OTOMATİK TÜRKÇE DUDAK OKUMA İÇİN BİLGİSAYARLI GÖRÜ VE DERİN ÖĞRENME MODELLERİNİN GELİŞTİRİLMESİ

Furkan SABAZ

2022
DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ

Tez Danışmanı
Doç. Dr. Ümit ATİLA

OTOMATİK TÜRKÇE DUDAK OKUMA İÇİN BİLGİSAYARLI GÖRÜ VE DERİN ÖĞRENME MODELLERİNİN GELİŞTİRİLMESİ

Furkan SABAZ

T.C.
Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalında
Doktora Tezi
Olarak Hazırlanmıştır

Tez Danışmanı
Doç. Dr. Ümit ATİLA

KARABÜK
Ekim 2022

Furkan SABAZ tarafından hazırlanan "OTOMATİK TÜRKÇE DUDAK OKUMA İÇİN BİLGİSAYARLI GÖRÜ VE DERİN ÖĞRENME MODELLERİNİN GELİŞTİRİLMESİ" başlıklı bu tezin Doktora Tezi olarak uygun olduğunu onaylarım.

Doç. Dr. Ümit ATİLA .....
Tez Danışmanı, Bilgisayar Mühendisliği Anabilim Dalı

Bu çalışma, jürimiz tarafından oy birliği ile Bilgisayar Mühendisliği Anabilim Dalında Doktora tezi olarak kabul edilmiştir. 14/10/2022

Ünvanı, Adı SOYADI (Kurumu)	İmzası
Başkan : Prof. Dr. Muhammet Ali AKÇAYOL (GÜ)	.....
Üye : Prof. Dr. Oğuz FINDIK (KBÜ)	.....
Üye : Doç. Dr. Ümit ATİLA (GÜ)	.....
Üye : Doç. Dr. Muhammed Kamil TURAN (KBÜ)	.....
Üye : Dr. Öğr. Üyesi Tuba ÇAĞLIKANTAR (GÜ)	.....

KBÜ Lisansüstü Eğitim Enstitüsü Yönetim Kurulu, bu tez ile, Doktora derecesini onamıştır.

Doç. Dr. Müslüm KUZU .....
Lisansüstü Eğitim Enstitüsü Müdürü

"Bu tezdeki tüm bilgilerin akademik kurallara ve etik ilkelere uygun olarak elde edildiğini ve sunulduğunu; ayrıca bu kuralların ve ilkelerin gerektirdiği şekilde, bu çalışmadan kaynaklanmayan bütün atıfları yaptığımı beyan ederim."

Furkan SABAZ

ÖZET

Doktora Tezi

OTOMATİK TÜRKÇE DUDAK OKUMA İÇİN BİLGİSAYARLI GÖRÜ VE DERİN ÖĞRENME MODELLERİNİN GELİŞTİRİLMESİ

Furkan SABAZ

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Doç. Dr. Ümit ATİLA
Ekim 2022, 243 sayfa

Son yıllarda özellikle derin öğrenme uygulamalarının yaygınlaşmasıyla birlikte önemi oldukça artan çalışmalardan biri de dudak okuma olmuştur. Araştırmacılar bu alanda, ses verisinin olmadığı sadece görüntünün olduğu verilerde kişinin ne söylediğini algılamaya çalışmaktadırlar. Daha önce yapılan çalışmalar incelendiğinde Çince, Korece, İngilizce, Almanca gibi çeşitli dillerde veri setleri üzerinden otomatik dudak okuma sistemleri geliştirildiği görülmektedir. Fakat yine bu çalışmalar sadece görüntü üzerinden dudak okumanın ışık, çekim mesafesi, kişinin cinsiyeti gibi birçok parametreye bağlı olmasından dolayı sistemin geliştirilmesinin zorlu olduğunu ortaya koymaktadır. Dudak okuma sistemleri ilk olarak klasik makine öğrenmesi yöntemleri kullanılarak geliştirilmiştir. Fakat özellikle son yıllarda derin öğrenme uygulamalarının gündeme gelmesiyle beraber bu konu tekrardan sıkça çalışılmaya başlanmıştır. Yapılan çalışmalarda, derin öğrenme modellerinin klasik makine öğrenmesi yöntemlerine göre çok daha başarılı sonuçlar verdiği gözlemlenmiştir.

Bu çalışmamızda Türkçe hazırlanmış dudak okuma veri seti üzerinde derin öğrenme modellerinin uygulanarak sonuçlarının kıyaslanması amaçlanmaktadır. Farklı dillerde bu alanda yapılmış çalışmalar olsa bile Türkçede bu alanda yapılmış güncel bir çalışma ve veri seti bulunmamaktadır. Literatürdeki farklı dillere ait veri setlerindeki kriterler göz önüne alınarak 111 kelimelik ve 113 cümlelik güncel görüntü teknolojileriyle oluşturulmuş bir veri seti oluşturulmuştur. Oluşturulan veri seti kullanılarak, "Türkçe otomatik dudak okuma sisteminin" geliştirilmesiyle beraber literatürdeki bu uygulama eksikliği de giderilmektedir. BiLSTM sınıflandırıcısı ve çeşitli CNN tabanlı modeller kullanılarak verilerin sınıflandırılması sağlanmaktadır.

Yaptığımız çalışmada kelime ve cümle veri setlerinin her ikisinde de Resnet-18-BiLSTM ikilisi en iyi sonucu vermektedir. Kelime veri seti için %84,5 ve cümle veri seti için %88,55 doğruluk değeri elde edilmiştir. Çalışma sonuçları incelendiğince neredeyse her modelde cümle tanımada, kelime tanımaya göre daha başarılı sonuçlar elde edildiği görülmektedir.

Anahtar Sözcükler : Dudak okuma, bilstm, derin öğrenme, veri seti, türkçe, makine öğrenmesi, bilgisayarlı görü.

Bilim Kodu **:** 92431

ABSTRACT

Ph. D. Thesis

DEVELOPMENT OF COMPUTER VISION AND DEEP LEARNING MODELS FOR AUTOMATIC TURKISH LIP READING

Furkan SABAZ

Karabuk University
Institute of Graduate Programs
Department of Computer Engineering

Thesis Advisor:
Assoc. Prof. Ümit ATİLA
October 2022, 243 pages

In recent years, lip-reading has been one of the studies whose importance has increased considerably, especially with the spread of deep learning applications. In this topic, researchers try to detect what a person says from video frames without sound. When the previous studies are analyzed, it is seen that automatic lip-reading systems have been developed for various languages such as Chinese, Korean, English and German. However, these studies reveal that the development of the system is difficult because lip-reading from video frame images without audio data depends on many parameters such as light, shooting distance, and the gender of the person. Lip-reading systems were first developed using classical machine learning methods. However, especially in recent years, with the popularity of deep learning applications, this subject has started to be studied more than before and studies reveal that in general, deep learning-based lip-reading gives more successful results.

Even though there are studies in this field in different languages, there is no current study and dataset in Turkish. Therefore, this study aims to investigate the performances of the state-of-the art deep learning models on Turkish lip-reading. To this aim, two new datasets, one with 111 words and other with 113 sentences were created using image processing techniques. The model used in this study to perform lip-reading extracts features from video frames using CNN based models and performs classification using Bidirectional Long Short-Term Memory (Bi-LSTM). Results of experiments reveal that, ResNet-18 and Bi-LSTM pair gives the best results in both word and sentence datasets with accuracy values 84.5% and 88.55%, respectively. It is also observed that, better performances are obtained in sentence recognition than word recognition in almost every model implemented.

Key Word **:** Lip reading, bilstm, deep learning, dataset, turkish, machine learning, computer vision.

Science Code : 92431

TEŞEKKÜR

Bu tez çalışmasının planlanmasında, araştırılmasında, yürütülmesinde ve oluşumunda ilgi ve desteğini esirgemeyen, engin bilgi ve tecrübelerinden yararlandığım, yönlendirme ve bilgilendirmeleriyle çalışmamı bilimsel temeller ışığında şekillendiren sayın hocam Doç. Dr. Ümit ATİLA'ya ve aileme sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

Sayfa

Sayfa

Sayfa

ÇİZELGELER DİZİNİ

Sayfa

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

SİMGELER

M : Mel Ölçeği.

f : Frekans

ln : Logaritmik fonksiyon

k : Videodaki kare sayısı

n : CNN modelinde bir görüntü için oluşturulan vektörün boyutu

ii : İntegral görüntüsü

I : İntegral işlemi ve görüntüsü

W : Derin ağlarda ve klasik ağlarda ağırlıklar

θ_m : m. Sınıflandırıcının karar eşik değeri

$\alpha_{m,i}$: Öznitelik çıkarıcı için sabit değer

$\beta_{m,i}$: Öznitelik çıkarıcı için sabit değer

$f_{m,i}$: 2 boyutlu integrallerin ağırlık toplamı

$f1$: Filtre boyutu

S : Stride değeri

p : Doldurma değeri

$\sigma()$: Sigmoid fonksiyonu

w_j' : Seyreltme katmanında uygulanan ağırlık çıktısı

$P(c)$: Seyreltme katmanı için hesaplanan olasılık değeri

W_{aa} : Bir önceki gizli durumun/hücrenin ağırlıkları

W_{ax} : Girdinin ağırlıkları

W_{ya} : Anlık işlem yapılan hücrenin çıktı ağırlıkları

$a^{(t)}$: t anındaki hücrenin gizli durumu

$a^{(t-1)}$: Önceki gizli durum

S : Stride değeri

b_a, b_y : Bias değerleri

h_t : t anındaki RNN hücresi çıktı değeri

h_{t-1} : Bir önceki hücrede yer alan çıktı değeri

X_t : RNN hücresine gelen giriş değeri

h : Gizli katmanlar

W_{hh} : Bir önceki gizli katmanın sahip olduğu ağırlık değerleri

W_{xh} : Şu anki gizli katmanın sahip olduğu ağırlık değerleri

W_{hy} : Çıktı katmanının ağırlık değerleri

$tanh$: Aktivasyon fonksiyonu

Γ_f : Unutma kapısının çıkış değeridir

W_f : Unutma kapısında kullanılan ağırlık değerleri

b_f : Unutma kapısı bias değeridir

Γ_u : Güncelleme kapısının çıkış değeridir

W_u : Güncelleme kapısında kullanılan ağırlık değerleri

b_u : Güncelleme kapısı bias değeridir

$c^{(t)}$: t anı için hesaplanan hücre durum değeri

$c^{(t-1)}$: Odaklanan hücreden bir önceki hücrenin durum değeri.

KISALTMALAR

AAM	: Active Appearance Model
ACC	: Accuracy – Doğruluk
ALR	: Automated Lip Reading
ASR	: Automated Speech Recognition
AV-ASR	: Audio Visual – Automated Speech Recoginiton
BiLSTM	: Bidirectional Long Short-Term Memory
CFI	: Concatenated Frame Image
CNN	: Convolutional Neural Network
DCT	: Discrete Cosine Transform
FFN	: Feed Forward Network
FPS	: Frame Per Second
FV	: Feature Vector
h	: Hour – Saat
IAC	: Inter-Application Communication
LDA	: Linear Discriminant Analysis
M	: Milyon
MLLT	: Maximum Likelihood Linear Transform
PCA	: Principal Component Analysis
RBM	: Restricted Boltzmann Machine
RDA	: Redundancy Analysis
ROI	: Region Of Interest
s	: Second – Saniye
SRR	: Sentence Recognition Rate
STCNN	: Spatiotemporal Convolutional Neural Network
WRR	: Word Recognition Rate

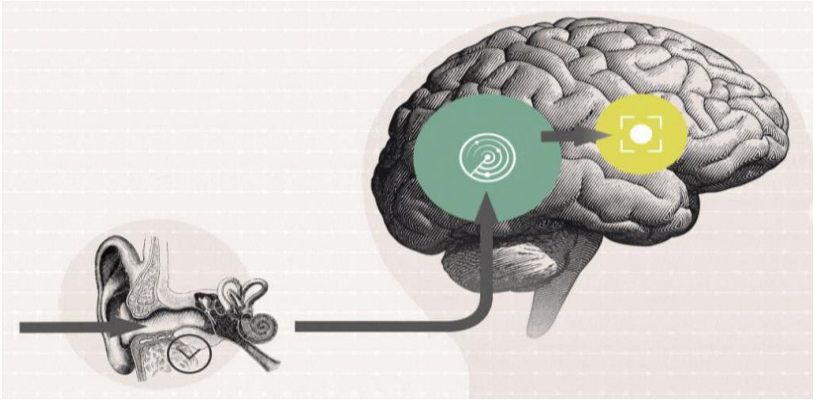
BÖLÜM 1

GİRİŞ

İnsanlar arasındaki iletişimin ve anlaşmanın en doğal yöntemi, konuşmadır. İnsanlar konuşurken oluşturduğu sesler üzerinden dil aracılığıyla anlaşmaya çalışırlar. Anlaşmada ve çevreyi algılamada ses kadar başka parametrelerde önemlidir. Örneğin görsel öğeler, koku veya tat bunlardan bazılarıdır. Bu durum sadece insanlar için geçerli değildir. Doğadaki pek çok canlı ses dışında farklı koşulları da değerlendirerek iletişime geçer veya çevresini algılamaya çalışır [65,66].

