle Bilgisayar Kontrolü.* Elazığ: Elektrik-Elektronik ve Bilgisayar Sempozyumu.

Ghai, W., & Singh, N. (2012, March). Literature Review on Automatic Speech Recognition. *International Journal of Computer Applications (0975-8887), 41*(8).

Gorthi, R., Joshi, C. G., & Shah, R. J. (2016). Call context metadata. International Business Machines Corporation.

Hasnat, A. M., Mowla, J., & Khan, M. (2007). *December 2007*. Hanoi, Vietnam: International Symposium on Natural Language Processing (SNLP).

Huang, X., Baker, J., & Reddy, R. (2014). A Historical Perspective of Speech Recognition. *Communications of the ACM, Vol. 57 No. 1, January* (s. 94-103). ACM.

IBM. (1960). *IBM Archives: Shoebox*. IBM special products (vol. 1): http://sysrun.haifa.il.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html adresinden alındı

Imtiaz, M. A., & Raja, G. (2016). Isolated word Automatic Speech Recognition (ASR) System using MFCC, DTW & KNN. *Multimedia and Broadcasting (APMediaCast), 2016 Asia Pacific Conference, 17-19 Nov.* Bali, Indonesia : IEEE.

Itakura, F. (1975). Minimum Prediction Residual Applied to Speech Recognition. *IEEE Transactions on Acoustic, Speech, Signal Processing, ASSP-23*(1), 67-72.

Joshi, S. C., & Cheeran, A. N. (2014). MATLAB Based Back-Propagation Neural Network for Automatic Speech Recognition. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 3, Issue 7, July* (s. 10498-10504). IJAREEIE.

Juang, B. H., & Rabiner, L. R. (2004). *Automatic Speech Recognition.* A Brief History of the Technology Development.

Jurafsky, D., & Martin, J. H. (2006). *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition.* http://stp.lingfil.uu.se/~santinim/ml/2014/JurafskyMartinSpeechAndLanguageProcessing2ed_draft%202007.pdf adresinden alındı

Karthikeyan, V., & Vijayalakshmi, V. J. (2016). Performance Compariosn Of Speech Recognition For Voice Enabling Applications. *American Journal of Engineering and Technology Research, Vol. 16, No.1*, (s. 48-56).

Kaur, G., Srivastava, M., & Kumar, A. (2017). Analysis of Feature Extraction Methods for Speaker Dependent Speech Recognition. *International Journal of Engineering and Technology Innovation, Vol 7, No 2.* IJETI.

Keerio, A., Mitra, B. K., Birch, P., Young, R., & Chatwin, C. (2009, January). On Preprocessing of Speech Signals. *Processing, International Journal of Signal, 5*(3).

Këpuska, V. Z., & Elharati, H. A. (2015, June). Robust Speech Recognition System Using Conventional and Hybrid Features of MFCC, LPCC, PLP, RASTA-PLP and Hidden Markov Model Classifier in Noisy Conditions. *Journal of Computer and Communications*(3), 1-9.

Kincaid, J. (2011, 2 13). *The Power Of Voice: A Conversation With The Head Of Google's Speech Technology*. Tech Crunch: https://techcrunch.com/2011/02/13/the-power-of-voice-a-conversation-with-the-head-of-googles-speech-technology/ adresinden alındı

Kinnunen, T. (2003). *Spectral Features for Automatic Text-Independent Speaker*. Finland: University of Joensuu, Department of Computer.

Koumpis, K., & Pavitt, K. (1999). Corporate Activities In Speech Recognition And Natural Language: Another "New Science"-Based Technology. *International Journal of Innovation Management*.

Kuş, P. (1998). *Ses Sinyallerinin Düşük Hızda İletimi.* Hacettepe Üniversitesi.

Larcher, A., Lee, K. A., Ma, B., & Li, H. (2014). Text-dependent speaker verification: Classifiers, databases and RSR2015. *Speech Communication 60* (s. 56–77). Available online at www.sciencedirect.com.

Lin, Y. L., & Wei, G. (2005). Speech Emotion Recognition Based on HMM and SVM. *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, (s. 4898-4901). Guangzhou.

