



T.C.

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ÇAĞRI MERKEZLERİ İÇİN DERİN ÖĞRENME TABANLI İNTERAKTİF KONUŞMA TANIMA

Mustafa Jumaah Ahmed AHMED

YÜKSEK LİSANS

Bilgisayar Mühendisliği Anabilim Dalı

OCAK-2020

KONYA

Her Hakkı Saklıdır

ÖZET

YÜKSEK LİSANS

ÇAĞRI MERKEZLERİ İÇİN DERİN ÖĞRENME TABANLI İNTERAKTİF KONUŞMA TANIMA

Mustafa Jumaah Ahmed AHMED

Selçuk Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç.Dr. H.Erdinç KOÇER

2020, 43 Sayfa

Jüri
Danışmanın Doç.Dr. H.Erdinç KOÇER
Prof. Dr. Harun UĞUZ
Doç.Dr. Halife KODAZ

Birçok ticari işletme müşterilerden gelen aramalarda çağrı merkezi sistemi kullanmaktadır. Bu merkezlere çoğunlukla santral görevlileri gelen çağrılara cevap vermekte ve müşterileri ilgili kişilere yönlendirmektedir. Aşırı yoğun çağrı trafiğinden dolayı müşteriler çok uzun süre bekletilmektedir. Bunun bir çözümü olarak santral görevlileri sayısının artırılması ihtiyacı doğmaktadır. Bu durum hem maliyetleri yükseltmekte hem de tam ve kesin bir çözüm olamamaktadır. Bu çalışmada Türkçe konuşma tanıma kullanılarak kişilerden gelen telefon çağrılarını otomatik olarak yönlendiren interaktif bir sistem tasarlanmış ve performans değerlendirmesi yapılmıştır. Çalışmanın ilk aşamasında ses tanıma problemlerinde özellik çıkarımı için literatürde sıkça kullanılan Mel-Frekansı Kepstral Katsayıları (MFKK) ve sesin sınıflandırılmasında kullanılan Saklı Markov Modeli (SMM) denenmiştir. Bu yaklaşımda gürültüsüz ortamda elde edilen ses dosyalarına MFKK ve ardından SMM uygulanarak sonuçlar kaydedilmiş ancak başarı oranı tatmin edici bulunmamıştır. Ardından derin öğrenme tabanlı bir yaklaşım ile Python programlama dili kullanılarak gürültülü ve gürültüsüz ortamlarda gelen çağrılarda girilen konuşmalar analiz edilmiştir. Geliştirilen yazılımın test aşamasında üniversite santral sistemi baz alınarak 10 bölüm ve 100 kişiden oluşan bir rehber veritabanı oluşturulmuştur. Yaş, cinsiyet, bölüm, süre ve ortam (gürültülü/gürültüsüz) parametreleri açısından tezde geliştirilen interaktif Türkçe konuşma tanıma sistemi değerlendirilmiştir.

Anahtar Kelimeler: Çağrı Merkezi Sistemi, Derin Öğrenme, Konuşma Tanıma.

ABSTRACT

MS THESIS

DEEP LEARNING BASED INTERACTIVE SPEECH RECOGNITION FOR CALL CENTERS

Mustafa Jumaah Ahmed AHMED

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE
INSTITUTE OF SELÇUK UNIVERSITY
THE DEGREE OF MASTER OF SCIENCE
IN COMPUTER ENGINEERING

Advisor: Assoc. Prof. Dr.H.Erdinç KOÇER

2020, 43 Pages

Jury
Advisor: Assoc. Prof. Dr.H.Erdinç KOÇER
Prof. Dr. Harun UĞUZ
Assoc. Prof. Dr. Halife KODAZ

Many commercial businesses use a call center system for incoming calls from customers. In these centers, mostly central personnel answer the incoming calls and direct the customers to the related persons. Due to excessive call traffic, customers are kept waiting for a very long time. As a solution, there is a need to increase the number of center personnels. This increases both costs and cannot be a complete and definitive solution. In this study, an interactive system that automatically directs phone calls from people using Turkish speech recognition is designed and performance evaluation is performed. In the first stage of the study, Mel-Frequency Cepstral Coefficients (MFCC) and Hidden Markov Model (HMM), which are used in the classification of sound, were used for feature extraction in voice recognition problems. In this approach, the results were recorded by applying MFCC and then HMM to the sound files obtained in a noiseless environment, but the success rate was not satisfactory. In this context, a deep learning based approach was used to analyze the conversations entered in incoming calls in noisy and noiseless environments by using Python programming language. During the test phase of the developed software, a database consisting of 10 departments and 100 people was created based on the university system. The interactive Turkish speech recognition system developed in the thesis was evaluated in terms of age, gender, department, duration and environment (noisy / noiseless) parameters.

Keywords: Call Center System, Deep Learning, Speech Recognition.

ÖNSÖZ

Bu tez çalışmasında Çağrı Merkezleri İçin Derin Öğrenme Tabanlı İnteraktif Konuşma Tanıma konusunun seçilmesinde ve tezin tamamlanması süreci içerisinde desteğini esirgemeyen değerli danışmanım sayın Doç.Dr. H.Erdinç KOÇER'ya sonsuz şükranlarımı sunarım. Öğrenim hayatım boyunca maddi ve manevi olarak sürekli yanımda olan ve yüksek lisans çalışmam süresince desteklerini her daim hissettiğim aileme ve özellikle değerli babama teşekkürlerimi bir borç bilirim.

Mostafa Jumaah Ahmed AHMED
KONYA-2020

İÇİNDEKİLER

ÖZET ..... **vii**

ABSTRACT ..... **v**

ÖNSÖZ ..... **vi**

İÇİNDEKİLER ..... **vii**

SİMGELER VE KISALTMALAR ..... **viii**

1. GİRİŞ **1**

2. KAYNAK ARAŞTIRMASI **3**

3. MATERYAL VE YÖNTEM **8**

3.1. Veri Tabanı 9
3.2. Mel-Frekansı Kepstral Katsayıları (MFKK) 9
3.3. Saklı Markov Modeli (SMM) 12
3.3.1.SMM'nin Unsurları 14
3.3.2.İleri Yönlü Değişkenler 15
3.3.3.Geri Yönlü Değişkenler 15
3.4. Yapay Sinir Ağları 16
3.4.1. Geri Yayılım 18
3.5. Derin Öğrenme 18
3.5.1. Tekrarlayan Sinir Ağları 19
3.5.2. Uzun – Kısa Süreli Bellek Sinir Ağları 19
3.6.Google Bulut ses tanıma 24

4. ARAŞTIRMA BULGULARI **26**

4.1. MFKK ve SMM yaklaşımı ile alınan sonuçlar 28
4.2. Derin öğrenme yaklaşımı ile alınan sonuçlar 31

5. SONUÇLAR VE ÖNERİLER **36**

5.1 Sonuçlar 36
5.2 Öneriler 37

KAYNAKLAR **39**

ÖZGEÇMİŞ **43**

SİMGELER VE KISALTMALAR

Simgeler

B_f, B_i Ve B_c: Bias
$\tilde{C}_T$: Hücre Durumu Güncelleme Adayı
C_t : Yeni Hücre Durumu
E_0 : Kayıp Fonksiyonu
H_t : Geçerli Zaman Adımı Çıkışı
H_{t-1} : Önceki Lstm Düğümünden Üretilen Çıktı
$İ_t$: Güncelleme Katmanı
N : Fft Sayısı
O_t : Çıktı Matrisi
P : Fft Çıkışı
P(Y |W) : Saklı Markov Modelindeki Olasılık
W : Saklı Markov Modelindeki Kelimeler
W_c : Aday Hücre Ağırlık Matrisi
W_f : Ağırlık Matrisi
W_i : Güncelleme Ağırlık Matrisi
W_i : İ İndeksindeki Ağırlık
X : Ses Sinyali
X_t : T Zamanında Giriş
Y : Vurgu Öncesi Çıktı
Y : Saklı Markov Modelinin Özellikleri
A : Filtre Katsayısı
Λ : Lamda Değeri
Σ : Sigma Değeri

Kısaltmalar

AKD : Ayrık Kosinüs Dönüşümü (Discrete Cosine Transform)
DN : Doğru Negatif(True Negatives)
DP : Doğru Pozitif (True Positives)
GPU : Grafik işlem birimi (Graphical Processing Unit)
HFD : Hızlı Foruier Dönüşümü (Fast Fourier Transform)
LSTM : Uzun – Kısa Süreli Bellek (Long Short-Term Memory)
MFKK : Mel-Frekans Kepstral Katsayıları (Mel Frequency Cepstral Coefficents)
OST : Otomatik Ses Tanıma (Automatic speech recognition)
RNN : Tekrarlayan sinir ağı (Recurrent Neural Network)
SMM : Saklı Markov Modeli (Hidden Markov Model)
SNR : Sinyal-Gürültü Oranını (Signal-To-Noise Ratio)
YN : Yanlış Negatif(False Negatives)
YP : Yanlış Pozitif (False Positives)
YSA : yapay sinir ağlarına (Artificial neural network)

1. GİRİŞ

Konuşma analitiği, insanların ses verisinden bilgi elde etmelerine imkân tanır. Duygu analizi, bir konuşmadan belli ürünler veya hizmetler hakkında anlam çıkarmaya çalışan konuşma analitiği kullanım alanlarından biridir. Bu analiz, telefonda müşterilerle konuşan müşteri temsilcilerinin müşteri ile ilişki kurmasına ve ortaya çıkabilecek sorunları çözmesine yardımcı olmak için de kullanılabilir. Ses duyarlılığı analizi için kullanılmakta olan iki ana yöntem vardır. Bunlardan ilki akustik modelleme diğeri ise dilsel modellemedir. Dilsel modelleme, ses kayıtlarının metin dosyalarına çevrilmesini ve sonrasında metin içeriğine dayalı analizler yapılmasını gerektirmektedir. Bu modelleme türünde belli başlı bazı kelime ve ifadelerin bazı ortamlarda daha yüksek bir olasılıkta kullanıldığı kabul edilmektedir. Bazı kelimeler ya da ifadelerle konuşma metinlerinde sık olarak karşılaşılması durumunda, bir şeylerin olmasını bekleyebilirsiniz. Birçok araştırmacı, dilsel modellemenin ses duygu analizinde iyi bir performansa sahip olduğuna yönelik güçlü kanıtlar ortaya koymuştur. Çalışmalar içerisinde araştırmacılar modellerini oluştururken sözcük özellikleri, bozukluklar ve anlamsal özelliklere önem vermiştir. Bununla birlikte, her bir ses kaydı için iyi bir konuşma metninin bir insan tarafından çıkarılması hem sıkıcı hem de maliyetli olabilir. Ses kayıtlarını metne doğru şekilde dönüştürmek için iyi bir otomatik konuşma metni çıkarma sisteminin oluşturulması ise zordur. Metindeki küçük bir yanlışlık bile dilsel modelde büyük bir farklılığa neden olabilir. Akustik modellemede ise, ses verisinin akustik özellikleri kullanılır. Bunlar arasında bilgisayar programları tarafından kolayca hesaplanabilen ses perdesi, yoğunluğu, konuşma hızı ve benzeri özellikler bulunmaktadır. Bu özellikler birlikte ele alındığında duygular hakkında bir dereceye kadar temel göstergeler sağlayabilirler. Akustik özelliklerin diğer birçok ticari ya da psikolojik analizde kullanılması, onları duygu analizinde de kullanmamız için bir diğer nedendir. Bununla birlikte, akustik modellemeyle ilgili sorun, ses kayıtlarının kalitesinin nihai sonuç üzerinde güçlü bir etkiye sahip olmasıdır. Ses verisinin akustik özelliklerini kullandığımızdan, ses kalitesi, sonunda iyi modeller oluşturmakta başarısız olacak olan bu akustik özellikler için doğru değerler elde etme yeteneğini önemli ölçüde etkileyebilir. Diğer problem, gerçek dünyada herhangi bir yerde, arka planda daima rastgele bir gürültünün varlığıdır. Model eğitim verileri bu rastgele sesleri ayırt edemeyebilir, dolayısıyla bu durum canlı bir konuşma sırasında duygu analizinin yapılması arzu edildiğinde üstesinden gelinmesi gereken bir zorluk olarak karşımıza çıkar. Ancak yine

de bu yöntem, kolay uygulanabilirliği ve etkinliği nedeniyle daha fazla inceleme ve araştırma yapmaya değmektedir.

Birçok ticari işletme müşterilerden gelen aramalarda çağrı merkezi sistemi kullanmaktadır. Bu merkezlere çoğunlukla santral görevlileri gelen çağrılara cevap vermekte ve müşterileri ilgili kişilere yönlendirmektedir. Aşırı yoğun çağrı trafiğinden dolayı müşteriler çok uzun süre bekletilmektedir. Bunun bir çözümü olarak santral görevlileri sayısının artırılması ihtiyacı doğmaktadır. Bu durum maliyet açısından işletmelere ek yükler getirmektedir. Ayrıca santral görevlilerinin uzun saatler boyunca çalışmalarından dolayı yorgunluk ve stres gibi etkenlerden dolayı çağrının yanlış yönlendirilmesi gibi sorunlarla da karşılaşılmaktadır.

Bu tez çalışmasında, ilk olarak ses tanıma problemlerinde özellik çıkarımı için literatürde sıkça kullanılan Mel-Frekansı Kepstral Katsayıları (MFKK) ve sesin sınıflandırılmasında kullanılan Saklı Markov Modeli (SMM) denenmiştir. Bu yaklaşımda gürültüsüz ortamda elde edilen ses dosyalarına MFKK ve ardından SMM uygulanarak sonuçlar kaydedilmiş ancak başarı oranı tatmin edici bulunmamıştır. Daha sonra konuşmadan metine ve metinden konuşmaya dönüştürme tekniklerinden derin öğrenme yaklaşımı kullanılarak çağrı merkezleri için interaktif bir yanıtlama ve yönlendirme sistemi önerilmiştir.

Tez kapsamında öncelikle kaynak araştırması yapılarak konu ile ilgili mevcut sistemler ve yöntemler incelenmiştir. Ardından ses tanımaya yönelik tekniklerin avantaj ve dezavantajları değerlendirilerek MFKK ve SMM yöntemi uygulanmış ancak başarı oranı düşük olduğundan derin öğrenme tabanlı yaklaşıma geçilmiştir. Son yıllarda görüntü ve sinyal işleme alanında sıkça adından söz edilen derin öğrenme tekniği temelde yapay zeka temelli bir sınıflandırma algoritması olan yapay sinir ağlarına (YSA) dayanmaktadır. Derin öğrenme tekniğinde, farklı filtre matrisleri kullanılarak giriş kümesinin (bu çalışmada ses sinyali) özellik vektörleri ağ içerisinde elde edilmekte ve YSA nın sınıflandırmadaki başarısı ile birleştirilmektedir. Dolayısıyla özellik çıkarımı işlemi ağ içerisinde tamamlanarak hızlı ve güvenilir bir sınıflandırma elde edilmektedir.

Phyton yazılım platformunda kodlanan yazılımın performans değerlendirmesi aşamasında yeni bir rehber veritabanı oluşturulmuştur. Üniversitenin santral sistemi temel alınarak 10 bölüm ve 100 kişiden oluşan veritabanı kullanılarak farklı kriterler bazında sistem test edilmiştir. Çalışmanın sonuçlar kısmında testlerden elde edilen bulgular yorumlanmış ve sistemin genel anlamda başarılı olduğu görülmüştür.

2. KAYNAK ARAŞTIRMASI

Ses tanıma ile ilgili ilk çalışmalar 1900 lü yılların başında yapılmıştır. İlk ticari ürün ise 1920'lerde piyasaya tanıtılan Radio Rex isimli bir ses tanıma oyuncağıdır. 1938 yılında Bell Laboratuvarındaki araştırmacılar bir konuşma sentez makinesi geliştirmiş ve New York'da yapılan Dünya fuarında tanıtmışlardır.

Otomatik Ses Tanıma (OST) için geliştirilen sistemlerde araştırmacılar çoğunlukla fonetik-akustik kavramlarını odaklanmıştır.

1952 yılında her bir basamağın/hanenin sesli harf bölümleri sırasında tahmin edilmiş biçimlendirici frekanslarını kullanarak tek bir konuşmacı için izole bir sayı tanıma sistemi geliştirilmiştir (Davis ve ark., 1952).

1959 yılında MIT Lincoln Lab'da yapılan çalışmada, sesli harfler için spektral rezonanslar ölçülerek konuşmacıdan bağımsız 10 sesli harf tanıyıcısı geliştirmiştir (Forgie ve Forgie, 1959).

1960-70'lerde Japon araştırmacılar ses ve konuşma tanıma alanına girmiştir. Araştırmacılar Bilgisayarlar yeterince hızlı olmadığı için, sistemlerinin bir parçası olarak özel amaçlı çip tasarlamışlardır. Tokyo'da, Nagata ve ark. bir H/W sesli harf tanıyıcı çipi kullanarak konuşma tanıma gerçekleştirmişlerdir (Nagata, 1963). Başka bir çalışmada ise, Kyoto Üniversitesinden Sakai H/W fonem tanıyıcısı çipi geliştirmiştir (Sakai, 1962). 1963 yılında, NEC Labs'da Nagata ve meslektaşları basamak/sayı tanıyıcısı geliştirmiştir. Bu, uzun verimli bir araştırma programına ön ayak olmuştur.