Beynin içinde her duyunun algılanması için farklı bölgeler bulunmaktadır. İşitme bölgesi, insan kulağından gelen ses sinyallerinin bilişsel olarak algıladığımız sese dönüştürülmesini sağlayan alandır. İşitme bölgesi; birincil işitme bölgesi, ikincil işitme bölgesi ve daha üst seviye işitme bölgeleri olarak birkaç farklı bölümden meydana gelmektedir. Birincil işitme bölgesi, kulaklardan iletilen sinyalleri değerlendirerek sesin tınısı, frekansı ve ses kaynağının uzaklığı gibi sinyale ait birçok parametrenin çözümlenmesini sağlamaktadır. Diğer bölgeler ise sesin sözel mesajları, duygusal içeriği, tını farklılığı gibi daha üst seviye mesajları çözümler. Tüm bu işitme alanları, temporal loblarının hemen üst kısmında yer almaktadır ve hem sol hem de sağ tarafta bulunur. Gelişen beyin görüntüleme deneyleri ve teknikleri, sol ve sağ beyindeki bölgelerin farklı görevler üstlenip buna yönelik işlemleri gerçekleştirdiğini tespit etmiştir. Örneğin müzikle pek de arası olmayan veya gelişmiş bir "müzik kulağı" olmayan bir insan müzik dinlediğinde beynin sağ tarafında bulunan işitme alanları çok daha fazla tetiklenip aktifleşmektedir. Bu insanlar normal bir iletişim sırasında konuşulan kelimeleri dinlerken de sol taraftaki beynin işitme bölgesinin daha fazla çalışıp aktifleştiği tespit edilmiştir. Bunun yanında, müzisyen veya müzik kulağı olan biri müzik dinlendiğinde ise beynin sol bölgesindeki işitme alanlarının daha fazla çalıştığı görülmüştür. Bu da beynin sol tarafındaki işitme bölgesinin ses analizinden, sağ taraftaki bölgeler de duygu ve

örüntüleri bilişsel olarak algılamaktan sorumlu olduğu teziyle ilişkilendirilmektedir [73-75].



Şekil 1.1. İşitme temel yapısı.

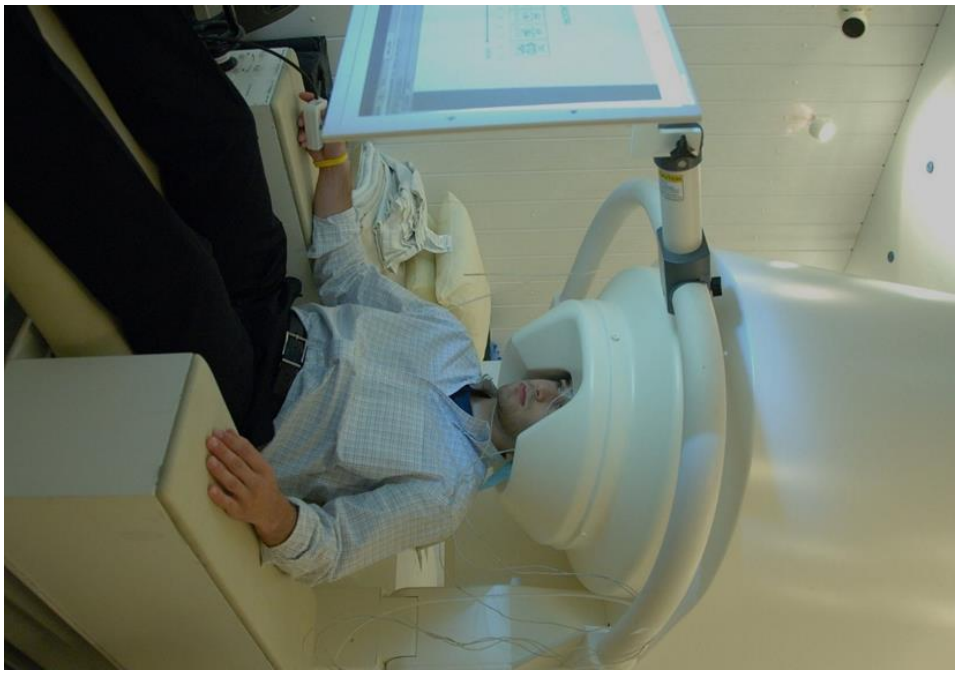
Dudak okuma davranışı da insanların karşıdaki kişilerin dudak hareketlerini analiz ederek söylenen kelimelerin ve cümlelerin ne olduğunu tahmin etmesidir. Ancak beynin sessiz (sesin olmadığı), görsel bir iletişimden nasıl anlam çıkardığı hâlâ tam olarak belirlenememiş olup tartışılmaktadır. Sessizlik içinde dudak okuma, beynin işitme bölgelerini harekete geçirir, ancak bu tür bir aktivasyonun normalde ona karşılık gelen işitsel uyaranın anında sentezini mi yoksa alakasız seslerin görüntülerini mi tamamladığı veya oluşturduğu bilinmemektedir. Bu da araştırmacıların beynin ses olmadan algıladığı verileri, tam olarak nasıl oluşturduğunu belirleyememesine sebep olmaktadır. Ayrıca bu işlemdeki doğruluk yüzdesinin ne olduğuna dair yapılmış kesin bilgiler veren bir çalışma da bulunmamaktadır.

Dudak okumaya dair ilk çalışmalar bilgisayar veya yazılım alanında değil de psikoloji ve tıp gibi alanlarda gerçekleştirilmiştir. Bunlardan ilki sayılabilecek McGurk ve McDonald görmenin konuşma algısındaki etkisini göstermek için yaptıkları deneysel çalışmada gözlemcilere birbirlerine uymayan işitsel ve görsel veriler sunmuşlardır. Deney sonucunda kişilerin kendilerine sunulanlardan farklı sesler algıladıkları görülmüştür. Çalışma sonucunda kişilerin bir şekilde görsel ve işitsel verileri ilişkilendirmeye çalıştığı değerlendirmesine varılmıştır. Bu çalışma dudak okumadan ziyade insanların iletişiminde, çevresini algılamada ve beynin bunu değerlendirmesinde tek başına sadece işitsel imgelerin veya sadece görsel imgelerin önemli olmadığını, beynin bunların tümünü değerlendirmeye aldığı sonucuna

varılmıştır [1]. Böyle bir çalışmadan sonra pek çok araştırma, insanlar farkında olmasa bile konuşma tanımada ve anlamlandırmada görsel bilgi kullanımının doğruluğu arttırdığını göstermiştir [2,3].

Skipper yaptığı tıbbi çalışmalarında konuşmacının ağzının izlenmesinin, dinlenilenler ve konuşma algısı üzerinde ciddi etkiler oluşturduğunu belirtmiştir [4]. Ses sinyalleri genel olarak video sinyallerinden çok daha bilgilendirici olmasına rağmen, çoğu insanın konuşmaları anlamak için dudak okumadan gelen ipuçlarını kullandığı kanıtlanmıştır. Yüz yüze iletişimde, özellikle zorlu dinleme koşullarında ve algının - sadece konuşma veya duyma değil- aşırı derecede zorlaştığı işitme engelli yaşlı popülasyonlarda konuşmayı anlamada konuşmacının artikülatör hareketlerinden (dudak hareketleri) rutin olarak görsel konuşma ipuçları çıkarmak o insanlar adına oldukça ciddi fayda sağlamaktadır [67,68]. Dudak okumanın insan beyni tarafından otomatik olarak yapıldığını gösteren çok kapsamlı bir başka çalışmada 17 kadın ve 11 erkek olmak üzere toplamda 28 kişiye önce hem ses hem de görsel verinin olduğu bir video izletilmiştir. İkinci aşamada görsel herhangi bir girdi olmadan sadece sözlü bir hikâye dinletilmiştir. Ardından da aynı konuşmacının ses verisi olmadan sadece videoyla anlattığı başka bir hikâye deneye katılanlara izletilmiştir. Böylece işitmeyle ilgili kortikal aktivitenin işitsel konuşmaya ve dudak hareketlerine nasıl dahil olduğu değerlendirilebilir. Değerlendirmeyi yapabilmek için Şekil 1.2'de bir örneği verilen manyetoensefalografi cihazı kullanmıştır. Bu cihaz sahip olduğu ekran vasıtasıyla izlettiği veya dinlettiği verilerin insan beyninde ne tür aktiviteler oluşturduğunu tespit eder ve MEG adı verilen farklı bir tür sinyal elde edilmesini sağlar. Çalışmada çok daha detaylı veriler tespit edilmiş olsa da dudak okumayı ilgilendiren kısmında, beynin erken işitsel korteksinde dudak okuma sinyallerinin de kullanılabildiği görülmüştür. Buna kanıt olarak da deneye katılan insanların dudak hareketlerini görmeleri kortekste nöronal aktiviteler oluşturup, beynin yan alt lobunda dudak hareketlerinin özelliklerini çıkararak bunlara karşılık gelen konuşma sesi özellikleriyle eşleştirir. Bu bilgi işitsel kortekslere iletilir ve konuşmanın ayrıştırılması, anlaşılması çok daha kolay hale gelmektedir [69]. Görsel ipuçları bulunduğu koşullara göre farklı oranlarda kullanılabilir. Örneğin gürültülü ortamlarda görsel kanal daha önemli hale gelebilir. Bunun yanında bazen kişilerde oluşan rahatsızlıklardan dolayı konuşma veya işitme işlevi tam olarak yerine getirilemeyebilir. Bu kişiler günlük hayatta başka insanlarla

olan iletişimlerindeki kaliteyi, yine dudak okuma yaparak arttırmaya çalışmaktadırlar. Ayrıca görüntülenen bir videoda konuşulanların anlaşılması için ses verisinin de olması gerekmektedir [5–10]. Eğer bu veri yoksa konuşulanlar anlaşılmamaktadır. Özellikle adli olaylarda böyle durumlarda konuşulanların tespiti için dudak okuma uzmanları görevlendirilir. Uzmanlar görüntülerde kişilerin dudak hareketlerini analiz eder ve söylenilenleri belirlemeye çalışarak gerekli mercilere tespitlerini raporlar [11].



Şekil 1.2. Manyetoensefalografi cihazı.

Özellikle nörolojik hastalıklardan dolayı kişiler konuşulanları anlamada problem çekmeye başlamaktadır. Afazi, insan beynindeki konuşma merkezlerindeki herhangi bir noktada meydana gelen bir hasar, darbe veya hastalık sonucunda konuşma, etrafta konuşulanları anlama, adlandırma, tekrar etme, yazma veya okuma gibi normalde temel sayılabilecek yetilerin kısmen ya da bütünüyle kaybıdır. Beyin kanamaları, beyin damar rahatsızlıkları, beynin bazı bölgelerindeki tümörler, enfeksiyon hastalıkları, kafaya alınan bir darbe travması gibi sebeplerle meydana gelebilmektedir. Afaziye sebep olacak durumlar moleküler nöro görüntüleme tekniklerinden MRI, FDG-PET, PiB-PET ile tespit edilebilmektedir. Başta bahsedilen problemlere ek olarak afaziye dikkat bozuklukları ve bellek sorunları (yakın veya uzak geçmişi hatırlamama) da eşlik edebilir. Afazi sonrası bireylerin normalde kolayca yapabildiği

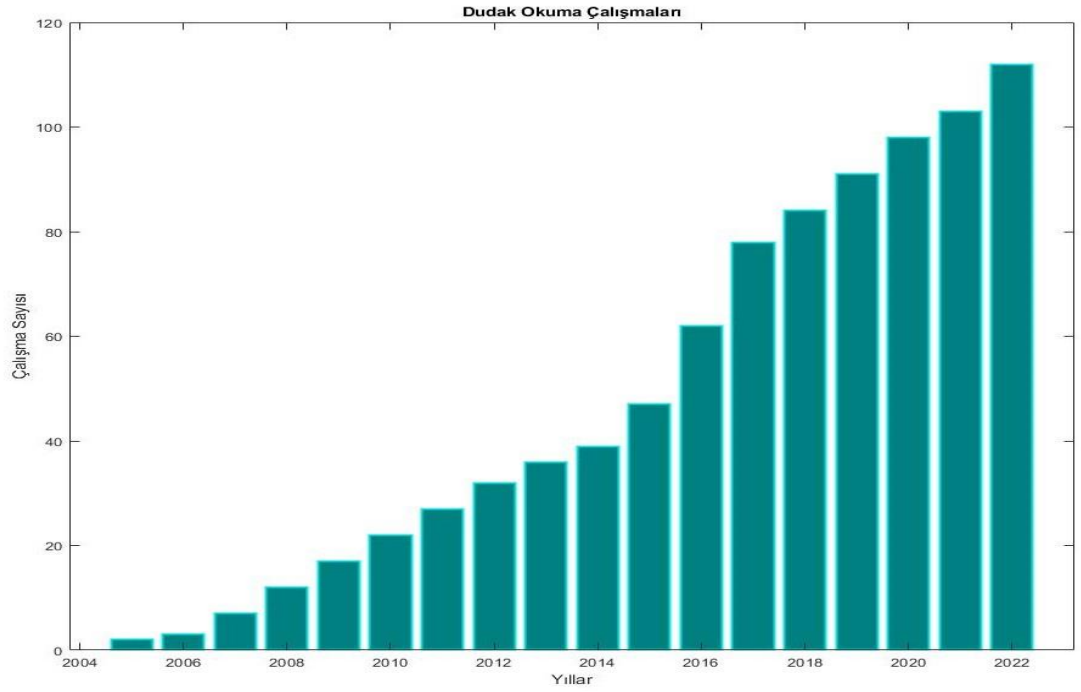
basit zihinsel hareketleri yerine getirme, plan yapma, karşılaştığı problemleri çözme ve seçenekler arasında karar verme yetenekleri de etkilenebilir. Afazisi olan bireyler, yapacakları bir işe nasıl ve nereden başlayacağını tespit etmekte, o işi yerine getirmek, problemi çözmek için gerekli adımları oluşturmakta ciddi sorunlar yaşayabilmektedir. Bu problem sadece gündelik hayat problemleri şeklinde değil de ses bazlı konuşmayı anlamada sorunlar oluşturur [69,75].