Mammone, R. J., Zhang, X., & Ramachandran, R. P. (1996). Robust Speaker Recognition: A Feature-Based Approach. *IEEE Signal Processing Magazine (Volume: 13, Issue: 5)* (s. 58). IEEE.

Manning, C. D., & Schütze, H. (1999). *Foundations Of Statistical Natural Language Processing.* Cambridge: Massachusetts Institute of Technology.

Mari, J. F., Fohr, D., & Junqua, J. C. (1996). A second-order HMM for high performance word and phoneme-based continuous speech recognition . *Acoustics, Speech, and Signal Processing, ICASSP-96. Conference Proceedings.* Atlanta, USA: IEEE International Conference.

Mengüşoğlu, E. (1999). *Bir Türkçe Sesli İfade Tanıma Sisteminin Kural Tabanlı Tasarımı ve Gerçekleştirimi.* Ankara: Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, s. 14-16, 22-26.

Moonasar, V., & Venayagamoorthy, G. K. (2001). A committee of neural networks for automatic speaker recognition (ASR) systems . *Missouri University of Science and Technology Scholars' Mine* (s. 2936-2940). IEEE.

Murty, K. R., & Yegnanarayana, B. (2006). Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE Signal Processing* (s. 52-55). IEEE.

Müller, M. (2007). Information Retrieval for Music and Motion. *Springer*, s. 65. doi:ISBN 978-3-540-74047-6

Neto, J., Almeida, L., Hochberg, M., Martins, C., Nunes, L., Renals, S., & Robinson, T. (1995). Speaker-Adaptation For Hybrid HMM-ANN Continuous Speech Recognition System. *4th European Conference on Speech Communication and Technology EUROSPEECH, September, 18-21* (s. 2171-2174). Madrid: ISCA.

Nicholson, J., Takahashi, K., & Nakatsu, R. (2000). Emotion Recognition in Speech Using Neural Networks. *Neural Computing & Applications, December, Volume 9, Issue 4* (s. 290–296). Springer.

Olson, K. H., & Belar, H. (1956). Phonetic Typewriter. *Journal of the Acoustic Sociecty of America, 28*(6), 1072-1081.

Oracle. (2017, 12 10). Oracle Help Center: https://docs.oracle.com/en adresinden alındı

Ostendorf, M., & Roukos, S. (1989). A stochastic segment model for phoneme-based continuous speech recognition. *IEEE Transactions on Acoustics, Speech, and*

Signal Processing (Volume: 37, Issue: 12, Dec) (s. 1857 - 1869). IEEE.

Öcal, K. (2005). *Otomatik Konuşma Tanıma Algoritmalarının Uygulamaları Yüksek Lisans Tezi.* Ankara: Ankara Üniversitesi Fen Bilimleri Enstitüsü.

Pallet, D. S., Fiscus, J. G., & Garofolo, J. S. (1990, June). DARPA Resource Management Benchmark Test Results. *National Institute of Standarts and Technology (NIST),* 298-305.

Pao, T.-L., Chen, Y.-T., Yeh, J.-H., & Li, P.-J. (2006). Mandarin Emotional Speech Recognition Based on SVM and NN. *Pattern Recognition, ICPR. 18th International Conference on, 18 September.* Hong Kong, China : IEEE.

Polur, P. D., Zhou, R., Yang, J., Fedra, A., & Hobson, R. S. (2001). Isolated Speech Recognition Using Artifical Neural Networks. *23rd Annual EMBS Conference.*

Prakoso, H., Ferdiana, R., & Hartanto, R. (2016). Indonesian Automatic Speech Recognition system using CMUSphinx toolkit and limited dataset. *Electronics and Smart Devices (ISESD), International Symposium, 29-30 Nov.* Bandung, Indonesia: IEEE.

Rabiner, L. R., Levinson, S. E., & Sondhi, M. M. (1983). On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition. *TOC, Volume 62, Issue 4, April* (s. 1075-1105). Bell Labs Technical Journal.