Özel şirketler de ses analizi konusunda çalışmalar yapmıştır. IBM araştırmacıları geniş kelime hazneli ses tanımasını incelemiştir. AT&T Bell Laboratuarlarında, araştırmacılar konuşmacıdan bağımsız ses tanıma testlerine başlamıştır (Rabiner ve ark., 1979). Ana dili olmayan ses tanıması için sözlüksel modele kelimeleri temsil etmek için gereken farklı yapı sayısını bulmak üzere çok sayıda kümeleme algoritması kullanılmıştır.

Carnegie Mellon Üniversitesinde geliştirilen Harphy sistemi, yüksek doğruluk ile 1011 kelimeden oluşan kelime verilerini tanımıştır (Lowrre, 1990). Hesaplamayı azaltmak ve en yakın eşleştirme dizilerini etkili bir şekilde belirlemek için ilk kez sonlu bir ağdan faydalanılmıştır.

1980'li yılların başında, araştırmanın odak noktası bağlantılı kelimeler ses tanıması olmuştur. 1980 yılında, Saklı Markov Modeli (SMM) yaklaşımı, geliştirilmiş kilit teknolojilerden biridir. IBM, Savunma Analizi Enstitüsü (IDA) ve Dragon Systems SMM'yi anlamıştır ancak 1980'li yılların ortalarına kadar ünü yayılmamıştır. Ses tanıma

sorunlarına bir çözüm olarak sinir ağları, 1980'lerin sonlarında yeniden tanıtılmış başka bir teknolojidir.

En önemli paradigma değişimi, istatistiksel yöntemlerin tanıtılması olmuştur. Özellikle 1970'leri başında Saklı Markov Model (SMM) ile stokastik işleme ses tanımada kayda değer katkılar yapmıştır. 30 yıl sonra, bu metodoloji halen kullanılmaktadır. 1980 li yıllarda ise konuşma tanıma araştırmalarında, şablona dayalı yaklaşımlar özellikle ön plana çıkmıştır (Ferguson, 1980; Rabiner, 1989).

Basitliklerine rağmen, Ngram dili modelleri önemli biçimde güçlü olduklarını kanıtlamıştır. Bugünlerde, en pratik ses tanıma sistemleri istatistiksel yaklaşıma dayalıdır ve ek gelişmeler ile sonuçları 1990'larda alınmıştır. 1993 yılında modele dayalı spektral tahmin algoritması geliştirilmiştir (Erell ve Weintraub, 1993). 1994 yılında Moshey J. Lasry, harflerin ve sayıların ses spektrogramını incelemiş ve özelliğe dayalı bir ses tanıması geliştirmiştir (Moore, 1994). SMM ses tanıma sistemi için, nörol ağ eğitiminde önemli bir yenilik olan vektör nicemleyicisi ağ kullanımına dayalıdır (Rigoll, 1994). Nam Soo Kim ve ark., SMM'ye dayalı sağlam bir çıktı olasılık dağıtımı öngörmek için çeşitli yöntemler tariflemiştir (Kim ve Un, 1995).

2000 yılında Juang ve Furui, geleneksel olarak Bayes'in çerçevesini izlemiştir ancak deneysel tanıma hatasının minimizasyonu ile bir optimizasyon sorununa değinilmiştir (Juang ve Furui, 2000). Bu çalışmada amaç, verilen verilere en iyi uyandan ziyade en az tanıma hatası olan bir tanıyıcı tasarlamaktır. Hata minimizasyonu için kullanılan teknikler, minimum sınıflandırma hatası ve maksimum karşılıklı bilgidir. Bu teknikler, ses tanıma performansına maksimum olabilirlik temelli yaklaşımı da beraberinde getirmiştir. Ses modeli ve telaffuzu arasında akustik uyumsuzluğu azaltmak için, maksimum olabilirlik stokastik eşleştirme yaklaşımı önerilmiştir (Sankar ve Lee, 1996).

1999 yılında gürültülü ortam için, sağlam ses tanıması için, işitsel modele yeni bir yaklaşım önerilmiştir (Kim ve ark., 1999). Bu yaklaşım, diğer modeller ile kıyaslandığı zaman sayısal olarak etkilidir. 2004 yılında, varyasyonel Bayesian tahmin tekniği geliştirilmiştir (Afify ve Siohan, 2004). Bu teknikte tahmin, parametrelerin posteriyor dağılımına dayalıdır. Giuseppe Richardi ve ark., OST'da uyumlayıcı öğrenme sorununu çözmek için bir teknik geliştirmiştir (Riccardi ve Hakkani-Tur, 2005). Yine 2005 yılında, performans ilerlemesi için geliştirilen geniş kelime bilgisi sürekli ses tanıma sisteminde önemli ilerlemeler kaydedilmiştir (Afify ve ark., 2005). 2005 yılında Furui, sesli harf

spektral rezonansını inceleyen bir araştırma yapmış ve telaffuza ilişkin analizler gerçekleştirmiştir (Furui, 2005).

Japonya'da 5 yıllık ulusal bir proje olan Corpus of Spontaneous Japanese (CSJ) projesi yine Furui tarafından yürütülmüştür. Bu projede kullanılan veritabanı yaklaşık 7 milyon kelimeden oluşmaktadır ve bu da 700 saatlik konuşmaya denk gelir. Bu projede kullanılan teknikler, akustik modelleme, cümle sınırı tespiti, telaffuz modellemesi, dil model adaptasyonu yanında akustik ve otomatik ses özetlemesidir (Furui ve ark., 2004).

Özellikle spontan konuşma için, ses tanıma sistemlerinin sağlamlığını daha da artırmak için telaffuz doğrulaması tekniği araştırılmıştır (Lleida ve Rose, 2000). İnsanlar birbiri ile konuştukları zaman, multimodal iletişim kurarlar. İletişim gürültülü bir ortamda gerçekleştiği zaman, ses tanıma başarısı düşmektedir. Ses tanımada, görsel yüz bilgisinin kullanılması, özellikle de dudak hareketi, araştırılmıştır ve sonuçlar her iki bilgi şeklinin kullanılmasının özellikle gürültülü ortamda sadece ses ya da sadece görsel bilgilerin kullanılmasına göre daha iyi performans sergilediğini göstermiştir.

(Patel ve Rao, 2010), SMM tabanlı bir tanıma yaklaşımına yönelik konuşma özelliğinin temsil anlamında geliştirilmesi için, Mel-frekanslı spektral bilgilerini kullanarak konuşma sinyallerinin tanınmasına yönelik bir yaklaşım sunmaktadır. Frekans spektral bilgisi, geleneksel Mel-frekanslı spektrum tabanı konuşma tanıma yaklaşımına dâhil edilir. Mel-frekanslı yaklaşım, konuşma sinyaline yönelik belirli bir çözünürlükteki frekans gözleminden yararlanır ve bu da çözünürlük özelliği çakışmasıyla sonuçlanan tanıma sınırına yol açar. Ayırma frekanslı çözünürlük ayrışımı, SMM tabanlı konuşma tanıma sistemine yönelik bir haritalama yaklaşımıdır. Simülasyon sonuçları, konuşma tanıma sistemi için öğrenme doğruluğu, öğrenme doğruluğu açısından konuşma tanıma kalite metriklerinde bir iyileşme göstermektedir.

(Ashraf ve ark., 2010), Urduca dilinde bir otomatik konuşma tanıma sistemi geliştirmek için bir yaklaşım geliştirmiştir. Önerilen sistem, mevcut sistemi geliştirmek için istatistiksel tabanlı bir yaklaşım (Saklı Markov Modeli) kullanan Sphinx4 adlı açık kaynaklı bir konuşma tanıma çerçevesine dayanmaktadır. Araştırmacılar, küçük boyutlu kelime dağarcığı için Konuşmacıdan Bağımsız bir konuşma tanıma sistemi, bir başka ifadeyle, Urducada en çok konuşulan elli iki izole kelimeden oluşan bir sistem ortaya konulmuşyut Ayrıca, bu araştırma çalışmasının orta ve büyük boyutlu dağarcık Urduca konuşma tanıma sisteminin geliştirilmesi için temel oluşturacağını öne sürmüşlerdir.

2011 yılında A.Poşul tarafından yapılan tez çalışmasında beceri bazlı yönlendirme algoritması tabanlı interaktif sesli yanıt sistemi geliştirilmiştir. Sistemin müşteri puanlamasına göre değerlendirmesi yapılmıştır (Poşul, 2011).

(Ranjan ve Dubey, 2016), Maithili (Hint dili) lehçesi için izole kelimelerin otomatik olarak tanınması sistemini geliştirmiştir. İzole kelime tanıma (IWR) sistemleri, bir kişi tarafından konuşulan bir kelimeyi bir mikrofon veya başka bir cihaz aracılığıyla tanır. Ranjan ve Dubey, telefon güvenliği ve internet araştırmacıları için kullanılan Maithili lehçesi kelimelerini veya bankacılık güvenlik çerçevelerini programlı bir şekilde arama ve bu sistemleri yönetmeyi algılamaya odaklanmışlardır. Maithili dilindeki kelimelerin konuşma sinyalleri vurgu vektörleri düzenlemesine dönüştürülür ve burada Mel-Frekansı Kepstral Katsayıları (MFCC) kullanılır. IWR, Saklı Markov Modeli (SMM) kullanılarak elde edilen eleman vektörleri üzerinden gösterilmiştir. SMM tabanlı lehçe modelleri ve kelime dağarcığına sahip akustik modeller, beş katlı çapraz doğrulama işlemi kullanılarak Maithili lehçesinin foneminin sesli harflerinin her birini kapsayan Maithili kelimelerinin yanlış sınıflandırma oranını hesaplamak için kullanılmıştır.

(Jiang ve ark., 2019) Derin Evrişimsel Sinir Ağlarına (DESA) ve Basit Tekrarlayan Birime (BTB) dayalı bir insan sesinden duygu çıkarma yöntemi geliştirmişlerdir. İlk olarak log Mel-spektrogramları statik, delta ve delta-delta ile ayarlanan akustik özelliklerden çıkarılır. Her bir ifadenin üç log Mel-spektrogramı kanalı, DESA girişi olarak zaman ekseninde çeşitli parçalara ayrılır. Ardından, ImageNet veri kümesi üzerinde önceden eğitilmiş AlexNet, ince ayarlar için her segmentte bu özellikleri öğrenmek için kullanılır. BTB modeli, öğrenilen bu segment düzeyindeki özellikleri bir araya getirir. Son olarak, insan sesinden duygu çıkarmaya yönelik türleri tanımlamak için bir SoftMax sınıflandırıcısı kullanılır. EMO-DB ve CASIA veritabanındaki deneysel sonuçlar, modellerinin konuşmada yer alan duyguları etkili bir şekilde tanıyabildiğini ve başka bir tür özelliğe dayalı olarak sınıflandırıcılardan daha iyi performans gösterdiğini göstermektedir.

(Zhao ve ark., 2019) saklı bir regresyon Bayesian ağı (LRBN) kullanarak uçtan uca konuşma tanıma modeline yönelik veri girişi için, paylaşılan konuşma özelliğini çıkarılmasına yönelik bir yöntem önermişlerdir. LRBN kompakt bir yapıya sahip olup, parametre öğrenmesi hızlıdır. Evrişimsel Sinir Ağları ile karşılaştırıldığında, daha basit ve anlaşılır bir yapıya ve daha az öğrenilecek parametreye sahiptir. Deneysel sonuçlar, Tibet çoklu lehçeli ses tanıma için hibrit LRBN/Çift Yönlü Uzun-Kısa Süreli Bellek-

Bağlantıcı Geçici Sınıflandırma mimarisinin avantajının ve LRBN'nin çoklu dil konuşma setlerini ayırt etmede yardımcı olduğunu göstermektedir.

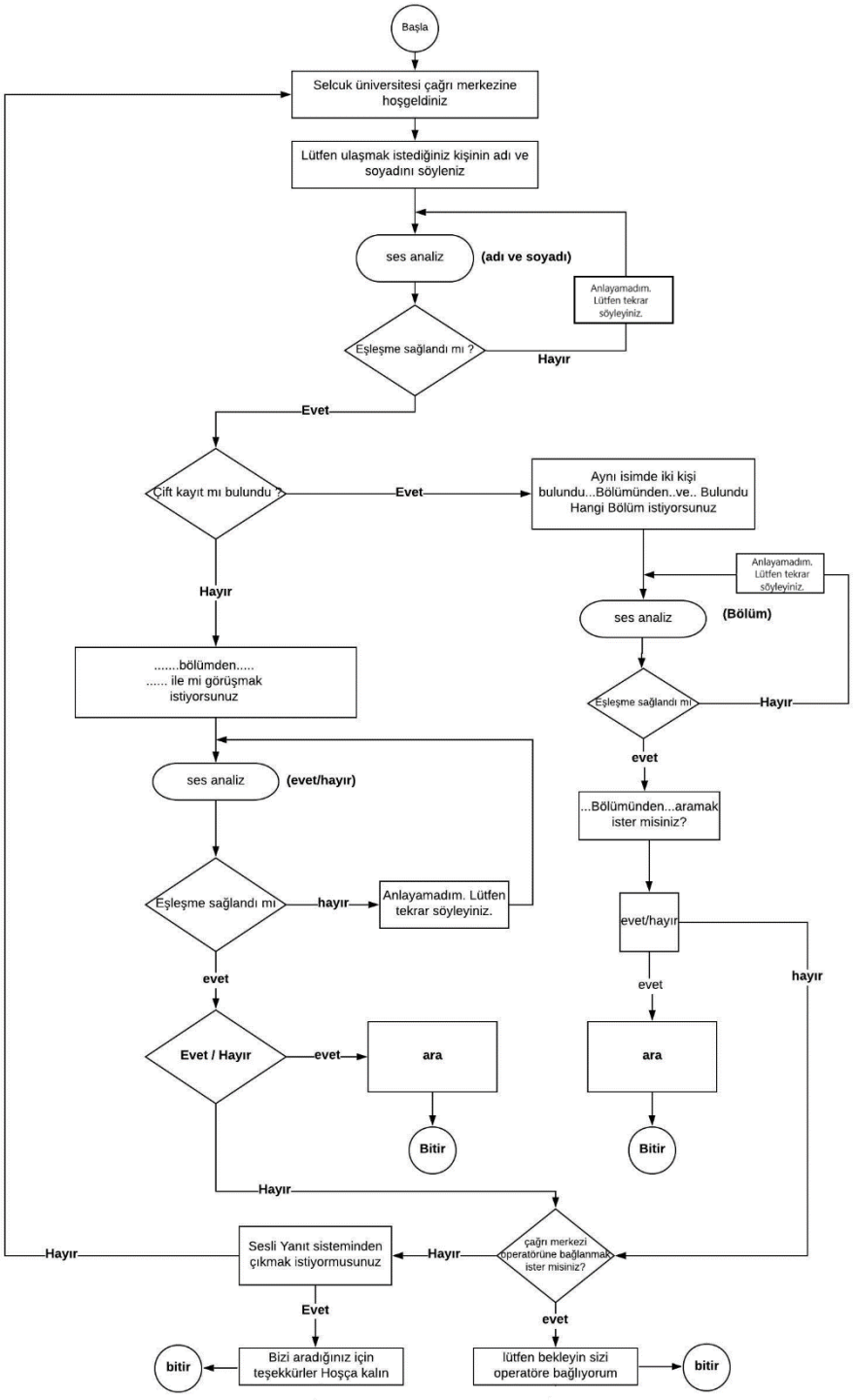
2019 yılında C.Akın tarafından yapılan tez çalışmasında ise skor seviyesinde birleştirme metodu kullanılarak gömülü donanım sistemi üzerinde ses tanıma yapılmıştır. Bu çalışmada İşitsel-Görsel İkizler Veritabanı (Audio-Visual Twins Database (AVTD)) kullanılmıştır. Bu veri setinin seçilmesinin amacı ise birbirine benzer karakteristiğe sahip olan ikizler üzerinde ses tanıma algoritmalarının başarı oranlarını incelemektir (Akın, 2019)

2019 yılında yapılan bir diğer tez çalışması U.Kimanuka tarafından geliştirilmiştir. Bu çalışmada derin sinir ağları tabanlı bir ses tanıma sistemi geliştirilmiş ve Gauss – Saklı Markov Model karışımı bir yaklaşımla elde edilen sonuçlarla karşılaştırma yapılmıştır (Kımanuka, 2019).

(Petridis ve ark., 2020) tamamen bağlı katmanlara ve küçük ölçekli veri kümelerine uygun Uzun-Kısa Süreli Bellek (LSTM) ağlarına dayanan uçtan uca görsel ses tanıma sistemi önermiştir. Model, özellikleri doğrudan ağız görüntülerinden çıkaran ve özellikleri farklı görüntülerden çıkaran olmak üzere iki akıştan oluşur. Her akıştaki zamansal dinamiğin modellenmesi için, sonradan başka bir BLSTM ile kaynaştırılan bir Çift Yönlü LSTM (BLSTM) kullanılır. OuluVS2, CUAVE, AVLetters ve AVLetters2 veri tabanlarında, sırasıyla, en son teknolojiye göre %0.6, %3.4, %3.9, %11.4 oranlarında mutlak bir gelişme bildirilmiştir.