Geçirilen hastalık her insanda farklı derecelerde sorunlar meydana getirdiğinden afazide standart bir rehabilitasyon ve tedavi sürecinden bahsetmek mümkün değildir. Uygulanacak tedavinin süreci ve derecesi; hasarın meydana geldiği yere, bölgedeki dağılma durumuna, şiddetine, kişinin motivasyonu, cinsiyeti, yaşı, eğitim seviyesi, afazinin tipine, yaşamını sürdürdüğü toplumsal çevreye, sebep olan duruma, hasarın oluşması üzerinden geçen zamana ve alınan terapiye göre bireysel değişkenlikler gösterir. Afazi hastalarında herhangi bir tedavi ve rehabilitasyon desteği almaksızın çeşitli derecelerde kendiliğinden iyileşme süreci gerçekleşebilmektedir. Fakat bu iyileşme ve düzelme çoğu zamanda oldukça sınırlı olmaktadır. Terapi desteğinin konuşmanın gelişiminde çok daha yararlı olduğu bilinmektedir. Bireysel özelliklerine göre hazırlanan tedavi programıyla dil, konuşma ve iletişim becerileri desteklenir. Afazili hastalarda yapılan rehabilitasyonun amacı kişilerin kaybettikleri konuşma veya konuşmayı anlama gibi becerileri yeniden öğrenmelerine yardımcı olmak ve kişilere kalıcı yetersizliklerinin üstesinden gelmelerini sağlayacak dudak okuma gibi yeni beceriler öğretmektir [69,70,75].

Tedavi sürecinin önemli basamaklarından biri, biraz da insan beyninin otomatik geliştirdiği dudak okuma becerisidir. Örneğin Shindo, yaptığı çalışmasında sağlıklı bireylerin, kelime sağırlığı (kelimeleri anlayamama, tekrarlayamama), işitsel agnozi ve afazi hastalarının olduğu bir deney grubunda hastaların, dudaklarını okuyarak konuşmayı ayırt etme becerilerini değerlendirmeye çalışmıştır. Hastaların konuşmayı anlama ve kelimeleri ayırt etme becerilerinin testi için anlamsız tek heceli kelimeler, dudak okuma testi, işitsel anlama için token testi (işitsel bir test türü), afazi testi kullanılmıştır. Elde ettiği sonuçlar, kelime sağırlığı veya afazisi olmayan işitsel agnozisi olan hastaların dudak okuma veya tek başına dinleme ile karşılaştırıldığında dinleme ile dudak okuyarak konuşmayı anlamayı iyileştirebileceğini göstermektedir.

Sağlıklı bireylere göre de bu genel hasta grubunun dudak okuma becerisinin daha yüksek olduğu görülmüştür. Sonuç olarak, dudak okumanın bu hastalarda konuşmayı anlamada faydalı olduğu tespit edilmiştir. Bunun yanında insanlar yaşlandıkça tıpta Presbiakuzi denilen "Yaşa Bağlı İşitme Azalması" durumuyla karşı karşıya kalmaktadır [71,72].

Son yıllarda parmak izi veya yüz tanıma gibi biyometrik bilgilerle birlikte dudak okumayla ilgili güvenliğe yönelik çalışmalar öne çıkmaktadır. Gürültülü ortamlarda işitsel verinin yanında görsel verileri de kullanarak akıllı telefonlara mesaj yazdırılması, görsel sessiz şifreler kullanarak çeşitli güvenlik önlemlerinin alınması [12-14], sessiz videolardan sadece görsel veriyi kullanarak ses oluşturma veya konuşma engelli insanlar için dudak hareketlerine göre ses sentezleme, işitme bozukluğu olan kişiler için dudak hareketlerinin takibini kolaylaştırma gibi çalışmalar bunlardan bazılarıdır [15-17]. Şekil 1.3'te ise bahsedilen bu çalışmaları da kapsayan dudak okuma çalışmalarının yıllara göre sayıları verilmiştir.



Şekil 1.3. Yıllara göre yapılan çalışma sayıları (2005-2022).

Dudak okumanın temel amacı, ses verisini kullanmaya gerek kalmadan sadece görüntü kullanılarak konuşulan ifadeleri tespit etmektir. Bu kapsamdaki uygulamalar harf, sayı, hece, kelime ve cümle tanıma gibi özel alanlara ayrılmıştır [18–20]. Literatürde

çeşitli dillerde belirlenen amaca yönelik veri kümeleri üzerinde çalışmalar yapılmaktadır. Verileri yüksek doğrulukla sıralı ilişkili ifadelerle sınıflandırmak veya tespit etmek, birçok farklı alanda önemli bir sorundur. Dudak okuma çalışmalarında karşılaşılan bir diğer temel zorluk, p, b ve m gibi farklı harflerin ağızdan çok benzer dudak hareketleriyle çıkmasıdır [21, 22]. Literatürde özellikle heceleri ayırt etmek için çeşitli yöntemler önerilmiştir [23-28]. Öte yandan, günlük hayatta bir kelimeyi veya cümleyi söylemek çoğu zaman saniyeler sürer. Aynı harf, hece veya kelimelerin aynı konuşmacı tarafından bile farklı dudak yapıları/şekli ile ifade edilebilmesi de bir diğer zor problemdir. Ayrıca, verilerden kaynaklanan üstesinden gelinmesi gereken başka sorunlar da vardır. Bu sorunlar kişinin cinsiyeti, dudak yapısı, çekim ortamı ve çekim açısı ile ilgilidir. Kişinin dudak açısının bozulması ve erkeklerde sakal veya bıyık bulunması da tanımlama sürecini zorlaştıran diğer faktörlerdir [7, 29]. Bu nedenle dudak okumada elde edilen her görüntü amaca uygun olarak yorumlanmalıdır.

Hazırlanan bu çalışma, 8 ana bölümden oluşmaktadır. Birinci bölümde dudak okuma işlemine, ilk olarak tıbbi şekilde yaklaşılmaktadır. Neden dudak okuma sistemlerinin geliştirildiğine ve güncel hayatta hangi problemlerin çözüldüğü verilmektedir. Daha sonra dudak okuma sistemlerinin literatürdeki durumu ve çalışma sayıları hakkında bir değerlendirme yapılmaktadır. Birinci bölüm olan giriş bölümünde tezin geri kalanı hakkında yol haritası belirlenmektedir.

İkinci bölümde ilk olarak dudak okuma işleminin temel amaçlarından söz edilmektedir. Daha sonra da literatür taraması sonucu dudak okumaya dair daha önce farklı diller için oluşturulmuş veri setleri değerlendirilmektedir. Ayrıca veri setlerini sınıflandırarak oluşturulan veri setinin hangi özelliklere sahip olması gerektiği de belirlenmektedir. Literatürdeki veri setlerinin eksiklikleri de belirlenerek oluşturulan veri setine dair nelerin düzenlemesi gerektiğiyle ilgili veriler yer alır.

Üçüncü bölümdeyse ikinci bölümde bahsedilen literatürdeki veri setleri üzerine geliştirilmiş modeller derlenmektedir. Sistemler bu bölümde iki sınıfa ayrılmaktadır. Geliştirilen sistemlerde hangi modeller ne aşamada kullanılmışsa bunların ayrıntıları verilmektedir. Hangi özellik çıkarımı yöntemlerinin ve sınıflandırıcılarının

kullanıldığına dair modeller detaylı bir şekilde açıklanmaktadır. Sistemler de kendi içinde veri setinin amacına göre ayrıca sınıflandırılmaktadır.

Dördüncü bölüm temelde iki ana alt bölümden oluşmaktadır. Birinci alt bölümde literatürdeki ilk Türkçe dudak okuma sisteminin detayları verilmektedir. Oluşturulan veri setini, ikinci bölümde listelenen veri setleriyle kıyaslayıp onlara göre avantaj ve dezavantajları listelenmektedir. İkinci alt bölümde de literatürdeki ilk Türkçe otomatik dudak okuma sistemine yönelik öneriler sunulmaktadır. Ayrıca ek olarak gerçek zamanlı bir dudak okuma sistemine dair model önerilmektedir.

Beşinci bölümde tez kapsamında geliştirilen veri seti ve modelin testleri yapılmaktadır. Test aşaması toplamda 6 bölümden oluşmaktadır. Her test bölümünde o testin hangi amaçla yapıldığı, test için kullanılan veri seti ve sonuçlar detaylı bir şekilde tablo halinde verilmektedir. Test bölümü, veri setinin ve modelin performansı hakkında birçok veriyle desteklenen raporlama yer almaktadır.

Altıncı bölümde, literatürdeki çalışmalar ve geliştirilen model ile literatürdeki veri setleri ve oluşturulan veri seti hakkında bütün karşılaştırmalar, analizler yorumlanmaktadır.

Yedinci bölümde ilerde yapılabilecek çalışmalar hakkında bazı önerilere yer verilmektedir.

Sekizinci ve son bölümdeyse tezin geneliyle ilgi sonuçlandırma yapılmıştır.

BÖLÜM 2

DUDAK OKUMA İŞLEMİ

Her alanda gelişen yazılımsal ve donanımsal teknolojilerle birlikte birçok problem çözülmektedir. Hatta daha önce düşünülmemiş problemler bile gelişen teknoloji sayesinde gündeme gelip bunlar için çözümler aranmaktadır. Bu problemlerden biri de dudak okumadır. Giriş bölümünde bahsedilen ihtiyaçlara yönelik çalışmalar 2000'li yılların başında hız kazanmıştır. Dudak okuma problemi farklı açılardan farklı dallara ayrılabilmektedir. Örneğin veri setinin oluşturulması veya dudak okumanın hangi amaçla yapılacağı bunlardan sadece birkaçıdır. İlerleyen kısımlarda probleme, tüm bu açılardan ayrıntılarıyla değinilecektir.

Literatürde ALR, ASR ve AV-ASR kelimeleri oldukça sık geçmektedir.

- Automated Lip Reading-Otomatik Dudak Okuma- (ALR): Sadece görseli kullanarak bir videoda konuşan kişinin ne dediğini anlamak üzerine kurulan sistemlerdir.
- Automated Speech Recognition-Otomatik Konuşma Tanıma-(ASR): Ses verisini kullanarak konuşan kişinin ne dediğini anlamaya çalışan sistemlerdir. Literatürde ayrıca "speech to text" sistemleri olarak da geçmektedir.
- Audio Visual-Automated Speech Recogniton-Görsel İşitsel Konuşma Tanıma- (AV-ASR): Konuşan kişinin ne dediğini anlamaya çalışırken hem görsel öğeleri hem de işitsel öğeleri kullanan sistemlerdir.

2.1. DUDAK OKUMA VERİ SETLERİ

Dudak okumayla ilgili veri setlerine ait literatür incelendiğinde çok çeşitli amaçlarla yapılmış veri setleri olduğu görülmektedir. Fakat tez kapsamında bunlardan en

yaygın kullanılanlara değinilmektedir. Tez kapsamında bu şekilde sınıflandırılmasa da veri setlerine genel amaçlarla bakıldığında 3'e ayrılmaktadır.

- Sadece görüntünün yer aldığı veri setleri (Visual Database)
- Sadece sesin yer aldığı veri setleri (Audio Database – dudak okuma amaçlı kullanılamaz. Dudak okumayla ilgili farklı amaçlar için oluşturulmuştur.)
- Hem görüntünün hem de sesin yer aldığı veri setleri (Audio-Visual Database)

Bunun yanında veri setleri içeriğine göre de sınıflandırılabilmektedir. Buna yönelik çalışmalar 90'ların sonunda başlamıştır. Tasarlanan ilk veri setleri harf, alfabe veya rakam tanıma gibi sınırlı kelime-harf havuzuna sahip özel ve basit tanıma görevlerine odaklanmıştır. Bu veri setleri, daha önceden tanımlanmış oldukça sınırlı sayıda bir kelime havuzuyla ve çok yüksek tekrarlar ile optimuma yakın ortamlardan (kısa mesafede dudakların yakından çekilmesi gibi) elde edilmiştir. Veri setlerinin bu şekilde oluşturulması eğitim sürelerinin ciddi derece kısalmasını sağladığından dolayı geniş çapta analiz edilmesine imkân sağlamıştır. Fakat bununla birlikte dudak okumaya dair oluşturulan ilk veri setleri düşük sayıda konuşmacı ve sınırlı miktarda veri barındırması sebebiyle daha gerçekçi uygulamaların yapılmasını zorlaştırmaktadır. Çünkü dudak okuma sistemlerinin geliştirilmesindeki en büyük zorluklardan biri de uygulamanın veya veri setinin genelleştirilebilmesi olarak görülmektedir. Eksikliklerin tespit edilmesiyle beraber sonraki veri setleri, kaydedilen veri miktarını arttırmaya ve daha karmaşık görevleri ele almaya odaklanarak, nispeten daha akıcı konuşmaları da çözmeye çalışan sistemlerin oluşturulmasını hedeflemiştir.

Büyük görsel, görsel ve işitsel veri setlerinin oluşturulmasında ele alınabilecek çeşitli faktörler bulunmaktadır. Bunlardan bazıları şunlardır;

- Konu
- Tekrar sayısı
- Işıklandırma
- Kafa duruşu
- Kelime seçimi
- Çözünürlük vb.

Oluşturulan ilk veri setlerinden sonra daha geniş çaplı veri setleri oluşturulma amacı ortaya çıkmıştır. Fakat yukarda bahsedilen tüm faktörleri ele alacak çapta geniş bir veri setini oluşturmak yerine bunlardan bir veya iki tanesini ele alarak orta büyüklükte görseller barındıran veri setlerini oluşturulmuştur. Alt bölümlerde veri setleri detaylı incelenmektedir fakat burada giriş anlamında bazı bilgiler vermek faydalı olacaktır. Örneğin dudak okumanın temelini oluşturan ve aynı zamanda bu konuyla ilgilenen araştırmacılar arasında en popüler veri setlerinden biri GRID (underGround Re-IDentification) veri setidir. GRID veri seti zamanla ufak değişiklikler göstermiştir fakat ilk çıktığı döneme göre çok fazla sayıda telaffuz içermekle beraber kelimeler birbirine anlamsal olarak oldukça benzer ve kısıtlı sayıda cümleler içermektedir. Çözünürlük, görüntüler arasında sabit değildir ve çözünürlükler teknolojinin durumu itibariyle düşüktür. RM-3000 ise tek bir erkek konuşmacı içermektedir fakat çok büyük bir kelime dağarcığına sahiptir. Şekil 2.1'de GRID veri setine Şekil 2.2'de RM-3000 veri setine ait görseller yer almaktadır.