Rabiner, L., & Juang, B. (1993). *Fundamental of Speech Recognition.* Englewood Cliffs, New Jersey: PTR Prentice Hall.

Reddy, D. R. (1967). Computer Recognition of Connected Speech. *Journal of the Acoustic Society of America, 42,* 329-347.

Sahidullah, M., & Saha, G. (2012, May). Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *4*, s. 543-565.

Saini, P., & Kaur, P. (2013). Automatic Speech Recognition: A Review. *International Journal of Engineering Trends and Technology- Volume4Issue2* (s. 132-136). ACE, Haryana, India: CSE Department, Kurukshetra University.

Sakai, T., & Doshita, S. (1962). The Phonetic Typewriter, Information Processing. *Proceedings ıf IFIP Congress*, (s. 445-450). Munich.

Sakoe, H., & Chiba, S. (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transaction on Acoustics, Speech and Signal Processing, ASSP-26*(1), 43-49.

Salloum, W., Edwards, E., Ghaffarzadegan, S., & David. (2017). Crowdsourced Continuous Improvement of Medical Speech Recognition. *In Proc. of the Workshop on Crowdsourcing, February.* San Francisco, USA: Deep Learning and Artificial Intelligence Agents at the Thirty-First AAAI Conference on Artificial Intelligence.

Scheme, E. J., Hudgins, B., & Parker, P. A. (2007). Myoelectric Signal Classification for Phoneme-Based Speech Recognition. *IEEE Transactions on Biomedical Engineering (Volume: 54, Issue: 4, April)* (s. 694 - 699). IEEE Engineering

in Medicine and Biology Society.

Seide, F., Li, G., Chen, X., & Yu, D. (2011). Feature engineering in context-dependent deep neural networks for conversational speech transcription. *in Proc. ASRU, pp.*, (s. 24-29).

Shipra, & Chandra, M. (2016). Hindi Vowel Classification using QCN-PNCC Features. *Indian Journal of Science & Technology, Volume 9, Issue 38, October.*

Shrawankar, U., & Thakare, V. M. (2013). Techniques for Feature Extraction In Speech Recognition System : A Comparative Study. *International Journal Of Computer Applications In Engineering, Technology and Sciences (IJCAETS),ISSN 0974-3596,2010, 6 May* (s. 412-418). Cornell University Library.

Silva, D. F., & Batista, G. E. (2016). Speeding Up All-Pairwise Dynamic Time Warping Matrix Calculation. *Proceedings of the 2016 SIAM International Conference on Data Mining.* PRDT16.

Singh, B., Kapur, N., & Kaur, P. (2012). Speech Recognition with Hidden Markov Model: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering, 2*(3), 400-403.

Singh, K. (2016). Speech Recognition: A Review of Literature. *International Journal of Engineering Trends and Technology (IJETT) - Volume 37 Number 6, July*, (s. 302-310). Patiala, India.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America, 8 (3),* (s. 185–190).

Sunny, S., Peter, S. D., & Jacob, K. P. (2012). A comparative study of parametric coding and wavelet coding based feature extraction techniques in recognizing spoken words. *CUBE International Information Technology Conference.* Semantic Scholar.

Suzuki, J., & Nakata, K. (1961). Recognition of Japanese Vowels-Preliminary to the Recognition of the Speech. *Journal of Radio Research Lab., 37*(8), 193-212.

Tabassum, M., Aziz Jahan, M. A., Rahman, M. M., Mohamed, S. B., & Rashid, M. A. (2017). Speaker Independent Speech Recognition of Isolated Words in Room Environment. *International Journal on Advanced Science, Engineering and Information Technology , Vol 7, No 2*, (s. 475-481).

Therrien, C. W. (1989). Decision estimation and classification: an introduction to pattern recognition and related topics. Naval Postgraduate School, Monterey, CA : ISBN:0-471-83102-6.

Tiwari, G., Pandey, M., & Shrestha, M. (2011). *Text-Prompted Remote Speaker Authentication.* Lalitpur, Nepal: Tribhuvan University Institute Of Engineering.