3. MATERYAL VE YÖNTEM

Bu çalışmada geliştirilen interaktif çağrı sisteminin akış diyagramı Şekil 3.1'de verilmiştir. Diyagramda gösterilen "ses analiz" blokları derin öğrenme tabanlı bir yaklaşımla çağrı merkezini arayan kişinin konuşmasını analiz ederek çıkarım sağlamaktadır.



Şekil 3.1. Önerilen sistemin akış şeması.

3.1. Veri Tabanı

Bu çalışmada geliştirilen ses tanıma uygulamasının performans değerlendirmesini yapmak amacıyla özel bir rehber veritabanı oluşturulmuştur. Bu veritabanı bir üniversitenin telefon santraliymiş gibi hazırlanarak bölüm ve akademisyen isimlerine göre 4 haneli dahili telefon numaraları kaydedilmiştir. 10 farklı bölümde kayıtlı 100 kişinin ismi rehbere kaydedilerek Türkçe konuşma tanıma ile otomatik çağrı yönlendirme gerçekleştirilmiştir. Rehbere kaydedilen bölüm ve isimler rasgele belirlenmiş ve tablo halinde Çizelge 3.1'de sunulmuştur. Kırmızı renkle belirtilenler (7 kişi x 2 = 14 kayıt) aynı isimde olup bölümleri farklı olan kişilerdir.

Çizelge 3.1. Oluşturulan Rehber Veri Tabanı

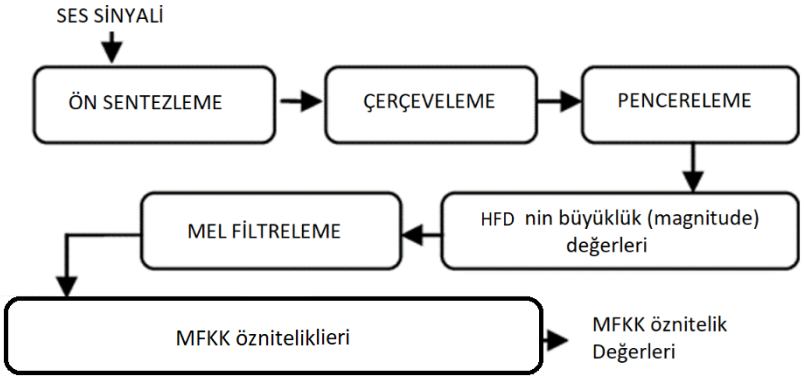
Ad-Soyad	Bölüm	Ad-Soyad	Bölüm	Adı-Soyad	Bölüm	Adı-Soyad	Bölüm
İb**DE**	EE	Kü**UY**	Biy.	Hu**KA**	Mak.	Ah**ER**	Mekat.
Er**KA**	EE	Er** AĞ**	Biy.	Hü**DO**	Mak.	Al* Nu* ŞE*	Mekat.
Ha**AR**	EE	Mü*Mu*SA*	Biy.	İb**UY**	Mak.	Ta**SA**	Mekat.
Ha**SÜ**	EE	Fe** SE**	Biy.	İl**ÖZ**	Mak.	Tu**DE**	Mekat.
Ha**SE**	EE	Os** ÖZ**	Biy.	Su**KO**	Mak.	Zi**ÖZ**	Mekat.
Er** KO**	EE	Ad**GÖ**	Kimya	Şe**AY**	Mak.	Ha**KO**	Uçak
Şa** TA**	EE	Ah*Al*SE*	Kimya	Mu**DO**	Mak.	Mu**ŞA**	Uçak
En** ÖZ**	EE	Al**YA**	Kimya	Em**AY**	Mak.	Ta**SA**	Uçak
Ha** AY**	EE	Es**ER**	Kimya	Ah**BO**	Mak.	Ke**CA**	Uçak
Ha*Hü* ÇE*	EE	Ak**AK**	Kimya	Gü**ÖZ**	Mak.	Al**SE**	Uçak
Ne**TA**	Bilg.	Se**TÜ**	Kimya	Ha*Al*BA*	Malz.	Me**ŞA**	Uçak
İs**SA**	Bilg.	Öz**AY**	Kimya	Ah**BO**	Malz.	Ci** NA**	Uçak
Al**ÖZ**	Bilg.	Gü** GÜ**	Kimya	Mu**KÖ**	Malz.	Bu** YA**	Uçak
Al**AL**	Bilg.	Os** AC**	Kimya	Se**KA**	Malz.	Mu** ŞA**	Uçak
Fa**BA**	Bilg.	Zü*Yı*AC*	Kimya	Ke**TÜ**	Malz.	Al*Nu*ŞE*	Uçak
Ba**SA**	Bilg.	Al** YA**	Hukuk	Ak**ŞA**	Malz.	Mu** Cİ**	Çevre
Bi**ÇE**	Bilg.	Hu*Ka*ÖR**	Hukuk	Fa** AY**	Malz.	Mu** AC**	Çevre
Bu**TE**	Bilg.	Sü** NE**	Hukuk	Se** SE**	Malz.	Ne** TA**	Çevre
Ci**NA**	Bilg.	Şe** EK**	Hukuk	Se** AL**	Malz.	Re** KU**	Çevre
Em**SA**	Bilg.	Kü** ZÜ**	Hukuk	Al**AK**	Malz.	Er** KO**	Çevre
Ba**SA**	Biy.	Mu** KU**	Hukuk	Ne**ÖR**	Mekat.	Ey** CA**	Çevre
Ha**SA**	Biy.	Os** AC**	Hukuk	Ab**SA**	Mekat.	Ha*Er*GÜ*	Çevre
Mu**ÇE**	Biy.	Öz** ÇA**	Hukuk	Mu**GÖ**	Mekat.	Mu** DO*	Çevre
Ha**IŞ**	Biy.	Rü** Bİ**	Hukuk	Uğ**TA**	Mekat.	Er** KA**	Çevre
Ha**DÜ**	Biy.	Mu** AC**	Hukuk	Ab**AS**	Mekat.	Me** ÇU**	Çevre

Biy.:Biyomedikal, Bilg.: Bilgisayar, EE: Elektrik-Elektronik, Mak.:Makine, Malz.:Malzeme, Mekat.:Mekatronik

3.2. Mel-Frekansı Kepstral Katsayıları (MFKK)

Mel-frekansı kepstral (MFK), bir ses sinyalinin güç spektrumunda kısa süreli gösterilmesidir. MFKK, doğrusal olmayan bir Mel-ölçeği frekansı üzerindeki bir log güç spektrumunun doğrusal kosinüs dönüşümüne dayanmaktadır. MFKK, bir ses sinyalinin, kepstral türdeki bir gösteriminden türetilen MFK analizi (Xu ve ark., 2004) kullanılarak toplanan katsayılardır. Mel-frekansı ve Kepstral arasındaki fark, insan işitme sisteminin

normal kepstraldeki doğrusal frekans bantlarından çok daha fazla tepkiden esinlenerek MFKK'de bulunan Mel-ölçeğindeki aralıklı frekans bantlarıdır. Şekil 3.2, MFKK öznitelik çıkarma işlemini göstermektedir.



Şekil 3.2. MFKK algoritması akış diyagramı

MFKK algoritması, ses sinyalinden öznitelikleri çıkarmaya yönelik aşağıdaki yedi adımdan oluşur:

1) **Ön sentezleme:** Ön sentezleme, ses sinyalini daha yüksek frekanslara yükseltmek için uygulanan bir filtredir. Düşük frekanslara kıyasla, yüksek frekansların daha küçük genliklere sahip olması nedeniyle, frekans spektrumunu dengelemekte fayda vardır. Bu, aynı zamanda, hızlı Fourier dönüşüm işlemi uygulandığında ortaya çıkan sayısal problemleri önlemek ve sinyal-gürültü oranını (SNR) iyileştirmek için kullanılır. Ön vurgu, x sinyaline denklem 3.1'deki gibi uygulanır:

$$y(t) = x(t) - ax(t-1) \tag{3.1}$$

Filtre α katsayısının tipik değerleri 0.95 ile 0.97 arasındadır. Ön vurgu filtresi, modern sistemler üzerinde önemli bir etkiye sahiptir, çünkü bu, sadece modern uygulamalarda HFD'nin sayısal sorunlarını önlemek için ortalama normalleştirme kullanılarak elde edilebilir.

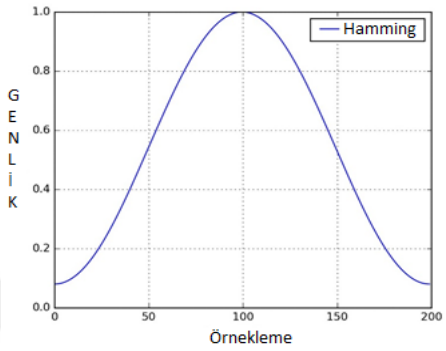
2) **Çerçeveleme:** Sinyali kısa süreli çerçevelere bölmek için, ön vurgunun ardından, çerçeveleme adımı kullanılır. Bu adım, sinyaldeki frekansların zamanla değişmesi nedeniyle kullanılır. Çerçevelemenin temel mantığı, frekans hatlarının zaman içinde sinyal kaybını önlemek için tüm konuşma sinyaline HFD uygulanmasının mantıklı olmamasıdır. Tipik çerçeveler periyodu 20ms-40ms arasındadır ve ardışık çerçeveler arasında% 50 çakışma vardır. Çerçevelemeyi kullanarak, ses sinyalindeki frekansların kısa bir süre boyunca sabitlendiğini varsayabiliriz,

böylece Fourier dönüşümü yaparak frekans hatlarına yönelik iyi bir tahmin elde edebiliriz.

3) **Pencereleme**: Çerçevelemenin ardından, her kareye, ses sinyali Hamming Penceresi, aşağıda belirtilen şekilde uygulanır:

$$w_n = 0.54 - 0.46_{cos}\left(\frac{2\pi n}{n-1}\right) \quad (3.2)$$

Burada, 0 ≤ n ≤ N−1 iken, N, Şekil 3.3'de gösterildiği gibi pencere uzunluğudur.



Şekil 3.3. Hamming penceresi

Hamming penceresinin uygulanmasının nedeni, hızlı Fourier dönüşümü (HFD) tarafından sinyalin sabit olduğu varsayımını önlemek ve spektral sızıntıyı azaltmaktır.

4) **hızlı Fourier dönüşümü:** Frekans spektrumunu hesaplamak için her karede bir N noktası HFD gerçekleştirilmiştir. Tipik olarak N değeri, 256 veya 512'dir. Her çerçeveden HFD çıkarıldıktan sonra, güç spektrumu denklem 3.3'te göstelildiği gibi hesaplanır:

$$p = \frac{|FFT(X_i)|^2}{N} \quad (3.3)$$

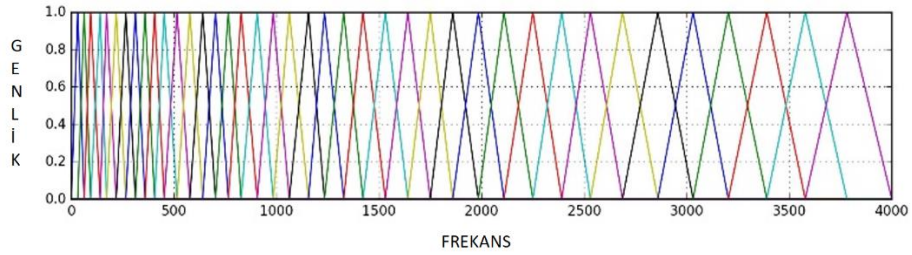
Yukarıdaki formülde x_i, x sinyalinin i^th çerçevesidir.

5) **Mel Filtre:** Mel Filtre son adımdır ve tipik olarak Mel-ölçekli 40 filtreye üçgen filtreler uygulanarak hesaplanır. Mel-ölçekli filtre yığını, insan kulağının sesleri algılayış biçimini taklit eder. Hertz ve Mel arasındaki dönüştürme, aşağıdaki denklem 3.4'te ve 3.5'te göstelildiği gibi hesaplanır:

$$m = 2595\log_{10}\left(1 + \frac{f}{700}\right) \quad (3.4)$$

$$f = 700\left(10^{m/2595} - 1\right) \quad (3.5)$$

Her filtre, merkez frekansta bir tepkiye sahip olan üçgen şeklindedir ve doğrusal olarak sıfıra düşer. Şekil 3.4, Mel-ölçekli bir filtre yığınını göstermektedir.



Şekil 3.4. Mel-ölçekli filtre yığını

6) **MFKK Öznitelikleri:** Filtre yığını çıktısı, yüksek derecede ilişkilidir, bu ise birçok makine öğrenimi algoritması için bir sorun teşkil eder, bu nedenle, filtre yığını katsayılarını ilintilemek ve filtrenin gösterimini sıkıştırmak için Ayrık Kosinüs Dönüşümü AKD (Discrete Cosine Transform (DCT)) uygulanır. Otomatik ses tanıma sisteminde genel olarak 2-13 katsayıları kullanılır.

3.3. Saklı Markov Modeli (SMM)

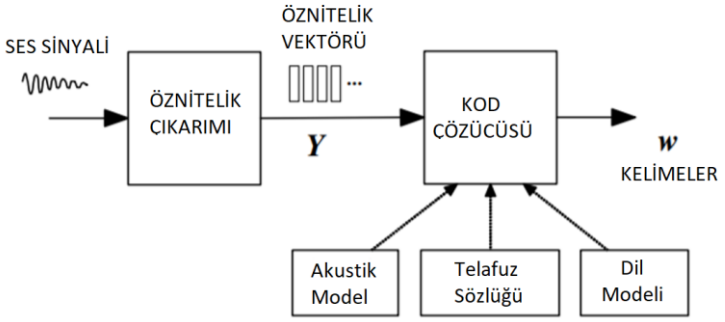
Saklı Markov modeli, modern ses tanıma sistemlerinde en çok kullanılan teknik olup, zamanla değişen spektral özellikler için kullanışlı bir modelleme yöntemidir. Son birkaç on yıllık zaman zarfında, neredeyse tüm OST sistemlerinde temel bir çerçeve olarak kullanılır (Evermann ve ark., 2004; Soltau ve ark., 2005; Matsoukas ve ark., 2006). Modern SMM tabanlı ses tanıma sistemleri, 1970'li yıllarda, Carnegie Mellon Üniversitesi ve IBM'deki araştırma grubu tarafından oluşturulmuştur.

Saklı Markov Modelleri, zaman serisi verilerini modellemek için kullanılan özel bir Bayes ağı türüdür. Bunlar, gözlem dizileri üzerindeki olasılık dağılımlarını temsil ederler (Ghahramani, 2001; Murphy, 2002).Saklı Markov Modeli, adını, iki tanımlayıcı özellikten alır. Bunlardan birincisinde, SMM, ikişerli stokastik bir model olarak düşünülebilir. Rasgele bir model olarak da bilinen stokastik bir model, zaman içinde bazı rastgele değişkenlerin veya sistemlerin gelişimini temsil eder. Sadece tek yönlü gelişebilen deterministik (*rasgele olmayan*) modellerin aksine, stokastik modeller bir miktar rasgele gelişir. Bir SMM'de stokastik bir model, bir dizi gözlem olarak gözlenirken, başka bir

stokastik model saklanmıştır (durumdadır). Saklı stokastik model, sadece gözlemlerle değerlendirilebilir (Cappé ve ark., 2006). İkinci tanımlayıcı özellik ise, saklı durumun birinci dereceden Markov özelliğini karşıladığını varsayar, bir başka ifadeyle, t zamanındaki durum, sadece t – 1 zamanındaki duruma bağlıdır ve ondan önceki tüm durumlardan bağımsızdır. Gözlemler aynı zamanda durumlarla ilgili birinci dereceden bir Markov özelliğini de yerine getirmektedir, bir başka ifadeyle, bir durum dikkate alındığında, bu duruma karşılık gelen gözlem, diğer tüm durumlardan ve gözlemlerden bağımsızdır (Ghahramani, 2001).