Şekil 2.1. GRID veri setinden örnekler.

Şekil 2.2. RM-3000 veri setinden örnek görseller.

GRID ve RM-3000 gibi veri setleri yayınlandıktan sonra daha ileri düzey veri setleri hazırlanmaya başlanmıştır. Bu amaçla televizyon programlarından elde edilen kesitlerle veri setleri oluşturulmak istenmiştir. Böylece farklı branşlardan kelimeler ve cümleler çok rahat bir şekilde veri setinde yer alabilir. Buna yönelik ilk çalışmalardan biri OXFORD üniversitesindeki Oxford-BBC adını verdikleri bir çalışma grubundan gelmiştir. Çalışma grubu, çoğu BBC spikerlerinden oluşan birçok farklı programdan kesitler alarak bunları veri seti haline getirmiştir. Veri setine de "Lip Reading in the Wild (LRW)" adı verilmiştir. Şekil 2.3'te de bu veri setine ait görseller verilmiştir. Çalışmaya ait oldukça fazla detay verilmiştir. Örneğin çalışmada eğitim için 01.01.2010 ve 31.08.2015 arasındaki veriler kullanılmıştır. Doğrulama için 01.09.2015 ve 24.12.2015 arasındaki veriler kullanılmıştır. Son olarak test için se 01.01.2016 ve 30.09.2016 arasındaki veriler kullanılmıştır.

Şekil 2.3. LRW veri setinden örnek görseller.

Oxford'daki çalışma grubu çalışmalarına LRS-2 yani Lip Reading Sentence-2 veri setini çıkararak devam etmiştir. Bu veri setinde cümlelerin maksimum karakter uzunluğu dahi verilmiştir. Şart olarak 100 karakterden fazla herhangi bir cümlenin bulundurulmadığı belirtilmiştir. Ayrıca o dönemlerdeki en büyük veri setidir. Tekrarlarla birlikte 100.00'den fazla veri barındırmaktadır. LRW veri setinde olduğu gibi hangi tarihe ait verilerin ne amaçla kullanıldığı tablo şeklinde verilmiştir. LRS-2 'den sonra çalışma grubu MV-LRS adında başka bir veri seti daha oluşturmuştur. Paylaşılan LRW, LRS-2 ve MV-LRS veri setlerinden sonra bu alanda yapılan çalışmalar daha fazla parametreye bağlı otomatik dudak okuma sistemlerinin geliştirilmesine ve daha da büyük veri setlerinin oluşturulmasına öncülük etmiştir.

Bir sonraki bölümde otomatik dudak okuma sistemlerinin geliştirilmesi konusunda en sık kullanılan veri setleri detaylı bir şekilde incelenecektir. Veri setleri, oluşturulma görevine (örneğin harfler, rakamlar, kelimeler ve cümleler) ve görüş açısına göre sınıflandırılarak karşılaştırılmaktadır. Görsel-işitsel (audio-visual) veri setlerinde genelde kişilerin karşıdan çekilmiş kayıtları hakimdir. Fakat ilerleyen dönemlerde otomatik dudak okuma sistemlerinin daha gerçekçi senaryolar üzerinden konuşmaların çözülmesi gerektiği düşünüldüğünden çoklu görünümlü dudak okuma veri setleri üzerinde çalışılmıştır ve çalışmalar günümüze kadar devam etmektedir. Bu durumda veri setleri çekim bakış açısı yönünden ikiye ayrılmaktadır;

- Karşıdan çekilmiş görüntüler (frontal view)
- Çok yönlü çekilmiş görüntüler (multiple view)

Bu sınıflandırılmaya ait veri setleri Bölüm 2.1.1, 2.1.2 ve 2.1.3'te tablo şeklinde verilmektedir. Bunun yanında veri setleri listelenirken oluşturulma yılları, dil, konuşmacı sayısı, amaçlanan tanımlama görevi, sınıf sayısı, ifade (telaffuz) sayısı, video çözünürlükleri, FPS (Frame Per Second) değerleri ve toplam süre gibi değerler karşılaştırma açısından verilmektedir. Telaffuz sayısı konusunda literatürde yer alan çalışmalarda ufak bir karışıklık söz konusudur. Çalışmaların bazılarında verilen telaffuz sayısı bir kişi için verilen ifade sayısının tekrarlanması olarak sunulmuşken, yine bazı çalışmalar için verilen telaffuz sayısı, o veri setinde bulunan herhangi bir ifadenin katılımcılar tarafından söylenen toplam telaffuz sayısını temsil etmektedir. Şekil 2.4.'te veri setlerine ait örnek görüntüler karışık bir şekilde yer almaktadır.

Şekil 2.4. Veri setlerine ait kesitler.

Literatür tarandığında sıklıkla bazı kelimelere denk gelinebilir. Bunlardan bir tanesi "phoneme" kelimesidir. Bu kelime İngilizcede bir kelimenin diğer bir kelimeden ayrılabilmesini sağlayan en küçük konuşma veya ses birimidir. "Pin" ve "pan" kelimelerindeki fark, sesli harflerden kaynaklanıyor. Türkçe 'ye fonem olarak çevrilmiştir. Türkçe 'de alfabedeki her harfe karşılık farklı bir fonem olduğu için

Türkçe okunduğu gibi yazılır veya yazıldığı gibi okunur. Türkçe için de “taş” ve “kaş” kelimeleri örnek verilebilir. Burada iki kelimeyi birbirinden ayırt eden fonem, “t” ve “k” harfleridir. Bu durumda /t/ ve /k/ fonem olmaktadır. Amerikan İngilizcesinde 25 harf ve 45 adet fonem bulunur. “Viseme” kelimesi de özellikle dudak okuma gibi alanlarda aynı gibi görünen veya birbirine çok benzeyen bir grup konuşma sesinden her birisine denir. Kısacası “Viseme” bir kelimeyi söylerken yüzün ve ağzın pozisyonunu temsil eder. Belirgin bir Türkçe karşılığı bulunamadığından dolayı tez kapsamında da “viseme” olarak kullanılmıştır [200].

2.1.1. Harf ve Rakam Tanıma

Otomatik dudak okuma sistemlerinin temelini oluşturan veri setleridir. Sistemlerin geliştirilmesine yönelik ilk çalışmalar, alfabe veya rakam tanıma gibi basit tanıma görevlerine yöneliktir. Mevcut veri setleri konuşmacı sayısı, dil, söylenen ifade sayısı ve çok kullanılmasa da uzamsal çözünürlük (spatial resolution) ve zamansal çözünürlük (temporal resolution) gibi yönlerden çeşitlilik gösterir. Zamansal çözünürlük aslında daha çok uydu görüntüleri üzerinden uzaktan algılama konularında kullanılmaktadır. Dünyanın etrafında çeşitli yörüngelerde dönen çok sayıda uydu bulunmaktadır. Zamansal çözünürlük, dönen bu uyduların dünya üzerinde bir bölgeye ait görüntüyü çekmesinden sonra aynı bölgeye ait uydu görüntüsünü yeniden alması arasında geçen süre olarak tanımlanmaktadır. Ek olarak aynı bölgeden yeniden görüntü alma süresinden söz edilirken o görüntünün ne tür bir açıyla alındığı da durumu belirleyen hususlardan bir tanesidir [77].

Harf veya alfabe tanıma işlemleri için en çok kullanılan veri seti 1998 yılında oluşturulan AVLetters [78] veri setidir. 376x288 çözünürlüğe ve 25 FPS oranına sahiptir. AVLetters veri setinde her harfi (A’dan Z’ye toplamda 26 harf) 3 defa tekrarlayan 10 konuşmacıdan gelen kayıtlar bulunur. Veri setinde hem ses hem de video verileri mevcuttur. Ayrıca 60 x 80 piksel boyutlarında dudak bölgelerine ait ROI verileri de sunulmuştur. Daha sonra AVLetters veri setindeki düşük çözünürlük ve az sayıda kişi içermesi gibi bazı zayıf yönler, AVLetters2 [79] ve AVICAR [80] veri setleriyle giderilmiştir. AVLetters2 veri setindeki tekrar sayısı 3 ve 7 arasında değişen değerlerde ayarlanmıştır. Çözünürlük ise 1920 x 1080 ve FPS değeri de 50 FPS olarak

arttırılmıştır. Fakat AVLetters' da 10 olan konuşmacı sayısı AVLetters2'de 5'e düşürülmüştür. Diğer yandan AVICAR, yüksek çözünürlüklü, çok konuşmacılı (multi-speaker) oldukça geniş bir veri setidir. 86 adet konuşmacıya sahiptir.

Rakam tanıma işlemi için XM2VTS 295 katılımcıyla oluşturulmuş en büyük çok konuşmacılı veri setlerinden birisidir. Bu veri seti aslında dudak okumanın yanında kişi tanımlama amacıyla da kullanılmıştır. Veri setinde her katılımcıdan 2 adet rakam dizisi ve fonetik olarak dengeli, düzgün 1 adet cümle söylenmesi istenmiştir. Bu rakam dizileri '0,1,2,3,4,5,6,7,8,9', '5,0,6,9,2,8,1,3,7,4' şeklindedir. VALID [82] veya BANCA [83] gibi diğer veri setleri, rakam tanıma amacıyla yayınlanmış olup XM2VTS veri setine benzer özellikleri takip etmiştir. VALID veri seti özellikle hem kontrollü hem de kontrol edilemeyen ışıklandırmada, akustik gürültü altında kişi tanımlama amacıyla tasarlanmıştır. VALID veri seti 5 farklı senaryoda 106 kişiden alınmış verileri barındırmaktadır. VALID veri setine benzer şekilde BANCA veri seti de özellikle kontrollü, az ışıklı ve olumsuz şartlı ortam olmak üzere 3 farklı senaryoda kişilerin kimliklerini tanımlayabilmek için oluşturulmuştur. 208 konuşmacıdan alınan örnekler, İngilizce, Fransızca, İtalyanca ve İspanyolca olmak üzere 4 farklı dilden veriler barındırır. BANCA veri setindeki konuşmacılardan 12 farklı oturumda, 12 haneli rastgele bir sayı, isimlerini, soy isimlerini, adreslerini ve doğum tarihlerini söylemeleri istenmiştir. Bu da yaklaşık olarak 35bin civarı veriye tekabül etmektedir.

Ancak otomatik dudak okuma sistemlerini rakam tanıma konusunda eğitmek için kullanılan en popüler veri tabanlarından birisi, XM2VTS ve VALID'den çok daha az konuşmacı içermesine rağmen CUAVE'dir [84]. CUAVE'nin popüler olma sebebi konuşmacı sayısı değildir. Veri seti 1 veya 2 kişinin konuşmacı olarak bulunduğu çok sayıda söz öbeği(ifade) ve veri sağlar. Verilerin tek konuşmacıdan alındığı görüntülerde konuşmacı kamera karşısında doğal bir şekilde dururken tekrarlarla birlikte 50 defa rakamları söylemiştir. Bundan sonra da konuşmacı farklı 2 profil görünümünde veya kamera açısında tekrarlarla birlikte 20 tane rakamı söylemiştir. Son olarak da hemen diğer kameraya bakarak 60 tane rakam daha söylemiştir. Bunlar CUAVE'nin tek konuşmacılı verilerinin özellikleridir. Çok konuşmacılı verilerindeyse aynı anda 2 kişiden veriler alınır. Fakat burada da 2 konuşmacı aynı anda konuşmamaktadır. Bir konuşmacı konuşurken diğer konuşmacı görüntülerde

bulunmasına rağmen konuşmamaktadır. Daha sonra konuşmacılardan sıralarını değiştirerek benzer sayıları 2 defa tekrarlayarak söylemeleri istenmiştir.

Bu veri setlerinden sonra rakam tanıma amacıyla İngilizce için AVICAR [80], AVOSES [85] ve AusTalk [86] veri setleri yayınlanmıştır. Aralarında dikkat çekici olanlardan biri olan AusTalk, Avustralya İngilizcesi için geliştirilmiştir. Bunun için ülkenin neredeyse her yerinden insanlar katılmıştır. 2011-2016 yılları arasında kaydedilmiştir. Yaşları 18-83 arasında değişen insanlarla yapılan bu veri setine toplamda 861 yetişkin birey katkı vermiştir. Bu insanları belirli bir kümeden seçmek yerine ülkeyi 15 farklı bölgeye ayırıp o bölgelerden katılımcılar seçmişlerdir. Her konuşmacıdan farklı zamanlarda yazılı ve spontane konuşma durumlarında ses ve dudak yapılarını analiz etmek için 3 farklı durumda 1'er saat veriler alınmıştır. Çalışmalarında belirttikleri üzere şu anda sadece yetişkinler üzerinde çalışılsa da daha sonra bu veri seti çocukları da dahil edecek şekilde daha fazla yaş grubu, daha fazla aksan ve daha fazla konuşma biçimini içerecek şekilde genişletilecektir. Farklı dillerde çalışmalar devam etmiştir. Örneğin Lehçe (Polonya dili) için AGH AV Corpus [87] veri seti oluşturulmuştur. Japonca rakam tanıma için CENSREC-1-AV [88] veri seti, İspanyolca için de AV@CAR [89] veri seti yayınlanmıştır. Bu veri setleri genel olarak orta seviye sayılabilecek uzaysal (spatial) ve zamansal (temporal) çözünürlüklere sahip olmakla beraber en az 15 konuşmacı ile kaydedilmişlerdir. Bazı veri setleriyse yayınlandıktan sonra bilimsel araştırmalara kapatılma durumları olmuştur. Bunlara örnek olarak IBMSR [91] ve IBMIH [90] gibi veri setleri verilebilir. IBMSR ve IBMIH çok sayıda kişinin katılımıyla yüksek sayılabilecek ifade kümesine sahip veri setleridir. Ancak şu anda kullanıma açık değillerdir. 2015 yılında da oldukça popüler hale gelen OuluVS2 [92] yayınlanmıştır. Yayınladıkları konferans çalışma metninde oldukça detaylı veriler sunulmaktadır. Bu çalışma metninde daha önceki veri setlerine atıf yapılarak yeterli konuşmacı sayısının olmaması, az sayıda telaffuz barındırması ve kameraların çekilme açısının hep benzer olması gibi sebepler sunularak veri setlerinin bu eksiklikleri giderdiğini öne sürmüşlerdir. Bünyesinde 50 farklı katılımcının söylediği 1600'e yakın ifade sunulmuştur. Bu veri setinin amaçlarından biri de esnek ağız hareketlerinin takibi ve analizini yaparak kişi tanımlama görevini gerçekleştirmektir. OuluVS2 birden fazla kamera ve açıyla çekildiğinden ötürü Bölüm 2.1.3'te de daha detaylı incelenmiştir. Biraz daha yakın zamana bakıldığında 2018