Utane, A. S., & Nalbalwar, S. L. (2013, April). Emotion Recognition Through Speech Using Gaussian Mixture Model And Hidden Markov Model. *International Journal of Advanced Research in Computer Science and Software Engineering, 3*(4), 742-746.

Uzunçarşılı, M. (2005). *Vektör Nicemleme Tekniklerine Dayalı Konuşmacı Tanıma Algoritmalarının İncelenmesi.* Ankara: Ankara Üniversitesi Fen Bilimleri Üniversitesi.

Valicek, J. (2017). *Language Models for Multilingual Continuous Speech Recognition.* Leden: Czech Technical University in Prague, Faculty of Electrical Engineering Department of Radioelectronics.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* New York: Springer-Verlag Publishing.

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

Vintsyuk, T. K. (1968). Speech Discrimination by Dynamic Programming. *Kibernetica, 4*(2), 81-88.

Wildstrom, S. (2011, 10 10). *Nuance Exec on iPhone 4S, Siri, and the Future of Speech*. Tech Pinions: https://techpinions.com/nuance-exec-on-iphone-4s-siri-and-the-future-of-speech/3307 adresinden alındı

Yalçın, N. (2008, Mart). Konuşma Tanıma Teorisi ve Teknikleri. *Kastamonu Eğitim Dergisi, 16*(1), 249-266.

Yaldır, A., & Baygün, M. K. (2006, Şubat). Linear Predictive Coding ve Dynamic Time Warping Teknikleri Kullanılarak Ses Tanıma Sistemi Geliştirilmesi. Denizli: Akademik Bilişim.

Yaşaroğlu, Y. (2003). *Multi-Modal Video Summarization Using Hidden Markov Models For Content-Based Multimedia Indexing.* Ankara: The Middle East Technical University.

Young, S. (1989). Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems. *University of Cambridge, October.* ResearchGate.

Zadeh, L. A. (1965). Fuzzy Sets. *Information And Control, 8*, s. 338-353.

Zeghidour, N., Usunier, N., Kokkinos, I., & Schatz, T. (2017). Learning Filterbanks From Raw Speech For Phone Recognition. *Cornell University Library.* Computation and Language.

Zhou, G., Hansen, J. J., & Kaiser, J. F. (2001). Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing, Volume: 9, Issue: 3, Mar* (s. 201 - 216). IEEE Signal Processing Society.

ÖZGEÇMİŞ

Mert Yılmaz Çakır, 1993 yılında Eminönü, İstanbul'da doğdu. İlk ve orta öğrenimini İstanbul'da tamamladı. 2015 yılında İstanbul Sabahattin Zaim Üniversitesi'nin ilk mezunlarından olarak Bilgisayar Mühendisliği bölümünden birinci, Mühendislik ve Doğa Bilimleri Fakültesi'nden en iyi ikinci dereceyle mezun olmuştur. 2014 yılında Matriks Bilgi Dağıtım Hizmetleri'nde stajını tamamlamış, 2015 yılında İstanbul Sabahattin Zaim Üniversitesi Bilgi İşlem Departmanı'nda çalışmış ve Aralık 2015'den itibaren Erişim Sağlayıcıları Birliği'nde Kaynak Kod Geliştirme Uzmanı olarak çalışmaya devam etmektedir. Ayrıca İstanbul Sabahattin Zaim Üniversitesi'nde Bilgisayar Bilimi ve Mühendisliği alanında tezli yüksek lisans programına devam etmektedir.

E-posta : mertyilmazcakir@gmail.com

Yayınlar

Çakır, M. Y. ve ark. (2017). Derin Sinir Ağları ile Konuşma Tespiti ve Cinsiyet Tahmini. *22.Türkiye'de İnternet Konferansı.* Beşiktaş, İstanbul: Bahçeşehir Üniversitesi.

Çakır, M. Y. ve ark. (2017). Yapay Sinir Ağları ve K-En Yakın Komşu Algoritmalarının Birlikte Çalışma Tekniği (Ensemble) ile Metin Türü Tanıma. *22.Türkiye'de İnternet Konferansı.* Beşiktaş, İstanbul: Bahçeşehir Üniversitesi.