SMM tabanlı ses tanıyıcısının çalışma prensibi Şekil 3.5'de gösterilmiştir. Burada girdi, konuşulan kelimenin, MFKK algoritması kullanılarak bir öznitelik sırasına dönüştürülen ses sinyalidir.

$$Y_{1:T} = \{y_1 \ldots\ldots y_T\} \tag{3.6}$$



Şekil 3.5. Saklı Markov Modeli

Kod çözücü, bir kelime dizisi üzerinde çalışır.

$$w_{1:L} = \{w_1 \ldots\ldots w_L\} \tag{3.7}$$

Y üretme olasılığı en yüksek olan denklem 3.8'da gösterilmiştir:

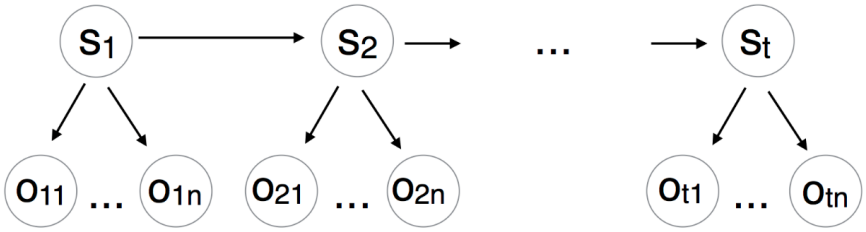
$$\hat{w} = \underbrace{argmax}_{w} \{p(Y|w)P(w)\} \tag{3.8}$$

p(Y |w) olasılığı, akustik model ile ve P (w) dil modeliyle belirlenir. Her bir ses birimi, akustik model sayesinde basit bir ses olarak gösterilir, örneğin, bir kedi, /k/, /e/, /d/, /i/ şeklinde gösterilir.

Herhangi bir w kelimesi için, karşılık gelen model sözcükleri sözlük olarak oluşturmak için sesbirimleri birleştirerek sentezlenir.

Modellerin parametreleri, söylenen kelimelerin ses sinyalini içeren eğitim verileri kullanılarak tahmin edilir. Genellikle dil modeli, her kelime olasılığı sadece N – 1 öncülüne bağlı olan bir N – gram modelinden oluşur.

Şekil 3.6'da bir SMM'deki durumlar ve gözlemler arasındaki bağımlılıkların grafiksel bir temsili gösterilmiştir. Burada, gözlemler O ile başlayan dairelerle ve saklı durumlar ise S ile başlayan dairelerle gösterilmiştir.



Şekil 3.6. Durumlar Arasındaki Bağımlılıklar ve Saklı Markov Modelindeki Gözlemler (Lison, 2014)

3.3.1.SMM'nin Unsurları

- Durum Seti: $S = \{S_1, S_2, ..., S_N\}$, burada $|S| = N$. t zamanındaki durum $q_t = S_i \in S$ gösterilir.
- Gözlem sembolleri kümesi: $V = \{v_1, v_2, ..., v_M\}$, burada $|V| = M$. t zamanındaki gözlem sembolü O_t ile gösterilir.
- Geçiş Olasılıkları: $A = \{a_{ij}\}$, burada $\{a_{ij}\} = P[q_t+1 = S_j \,|q_t = S_i\,]$, $1 \leq i, j \leq N$ (S_i durumundan S_j durumuna geçiş olasılığı).
- Emisyon Olasılıkları: $B = \{b_j(k)\}$, burada $b_j(k) = p[O_t = v_k|q_t = S_j\,]$, $1 \leq j \leq N$, $1 \leq k \leq M$ (S_j durumundan v_k sembolü
- Başlangıç Durumu Olasılıkları: $\pi = \{\pi_i\}$, burada $\pi_i = P(q_1 = S_i)$, $1 \leq i \leq N$ (Başlangıç durumunun S_i olma olasılığı)

Bir SMM'nin unsurları göz önüne alındığında, SMM, $\lambda = (A, B, \pi)$ olarak ifade edilebilir, burada N ve M, dolaylı olarak A ve B tarafından temsil edilir. Bir SMM, ayrıca, bir dizi gözlem sembolü gerektirir: O = O1O2...OT. Bir SMM λ ve bir gözlem dizisi, O, dikkate alındığında, durumların ve gözlemlerin birinci dereceden Markov özellikleri, bir dizi durum ve gözlemin ortak dağılımını denklem 3.9'da gibi yeniden düzenlememizi sağlar:

$$P(S_{1:T}, O_{1:T}|\lambda) = P(S_1)P(O_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(O_t|S_t) \qquad (3.9)$$

3.3.2.İleri Yönlü Değişkenler

İleri olasılık (αt(i)), λ modeli dikkate alındığında, t zamanındaki kısmi gözlem dizisi O1O2...Ot ve Si durumunun olasılığıdır. Bir başka ifadeyle, durum dizisinin, t zamanına kadar gözlenmiş, t zamanındaki bir Si durumunda sona erme olasılığıdır.

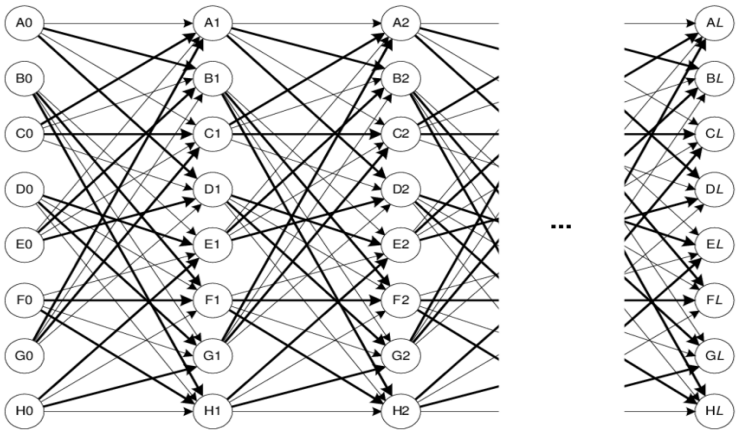
$$\alpha_t(i) = P(O_1 O_2 O_3 \ldots O_t, q_t = S_i \mid \lambda) \tag{3.10}$$

Tüm zaman adımlarında, tüm durumların ileri olasılığını hesaplamak için dinamik bir programlama yaklaşımı kullanılır. İleri olasılık (αt(i)), aşağıdaki gibi, dinamik programlama kullanılarak indüktif olarak hesaplanır:

- Başlatma: $\alpha_1(i) = \pi_i b_i(O_1), 1 \leq i \leq N$ (3.11)
- Tümevarım:

$$\alpha_{t+1}(i) = \left[\sum_{j=1}^{N} \alpha_t(j)\alpha_{ij}\right] b_i(O_{t+1}), 1 \leq \text{t} \leq \text{T} - 1, 1 \leq \text{j} \leq \text{N} \tag{3.12}$$

İleri değişken hesaplaması için bu dinamik programlama yaklaşımı, T uzunluğu ve N boyutu halindeki bir durum kafesi olarak görülebilir. Şekil 3.7'da durum kafesini görsel şekilde sunmaktadır. Dinamik programlama yaklaşımı, durum kafesindeki ileri değişkeni O (TN2) zamanında hesaplamamızı sağlar.



Şekil 3.7. Durum Kafesi Örneği.(Wang ve ark., 2007)

3.3.3.Geri Yönlü Değişkenler

Geri yönlü olasılık (βt(i)), i durumu ve t zamanı ve λ modeli göz önüne alındığında, gözlem dizisinin t + 1 zamanından sona kadar olan olasılığıdır. Bu, t + 1

zamanından sonuna kadar bir gözlem dizisini, t zamanındaki bir Si durumunda başlayarak gözlemleme olasılığıdır.

$$\beta_t \mathrm{t(i)} \; = \; \mathrm{P}(\mathrm{O}_{t+1} \, \mathrm{O}_{t+2} \, \mathrm{O}_{t+3} \ldots \mathrm{O}_T | \mathrm{q}_t \mathrm{t} = \; \mathrm{S}_i \; , \lambda) \qquad (3.13)$$

İleri yönlü değişkene benzer şekilde, tüm zaman adımlarında tüm durumların geri yönlü olasılığını hesaplamak için dinamik bir programlama yaklaşımı kullanılır. Geri yönlü olasılık (βt(i)), aşağıdaki gibi dinamik programlama kullanılarak indüktif olarak hesaplanır:

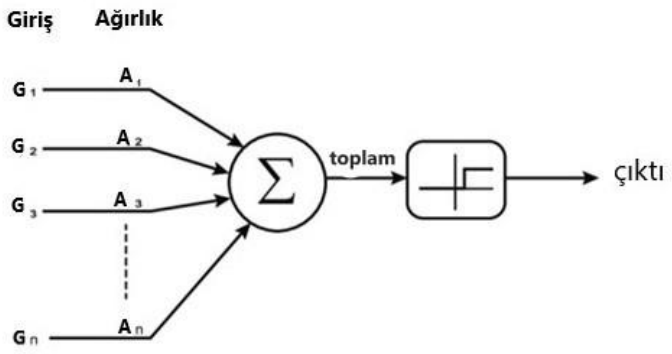
- Başlatma: $\beta_T(i) = 1, \quad 1 \leq \mathrm{i} \leq \mathrm{N}$ (3.14)
- Tümevarım:

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} b_j(O_{t+1}), \; \mathrm{T} - 1 \geq \mathrm{t} \geq 1, 1 \leq \mathrm{j} \leq \mathrm{N} \qquad (3.15)$$

T uzunluğu ve N genişliğindeki bir kafes durumunun geri yönlü olasılık hesaplamasının karmaşıklığı, O (TN2)'dur.

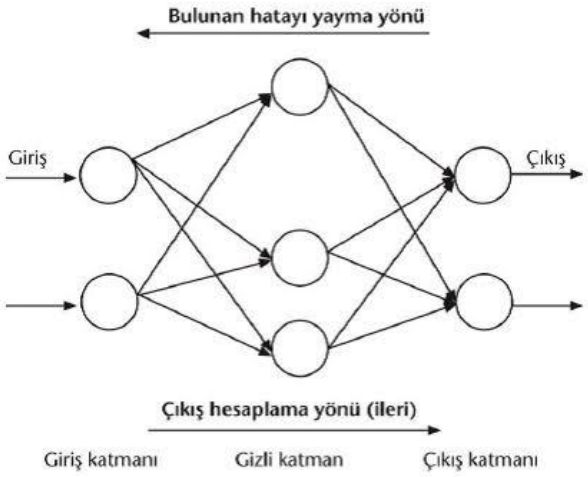
3.4. Yapay Sinir Ağları

Bir sinir ağının genel yapısı hakkında genel bir bilgilendirme yapılması amacıyla sinir ağlarına bu bölümde yer verilecektir. 1989'da Dr. Robert Hecht-Neilsen sinir ağlarını "… harici girdilere karşı dinamik durum tepkisi vererek bilgiyi işleyen bir dizi basit, birbirine yüksek seviyede bağlı işlem elemanlarından oluşan bir bilgisayar sistemi" olarak tanımlamıştır (Caudill, 1989). Sinir ağındaki klasik fakat bir o kadar da basit bir düğüm tipi, McCulloch-Pitts düğümüdür (McCulloch ve Pitts, 1943). Bu düğümün veya McCulloch ve Pitt'in onları ifade etmeyi tercih etiği ismiyle bir nöronun gösterimi Şekil 3.8'da görülebilir.



Şekil 3.8. Klasik bir sinir ağı düğümü olan McCulloch-Pitts nöronu (Chakraborty ve Sarddar)

Bir sinir ağında, Şekil 3.9'de görüldüğü gibi üç çeşit katman vardır. Giriş katmanı, giriş verilerini beslediğimiz düğümlerdir. Çıkış katmanı, giriş katmanında verilen giriş verisi doğrultusunda ağ tarafından üretilen çıktıdır. Yeni frekansları tahmin etmek istediğimiz için projemizde, çıktı katmanının çıktı boyutu girdi katmanının boyutuna, yani bu katmandaki düğüm sayısına eşit olmalıdır. Orta katman, gizli katman olarak adlandırılmaktadır. Basit şekilde ifade etmek gerekirse gizli katman boyutu ve sayısı, ağın ne kadar bilgi ayırt edebileceğini belirler. Bu sinir ağının temel yapısıdır.



Şekil 3.9. Basit bir ileri beslemeli sinir ağının örnek yapısı.(SA) (derinogrenme.com, 2017)

Şimdi ise ağın gerçekte nasıl öğrenebildiğini anlatacağız. Her katman bir düğüm dizisi ile ilgilidir. Bir katmandaki her düğümden, Şekil 3.9'de gösterildiği gibi, onu bir sonraki katmanda bulunan düğümlerin her birine bağlayan bir uç vardır. Her düğüm, uçlardan gelen girdilere dayanarak önceden tanımlanmış bir hesaplama yapar. McCulloch-Pitts düğümünde yapılan hesaplamalar temel olarak bir sigmoid fonksiyonudur. Girdileri toplarız ve eğer toplam girdi belirli bir eşiğin üzerindeyse çıktı olarak 1, değilse 0 üretiriz. Daha sonra göreceğimiz gibi, bu düğümlerin daha karmaşık gösterimleri vardır, ancak McCulloch-Pitts neuronu, sinir ağlarının temellerini anlamak için iyi bir başlangıç noktasıdır. Şekil 3.9'de, bu düğümlerle kombinasyon halinde olan uçlar oklar ile gösterilmiştir. Bu uçlara daha yaygın tabiriyle ağın ağırlıkları denir.

Temel olarak yapılan iş, bir nöronun çıktısını kendi değeri ile çarpmak ve ardından çıkan sonucu ucunun bağlı olduğu bir sonraki düğüme girdi yaparak onu beslemektir. Üretilen çıktı doğrultusunda bu ağırlıkları güncelleyerek ağa hangi girdi verilerinin hangi çıktı verilerini üretmesi gerektiğini ayırt etmeyi öğretebiliriz. Ağırlıkları güncelleme işlemine geri yayılım (back propagation) adı verilir.

3.4.1. Geri Yayılım

Bir sinir ağının eğitiminin geri yayılım kısmı aslında öğrenmenin gerçekleştiği yerdir. Burası ağın, verilen bir girdiye karşılık doğru çıktıyı elde etmek için bütün ağ boyunca ağırlıklarını güncellediği yerdir. Geri yayılım algoritmasının iç yapısına yönelik çalışmalar (Rumelhart ve ark., 1985) tarafından açıklanmış olup, bu bölümde konsept hakkında genel bir açıklama yapacağız. Geri yayılım, her bir çıkış düğümünü beklenen çıktı ile karşılaştırdığımız çıktıdan başlar. Elde ettiğimiz çıktı ile beklenen çıktı arasındaki fark birçok şekilde hesaplanabilir. Bu hesaplama için kullanılan klasik bir yöntem, ortalama karesel hatadır. Hatayı hesaplamak için kullanılan fonksiyon bizim kayıp/yitim fonksiyonumuz olarak adlandırılır. Kayıp fonksiyonunun geri yayılım algoritmasıyla çalışabilmesi için türevlenebilir olması gerekmektedir. Bir sonraki sefer girdi ağa tekrar verildiğinde beklenen çıktıya daha yakın çıktı değerleri elde etmek için şimdi ağdaki tüm ağırlıkları güncellemek istiyoruz. İşe kayıp fonksiyonunun çıkış düğümüne gelen uçlara göre kısmi türevini alarak başlıyoruz. Her türev, kayıp fonksiyonu çıktısının her bir giriş ağırlığına ne kadar bağlı olduğunu ifade etmektedir. Ağırlıklar artık güncellenmiştir ve bir önceki katmanın düğümlerindeki aktivasyon fonksiyonlarının hatası aynı yöntemle güncellenebilir. Bu güncelleme, tüm ağı güncellemek için giriş katmanına kadar tekrar tekrar yapılır.

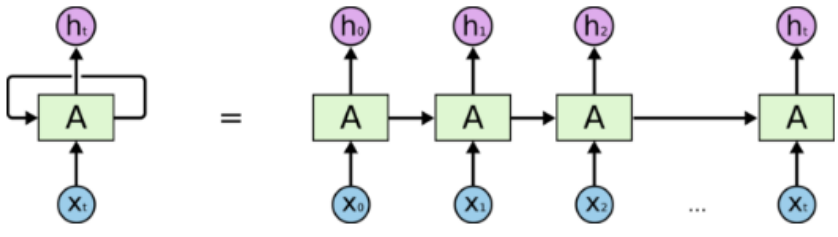
3.5. Derin Öğrenme

Son 10-15 yıl içerisinde hesaplama kapasitesinde yaşanan ilerlemeler sayesinde, makine öğrenmesinde yeni bir alan ortaya çıkmıştır. Bu alana derin öğrenme denilmektedir. Kavram olarak derin öğrenme, makine öğrenmesi içindeki birçok algoritmayı kapsayabilir. Ancak, sinir ağları açısından derin öğrenme, genellikle çıktı ve girdi katmanı arasında her katmanın çok sayıda düğüm setinden oluştuğu çok fazla gizli katmandan oluşan çok büyük modeller olarak görülmektedir. Derin öğrenme alanında yaşanan ilerlemenin asıl nedeni, derin öğrenmenin grafik işlem birimleri (GPU'lar) için genel amaçlı hesaplama maksadıyla uygulanmasıdır (Fung ve Mann, 2008). Sinir ağları açısından değerlendirildiğinde GPU hesaplamalarının CPU'lara karşı iki ana üstünlüğü bulunur. Bunlardan ilki, geniş miktardaki verinin işlem ünitesine hızlı erişimdir, böylece hareketli verilerin ek yükü azalır. Diğer üstünlüğü ise GPU'nun yaptığı hesaplama

türlerinin optimize edilmiş olmasıdır. GPU'lar, hesaplama matrisleri için iyi optimize edilmiş fonksiyonlara sahiptir. Yapay sinir ağlarının eğitilmeleri için büyük ölçüde matris işlemlerine ihtiyaç duyulmasından dolayı, GPU daha hızlı işlem yapılmasına imkân verir (Fung ve Mann, 2008).