yılında yayınlanmış AVDigits [86] veri seti 16 farklı milletten 53 farklı katılımcıdan (41 erkek ve 12 kadın), 3 farklı açıdan çekilmiş (frontal, 45 derece açıyla ve profil görünümüyle) görüntülerle 800'e yakın tekrar sayılarına sahip veriler barındırmaktadır. AVDigits veri seti rakamlar ve kısa sözler olmak üzere 2 parçadan oluşur. İlk parçada katılımcılar rakam tanıma veri setini oluşturmak için 0'dan 9'a kadar olan rakamları rastgele sırayla 5 defa söylemişlerdir. İkinci parçasında katılımcılardan 10 kısa ifadenin söylenmesi istenmiştir. Bunlar Türkçe karşılıklarıyla 'Pardon!', 'Merhaba', 'Nasılsın?', 'Görüşürüz' gibi gündelik hayatta sık kullanılan ifadelerdir. Kullanıcılara bu sözler 5 defa tekrarlamayla 3 farklı tonda söyletilmiştir. Veri setine katılanların yaş ortalaması da 26.3 olarak belirtilmiştir. Veri setinin çıkış amaçlarından biri her ne kadar dudak okuma olsa da bunun yanında sessiz konuşma (silent speech) görevini yerine getirmek için de yayınlanmıştır.

Stanford veri seti [206] de rakam, harf ve cümle barındıran orta büyüklükte bir veri setidir. Diğer birçok veri setinde olduğu gibi tüm rakamları barındırır (0-9 arasındaki). Harf olarak da İngiliz alfabesindeki tüm harfleri (toplamda 26 tane) ve günlük hayatta kullanılabilen 49 farklı cümleyi barındırır. Rakamlar, harfler ve cümleler 23 konuşmacıdan alınarak hem işitsel hem de dudak okuma video verilerinden oluşturulmuştur. Çizelge 2.1'de literatürde yer alan dudak okuma veri setleri listelenmektedir. Çizelge 2.1'de ayrıca çeşitli başlıklar altında veri setlerine ait sayısal detaylar da verilmektedir ve üs ifadesi olarak belirtilen 'k', o sayının kelime sayısı olduğunu belirtmektedir.

Çizelge 2.1. Veri setlerine ait detaylar.

İsim	Yıl	Dil	Konuşmacı Sayısı	Görev	Sınıf Sayısı	Tekrar Sayısı	Çözünürlük	FPS
AVLetters	1998	İngilizce	10	Harf	26	780	376×288	25
XM2VTS	1999	İngilizce	295	Rakam	10	885	720×576	25
IBMViaVoice	2000	İngilizce	290	Cümle	10,500	24,325	704×480	30
VIDTIMIT	2002	İngilizce	43	Cümle	346	430	512×384	25
BANCA	2003	Çoklu	208	Rakam	10	29,952	720×576	25
IBMIH	2004	İngilizce	79	Rakam	10	16,197	720×480	30
AVOZES	2004	İngilizce	20	Rakam	10	200	720×480	30
				Cümle	3	60		
AV@CAR	2004	İspanyolca	20	Harf	26	800	768×576	25
				Rakam	10	600		
				Cümle	250	6,000		
AVICAR	2004	İngilizce	86	Harf	26	59,000	720×480	30
				Rakam	13			
CUAVE	2004	İngilizce	36	Rakam	10	7,000	720×480	30
AV-TIMIT	2004	İngilizce	233	Cümle	510	4,660	720×480	30
VALID	2005	İngilizce	106	Rakam	10	1,590	576×720	25
GRID	2006	İngilizce	34	Deyim(Phrase)	51	34,000	720×576	25
IBMSR	2008	İngilizce	38	Rakam	10	1,661	368×240	30
AVLetters2	2008	İngilizce	5	Harf	26	910	1920×1080	50
I V 2	2008	Fransızca	300	Cümle	15	4,500	780×576	25
UWB-07-	2008	Çekçe	50	Cümle	7,550	10,000	720×576	50
OuluVS	2009	İngilizce	20	Deyim(Phrase)	10	1,000	720×576	25
CENSREC-1-	2010	Japonca	42	Rakam	10	3,234	720×480	30
QuLips	2010	İngilizce	2	Rakam	10	3,600	720×576	25
NDUTAVSC	2010	Almanca	66	Rakam	6,907	6,907	640×480	100
				Kelime				
				Cümle				
WAPUSK20	2010	İngilizce	20	Deyim(Phrase)	52	2,000	640×480	32
LILiR	2010	İngilizce	12	Cümle	200	2,400	720×576	25

Çizelge 2.1. (devam ediyor).

İsim	Yıl	Dil	Konuşmacı Sayısı	Görev	Sınıf Sayısı	Tekrar Sayısı	Çözünürlük	FPS
BL	2011	Fransızca	17	Cümle	238	4046	640×480	30
UNMC-	2011	İngilizce	123	Cümle	12	2460	708×640	29
MOBIO	2012	İngilizce	150	Cümle	N/A	N/A	640×480	16
AGH AV	2012	Lehçe	20	Rakam	N/A	N/A	1920×1080	50
MIRACL-VC	2014	İngilizce	15	Kelime	10	1500	640×480	15
				Deyim	10	1500		
AusTalk	2014	İngilizce	1000	Rakam	10	24000	640×480	Karışık
				Kelime	966	966000		
				Cümle	59	59000		
MODALITY	2015	İngilizce	35	Kelime	182	231	1920×1080	100
OuluVS2	2015	İngilizce	53	Rakam	10	1590	1920×1080	30
				Deyim				
				Cümle	530	530		
RM-3000	2015	İngilizce	1	Cümle	1,000[k]	3,000	360×640	60
IBM AV-ASR	2015	İngilizce	262	Cümle	10,400[k]	N/A	704×480	30
TCD-TIMIT	2015	İngilizce	62	Cümle	5,954	6913	1920×1080	30
HAVRUS	2016	Rusça	20	Cümle	1,530	4000	640×460	200
LRW	2016	İngilizce	1,000+	Kelime	500	400000	256×256	25
LRS	2017	İngilizce	1,000+	Cümle	17,428[k]	118116	160×160	25
VLRF	2017	İspanyolca	24	Cümle	1,374[k]	10200[k]	1280×720	50
MV-LRS	2017	İngilizce	1,000+	Cümle	14,960	74564	160×160	25
AVDigits	2018	İngilizce	53	Rakam	10	795	1280x780	30
			39	Deyim		5850		

2.1.2. Kelime ve Cümle Tanıma

Alfabe ve rakam tanıma üzerine oluşturulan veri setleri çalışmalarda oldukça sık kullanılmıştır. Çünkü genelde bahsi geçen veri setleri sınırlandırılmış ortamlarda, sınırlı bir kelime dağarcığı ve sınıf başına çok sayıda örnekler ile çalışılmasını sağlar. Bu durum dudak okumaya yönelik tasarlanan yöntemlerin etkinliğini analiz etmek için faydalı olsa da ortaya çıkan modeller sınırlı kapsamda veya verileri ezberleme eğiliminde olur. Bu sebeple de kelime veya cümle tanıma gibi daha karmaşık işlemleri gerçekleştirmesi zordur. Ancak otomatik dudak okuma sistemlerinin asıl nihai amacı, temel olarak cümle terimlerinden, kelimelerinden veya eklerinden doğal konuşmayı anlamaktır. Bu da kelime, kelime grubu ve fonetik olarak dengeli cümleler içeren veri setlerinin oluşturulmasını, bunlar üzerine çalışılmasını gerekli kılmıştır. Açıklanan sebeplerden dolayı da harf ve rakam tanıma gibi daha kolay işlemler için tasarlanan veri setlerine ek olarak bir yandan da kelime ve cümle tanımaya yönelik veri setlerinin oluşturulması için de çalışmalar başlamıştır.

Cümleleri de kapsayan en eski görsel ve işitsel veri setlerinden biri, IBM tarafından yaklaşık olarak 10400 farklı kelime ve 24325 cümle içeren IBMViaVoiceT [94] veri setidir. IBMViaVoiceT veri setinin en önemli özelliği LVCSR (Large-Vocabulary Continous Speech Recognition Systems) sistemlerine yönelik en büyük veri seti olmasıdır. Ek olarak veri seti üzerinden IBMViaVoice adında konuşma modelleme yazılımı da geliştirilmiştir. Bu ürün de ticari bir ürün olduğu için IBMViaVoiceT veri seti bilimsel araştırmalara açık değildir. Benzer yıllarda 2002'de VIDTIMIT [95] veri seti oluşturulmuştur ve bilimsel araştırmalara açıktır. Bu veri seti her ne kadar cümleler içerse de asıl amacı kişi doğrulama uygulamaları geliştirmektir. Kayıt işlemi 3 oturumda yapılmıştır. Bu da diğer veri setlerinin çoğunda olmayan bir özelliktir. Böyle yapılmasının sebebi, insanların farklı zamanlarda ve şartlarda verilerin alınmasıyla birlikte sistemin doğruluğuna yönelik bir ayarlamadır. Birinci ve ikinci oturum arasında 7 gün beklemişlerdir. İkinci ve üçüncü oturum arasında da 6 gün beklenmiştir. 43 konuşmacı 346 farklı cümleyi 10'ar defa söylemiştir. Cümleler TIMIT adı verilen derlemden (corpus) alınmıştır. Aynı grup tarafından da yine AV-TIMIT [96] adı verilen veri seti 2004 yılında yayınlanmıştır. AV-TIMIT veri seti 233 katılımcı ve 510 farklı cümleden oluşur.

Harf ve rakam tanıma bölümünde anlatılan bazı veri setleri de aynı zamanda kelime ve cümleler barındıran kısımlara sahiptir. AV@CAR 250 adet fonetik olarak dengeli cümle, AVICAR 1300'den fazla farklı kelime ve AVOZES ise Avustralya İngilizcesi için kelimeler barındırmaktadır.

2007 ve 2015 yılları arasında birçok veri seti yayınlanmıştır. Bunların çoğu dil olarak İngilizce içeriklere sahiptir. Fakat farklı dillerde de çalışmalar yapılmıştır. Örneğin 2007 yılında Fransızca oluşturulmuş veri setinde konuşan yüz verisi (TFD- Talking Face Data) oluşturmuşlardır. Veri seti aslında dudak okuma amacıyla yayınlanmamıştır. Yayınlanma sebebi insanların kamera karşısında konuşmalarını onların dudak ve konuşmalarını analiz ederek bu konuşmacıları taklit edecek yapay bir yüz oluşturmaktır. Katılımcılardan alınan 2 boyutlu görüntüleri kullanarak stereoskopi yöntemiyle birbirine benzeyen 2 boyutlu görüntülere üçüncü boyut kazandırılmıştır. Fakat veri setinin içeriği itibariyle dudak okumaya müsaittir [97]. Farklı dillere örnek olarak bir başka veri seti de 2008 yılında Çekçe dilinde yayınlanmıştır [98].

İngilizce tabanlı korporalar arasında OuluVS [99] veri seti, otomatik dudak okuma sistemlerini değerlendirmek için en çok kullanılan veri setlerinden biridir. OuluVS, her ifadenin aynı konuşmacı tarafından 5 defaya kadar tekrarlandığı, İngilizcedeki günlük hayatta kullanılan 10 tane kısa cümlenin 20 konuşmacı tarafından söylendiği verilerden meydana gelmektedir.

- MIRACL-VC [101],
- UNMC-VIER [102],
- LILiR [100],
- AusTalk [80]

Veri setleri sırasıyla 15, 123, 12 ve 1000 konuşmacı içerir. Bununla birlikte MIRACL-VC ve UNMC-VIER oldukça az sayıda cümle (10 ve 12) içerirken, LILiR ve AusTalk sırasıyla 200 ve 59 adet birbirinden farklı cümle içermektedir.

İngilizce veri setlerinden biri de MOBIO'dur. MOBIO veri seti, daha önce bahsedilenlerden farklı olarak, bir cep telefonundan otomatik yüz ve konuşmacı tanımayı gerçekleştirmek için tasarlanmıştır. 150 konuşmacının katıldığı bu veri seti, 2 kısımdan oluşmaktadır. İlk kısımda konuşmacılar kısa ve rastgele gelen soruları yanıtlamışlardır. İkinci oturumda da önceden yazılan metinleri okumuşlardır. Tüm bu videolar, konuşmacıların kendi başına cep telefonlarıyla kaydettiği görüntülerden oluşur.

Benzer yıllarda diğer dillerde kaydedilen görsel-işitsel veri setleri, İngilizcedekilere göre sayıca çok daha azdır. Örneğin IV2 [97] ve BL [103] veri setleri Fransızcada yayınlananlara örnek verilebilir. IV2, 15 cümle söyleyen çok sayıda konuşmacı (300) sağlarken, BL de 17 konuşmacı 238 cümle sağlar. Diğer örnekler arasında Çekçede 50 katılımcıdan 10000 ifade sağlayan UWB-07-ICAVR veri seti [68] bulunur. NDUTAVSC veri seti de Almanca için yayınlanmış 66 konuşmacının yer aldığı verilere sahiptir. İspanyolca için daha önce bahsedildiği üzere AV@CAR [63] veri seti ve VLRF [37] veri seti de 24 konuşmacıdan toplamda 1507 söz öbeği (ifade) barındırmaktadır.

Günümüze daha yakın zamanlara gelindiğinde veri setleri yayınlanmaya devam etmektedir. Bunların arasında tek konuşmacılı, 979 farklı kelime ve toplam 3000 söylenen ifadeden oluşan RM-3000 [105] veri seti bulunmaktadır. RM-3000 veri setinde toplamda 26114 tane kelime,105561 tane fonem, cümle başına düşen kelime sayısı 8.70, cümle başına düşen fonem sayısı 35.19, kelime başına düşen fonem sayısı 4.04'tür. Tek konuşmacılı veri setlerinin yanında çok konuşmacılı veri setleri de yayınlanmıştır. Bunlardan bazıları aşağıda verilmiştir.

- HAVRUYS [92]
- VLRF [106]
- OuluVS2 [107]
- IBM AV-ASR [108]
- TCD-TIMIT [104]
- AVDigit [93]
- Stanford [206]

Yukarıda sayılan 7 adet veri seti başta olmak üzere birçok çok konuşmacılı veri seti yayınlanmıştır. Sıralanan 7 veri seti sırasıyla 20, 24, 53, 262, 62, 53, 23 konuşmacıya sahiptir. Stanford veri setinde toplamda 49 farklı cümleyi söyleyen 23 konuşmacı bulunmaktadır. OuluVS2, söz öbekleri ve cümleler söyleyen konuşmacıların kayıtlarını içerir. Her bir konuşmacı, OuluVS veri setine benzer olarak günlük hayatta sıkça kullanılan 10 cümleyi üçer defa tekrarlamıştır ve TIMIT derlemindeki toplam 530 cümle arasından rastgele seçilen 10 tane cümle yine konuşmacılar tarafından 1 defa okunmuştur. Öte yandan, TCD-TIMIT veri seti 7000'e yakın birbirinden farklı cümle ve yaklaşık 14000 söz öbeği içerirken, HAVRUYS [92] veri seti Rusça 'da düzenlenmiş olup 20 konuşmacıdan alınmış toplam 4000 ifadeye sahiptir. TCD-TIMIT veri setinin önemli bir özelliği de konuşmacılara söylettikleri sözlerin duygu durumu veya tür olarak sınıflandırılmasıdır. Örneğin common (yaygın), anger (kızgınlık), disgust (iğrenç), fear (korku), happiness (mutluluk), sadness (üzüntü), suprise (şaşkın), neutral (doğal) bu sınıflardan bazılarıdır. HAVRUYS veri seti her ne kadar otomatik dudak okuma sistemlerini geliştirmek için kullanılabilse de asıl oluşturulma amacı "Görsel-İşitsel Konuşma Tanıma- AVSR" projelerine kaynaklık etmektir. IBM tarafından IBMViaVoiceT veri setinden yukarda bahsedilmiştir. Buna ek olarak IBM yine IBM AV-ASR veri setini yayınlamıştır. IBM AV-ASR veri setinde hem kelimeler için hem de cümleleri için ayrı bölümler yer almaktadır. Veri setindeki cümleler 262 katılımcıdan alınmış toplam uzunluğu 40 saate yakın olan veriler barındırır. Veriler düzgün bir stüdyo ortamında kaydedilmiştir. Ses verisi 16 KHz frekans değerine sahipken videoların tamamı 30 FPS ve 704 x 480 çözünürlüğe sahiptir. Kelimeler bölümü ise 10400 tane farklı kelimeden oluşur. Fakat bu veri seti de tıpkı IBMViaVoiceT veri setinde olduğu gibi ticari amaçlarla kullanıldığı için henüz kullanıma açılmamıştır. İspanyolca için oluşturulan VLRF veri seti, fonetik olarak düzgün 500 cümleden (bu cümleler de toplamda 10000 'den fazla kelime havuzundan oluşturulmuş) oluşan bir havuzdan 24 konuşmacının her birisinin rastgele olarak seçtiği 25 tane cümlenin üç kez tekrarlanmasıyla oluşturulmuştur. Bu veri setinin diğer çalışmalardan farkı, değişik duyma problemlerine sahip katılımcılara yer vermesidir. 24 konuşmacıdan 15'inin işitme problemi bulunmazken 9'unun ciddi işitme problemi veya işitme engeli bulunmaktadır. Bu 9 kişi aynı zamanda gerçek hayatta da dudak okuması yapan kişilerdir. En yenilerden biri olan AVDigits veri seti,

tıpkı OuluVS ve OuluVS2 ‘de olduğu gibi günlük hayatta sıkça kullanılan 10 adet cümleyi söyleyen 39 konuşmacının videolarını içerir. Her cümle 3 farklı konuşma modunda cümlelerin 5 kez tekrarlanmasıyla söylenmiştir. 3 farklı konuşma modu ise sırasıyla “normal”, “fısıltılı” ve “sessiz” olacak şekildedir.

Göz önünde bulundurulması gereken bir diğer önemli unsur da dudak okuma sistemleri de dahil olmak üzere bilgisayarlı görünün birçok alanında önemli ilerlemeler sağlayan Derin Sinir Ağlarının (Deep Neural Networks) son birkaç yılda daha da yaygın olarak kullanılmasıdır. Derin ağlar, sınıflandırma performansında ciddi gelişmeler göstermiş olsa da bu durum aslında sadece eğitim için uygun veriler mevcutsa mümkündür. Başka bir deyişle derin sinir ağların mantığında özellikle Bölüm 4.2’de detaylarının verildiği şekliyle çok büyük miktarlarda veriyle eğitim ihtiyacı bulunur. Tez kapsamında şimdiye kadar otomatik dudak okuma sistemleri için çok sayıda görsel-işitsel veri setlerinden bahsedilmiş olsa da çoğu yeterli sayıda örnek içermemekte veya iyi genelleme yapan derin öğrenme modellerini eğitmek için yeterli kelime, cümle, harf vs. çeşitliğine sahip değillerdir. Derin öğrenme modellerinin diğer alanlarda da karşılaştığı problem olan veri setinin azlığı veya genelleştirme problemi otomatik dudak okuma sistemleri için de ortaya çıkmıştır. Bahsedildiği üzere bu problem otomatik dudak okuma problemine özel bir şey olmayıp derin öğrenme modellerine özel bir durumdur. Bu yüzden de derin öğrenme modellerinin popüler hale gelmesiyle birlikte mevcut veri setleri, korporalar arasında sınıf başına daha fazla sayıda söz öbeğine, kelimeye, cümleye veya ifadeye sahip olanlar daha sık kullanılmaya başlanmıştır. Bu tarz veri setleri eskiden var olsalar da büyüklük veya çalışma zorluğu gibi çeşitli sebeplerle pek tercih edilmemiştir. Aynı zamanda büyük veri seti ihtiyacıyla birlikte başka birkaç yaklaşım daha oluştu. Bunlardan biri belirli bir cümle kalıbı verilerek o cümle kalıbının kombinasyonlarını üreterek veri setini oluşturmaktır. Stillerden bir diğeri de özellikle cümleler için maksimum bir süre sınır koymaktır. Son olarak sık kullanılmaya başlanan bir başka yöntem de konuşmacı hızının ayarlanmasıdır. Örneğin GRID [109] veri seti dudak okuma sistemlerine ait çalışmalarının yeni yeni hız kazandığı dönemlerde yani 2006 yılında yayınlanmıştır. Ancak kullanımı son 3-4 senedir önemli ölçüde artmıştır. GRID veri setindeki 34 konuşmacı, her biri 3 saniyelik uzunluğa sahip 1000 tane cümle söylemektedir. Videolar 25 FPS’e sahip olduğundan dolayı da her bir cümle 75 kareden oluşur. GRID

veri setindeki tüm videolar aynı kaliteye sahip değillerdir. “Normal” ve “Yüksek” olmak üzere 2 farklı kalitede videolar mevcuttur. Örneğin normal kalitedeki videoların çoğu 360 x 288 piksel çözünürlüğe sahiptir. Toplam video uzunluğu 28 saattir. Her konuşmacı tüm “renk”, “rakam” ve “harf” kombinasyonlarını söyleyerek sırasıyla aşağıda belirtilen cümle yapısını oluşturmuşlardır.

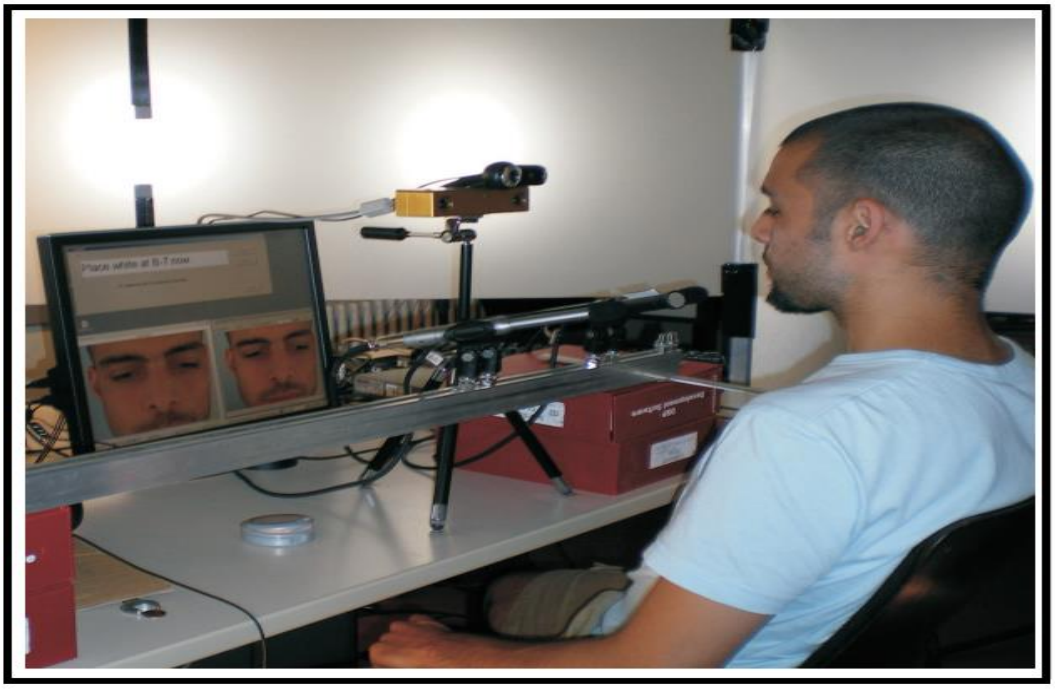
- <Fiil>
- <Renk>
- <Edat>
- <Rakam>
- <Harf>
- <Zarf>

Örnek bir GRID cümlesi “Place brown in A 7 please” şeklindedir. Bu şekilde toplamda 34000 farklı kombinasyonda söz öbeği, 51 farklı kelimeye sahip kısmen birbirine benzeyen cümlelerden oluşur. Böylece aslında GRID veri setiyle kelime, cümle, harf, rakam gibi birçok yapının da tanınmasına imkân sağlanmıştır. GRID’in ilginç özelliklerinden biri de İngilizce olmasına rağmen ‘w’ harfini harf listesinde bulundurmamasıdır. Kalıplar üzerinden oluşturulan GRID tarzı veri setlerinin çok önemli bir problemi vardır. O da kalıplara dayalı oluşturulduğu için cümle yapılarında kullanılan kelimeler kısıtlı ve dengesiz olmaktadır. Örneğin bu kalıplarda kullanılan rakam sınıf sayısıyla, harf sınıf sayısı birbirinden çok farklıdır. Çünkü rakam sayısı 10 iken harf sayısı 26’dır. Bu sayısal farklılık, beraberinde cümle içinde geçme sayısının ve sıklığının değişmesini de getirmektedir. Bir ifadenin cümlelerde geçme sıklığı büyük bir farklılık gösterirse geliştirilen sistemlerin buna adaptasyonu veya modelin genelleştirilebilmesi daha da zorlaşır.

MODALITY [111] ve WAPUSK20 [110] gibi GRID veri setindekine benzer bir cümle yapısını izleyen başka veri setleri de bulunmaktadır. Örneğin WAPUSK20 veri setinin cümle yapısı yayınlanan çalışma incelendiğinde aşağıdaki gibidir;

- Fiil: {‘Bin’ ‘Place’ ‘Lay’ ‘Set’}- Fiiller çok sınırlı sayıdadır.
- Renk: {‘white’ ‘green’ ‘red’ ‘blue’}- Renkler 4 adet ile sınırlıdır.
- Edat: {‘by’ ‘at’ ‘with’ ‘in’}- Edatlar 4 adet ile sınırlıdır.
- Harf: {‘A’ dan ‘Z’ ye kadar}
- Rakam: {‘0’- ‘9’ arası}
- Zarf: {‘soon’, ‘again’, ‘please’ ‘now’}- Zarflar 4 adet ile sınırlıdır.

WAPUSK20‘nin GRID’den ufak da olsa bir farkı vardır. O da ‘w’ harfine harf listesinde yer vermesidir. WAPUSK20 veri setinde her konuşmacı 100 tane cümle söylemiştir. Kayıt süresi için her konuşma, 3 saniye ile sınırlandırılırken konuşma hızı için herhangi bir şart koşulmamıştır. 2 konuşmacı hariç tüm veriler, kuruluma giriş, tekrarlar vb. aşamalar da dahil olmak üzere yaklaşık 1 saatlik tek bir oturumda kaydedilmiştir. Veri setindeki konuşmacılar tarafından söylenen tüm cümleler birbirinden farklıdır. Bu sebeple bir konuşmacının söylediği cümleleri daha önce başka bir konuşmacı söylememiştir. WAPUSK20 veri seti konuşmacıların kameraya çok yakın durduğu ve dudaklarının çok net belli olduğu bir veri setidir. Şekil 2.5’te WAPUSK20 veri setine ait bir kayıt örneği verilmiştir.