3.5.1. Tekrarlayan Sinir Ağları

Derin öğrenme konusunda ilerleme sağlandıkça, daha önce denenmemiş teorik algoritmalar, karmaşık problemlerin çözümü için test edilebilir ve kullanılabilir hale gelmiştir. Bu tür algoritmalardan biri tekrarlayan sinir ağlarıdır (Recurrent Neural Networks - RNN). Jürgen Schmidhuber, RNN'lerin erken dönem benimseyicisi ve uygulayıcısıydı ve RNN'ler açısından birçok büyük keşfe imza atan bir araştırma ekibine öncülük etti (Schmidhuber, 2015). Tekrarlayan sinir ağının uygulanmasıyla kaydedilen en büyük ilerleme, bir sonraki katmandaki düğümler ile önceki katmanda bulunan düğümler arasında bağlantı kurmaya elverişlilik kabiliyetidir. Yani bu kabiliyet, sinir ağlarının yönlendirilmiş çevrimler olarak ifade edilmesine izin verir. Bu sinir ağları artık eğitilmiş girdilerin döngülere bağlı olmasına izin vermektedir. Girişin durağan olması gerektiği önceki sinir ağı algoritmalarına kıyasla bu durum, tekrarlayan sinir ağlarının giriş ve çıkış verilerini, örneğin metin cümlelerini birbiri ardına, sıralı şekilde modellemesini ve aynı zamanda bu silsileler arasındaki bağımlılığı modellemesini mümkün kılmıştır. Şekil 3.10'de tekrarlayan bir sinir ağı mimarisine ait örnek gösterilmektedir.

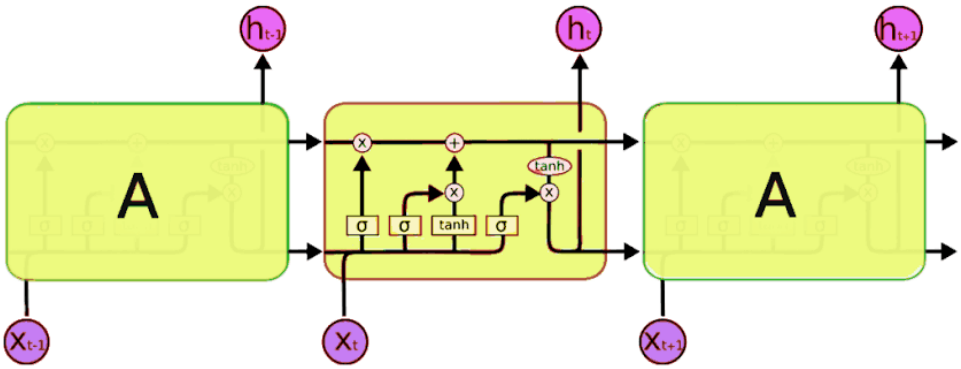


Şekil 3.10. Tekrarlayan sinir ağının örnek yapısı (Mittal, 2019)

3.5.2. Uzun – Kısa Süreli Bellek Sinir Ağları

Tekrarlayan sinir ağları uygulamaları sayesinde verilerimizi zaman içinde önceki adımlara bağımlı diziler halinde modellemeye başlayabiliriz. Daha önce açıklanan

tekrarlayan sinir ağı, uzun vadeli bellek açısından mahzurludur. Tekrarlayan sinir ağları kısa zaman adımları içindeki bağımlılıklar için iyi çalışır. Fakat tekrarlayan sinir ağlarının teoride uzun vadeli bağımlılıkları modellemesi mümkün gözükürken, gerçek durumda bunu başarması çok zordur (Hochreiter, 1991; Bengio ve ark., 1994). İşte bu yetersizlik, uzun-kısa süreli bellek (Long-Short Term Memory - LSTM) sinir ağının ortaya çıkmasının arkasındaki asıl sebep ve gereklilik olmuştur (Hochreiter ve Schmidhuber, 1997). LSTM ağları düğümlerin geçmiş bilgileri nasıl koruduğunu yeniden tanımlayarak uzun vadeli bağımlılıkları daha iyi sürdürebilir. Böylece uzun dönem bağımlılıkları idare etmek LSTM ağı için öğrenmeye uğraştığı bir şeyden çok doğal bir şey haline gelir. Daha önce de ifade edildiği gibi LSTM, uzun vadeli bağımlılıkları korumak için düğümün yapısını yeniden tanımlar. Düzenli RNN'ler, doğrusal olmayan tek bir aktivasyon fonksiyonu, örneğin *tanh* fonksiyonu gibi çok basit bir düğüm yapısını kullanabilir. LSTM ağının yapısı çok daha karmaşıktır. Bir LSTM düğümü çizimi Şekil 3.11'te görüldüğü gibidir.



Şekil 3.11. Bir LSTM düğümünün yapısını göstermektedir (Srıvastava, 2017)

LSTM düğümü dört etkileşim katmanından oluşur. LSTM düğümünün çekirdek kısmı, Şekil 3.9'te görüldüğü gibi düğümden geçen düz çizgidir. Bu çizgi hücre durumunu gösterir. Hücre durumu, büyüklüğü LSTM düğümü için ayarlanan nöronların sayısıyla belirlenmekte olan bir matris olarak modellenmiştir. Bir sonraki zaman adımına kaydedilecek bilgileri ekleyebileceğimiz veya kaldırabileceğimiz yer hücre durumudur. Şimdi LSTM düğümünün her katmanından geçeceğiz ve her katmanın hücre durumunu nasıl etkilediğini açıklayacağız. LSTM düğümünde toplam dört katman vardır. Tıpkı hücre durumu gibi, katmanların her biri, büyüklüğü bir LSTM düğümü için nöronların sayısı ile belirlenen bir matris olarak modellenmiştir. Her katman aynı girişe, yani önceki LSTM düğümünün çıkışına ve mevcut zaman adımındaki girişe sahiptir. İlk katman, Şekil 3.9'te en soldaki sarı kutu olarak görülen sigmoid katmandır. Bu katman unutma katmanı

olarak adlandırılır. Bu katmanın görevi, önceden bilinen fakat şimdi unutmak istediğimiz bilgileri silerek hücre durumunu güncellemektir. Bunu görselleştirmenin bir yolu, bir odanın içerisinde bulunan kişileri düşünmektir. Aynı kişiler odada olduğu sürece, her bir kişiye atıfta bulunmak istiyoruz. Ancak bir kişi odadan ayrılırsa, bu kişiyi odanın mevcudundan çıkarmak istiyoruz ki böylece mevcut olmayan bir kişiye atıfta bulunmayalım. Unutma katmanının yaptığı iş budur. Unutma katmanında gereken hesaplamalar için kullanılan formül denklem 3.16'daki gibidir.

$$f_t = \sigma(W_f \,.\, [h_{t-1}, x_t] + b_f) \tag{3.16}$$

Burada σ sigmoid fonksiyonu, W_f unutma katmanının ağırlık matrisini, · çarpım, h_{t-1} önceki LSTM düğümünden üretilen çıktı, x_t anında ağa gelen girdi ve b_f unutma katmanının yanlılık terimidir. Sonraki iki katman, hücre durumuna hangi bilgilerin ekleneceğine karar verir. İlk katman sigmoid bir katmandır, bu katman durumun hangi bölümlerini güncellemek istediğimize karar verir. Diğer katman ise tanh katmanıdır. Tanh katmanı, mevcut hücre durumuna eklemek istediğimiz yeni bir hücre durumu üretir. Bu iki katmanın işbirliği yapması, tanh katmanının yeni bir hücre durumu adayı üretmesinden geçer. Bu aday hücre durumu daha sonra, hangi bölümlerinin mevcut hücre durumuna eklenmesi gerektiğine karar veren bir kapı bekçisi olarak görev yapan sigmoid katmandan geçirilir. Bu yöntem, hücre durumuna yeni içerik eklemek için bir yol sunar. Odadaki kişiler benzetmesini kullanmaya devam edersek, bu yöntem odaya yeni bir kişi eklememize olanak sağlar. Hücre durumunun mevcut durumu, girdi olarak bir sonraki zaman adımına aktarılır. Mevcut hücre durumu, mevcut zaman adımının çıktısını üretmek için de kullanılacaktır. Hücre durumu güncelleme işlemi denklem 3.17, 3.18 ve 3.19'deki formüllerle ifade edilmiştir.

$$i_t = \sigma(w_i.\,[h_t - 1.\,x_t] + bc) \tag{3.17}$$
$$\tilde{c}_t = tanh(w_c.\,[h_t - 1.\,x_t] + bc) \tag{3.18}$$
$$c_t = f_t \odot c_t - 1 + i_t \odot \tilde{c}_t \tag{3.19}$$

Yukarıdaki formüllerde i_t güncelleme katmanını temsil eder ve $\tilde{C}_t$ hücre durumu güncelleme adayıdır. $\odot$ gösterimi noktasal çarpma işlemini ifade eder. W_i ve W_C sırasıyla güncelleme katmanı ağırlık matrisi ve hücre durumu adayıdır. b_i ve b_C ağırlık matrisleriyle aynı adlandırma kuralını izleyen yanlılık terimidir. C_t yeni hücre durumudur. İlk önce unutma katmanı adı verilen matris ile çarpılır ve ardından güncelleme katmanı adı verilen matris ile toplanır. Son katman çıktı katmanıdır. Bu katmanda çıktı olarak neyi

üretmek istediğimize karar veririz. Çıktı, güncellenmiş hücre durumuna bağlı olacak, ancak bir sigmoid katman tarafından filtrelenecektir. Hücre durumu, hücre değerlerinin -1 ve 1 arasında oluşmasını sağlamak için tanh fonksiyonundan geçirilir ve sonuç, çıktıyı üretmek için sigmoid katmanı adı verilen matris ile çarpılır. Bu çıktı aynı zamanda, hücre durumunu güncellemek için LSTM katmanının bir sonraki zaman adımında girdi olarak da kullanılacaktır. Çıkış katmanını hesaplamak için kullanılan formüller denklem 3.20 ve 3.21'te ifade edilmiştir.

$$O_t = \sigma\left(w_o\left[h_{t-1,}x_t\right] + b_o\right) \quad (3.20)$$

$$h_t = o_t \odot tanh(C_t) \quad (3.21)$$

Burada o_t, geçerli zaman adım çıktısını üretmek için tanh eşlenmiş hücre durumu, C_t, ile çarpılan çıktı matrisi h_t dir.

Dizi Sırasıyla Öğrenme: Sorunumuz temelde bir dizinin diğer diziye eşlenmesi olduğundan, bunu ağ için ifade edebilmenin bir yoluna ihtiyaç duyarız. Bu amaçla ortak ve etkili bir yaklaşım, Sutskever ve diğ. tarafından tanımlanmıştır (Sutskever ve ark., 2014). Makalelerinde İngilizce'den Fransızca'ya, metinden metine çevirme görevini üstlenmiş ve harika sonuçlar elde etmişlerdir. Kullandıkları düzenek iki derin LSTM ağından oluşmaktaydı. Birisi girdi dizisini kodlamak ve diğeri çıktı dizisini çözmek için. LSTM'leri ayırmanın iki yararı bulunmaktadır. İlk olarak model parametrelerinin sayısı artar ve sadece ihmal edilebilir seviyede hesaplama maliyeti doğurur. İkinci olarak ise LSTM'yi aynı anda çoklu dil çiftinde eğitmeye uygun hale getirir (Kalchbrenner ve Blunsom, 2013).

Düzenlileştirme: Yapay sinir ağları ve makine öğrenme algoritmaları oluştururken, ezberleme sorunlarının farkında olmak önemlidir. Ezberleme, modelimizin eğitim seti bünyesinde kendine özgü özellikleri öğrenmeye başladığında gerçekleşir. Yani model ya girdi verisinden çıktı verilerini oluşturan genel kuralları öğrenmez, ya da toplam veri seti için genel kurallardan daha fazla kural öğrenir. Bu durum, değerlendirme hatamız artarken eğitim hatamızın azalması ile sonuçlanır. Sonunda, modelin öğrendiği eğitime özgü belli kurallar nedeniyle bilinmeyen veriler üzerinde daha kötü performans gösteren bir model ile karşı karşıya kalırız. Modelimizde, bu ezberleme sorununun önüne geçmek için iki teknik kullanıyoruz. İlk yöntem ağırlık azalımı olarak adlandırılmaktadır. İkincisi ise seyreltme olarak adlandırılmaktadır. Bu iki teknik aşağıdaki alt bölümlerde açıklanmıştır.

Ağırlık Azalımı: Tüm düzenlileştirme yöntemlerinin hedefi, sinir ağının öğrenmeye çalıştığı veriler üzerinde daha iyi genelleştirilme yapmasını sağlamaktır. Bu yöntem ezberleme potansiyelini azaltmak için kullanılan bir mekanizmadır. Ağırlık azalımının arkasındaki fikir oldukça basittir ve uzun zamandır bilinmektedir (Hinton, 1987). Bir ağın karmaşıklığını modellemenin bir yolu, ağın serbest parametrelerinden, yani eşikler ve ağırlıklar gibi parametrelerden geçmektedir. Bu amaçla yaygın bir yaklaşım olarak eğitim hatasını hala küçük tutarken, serbest parametrelerin sayısı en aza indirilir (Baum ve Haussler, 1989; Tishby ve ark., 1989; Schwartz ve ark., 1990). Ağırlık azalımı yöntemi, serbest parametrelerin sayısını en aza indirgemek yerine onlardan faydalandığımız başka bir yaklaşımı bu soruna çözüm olarak getirmektedir. Ağırlık azalımı yöntemi basitçe ağın kayıp fonksiyonuna bir ektir. Denklem 3.22 doğrultusunda açıklanabilir.

$$E(w) = E_0(w) + \frac{1}{2} + \lambda \sum_i w_i^2, \tag{3.22}$$

Bu denklemde E_0 seçilen kayıp fonksiyonumuzdur ve λ, ağırlık azalımı faktörünün hatayı ne kadar etkileyeceğine karar veren ağırlık azalımı skalarıdır. Çok küçük bir λ değerine sahipsek, ağırlık azalımı, ağı düzenlileştirmeye yetmez. Buna karşılık eğer λ çok büyükse hata fonksiyonumuz azalacak ve ağımız ağ ağırlığını 0'da tutma eğilimine girecektir. Denklemde w_i ağdaki i indeksindeki bir ağırlıktır ve ağırlık azalımı faktörü temelde bütün ağın ağırlık karelerinin toplamıdır. Ağırlık azalımı yönteminin ağ performansına katkısı, hata oranını düşük tutan en küçük ağırlık setini seçerek ağırlık vektörünün ilgisiz bileşenlerini en aza indirmesidir. Bu yöntem iyi seçildiği takdirde, hedeflerden bazı statik gürültüyü de kaldırabilir (Krogh ve Hertz, 1992).

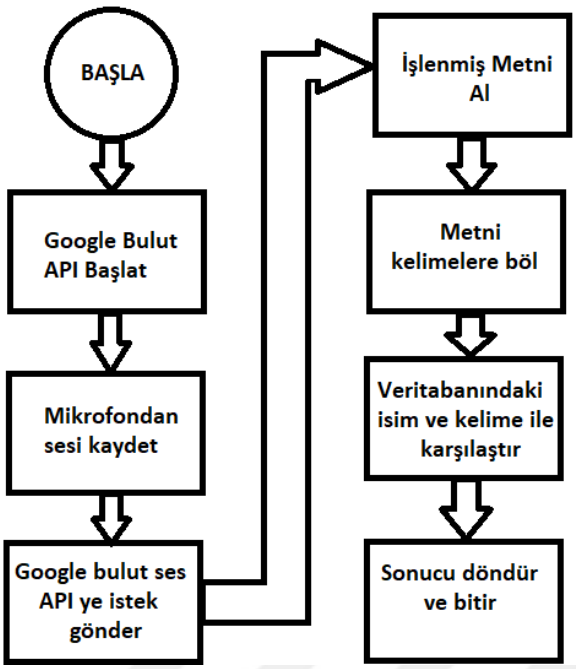
Seyreltme: LSTM ağı gibi doğrusal olmayan düğümleri olan bir sinir ağını eğitirken, düğümler doğru çıktıyı tahmin edebilmek için özellikleri tanımlamayı öğrenir. Daha büyük ağlarda ağırlıkların düğümleri birbirine bağlayan çeşitli ayarları vardır ve bu ayarlar eğitim verilerinin neredeyse kusursuz bir şekilde tahmin edilmesini sağlayabilirler. Genelleştirme amacıyla, eğitim setinin az sayıdaki güçlü özelliğine odaklanan ağırlık sayısını azaltmak istiyoruz (Srivastava ve ark., 2014). Bunun için seyrekleştirme güçlü bir yaklaşımdır. Seyrekleştirmenin altında yatan mekanizma, eğitim sırasında ağdaki düğümleri rasgele devre dışı bırakmaktır. Bunu yaparak, birimlerin

birbirlerine çok fazla uyum sağlamasının önüne geçiyoruz. Eş-uyum, iki nöronun aynı özelliği tekrar tekrar algılamaya başladığı zaman gerçekleşir. Aynı özelliği tespit eden nöronların sayısını azaltarak, eğitim setinde daha fazla özellik öğrenebilir ve sonunda daha iyi test sonuçları elde edebiliriz. Çeşitli karşılaştırmalar sonucunda, seyreltmenin uygulanmasının küçük bir eğitim setinde eğitilmiş karmaşık sinir ağları üzerinde büyük bir gelişme sağlayabileceği gösterilmiştir (Srivastava ve ark., 2014). Seyreltme yöntemini kullanan kişi, eğitim sırasında her bir düğümün hariç tutulma olasılığını ifade eden seyreltme olasılığına karar vermelidir.