Şekil 2.5. WAPUSK20 veri setinin çekimlerine ait görsel.

GRID, WAPUSK20 gibi veri setleri derin öğrenme modellerini eğitmek için diğer veri setlerinden daha iyi olmanın yanında genelleştirme amacına bir temel oluşturmuş olsalar da bu veri setlerinin kapsadıkları oldukça az kelime havuzundan dolayı genelleştirme düşünüldüğü kadar geniş çapta değildir.

Bu nedenle, hem çok sayıda söz öbeği ve ifade hem de daha geniş bir çeşitlilik sağlamak amacıyla son zamanlarda yeni veri setleri oluşturulmuştur. Bu konuya ilişkin göze çarpan çalışmalar, MV-LRS [57], LRW [55], LRS [57] bunlardan bazılarıdır. Harf ve rakamların bahsedildiği bölümde LRW veri setinden ayrıca bahsedilmiştir. Kelimeler için hazırlanan LRW (Lip Reading Kelimeler) ve cümleler için hazırlanan LRS (Lip Reading Cümleler) veri setleri, 2010 ve 2016 yılları arasında yayınlanan BBC televizyon programlarından alınan kayıtlara dayanmaktadır. LRW ve LRS'de 1000'den fazla konuşmacıdan alınan cümleler ve her biri en az 800 kez geçen 500 kelimelik bir kelime havuzu bulunur. Toplamdaysa LRW, 400.000'e yakın söz öbeği barındırır. Tıpkı LRW ve LRS veri setlerinde olduğu gibi MultiView-LRS (MV-LRS) veri seti de BBC televizyon programlarından kaydedilmiştir. LRW ve LRS sadece karşıdan çekilmiş görüntüler içerirken MV-LRS 0 - 90 derece arasındaki çeşitli görüş açılardan çekimler barındırır.

Bahsedilen veri setleri dışında çok fazla kullanılmayan veri setleri de bulunmaktadır. Bunlar her ne kadar yayınlanmış olsalar da literatürde çeşitli sebeplerden dolayı kullanılmamaktadır. United Nations 1 (Birleşmiş Milletler 1) adı verilen veri seti İngilizce, Arapça, Çince, Fransızca, Rusça ve İspanyolca için tasarlanmış çok dilli konuşmacılar kullanarak konuşmacıya bağlı dil tanıma için tasarlanmıştır. United Nations 2 (Birleşmiş Milletler 2) olarak bilinen ikinci veri seti de belirtilen bu dillerdeki konuşma arasında belirli parametrelere göre ayrım (dil açısından) yapmak için oluşturulmuştur. United Nations 1 ve United Nations 2 veri setlerinin ikisi de "Birleşmiş Milletler İnsan Hakları Bildirgesi"ni okuyan konuşmacıların ses ve görüntülerini içerir. Hatta ses ve görüntüye ek olarak veri setleri, başta dudaklar olmak üzere konuşmacıların yüz özellikleriyle ilgili bir dizi veriler olarak yüzün kenar, sınır, x ve y koordinatlarına karşılık gelen bazı bilgiler içerir. 2 veri setinin önemli özelliklerinden bir diğeri kaydedilen konuşmacıların hepsi en az 3 dilde akıcı konuşabilen çok dilli kişilerdir. Seçilen bu kişiler rastgele seçilmiş kişiler değil de

çocukluğundan itibaren çok dil konuşarak yetişen kişilerdir [160]. Bu tarz veri setlerinin oluşturulmasındaki asıl amaç dudak okutulmasını sağlayarak ne söylendiğinin anlaşılması değildir. Genelde amaç, her iki dile ait kelime ve cümleleri daha önce görmemiş sistemlerin konuşmacıların hangi dili konuştuğunun tespit edebilmesidir. Eğer veri seti sadece bir konuşmacının aynı şeyleri çoklu dillerde söylemesiyle oluşturulduysa genelde bu tarz veri setleri Konuşmacıya Bağlı Görsel Dil Tanımlama (Speaker Dependent Visual Language Identification) adı verilen bir test yöntemi için kullanılmaktadır [161].

Çizelge 2.2. Veri setlerinden örnek ifadeler.

Veri Setinin Adı	**Dil**	**Yıl**	**Örnek İfadeler**
GRID	İngilizce	2006	Lay red at B 2 again.
			Bin green in C 3 now.
			Place yellow at D 9 please.
OuluVS	İngilizce	2009	I am sorry.
			How are you?
			Excuse me!
AVICAR	İngilizce	2004	Are holiday aprons available to us?
			Movies never have enough villains.
			Cut a small corner off each edge.
VIDTIMIT	İngilizce	2009	He took his mask from his forehead and threw it, unexpectedly, across the deck.
			Make lid for sugar bowl the same as jar lids, omitting design disk.
			Grandmother outgrew her upbringing in petticoats.
UNMC-VIER	İngilizce	2011	Don't ask me to carry an oily rag like that
			She had your dark suit in greasy wash water all year.
			Joe took fathers green shoe bench out.

Çizelge 2.2. (devam ediyor).

Veri Setinin Adı	**Dil**	**Yıl**	**Örnek İfadeler**
OuluVS2	İngilizce	2015	1 7 3 5 1 6 2 6 6 7
			Have a good time.
			Chocolate and roses never fail as a romantic gift.
VLRF	İspanyolca	2017	A las ocho de la manana ya estaba haciendo pasteles.
			Eligieron una casa all´ı con las mismas condiciones.
			Los gusanos son animales invertebrados sin extremidades.
TCD-TIMIT	İngilizce	2015	Anger: Who authorized the unlimited expense account?
			Neutral: The best way to learn is to solve extra problems.
			Surprise: The carpet cleaners shampooed our oriental rug.
LRS	İngilizce	2017	When you’re cooking chips at home.
			Through what they call a knife block.
			The traditional chip pan often stays on the shelf.

2.1.3. Çoklu Görünümlü Veri Setleri (Multiview Dataset)

Otomatik dudak okuma sistemleri genellikle önden görünümlü veri setlerinin kullanılmasına dayalıdır. Fakat günlük hayat pratiklerinde kullanılacak bir sistemde giriş olarak verilen görüntülerin sadece önden çekilmiş görüntüler olmasını sağlamak her zaman mümkün değildir. Örneğin, tek kamera ile bir konuşmada birden fazla konuşmacıyı görüntüleme durumunda her konuşmacı için farklı açılardan görüntülerle çalışılması gerekmektedir. Bu sebeple de uygulanabilir pratik dudak okuma sistemleri, gerçekçi senaryolarda konuşmayı anlayabilmesi için çoklu görünümlü dudak okuma veri setleri ile çalışmalıdır. Buna dair yapılmış çalışmaların birinde 3 soruya cevap aranmıştır. (1) “Bilgisayarlar, yüzün tam karşıdan mı yoksa tam karşıdan olmayan bir görünümünü kullanarak mı daha iyi dudak okur?”. (2) “Bir bilgisayar dudak okuma sistemi için en iyi görüş açısı nedir?”. (3) “Bir bilgisayar dudak okuma sistemi, görüş açısından bağımsız olarak nasıl çalışır hale getirilebilir?”. Meslek olarak dudak okuma uzmanlarıyla yapılan bir araştırmada tam karşıdan çekilmiş görüntülerle dudak okuma

işlemini gerçekleştirmenin en iyi sonucu vermediği saptanmıştır. Bunun sebepleri olarak da tam olarak önden çekildiğinde veya bu insanlar dudak okurken tam karşıya geçtiğinde, dudağı okunmak istenen kişinin dudak çıkıntıları, yuvarlanması ve hareketlerinin tam olarak gözlenememesi gibi durumlar gösterilmiştir. Bu gibi sorunlardan dolayı da tam olarak karşıdan çekilmiş görüntülerden ziyade 15- 20 derecelik açılarla çekilmesinin daha iyi sonuç vereceği sonucuna varılmıştır. Varılan sonucu da 5 farklı açıdan çekilmiş bir veri seti üzerinde çalışmalar yaparak da desteklemişlerdir [113].

Çok açılı veya çoklu görünümlü veri setlerinde en iyi hangi açıdan sonuç vereceğine dair çalışmalardan bir diğerinde Bowden ve arkadaşları 1 konuşmacının 1920 x 1080 çözünürlükte 59.94 FPS değerine sahip 3 adet Sony Xacti FH1 kamera ile kaydetmişlerdir. Üç kamera, konuşmacıya yaklaşık olarak 0°, 30° ve 45° derecelik açılarda tripodlar üzerine monte edilmiştir. Konuşmacıya cümleyi tekrar etmesi dışında başka bir talimat verilmemiştir. Veri setini oluşturma 2 farklı kısımda gerçekleşmiştir. Birinci kısımda veri setindeki amaç, ABD şehirlerini belirlemektir. Dallas, Los Angeles, Chicago, New York, Philadelphia, Phoenix, Houston, San Antonio, San Diego ve San Jose gibi şehir isimlerini taşıyıcı cümlesine gömerek: “I am going to (see) XXX soon / again” (Yakında / tekrar XXX'e gidiyorum)” şeklinde söyletmektedirler. Bu şekilde veri setinde toplamda 120 tane farklı cümle oluşturulmuştur. Örnek cümlelerden bazıları şunlardır: “I am going to Houston again”, “I am going to see San Antonio soon”. Veri setini oluşturmada ikinci kısımda 637 saniye (yaklaşık 10 dakika) süren bir konuşma yapan 2 kişi kaydedilmiştir. Bu aşamada daha gerçekçi bir senaryoda işlerin kısmen zorlaştırılması sağlanmıştır. Her iki konuşmacı da birinci kısımdaki şehirleri konuşmalarına dahil etmiştir. Konuşma, önceden hazırlanmış bir metne bağlı olmadığı için 2 kişinin arasında geçen sohbette şehir adlarının söylenmesi gereken yerler kısıtlanmamış ve şehir adlarını kullanacakları yeri her iki konuşmacının kararına bırakmışlardır. Bu bölümde, kamera bir tripoda takılmak yerine elde tutulmuştur. Bunun sebebi de sık sık kamerada doğal sarsıntılar meydana gelecektir ve böylece konuşmacıların videodaki konumları büyük ve hızlı bir şekilde değişecektir. Ek olarak, konuşma sırasında her iki konuşmacı da (kameradan uzakta) yüzlerini birbirine çevirdikçe mevcut olan baş duruşu nedeniyle video daha zorlu hale gelmiştir. Bu iki parçalı veri seti, son derece doğal ve zorlu

koşullarda kaydedilen düşük kaliteli videodan oluştuğu için bugüne kadarki en zorlu dudak okuma veri setlerinden biridir. Bu tarz zorlu ver setlerinin oluşturulmasındaki amaç, dudak okuma sistemlerinin hangi bileşenlerinin sağlam veya verilere bağımlı olabileceğini belirlemektir. Geliştirdikleri sistemi oluşturdukları bu veri seti üzerinde ve LILiR veri setinde kullandıklarında en iyi sonuçlara 30^o ile elde edilmiş görüntülerde ulaşılmıştır [11]. Şekil 2.6'da Bowden'ın oluşturduğu veri setine ait örnek bir görüntü yer almaktadır.

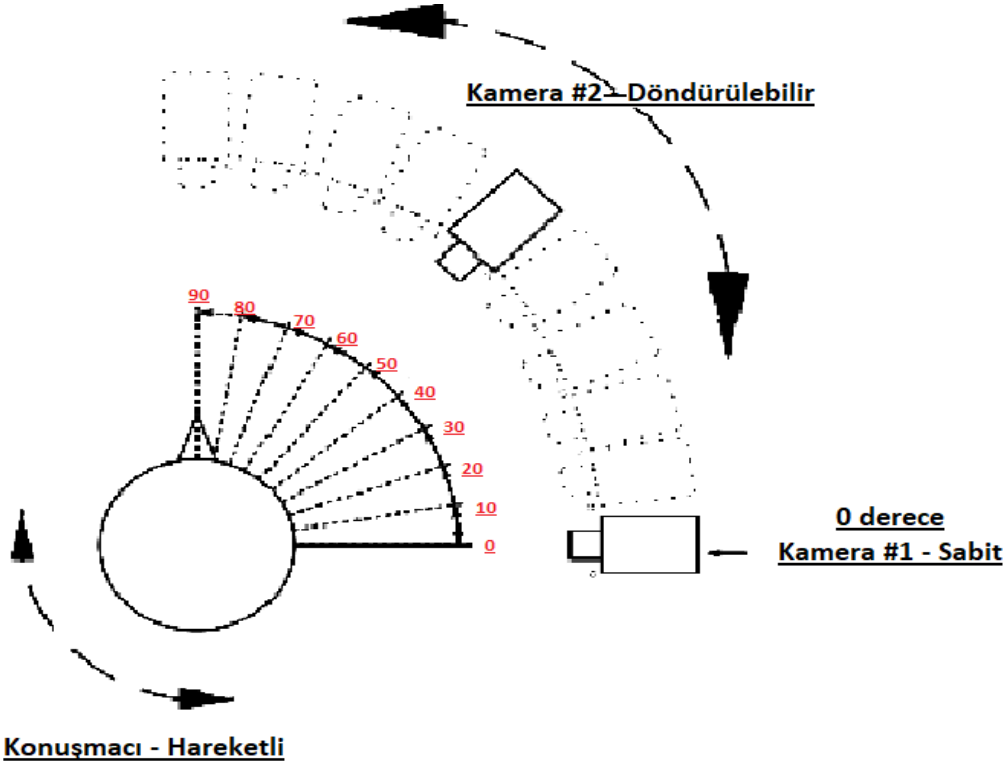
Şekil 2.6. Bowden veri setinin 2. kısmına ait görüntü.