3.6.Google Bulut ses tanıma

Google Bulut, kullanımı kolay bir uygulama arayüzü ile (API- Application Interface) derin öğrenme tabanlı sinir ağı modelleri uygulayarak geliştiricilerin sesi metne dönüştürmesini sağlar. Bu API, global anlamda 120 dili ve varyantı tanır. Önerdiğimiz sistemde, daha fazla işlem yapılması adına, konuşulan cümleleri metne dönüştürmek için Google bulut ses tanıma API (Uygulama programı ara yüzü) kullandık. Google tarafından kullanılan algoritma, bu bölümde daha önce açıklanan Uzun-Kısa Süreli Bellek ve Tekrarlayan Sinir Ağlarıdır. Aşağıdaki Akış Şeması, bu işlevi sistemimizde nasıl uyguladığımızı göstermektedir.

Google Bulut yapısı gerçek zamanlı akış üzerinden de işlem yapabilmektedir. Bu mimaride FLAC, AMR, PCMU ve Linear-16 dahil olmak üzere birçok ses kodlaması desteklenir. Ayrıca otomatik dil algılama seçeneği de mevcuttur. Bununla birlikte gürültüye karşı dayanıklılık sistemi kapsamında ek ses giderme işlemleri gerekmeden, birçok ortamda gürültülü ses verilerini işleyebilmektedir. Uygunsuz içeril filtrleme özelliği sayesinde bazı dillerde metin sonuçlarındaki uygunsuz içerik filtrelenebilmektedir. Otomatik noktalama kapsamında metin içerisine nokta, virgül vb noktalama işretleri de konabilir. Model seçimi özelliğine sahip Google Bulut yapısı önceden eğitilmiş modellerin yanı sıra kendi eğittiğimiz modelleri de dahil edebilir. Konuşmacı ayrıştırma modülü sayesinde her ifadeyi konuşmadaki kişilerden hangisinin sarf ettiğine dair otomatik tahminler vermektedir.



Şekil 3.12. Google bulut konuşma API ara yüzünün sistemimizdeki kullanım akış şeması.

Akış şemasında da gösterildiği gibi, program, öncelikle, yetkilendirme kodunu kullanarak Google bulut ses API'sına bağlanır ve Başlatma işlemini gerçekleştirir. Ardından, mikrofonu kullanarak sesi kaydeder ve bunu Google Bulut'a bir istek olarak gönderir. Önceden eğitilmiş yapıdaki Google Bulut işlenen metni alarak önce bölme işlemi yapar. Daha sonra bölünen metni veritabanındaki isim ve kelimlerle karşılaştırarak sonucu döndürür.

4. ARAŞTIRMA BULGULARI

Bu çalışmada çağrı merkezleri için Türkçe ses tanıma tabanlı otomatik sesli yanıt ve yönlendirme sistemi tasarlanmıştır**.** Öncelikle MFKK ve SMM yaklaşımı kullanarak ses tanıma uygulaması gerçekleştirilmiş ve sonuçları irdelenmiştir. Başarı oranının yeterli seviyede olmadığı görülmüş ve derin öğrenme yaklaşımı ile ses tanıma sistemi geliştirilmiş ve uygulanmıştır. Bu bölümde her iki yaklaşıma ilişkin araştırma bulguları sunulmuştur.

Tasarlanan sistemin ses/konuşma tanıma yazılımı Phyton programlama dilinde kodlanmıştır. Yazılımın amacı, telefon görüşmelerinden alınan sesleri analiz ederek arayan kişinin söylediklerini anlamak ve buna göre yönlendirme yapmaktır. Yazılım, hem konuşmayı metne hem de metni konuşmaya dönüştürmektedir.

Sistemin işleyişine ilişkin akış diyagramı Materyal ve Yöntem bölümünün başında verilmiştir. Bu diyagramda da görüldüğü üzere, önceden hazırlanmış metinlerin sese dönüştürülerek arayan kişiden bilgi talep edilmekte ve telefondaki kişinin söyledikleri derin öğrenme tabanlı yaklaşımla tanımlanarak metne dönüştürülmekte ve yönlendirme yapılmaktadır. Sistem içerisinde bazı kısımlarda aranmak istenen kişinin adı ve soyadı istenmekte ve devamında bölümü de istenerek onay alınmaktadır. Onay kısmında “evet” ve “hayır” şeklinde cevap istenmekte ve tanınan cevap metnine göre yön seçenekleri belirlenmektedir. Sistem eğer karşıdaki kişinin ulaşmak istediği kişiyi tanıma işleminde bir sonuca ulaşamazsa en başa dönme veya konuşmayı sonlandırma seçeneklerini sunmaktadır. Aynı isimde çift kayıt çıkması durumunda ise bölüm isimleri okunarak seçim yapması istenmektedir.

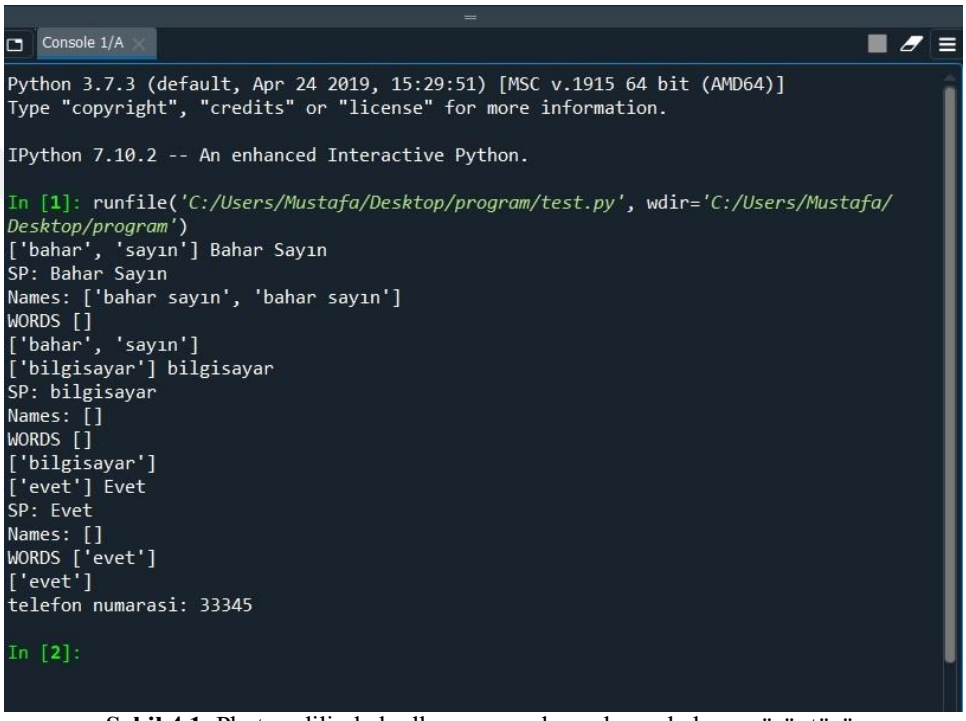
Geliştirilen interaktif sesli yanıt ve yönlendirme sisteminin performansını test etmek amacıyla oluşturulan veritabanı rehber niteliğinde olup 100 kişi ve 10 bölüm ihtiva etmektedir.

Otomatik sesli yanıt ve çağrı sistemlerinde karşılaşılabilecek durumlar göz önünde bulundurularak sistemin performans kriterleri belirlenmiş ve akış şeması buna göre şekillendirilmiştir. Örneğin aynı isimde 2 kişi olması ihtimaline karşı veritabanına kaydedilen 100 isim arasında 14 adet aynı isimde kayıt oluşturulmuştur. Seçim yapılabilmesi adına bu kişiler ayrı bölümlerde tanımlanarak bölümlerin seçimine yönelik ses tanıma seçenekleri eklenmiştir. Bununla birlikte telefonla konuşulması sırasında ortam gürültülerinin de olacağı ön görülerek testler gürültülü ve gürültüsüz ortam testi

olarak ayrılmıştır. Gürültüsüz ortam testlerinde arka plan sessiz olurken gürültülü ortam testlerinde telefon görüşmesi sırasında aşağıdaki durumlar oluşturulmuştur:

- Arka planda iki kişinin normal sesle konuşması,
- Arka planda normal ses düzeyinde müzik çalınması
- Arka planda kalabalık gürültüsü (uğultu şeklinde karmaşık sesler) olması

Şekil 4.1. de test sürecine ilişkin konsol ekranı verilmiştir. Bu ekranda "Bahar Sayın" ismi iki farklı bölümde olduğu için kullanıcıdan bölüm seçimi yapması istenmiş ve yönlendirme gerçekleştirilmiştir.



Şekil 4.1. Phyton dilinde kodlanmış yazılımın konsol ekran görüntüsü.

Ses tanıma sistemi test edilirken farklı ses ve tınıya sahip insanların sisteme giriş yapması sağlanmıştır. Bu amaçla telefon görüşmelerini yapmak amacıyla farklı yaşlarda ve cinsiyette 50 kişi belirlenmiş ve gürültüsüz/gürültülü olmak üzere 100 adet deneme yapılması sağlanmıştır. Çizelge 4.1'de kullanıcılara ilişkin yaş ve cinsiyet bilgileri verilmiştir.

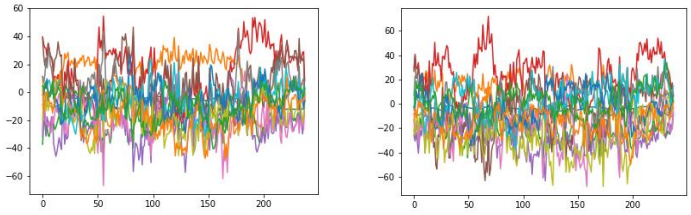
Çizelge 4.1. Test kullanıcı bilgileri.

CİNSİYET/YAŞ	17-24	25-34	35-44	45 VE ÜSTÜ	Toplam
ERKEK	6	21	1	0	28
KADIN	14	7	0	1	22
TOPLAM	20	28	1	1	50
Test Sayısı	40	56	2	2	100

4.1. MFKK ve SMM yaklaşımı ile alınan sonuçlar

Öznitelik çıkarmada en çok kullanılan yöntemlerden biri, otomatik ses tanıma sistemlerine yönelik standart bir yöntem olarak kabul edilen Mel-Frekansı Kepstral Katsayıları (MFKK) algoritmasıdır (Motlıcek, 2002) .Literatürde yer alan MFKK tabanlı sistemlerin birçoğu, etkili ses sinyali kodlamasına yönelik olarak yaklaşık 20 katsayı ve sıklıkla da 10-12 katsayı kullanmıştır (Hagen ve ark., 2003). MFKK algoritması, spektral forma bağlı olması nedeniyle gürültüye duyarlıdır, bu nedenle, bu problemden kaçınmak için, yöntemlerin çoğu, ses sinyallerinin periyodikliğinde bilgi kullanır ve aynı zamanda periyodik olmayan sinyaller içerir (Ishizuka ve Nakatani, 2006).MFKK algoritmasının temel mantığı, doğrusal olmayan frekans ölçeğinin, 1 kHz altındaki frekanslar için doğrusal ve yukarıda belirtilen logaritmik Mel-Frekans ölçeğine yönelik bir tahminde bulunulmasıdır, bu kavram biyolojiden esinlenmiştir. Bunun nedeni, insan işitme sisteminin, 1 kHz üzerindeki frekanslarda daha az seçici davranmasıdır. MFKK kullanılarak çıkarılan öznitelikler, log filtre yığını enerjilerinin kepstrale karşılık gelir. Diğer yandan, Hızlı Fourier dönüşümü (HFD) gibi bazı güç spektral analiz teknikleri, bazen sinyalin zaman içindeki frekans içeriğinin güç spektrumu tarafından tanımlandığı birçok otomatik ses tanıma sisteminde uygulanır. Bu çalışmada öncelikle literatürde ses sentezinde sıklıkla kullanılan MFKK öznitelik ve Saklı Markov Modeline dayanan bir ses tanıma sistemi uygulanmış ve sonuçlar analiz edilmiştir.

Literatüre bakıldığında otomatik ses tanıma sisteminde genel olarak 2-13 arası MFKK katsayısı elde edildiği görülmektedir. Bu çalışmada 13 adet MFKK katsayısı elde edilerek öznitelik çıkarma işlemi yapılmıştır. MFCC algoritmasının bir kanal sesi alması ve daha hızlı işlem için kaydedilen veri kümesinin örnekleme hızının 8000Hz olması nedeniyle, çalışmamızda, stereo (çiftli) yerine mono (tekli) ses kullanmayı tercih ettik. Şekil 4.2'de "*Humar Kahramanlı Örnek*" ve "*Mücahit Mustafa Sarıtaş*" ifadelerine ilişkin sinyallerden elde edilen MFKK katsayı grafiği gösterilmiştir.



(a) (b)

Şekil 4.2. MFKK katsayı grafiği **a)** *Humar Kahramanlı Örnek* **b)** *Mücahit Mustafa Sarıtaş*

MFKK – SMM yaklaşımından elde edilen sonuçların değerlendirilmesinde Doğru Pozitif (DP), Doğru Negatif (DN), Yanlış Pozitif (YP) ve Yanlış Negatif (YN) verilerine dayanarak elde edilen keskinlik (Precision), Hassasiyet (Recall), F1 puanı (F1 Score) ve doğruluk (accuracy) değerleri kullanılmıştır.

Doğru pozitif (DP): Bunlar doğru tahmin edilen pozitif değerlerdir, yani gerçek sınıfın değeri evettir ve tahmin edilen sınıfın değeri de evettir.

Doğru negatif (DN): Bunlar doğru tahmin edilen negatif değerlerdir, yani gerçek sınıfın değeri hayırdır ve tahmini sınıfın değeri de hayırdır.

Yanlış pozitifler ve yanlış negatifler, gerçek sınıfınız tahmin edilen sınıfla çeliştiğinde ortaya çıkar.

Yanlış pozitif (YP): Gerçek sınıf hayır ve tahmin edilen sınıf evet olduğunda ortaya çıkar.

Yanlış negatif (YN): Gerçek sınıf evet olduğunda, tahmin edilen sınıf hayır olduğunda ortaya çıkar.

Kesinlik: Doğru tahmin edilen pozitif gözlemlerin toplam tahmin edilen pozitif gözlemlere oranıdır (DP/DP+YP).

Hassasiyet: Doğru tahmin edilen pozitif gözlemlerin "evet" gerçek sınıfındaki tüm gözlemlere oranıdır (DP/DP+YN).

Doğruluk: Bu ölçme tekniği en sezgisel performans ölçüsüdür ve basitçe doğru tahmin edilmiş gözlemin toplam gözlemlere oranıdır. Eğer yüksek doğruluk varsa, modelimizin en iyi olduğunu düşünebiliriz. Doğruluk büyük bir ölçüdür, ancak yalnızca yanlış pozitif ve yanlış negatif değerlerin neredeyse aynı olduğu simetrik veri kümeleri olduğunda iyi sonuç verir.

F1 puanı: Hassasiyet ve Keskinlik'in ağırlıklı ortalamasıdır. Bu nedenle, bu puan hem yanlış pozitifleri hem de yanlış negatifleri dikkate alır. Sezgisel olarak doğruluk (accuracy) kadar anlaşılması kolay değildir, ancak F1 puanı, özellikle eşit olmayan bir sınıf dağılımınız varsa, doğruluktan (accuracy) genellikle daha kullanışlıdır. Yanlış pozitifler ve yanlış negatifler benzer değere sahipse doğruluk en iyi sonucu verir. Yanlış pozitiflerin ve yanlış negatiflerin değerleri çok farklıysa, hem Hassaslığa hem de Keskinlik'e bakmak daha iyidir. F1 puanı şu şekilde bulunur:

F1 Score = 2*(Keskinlik * Hassasiyet) / (Keskinlik + Hassasiyet)

MFKK – SMM yaklaşımı ile veri setinin içerdiği tüm isimler için uygulama 10 defa çalıştırılarak test yapılmıştır. Çizelge 4.2'de yapılan ilk testin sonuçları verilmiştir.

Çizelge 4.2. MFKK-SMM test sonuçları

Adı soyadı	Keskinlik	Hassasiyet	F1-puan	Adı soyadı	Keskinlik	Hassasiyet	F1-puan
Me**ÇU**	%33	%100	%50	Gü**GÜ**	%0	%0	%0
Mu**Ci**	%100	%100	%100	Os**AC**	%100	%100	%100
Mu**AC**	%0	%0	%0	Zü*YI*AC*	%100	%100	%100
Ye *SA**	%100	%100	%100	Ak**AK**	%100	%100	%100
Re**KU**	%0	%0	%0	Se**TÜ**	%0	%0	%0
Er**KO**	%100	%100	%100	Öz**AY**	%100	%100	%100
Ey**CA**	%100	%100	%100	Gü**ÖZ**	%100	%100	%100
Ha*ER*GÜ*	%100	%100	%100	Hu**KA**	%100	%100	%100
Mu**DO**	%100	%100	%100	Hü**DO**	%100	%100	%100
Er**KA**	%0	%0	%0	İb**UY**	%100	%100	%100
Er**KO**	%100	%100	%100	İk**ÖZ**	%100	%100	%100
Şa**TA**	%0	%0	%0	Su**KO**	%50	%100	%67
En**ÖZ**	%100	%100	%100	Şe**AY**	%100	%100	%100
Ha**AY**	%100	%100	%100	Mu**DO**	%100	%100	%100
Ha**AR**	%100	%100	%100	Em**AY**	%100	%100	%100
Ha**SÜ**	%100	%100	%100	Ah**BO**	%100	%100	%100
Ha**SE**	%100	%100	%100	Ke**TÜ**	%100	%100	%100
Ha*HÜ*ÇE*	%100	%50	%67	Ak**ŞA**	%100	%100	%100
İb**DE**	%50	%100	%67	Mu**KÖ**	%100	%100	%100
Er**KA**	%0	%0	%0	Se**AL**	%100	%50	%67
Ne**TR**	%0	%0	%0	Se**SE**	%50	%100	%67
İs**SA**	%0	%0	%0	Fa**AY**	%100	%100	%100
Al**ÖZ**	%0	%0	%0	Se**KA**	%0	%0	%0
Al**AL**	%0	%0	%0	Al**AK**	%100	%100	%100
Fa**BA**	%0	%0	%0	Ha*AL*BA*	%100	%100	%100
Ba**SA**	%100	%100	%100	Ah**BO**	%100	%100	%100
Bi**ÇE**	%100	%100	%100	Ta**SA**	%100	%100	%100
Bu**TE**	%50	%100	%67	Tu**DE**	%100	%50	%67
Ci**NA**	%100	%50	%67	Zi**ÖZ**	%100	%100	%100
Em**SA**	%33	%100	%5	Ne**ÖR**	%100	%100	%100
Ba**SA**	%0	%0	%0	Ab**SA**	%100	%50	%67
Ha**SA**	%50	%100	%67	Mu**GÖ**	%50	%100	%67
Mu**ÇE**	%100	%100	%100	Uğ**TA**	%100	%100	%100
Ha**İŞ**	%100	%100	%100	Ab**SA**	%100	%100	%100
Ha**DÜ**	%100	%100	%100	Ah**ER**	%8	%100	%15
Kü**UY**	%100	%100	%100	Al*NU*ŞE*	%100	%50	%67
Er**AĞ**	%100	%100	%100	Al*NU*ŞE*	%100	%100	%100
Mü*MU*SA*	%100	%100	%100	Mu**AH**	%100	%100	%100
Fe**SE**	%100	%100	%100	Ha**KO**	%17	%100	%29
Os**ÖZ**	%100	%100	%100	Ci**NA**	%0	%0	%0
Mu**AC**	%0	%0	%0	Bu**YA**	%0	%0	%0
Ah**YŞ**	%0	%0	%0	Mu**ŞA**	%0	%0	%0
Hu*KA*ÖR*	%0	%0	%0	Ta**SA**	%100	%100	%100
Sü**NE**	%100	%100	%100	Ke**CA**	%100	%100	%100
Şe**EK**	%100	%100	%100	Al**SE**	%100	%100	%100
Kü**ZÜ**	%0	%0	%0	Me**ŞA**	%100	%100	%100
Mu**KU**	%20	%100	%33	Ad**GÖ**	%100	%100	%100
Os**AC**	%0	%0	%0	Ah*Aİ*SE*	%0	%0	%0
Öz**ÇA**	%100	%100	%100	Al**YA**	%100	%100	%100
Rü**Bİ**	%100	%100	%100	Es**ER**	%100	%100	%100
				ORTALAMA	**%70,11**	**%74,00**	**%69,36**
				DOĞRULUK	**%72,01**		

MFKK – SMM yaklaşımında yapılan ilk teste ilişkin elde edilen ve Çizelge 4.2'de verilen F1 puanı ve Doğruluk oranının sırasıyla %69,36 ve %72,97 çıktığı görülmektedir. Yapılan 10 denemeye ilişkin doğruluk (accuracy) sonuçları Çizelge 4.3'de verilmiştir.

Çizelge 4.3. MFCC-HMM yaklaşımı doğruluk oranları

Deneme No	Doğruluk Oranı (%)
1	72,97
2	69,36
3	68,46
4	70,27
5	71,17
6	70.90
7	71.81
8	72.72
9	73.87
10	75.45
Ortalama	**72.95**

4.2. Derin öğrenme yaklaşımı ile alınan sonuçlar

MFKK-SMM yaklaşımı ile yapılan ilk uygulamada doğruluk oranı ortalama % 72.95 çıkmıştır. Son yıllarda daha çok rağbet edilen ve çok sayıda sinyal ve görüntü işleme uygulamasında kullanılan derin öğrenme metodu bu çalışma kapsamında da uygulanarak test edilmiştir. Materyal ve yöntem kısmında da belirtildiği üzere testler gürültülü/gürültüsüz ortam testleri olarak 2 farklı ortamda alınan seslere göre gerçekleştirilmiştir.

Test sonuçları verilirken cinsiyet ve yaş aralıkları dikkate alınmış olup test sayısı, doğru/yanlış algılama sayısı ve doğruluk oranı (%) çizelgelerde verilmiştir. Bu bölümün son kısmında verilen çizelgelerde ise sonuca ulaşılmış durumlara ilişkin süreler (sn) verilmiş ve tek kayıt – çift kayıt (aynı isimde olan kişilerin bölüm verileri de istenmesi durumu) durumuna göre analiz edilmiş ve yorumlanmıştır.

Çizelge 4.1'de bilgileri verilen 50 kişi gürültülü ve gürültüsüz ortam testi olmak üzere 2 kez sisteme giriş yapmış ve toplamda 100 adet deneme yapılmıştır. Bu denemelerde alınan başarı durumlarına ilişkin sonuçlar Çizelge 4.2'de verilmiştir.

Çizelge 4.4. Derin öğrenme ile elde edilen Doğru-Yanlış tanıma verileri.

CİNSİYET	Yaş Aralığı	Test Sayısı	Doğru Algılama	Yanlış Algılama	Doğruluk Oranı %
ERKEK	17-24	12	11	1	91.67
	25-34	42	32	10	76.19
	35-44	2	2	0	100
	45 ve üstü	0	0	0	-
TOPLAM		**56**	**45**	**11**	**89.29**
KADIN	17-24	28	25	3	89.29
	25-34	14	10	4	71.43
	35-44	0	0	0	-
	45 ve üstü	2	2	0	100
TOPLAM		**44**	**37**	**7**	**86.90**
GENEL TOPLAM		**100**	**82**	**18**	**88.10**

Çizelge 4.4'ye bakıldığında kadınların genel doğruluk oranının % 86.9 ve erkeklerin genel doğruluk oranının % 89.29 olduğu görülmektedir. Dolayısıyla ses tanıma açısından sistem erkeklerin sesini kadınlara göre daha başarılı tanımıştır. Yaş aralığına göre genel doğruluk oranını analiz edersek sistemin 35 yaş ve üzeri kullanıcıların seslerini daha başarılı (%100) tanıdığını, 25-34 yaş aralığının sesini tanımada ise daha az başarılı (%76.19 ve %71.43) olduğu gözlenmiştir.

Arka plan gürültü faktörünün ses tanımaya olan etkisinin araştırılması amacıyla gürültülü ve gürültüsüz durum test sonuçları incelenmiştir. Öncelikle Çizelge 4.5'de gürültüsüz ortam test sonuçları ve ardından Çizelge 4.6.'de gürültülü ortam test sonuçları verilmiştir.

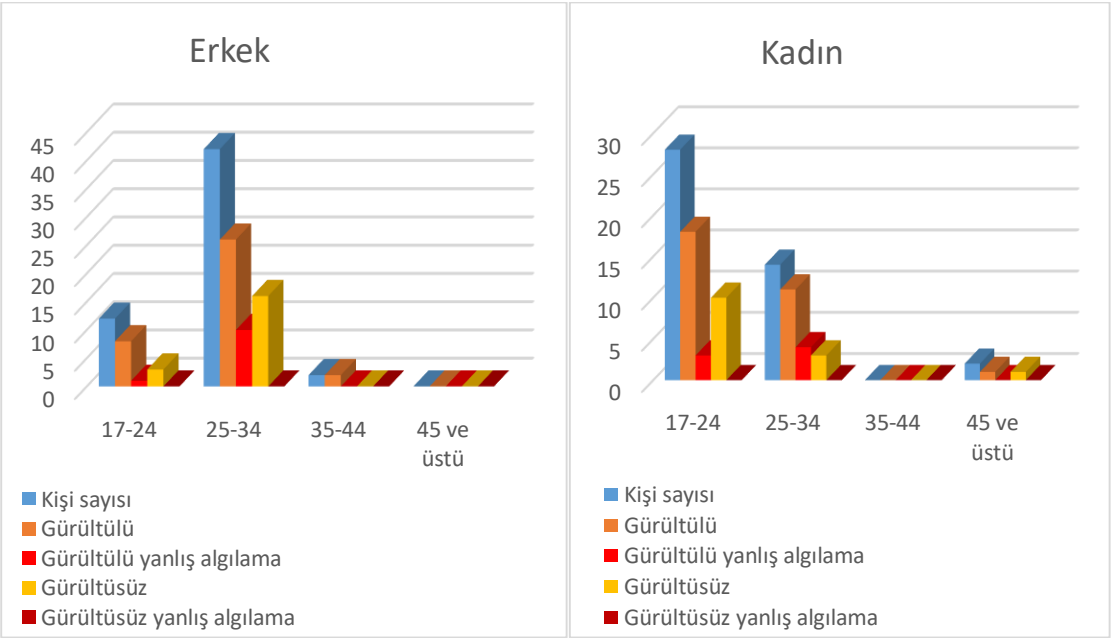
Çizelge 4.5. Gürültüsüz ortam test sonuçları.

	GÜRÜLTÜSÜZ ORTAM TESTLERİ				
CİNSİYET	Yaş Aralığı	Test Sayısı	Doğru Algılama	Yanlış Algılama	Doğruluk Oranı %
ERKEK	17-24	3	3	0	100
	25-34	16	16	0	100
	35-44	0	0	0	-
	45 ve üstü	0	0	0	-
TOPLAM		**19**	**19**	**0**	**100**
KADIN	17-24	10	10	0	100
	25-34	3	3	0	100
	35-44	0	0	0	-
	45 ve üstü	1	1	0	100
TOPLAM		**14**	**14**	**0**	**100**
GENEL TOPLAM		**33**	**33**	**0**	**100**

Çizelge 4.6. Gürültülü ortam test sonuçları.

	GÜRÜLTÜLÜ ORTAM TESTLERİ				
CİNSİYET	**Yaş Aralığı**	**Test Sayısı**	**Doğru Algılama**	**Yanlış Algılama**	**Doğruluk Oranı %**
ERKEK	17-24	8	7	1	87.50
	25-34	26	16	10	61.54
	35-44	2	2	0	100
	45 ve üstü	0	0	0	-
TOPLAM		**36**	**25**	**11**	**83.01**
KADIN	17-24	18	15	3	83.33
	25-34	11	7	4	63.64
	35-44	0	0	0	-
	45 ve üstü	1	1	0	100
TOPLAM		**30**	**23**	**7**	**82.32**
GENEL TOPLAM		**66**	**48**	**18**	**82.67**

Çizelge 4.5 incelendiğinde gürültüsüz ortam testlerin sistemin tam başarı (%100) oranı ile çalıştığını görüyoruz. Çizelge 4.6'e baktığımızda ise gürültü faktörünün başarıyı olumsuz etkilediği görülmektedir. Erkeklerin gürültülü ortamlarda yaptığı testlerde %83.01 başarı oranını, kadınların ise gürültülü ortam testlerinde %82.32 başarı oranını elde ettiği gözlenmiştir. Şekil 4.3'de Erkek ve Kadın katılımcıların gürültülü ve gürültüsüz ortamlara göre yaptıkları testlerin sonuçlarını veren grafik görülmektedir.



Şekil 4.3. Erkek ve Kadın karşılaştırmalı genel başarı durumları grafiği.

Çalışmada yapılan testlerin sürelerine ilişkin değerlendirme yapıldığında 2 farklı duruma göre değerlendirme yapılması daha doğru olacaktır. Öncelikle hata oluşan durumlara ilişkin süreler (Çizelge 4.7), ardından gürültüsüz ve hatasız sonuca ulaşılan durumlara ilişkin süreler (Çizelge 4.8) tablo halinde verilecektir.

Çizelge 4.7. Gürültülü ortam + Hatalı sonuç süreleri.

Ad Soyad	Bölüm	Süre (sn)	Hata	Gürültü
Ah**BO**	Makine	78,1	**3 tekrar (bölüm adı)**	var
Ah**ER**	Mekatronik	44,6	**2 tekrar (kişi adı)**	var
Ak**ŞA**	Malzeme	48,1	**2 tekrar (evet/hayır)**	var
Ci**NA**	Bilgisayar	69,8	**2 tekrar (kişi adı)**	var
Es**ER**	Kimya	63,5	**3 tekrar (kişi adı)**	var
Fa**BA**	Bilgisayar	69,2	**3 tekrar (kişi adı)**	var
Ha**DÜ**	Biyomedikal	43,9	**3 tekrar (kişi adı)**	var
Ha**IŞ**	Biyomedikal	83,1	**4 tekrar (kişi adı)**	var
Kü**ZÜ**	Hukuk	84,1	**3 tekrar (kişi adı)**	var
Ne**TA**	Bilgisayar	47,8	**2 tekrar (kişi adı)**	var
Os**AC**	Hukuk	41,1	**2 tekrar (bölüm adı)**	var
Rü**Bİ**	Hukuk	64,6	**3 tekrar (kişi adı)**	var
Al**ÖZ**	Bilgisayar	46,7	**1 tekr (kişi adı) - 2 tekr (evet/hayır)**	var
	Ortalama Süre:	60,35		

Çizelge 4.8. Gürültüsüz ortam + Hatasız sonuç süreleri.

Ad Soyad	Bölüm	Süre (sn)	Hata	Gürültü
Al**YA**	Kimya	41,5	0	yok
Em**AY**	Makine	29,6	0	yok
Ha**AR**	Elektrik-Elektonik	27,8	0	yok
Ha**KO**	Uçak	32,7	0	yok
Ha**SA**	Biyomedikal	30,9	0	yok
Ha**SE**	Elektrik-Elektonik	29,7	0	yok
Ha**SÜ**	Elektrik-Elektonik	31,3	0	yok
İb**DE**	Elektrik-Elektonik	29,7	0	yok
Ke**CA**	Uçak	46,2	0	yok
Ke**TÜ**	Malzeme	29,4	0	yok
Kü**UY**	Biyomedikal	37,6	0	yok
Mu**ÇE**	Biyomedikal	30,9	0	yok
Mu**KU**	Hukuk	30,1	0	yok
Öz**ÇA**	Hukuk	32,5	0	yok
Sü**NE**	Hukuk	36,7	0	yok
Şe**EK**	Hukuk	30,9	0	yok
Ta**SA**	Uçak	32,8	0	yok
	Ortalama Süre:	32,95		

Çizelge 4.7'te gürültülü ortam ve hatalı sonuç elde edilmesi durumlarına ilişkin süreler verilmiştir. Süreler incelendiğinde hata durumlarına göre işlem süresinin oldukça uzadığı ve ortalama olarak 60,35 sn sürede işlemin tamamlandığı görülmektedir. Çizelge 4.8'da ise gürültüsüz ortam ve hatasız duruma ilişkin süreler verilmiştir. İşlem sırasında hata olmamasından dolayı işlem süresi oldukça kısalmış ve 32,95 sn gibi bir ortalama süre hesaplanmıştır.

İşlem sırasında oluşan hatalara bakıldığında "kişi adı" kısmında çok fazla hata oluştuğu görülmektedir. Bunu "bölüm adı" ve "evet/hayır" kısımları takip etmektedir. Gürültüsüz ortamda yapılan testlerde ses tanıma kısmında hiç hata oluşmamış ve kullanıcıdan tekrarlaması istenmemiştir.

5. SONUÇLAR VE ÖNERİLER

5.1 Sonuçlar

Bu çalışmada Türkçe konuşma tanıma kullanılarak kişilerden gelen telefon çağrılarını otomatik olarak yönlendiren interaktif bir sistem tasarlanmış ve performans değerlendirmesi yapılmıştır. Bu kapsamda öncelikle Mel-Frekans Kepstral Katsayıları (MFKK) – Saklı Markov Modeli (SMM) yaklaşımı ile ses tanıma gerçekleştirilmeye çalışılmıştır. MFKK zamana göre değişen sinyallerin frekans ve genlik değerlerine göre öznitelik çıkarımı gerçekleştiren tekniktir. Bu teknikte ses örneği ön sentezleme yapıldıktan sonra çerçeveleme ve pencereleme işlemine tabi tutularak HFD büyüklükleri elde edilmektedir. Ardından elde edilen HFD büyüklükleri Mel-filtreleme işlemi uygulanarak sırasıyla LOG (Laplacian of Gaussian) filtreleme ve Ayrık Sürekli Dönüşüm (Discrete Continuous Transform – DCT) işleminden geçirilmiş ve MFKK öznitelikleri elde edilmiştir. SMM ise öznitelikleri belirlenmiş veri kümelerini kullanarak sınıflandırma yapabilen bir yöntemdir. Bu yöntemde bir kod çözücüsü işlem görür. Bu kod çözücüsü özellik vektöründen gelen verileri akustik ve dil modellerine göre telaffuz sözlüğünden karşılaştırma yaparak hangi kelime olduğunu bulmaktadır. MFKK-SMM yaklaşımını kullanarak 100 isimden oluşan veri setine uygulama yapılmış ve % 72.95 başarı oranı elde edilmiştir. Bu oran incelendiğinde bir çağrı merkezi yönlendirmesi için yeterli bir oran değildir.

Tez çalışması kapsamında yapılan bir diğer uygulamada ise derin öğrenme tabanlı bir yaklaşım ile Python programlama dili kullanılarak gürültülü ve gürültüsüz ortamlarda gelen çağrılarda girilen konuşmalar analiz edilmiştir. Geliştirilen yazılımın test aşamasında üniversite santral sistemi baz alınarak 10 bölüm ve 100 kişiden oluşan bir rehber veritabanı oluşturulmuştur. Yaş, cinsiyet, bölüm, süre ve ortam (gürültülü/gürültüsüz) parametreleri açısından tezde geliştirilen interaktif Türkçe konuşma tanıma sistemi değerlendirilmiştir.

Otomatik sesli yanıt ve çağrı sistemlerinde karşılaşılabilecek durumlar göz önünde bulundurularak sistemin performans kriterleri belirlenmiş ve akış şeması buna göre şekillendirilmiştir. Örneğin aynı isimde 2 kişi olması ihtimaline karşı veritabanına kaydedilen 100 isim arasında 14 adet aynı isimde kayıt oluşturulmuştur. Seçim yapılabilmesi adına bu kişiler ayrı bölümlerde tanımlanarak bölümlerin seçimine yönelik ses tanıma seçenekleri eklenmiştir. Bununla birlikte telefonla konuşulması sırasında

ortam gürültülerinin de olacağı ön görülerek testler gürültülü ve gürültüsüz ortam testi olarak ayrılmıştır.

Ses tanıma sistemi test edilirken farklı ses ve tınıya sahip insanların sisteme giriş yapması sağlanmıştır. Bu amaçla telefon görüşmelerini yapmak amacıyla farklı yaşlarda ve cinsiyette 50 kişi belirlenmiş ve gürültüsüz/gürültülü olmak üzere 100 adet deneme yapılması sağlanmıştır.

Test sonuçları irdelendiğinde, kadınların genel doğruluk oranının % 86.9 ve erkeklerin genel doğruluk oranının % 89.29 olduğu görülmüştür. Dolayısıyla ses tanıma açısından sistem erkeklerin sesini kadınlara göre daha başarılı tanımıştır. Yaş aralığına göre genel doğruluk oranını analiz edersek sistemin 35 yaş ve üzeri kullanıcıların seslerini daha başarılı (%100) tanıdığını, 25-34 yaş aralığının sesini tanımada ise daha az başarılı (%76.19 ve %71.43) olduğu gözlenmiştir.

Arka plan gürültü faktörünün ses tanımaya olan etkisinin araştırılması amacıyla gürültülü ve gürültüsüz durum test sonuçları incelenmiştir. Buna göre gürültüsüz ortam testlerin sistemin tam başarı (%100) oranı ile çalıştığı görülmüştür. Gürültülü ortam testlerinde ise erkeklerin %83.01 başarı oranı, kadınların ise %82.32 başarı oranını elde ettiği gözlenmiştir.

Çalışmada yapılan testlerin sürelerine ilişkin değerlendirme 2 farklı duruma göre yapılmıştır. Bunlar gürültülü ortam + hatalı durum ve gürültüsüz ortam + hatasız durum olarak sıralanmıştır. Süreler incelendiğinde hata durumlarına göre işlem süresinin oldukça uzadığı ve ortalama olarak 60,35 sn sürede işlemin tamamlandığı görülmüştür. Hatasız duruma ilişkin sürelere bakıldığında ise işlem sırasında hata olmamasından dolayı işlem süresi oldukça kısalmış ve 32,95 sn gibi bir ortalama süre hesaplanmıştır.

İşlem sırasında oluşan hatalara bakıldığında “kişi adı” kısmında çok fazla hata oluştuğu görülmüştür. Daha sonra “bölüm adı” ve “evet/hayır” kısımlarında hata oluştuğu ve tekrarlanma istendiği gözlenmiştir.

5.2 Öneriler

Çalışma Türkçe konuşma tanımaya yönelik gerçekleştirilmiştir. Derin öğrenme yaklaşımı ile oldukça yüksek başarı oranı elde edilmiştir. Ancak farklı ses sentez yöntemlerinin de (MFCC, HMM vb) denenerek bu başarının artırılması bu alandaki literatüre katkı anlamında olumlu sonuçları olacaktır.

Bununla birlikte gürültülü ortamlarda oluşan hataları en aza indirmek amacıyla sesin sentezlenmesinden önce bir filtreleme yapılması gerekebilir. Ancak bu durum işlem sürelerini uzatacak ve telefon trafiğinde yığılmalara sebep olacaktır.

Sistem üniversite santraline yönelik tasarlanmış olup farklı kurumlara uygulanması mümkündür. Bu açıdan bu çalışmanın literatüre önemli katkı yaptığı aşikardır.

KAYNAKLAR

Afify, M. ve Siohan, O., 2004, Sequential estimation with optimal forgetting for robust speech recognition, *IEEE transactions on speech and Audio Processing*, 12 (1), 19-26.

Afify, M., Liu, F., Jiang, H. ve Siohan, O., 2005, A new verification-based fast-match for large vocabulary continuous speech recognition, *IEEE transactions on speech and Audio Processing*, 13 (4), 546-553.

Akın, c., 2019, voıce recognıtıon system wıth score level fusıon methods and embedded system desıgn, *ıstanbul technıcal unıversıty*.

Ashraf, J., Iqbal, N., Khattak, N. S. ve Zaidi, A. M., 2010, Speaker independent Urdu speech recognition using HMM, *2010 The 7th International Conference on Informatics and Systems (INFOS)*, 1-5.

Baum, E. B. ve Haussler, D., 1989, What size net gives valid generalization?, *Advances in neural information processing systems*, 81-90.

Bengio, Y., Simard, P. ve Frasconi, P., 1994, Learning long-term dependencies with gradient descent is difficult, *IEEE transactions on neural networks*, 5 (2), 157-166.

Cappé, O., Moulines, E. ve Rydén, T., 2006, Inference in hidden Markov models, Springer Science & Business Media, p.

Caudill, M., 1989, Neural networks primer, part VIII, *AI Expert*, 4 (8), 61-67.

Chakraborty, S. ve Sarddar, D., Optimized Routing Approach using McCulloch-Pitts Neuron model.

Davis, K., Biddulph, R. ve Balashek, S., 1952, Automatic recognition of spoken digits, *The Journal of the Acoustical Society of America*, 24 (6), 637-642.

derinogrenme.com, 2017, Yapay Sinir Ağları, *Web*.

Erell, A. ve Weintraub, M., 1993, Filterbank-energy estimation using mixture and Markov models for recognition of noisy speech, *IEEE transactions on speech and Audio Processing*, 1 (1), 68-76.

Ferguson, J., 1980, Hidden Markov Models for Speech, *IDA, princeton, NJ*.

Forgie, J. W. ve Forgie, C. D., 1959, Results obtained from a vowel recognition computer program, *The Journal of the Acoustical Society of America*, 31 (11), 1480-1489.

Fung, J. ve Mann, S., 2008, Using graphics devices in reverse: GPU-based image processing and computer vision, *2008 IEEE international conference on multimedia and expo*, 9-12.

Furui, S., Kikuchi, T., Shinnaka, Y. ve Hori, C., 2004, Speech-to-text and speech-to-speech summarization of spontaneous speech, *IEEE transactions on speech and Audio Processing*, 12 (4), 401-408.

Furui, S., 2005, 50 years of progress in speech and speaker recognition research, *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, 1 (2), 64-74.

Ghahramani, Z., 2001, An introduction to hidden Markov models and Bayesian networks, In: Hidden Markov models: applications in computer vision, Eds: World Scientific, p. 9-41.

Hagen, A., Connors, D. A. ve Pellom, B. L., 2003, The analysis and design of architecture systems for speech recognition on modern handheld-computing devices, *Proceedings of the 1st IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis*, 65-70.

Hinton, G. E., 1987, Learning translation invariant recognition in a massively parallel networks, *International Conference on Parallel Architectures and Languages Europe*, 1-13.

Hochreiter, S., 1991, Untersuchungen zu dynamischen neuronalen Netzen (Master's thesis), *Institut für Informatik, Technische Universität München*.

Hochreiter, S. ve Schmidhuber, J., 1997, Long short-term memory, *Neural computation*, 9 (8), 1735-1780.

Ishizuka, K. ve Nakatani, T., 2006, A feature extraction method using subband based periodicity and aperiodicity decomposition with noise robust frontend processing for automatic speech recognition, *Speech communication*, 48 (11), 1447-1457.

Jiang, P., Fu, H. ve Tao, H., 2019, Speech Emotion Recognition Using Deep Convolutional Neural Network and Simple Recurrent Unit, *Engineering Letters*, 27 (4).

Juang, B.-H. ve Furui, S., 2000, Automatic recognition and understanding of spoken language-a first step toward natural human-machine communication, *Proceedings of the IEEE*, 88 (8), 1142-1165.

Kalchbrenner, N. ve Blunsom, P., 2013, Recurrent continuous translation models, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700-1709.

Kim, D.-S., Lee, S.-Y. ve Kil, R. M., 1999, Auditory processing of speech signals for robust speech recognition in real-world noisy environments, *IEEE transactions on speech and Audio Processing*, 7 (1), 55-69.

Kim, N. S. ve Un, C. K., 1995, On estimating robust probability distribution in HMM-based speech recognition, *IEEE transactions on speech and Audio Processing*, 3 (4), 279-285.

Kımanuka, u., 2019, turkısh speech recognıtıon based on deep neural networks, *kocaelı unıversıty*.

Krogh, A. ve Hertz, J. A., 1992, A simple weight decay can improve generalization, *Advances in neural information processing systems*, 950-957.

Lison, P., 2014.

Lleida, E. ve Rose, R. C., 2000, Utterance verification in continuous speech recognition: Decoding and training procedures, *IEEE transactions on speech and Audio Processing*, 8 (2), 126-139.

Lowrre, B., 1990, The HARPY speech understanding system, Trends in Speech Recognition, W, *Lea, Ed., SpeechScience Pub*, 576-586.

McCulloch, W. S. ve Pitts, W., 1943, A logical calculus of the ideas immanent in nervous activity, *The bulletin of mathematical biophysics*, 5 (4), 115-133.

Mittal, A., 2019, Understanding RNN and LSTM.

Moore, R. K., 1994, Twenty things we still don't know about speech, *Proc. CRIM/FORWISS Workshop on Progress and Prospects of speech Research an Technology*.

Motlıcek, P., 2002, Feature extraction in speech coding and recognition, *Technical Report of PhD research internship in ASP Group, OGI-OHSU,< http …*.

Murphy, K. P., 2002, Dynamic bayesian networks, *Probabilistic Graphical Models, M. Jordan*, 7, 431.

Nagata, K., 1963, Spoken digit recognizer for Japanese language, *NEC research & development* (6), ????

Patel, I. ve Rao, Y. S., 2010, Speech recognition using HMM with MFCC-An analysis using frequency specral decomposion technique, *Signal & Image Processing: An International Journal (SIPIJ)*, 1 (2), 101-110.

Petridis, S., Wang, Y., Ma, P., Li, Z. ve Pantic, M., 2020, End-to-End Visual Speech Recognition for Small-Scale Datasets, *Pattern Recognition Letters*.

Poşul, a., 2011, telefon bankacılığında interaktif sesli yanıt sistemleri ve beceri bazlı yönlendirme algoritması, *beykent üniversitesi*.

Rabiner, L., Levinson, S., Rosenberg, A. ve Wilpon, J., 1979, Speaker-independent recognition of isolated words using clustering techniques, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27 (4), 336-349.

Rabiner, L. R., 1989, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, 77 (2), 257-286.

Ranjan, R. ve Dubey, R. K., 2016, Isolated word recognition using HMM for Maithili dialect, *2016 International Conference on Signal Processing and Communication (ICSC)*, 323-327.

Riccardi, G. ve Hakkani-Tur, D., 2005, Active learning: Theory and applications to automatic speech recognition, *IEEE transactions on speech and Audio Processing*, 13 (4), 504-511.

Rigoll, G., 1994, Maximum mutual information neural networks for hybrid connectionist-HMM speech recognition systems, *IEEE transactions on speech and Audio Processing*, 2 (1), 175-184.

Rumelhart, D. E., Hinton, G. E. ve Williams, R. J., 1985, Learning internal representations by error propagation, *California Univ San Diego La Jolla Inst for Cognitive Science*.

Sakai, T., 1962, The phonetic typewriter, *Proc. IFIP Congress 62*, 445-450.

Sankar, A. ve Lee, C.-H., 1996, A maximum-likelihood approach to stochastic matching for robust speech recognition, *IEEE transactions on speech and Audio Processing*, 4 (3), 190-202.

Schmidhuber, J., 2015, Deep learning in neural networks: An overview, *Neural networks*, 61, 85-117.

Schwartz, D. B., Samalam, V. K., Solla, S. A. ve Denker, J. S., 1990, Exhaustive learning, *Neural computation*, 2 (3), 374-385.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. ve Salakhutdinov, R., 2014, Dropout: a simple way to prevent neural networks from overfitting, *The journal of machine learning research*, 15 (1), 1929-1958.

Srıvastava, P., 2017, Essentials of Deep Learning : Introduction to Long Short Term Memory.

Sutskever, I., Vinyals, O. ve Le, Q., 2014, Sequence to sequence learning with neural networks, *Advances in NIPS*.

Tishby, N., Levin, E. ve Solla, S. A., 1989, Consistent inference of probabilities in layered networks: Predictions and generalization, *International Joint Conference on Neural Networks*, 403-409.

Wang, C. K., Doërr, G. ve Cox, I., 2007, Trellis coded modulation to improve dirty paper trellis watermarking, *Security, Steganography, and Watermarking of Multimedia Contents IX*, 65050G.

Xu, M., Duan, L.-Y., Cai, J., Chia, L.-T., Xu, C. ve Tian, Q., 2004, HMM-based audio keyword generation, *Pacific-Rim Conference on Multimedia*, 566-574.

Zhao, Y., Yue, J., Song, W., Xu, X., Li, X., Wu, L. ve Ji, Q., 2019, Tibetan Multi-Dialect Speech Recognition Using Latent Regression Bayesian Network and End-To-End Mode.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı	**:**	Mustafa jumaah ahmed AHMED
Uyruğu	**:**	IRAK
Doğum Yeri ve Tarihi	**:**	KERKÜK – 04.02.1989
Telefon	**:**	5316633033
Faks	**:**	
e-mail	**:**	mustafa.jumaahahmed@yahoo.com

EĞİTİM

Derece		**Adı, İlçe, İl**	**Bitirme Yılı**
Lise	:	**Almas lisesi, Kerkük, Merkez**	**2010**
Üniversite	:	**Al Qalam Üni. Mühendislik Fak.**	**2016**
Yüksek Lisans	:	**Selçuk Üniversitesi, Konya**	**2020**
Doktora	:		

İŞ DENEYİMLERİ

Yıl	**Kurum**	**Görevi**

UZMANLIK ALANI

Bilgisayar Mühendisliği

YABANCI DİLLER

Arapça ve İngilizce

BELİRTMEK İSTEĞİNİZ DİĞER ÖZELLİKLER

YAYINLAR

Dergiler

Hasan Erdinç KOÇER, **Mustafa Jumaah Ahmed AHMED**, Turkish Speech recognition using Mel-frequency cepstral coefficients(MFCC) and Hidden Markov Model (HMM), DergiPark , Vol 2, Issue 2 (2019), pp. 39-44.