Görsel-işitsel sistemler için çoklu görünüm veri setlerini oluşturmak amacıyla kullanılan araç ve ortamlarda önemli değişkenlikler vardır. Bunlardan en önemlisi açıdır. Bazı veri setleri yalnızca tam önden ve tam profilden görüntüler içerirken, bazıları da -90, 0 ve 90 dereceler arasında değişecek açılarda veriler barındırır. Yine bunu yaparken de tekli kamera veya çoklu kamera özelliklerinin de belirtilmesi gerekir. Bazı çalışmalarda çoklu görünümü sağlamak için tek kamera kullanılarak kişinin farklı zamanlarda farklı açılardan görüntüleri alınır. Bazı çalışmalarda da konuşmacının kaydı farklı açılarda konumlandırılan birden fazla kamera aracılığıyla yapılır.

Bölüm 2.1.1'de açıklanan AVICAR veri seti, dört kamera ve sekiz mikrofon kullanılarak, hareket halindeki bir otomobilde kaydedilmiştir. Kameralar aracın ön paneline yerleştirilmiştir. Bu 4 kamera sürücüyü farklı açılarda yakından kaydetmiştir. Bunun gibi CUAVE, CMU AVPFV (CMU Audio-Visual Profile and Frontal View) [114] ve IBMSR veri setleri de hem karşıdan hem de profil görünümlerinden kayıtlar içerir. CUAVE veri seti, tek kameradan hem önden hem de profilden alınmış görüntüleri kullanır. Konuşmacılar tarafından arka arkaya rakamlar söylenmiştir.

CUAVE veri seti Bölüm 2.1.1‘de detaylarıyla zaten bahsedilmektedir. CMU AVPFV veri seti, ses geçirmez bir IAC (Inter-Application Communication) stüdyosunda eş zamanlı olarak kaydedilen profil ve önden görünümlü sesi de barındıran videolardan oluşur. Videolar, 30 FPS VGA (640*480) çözünürlükte kaydedilmiştir. Kelimeleri, MRT (Modified Rhyme Test – Modifiye Kafiye Testi) adı verilen konuşma anlaşılırlığı testinde yaygın olarak kullanılan bir testten seçmişlerdir [115]. MRT listesindeki tüm kelimeler ünsüz – ünlü – ünsüz kalıbındadır. Veriler, her biri 150 kelimelik MRT listesini 10 kez tekrarlayan 10 konuşmacıdan toplanmıştır. Benzer şekilde IMBSR veri seti, 38 konuşmacının belli rakam dizilerini, 3 farklı açıdan (önden ve diğer ikisi yan açılardan) görüntülerini aynı anda kaydeden 3 kameranın önünde söylemesiyle oluşturulmuştur. Fakat IMBSR veri seti bilimsel araştırmalara açık değildir.

IMBSR veri setiyle yakın zamanlarda 0 ile 90 derece arasında değişen görüş açılarına sahip görüntüleri barındıran birkaç veri seti daha yayınlanmıştır. Bunlardan ikisine rakam tanıma işlemi için QuLips [116] ve LTS5[117] veri setleri örnek verilebilir. QuLips veri seti çeşitli aşamalardan oluşmaktadır. Veri toplamanın ilk aşamasında 2 kamera ve 2 konuşmacı kullanılmıştır. Kamera 25 FPS ve 720x576 çözünürlüğe ayarlanmıştır. Sesler kameranın dahili mikrofonuyla kaydedilmiştir. Şekil 2.7’de QuLips’e ait kurulumlarına ait bir örnek görüntü verilmiştir. Konuşmacıdan kameralara kadar olan alan, görsel olarak 0 ile 90 derecelik (0 ve 90 dahil) bölümleri 10’ar derecelik artışlarla alt alanlara bölünmüştür. 2 kameradan bir tanesi 0°’de sabit bir şekilde çekim yaparken 2. kamera 10’ar derecede bölünmüş alanlarda hareket ettirilmiştir. Kişi her söylediği ifadeyi 10’ar derecelik bölünmüş alanlara dönerek yeniden tekrarlamıştır. Böylece her açıdan her ifadenin görüntüleri elde edilmiştir. İkinci kameranın her konumu için konuşmacı aynı rakam dizisini 3 defa tekrarlamıştır. Konuşmacının arkasında mavi bir arka plan kullanılmıştır. QuLips’te de tıpkı XM2VTS veri setinde olduğu gibi rakam tanıma amacıyla konuşmacılar ‘0-1-2-3-4-5-6-7-8-9’ ve ’5-0-6-9-2-8-1-3-7-4’ sayı dizilerini söylemiştir. Bu veri setinin önemli özelliği yalnızca 2 kamera kullanılmış olmasına rağmen ifadelerin, açılar arasında kontrollü bir şekilde karşılaştırılabilmesine imkân vermesidir. Çünkü bir ifade aynı kişi tarafından farklı açılarda defalarca söylenmiştir. 10 açının her birisinde her konuşmacı için 180 ve toplamda ise 3600 adet veri toplanmıştır.



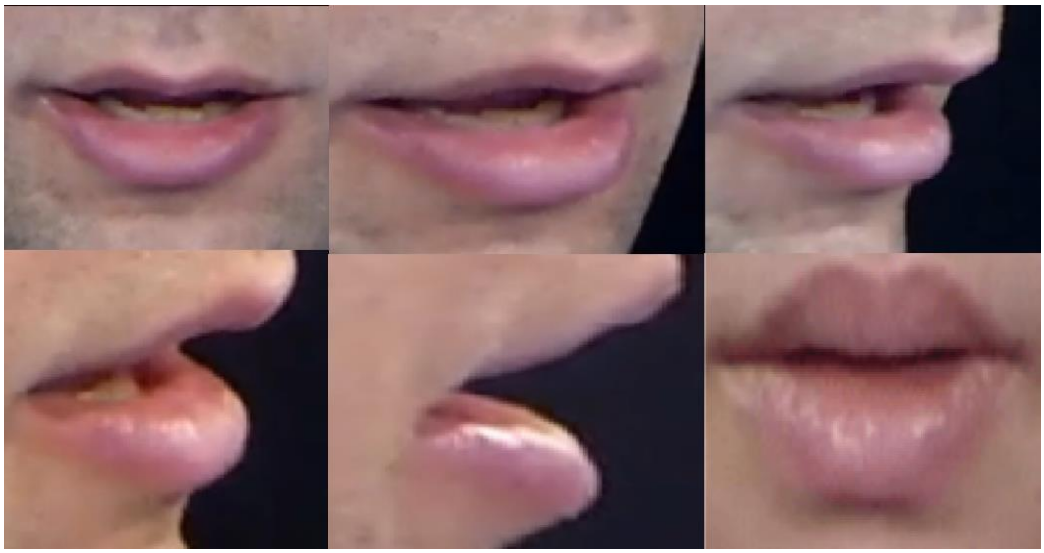
Şekil 2.7. QuLips veri seti için kullanılan düzeneğin planı.

LTS5 veri seti, İsviçre'de "Swiss Federal Institute of Technology Lausanne" üniversitesindeki bir laboratuvar grubu tarafından Fransızcada geliştirilmiştir. Bu veri setinde de yine biri sabit olmak üzere 2 kamera kullanılmıştır. Hareket halindeki kamera QuLips'tekine benzer olarak belirli derecelerde dönmektedir. LTS5'te 30'ar derecelik açılarla konuşmacının önünde döner. Yine QuLips'te olduğu gibi konuşmacılar belirlenen her açıda rakam dizisini 3 defa tekrarlamıştır.

Cümle tanıma için ise İngilizcede çeşitli çoklu görünüm veri setleri yayınlanmıştır: TCD-TIMIT [106], LILiR [100], OuluVS2 [92], HIT-AVDB-II [118], MV-LRS [57] ve AVDigits [93]. Bu veri setlerin çoğu birden fazla kamerayla tek seferde kaydedilmiştir. Böylece farklı açılardan çekilen senkronize görüntüler elde edilmiştir. Örneğin LILiR konuşmacının karşısında 0, 30, 45, 60 ve 90 derecelerinde bulunan 5 kameradan alınmış görüntüleri içerirken, OuluVS2 veri seti de LILiR ile kamera konumları açısından aynıdır. Fakat OuluVS2 veri setinin LILiR'den farkı bunu 5

kamera ile değil 2 kamera ile yapmasıdır. Ayrıca bu iki kameranın çözünürlükleri de farklıdır.

OuluVS2 6 farklı (5 tanesi GoPro Hero 3 ve 1 tanesi de PuxeLink PL-B774U) kamerayla 5 farklı açıdan çekilmiş görüntülerden oluşmaktadır. 5 GoPro kamerasıyla çekilmiş video çözünürlükleri HD (High Definition), 1920 x 1080 piksel ve 30 FPS değerlerine sahiptir. Diğer yandan 1 tane PuxeLink kamerasıyla çekilmiş videolar 640 x 480 piksel çözünürlüğe ve 100 FPS'e sahiptir. 5 adet olan HD çözünürlüğe sahip kameralar HD1, HD2, HD3, HD4, HD5 adlarına sahip olup konuşmacının sırasıyla 0^o (tam karşıdan), 30^o, 45^o, 60^o ve 90^o (profilden) konumlarında bulunur. Yapılan kayıtlarda konuşmacının yüzünü aydınlatmak için kameranın arkasına yerleştirilmiş üç ekstra ışık kaynağı bulunmaktadır. Çekim mekanı, sıradan bir ofis ortamıdır. Rakamlar ve ifadeler slaytlar kullanılarak konuşmacının önündeki bir dizüstü bilgisayar ekranında görüntülenmiştir. Ayrıca slayta konuşmacının hangi modda (normal, fısıldayarak, hızlı, yavaş gibi) konuşması gerektiğine dair ek bilgiler de eklenmiştir. Konuşmacılara her ifadeden sonra bir sonraki slayta geçmek için boşluk tuşuna bastırılmıştır. Boşluk çubuğu vuruş sesi, farklı rakamları ve ifadeleri otomatik ayırmak için kullanılmıştır. OuluVS2'nin orijinal görüntülerinde konuşmacının tüm yüzü görülebilmektedir. Ancak OuluVS2 ayrıca konuşmacının ağız çevresinden ROI görüntüsünü de paylaşmaktadır. Şekil 2.8'de her kare için 5 HD kameradan çıkarılarak sağlanan örnek bir ROI görüntüler kümesi verilmiştir. Her cümlenin veya karenin ROI boyutları farklılık gösterebilir. Yakından ve tam karşıdan çekilmiş görüntülerin ROI'si genelde yatay ve uzundur. Yakın profilden alınmış görüntülerin ROI'si dikey ve uzundur. OuluVS2'deki ROI görüntüleri için istatistikler Çizelge 2.3'te verilmiştir. Bu istatistiklere bakıldığında cümlelere ait ortalama kare sayısı, rakamlara ait olanlardan daha azdır [56,92,93].



Şekil 2.8. OuluVS2 veri setinden örnek görüntüler.

Çizelge 2.3. OuluVS2 veri setine ait istatistikler.

		Rakam Seti			Cümle Seti		
		Min.	Maks.	Ort.	Min.	Maks.	Ort.
HD1	Genişlik	162	294	208	158	262	201
	Yükseklik	64	218	125	64	196	116
HD2	Genişlik	138	240	183	130	214	175
	Yükseklik	64	190	122	64	186	114
HD3	Genişlik	130	230	180	124	210	171
	Yükseklik	64	198	123	64	184	114
HD4	Genişlik	100	204	146	94	184	137
	Yükseklik	64	186	118	64	168	109
HD5	Genişlik	64	164	97	64	136	89
	Yükseklik	86	190	133	78	188	126
Kare Sayısı		83	297	161	8	20	36

Benzer şekilde TCD-TIMIT ve HIT-AVDB-II veri setleri, biri önden görünümde (frontal view) sabitlenmiş ve diğeri TCD-TIMIT için 30^o de sabitlenmiş, HIT-AVDB-II için ise 30^o, 60^o ve 90^o'de dönebilen 3 kameralı kayıtlar içerir. Çinlilerin yapmış olduğu HIT-AVDB-II veri setinin farklı bir yönü bulunmaktadır. O da çoklu dilde oluşturulmasıdır. Çince ve İngilizce şiirler, tekerlemeler, rakamların yanı sıra Yunan alfabesi ve müzikleriyle ilgili videolar da bulunur. AVDigits veri seti, biri önden görünümde, diğeri 45^o'de sabitlenmiş ve üçüncüsü de tam profil görünümünde

sabitlenmiş toplamda üç kamera ile yüksek çözünürlüklü kayıtlar içerir. Son olarak, MV-LRS veri seti, insanların birbirleriyle sohbet ettiği ve bu nedenle yandan çekilmiş görüntülerin oransal olarak daha yüksek olduğu konu olarak birbirinden çok farklı BBC televizyon programlarına ait kesitlerden oluşur. Bu kesitlerden görüntüler rastgele olduğu için açılar sabit değildir, 0^o ve 90^o arasında herhangi bir değer olabilir.

Çizelge 2.4. Çoklu görünümlü veri setlerinin listesi ve özellikleri.

Veri Setinin Adı	**Dili**	**Görevi**	**Konuşmacı**	**Sınıf**	**İfade Sayısı**	**Görüş Açısı (°)**
CUAVE	İngilizce	Rakam	36	10	7,000	-90, 0, 90
AVICAR	İngilizce	Cümle	100	1,317	59,000	4 farklı açı
CMU	İngilizce	Kelime	10	150	15,000	0, 90
IBMSR	İngilizce	Rakam	38	10	1,661	-90, 0, 90
HIT-	İngilizce,	Cümle	30	11	1,980	0, 30, 60, 90
QuLips	İngilizce	Rakam	2	10	3,600	0, 10, 20, ..., 90
LILiR	İngilizce	Cümle	12	200	2,400	0, 30, 45, 60, 90
LTS5	Fransızca	Rakam	20	10	180	0, 30, 60, 90
OuluVS2	İngilizce	Cümle	53	540	2,120	0, 30, 45, 60, 90
TCD-	İngilizce	Cümle	62	6,913	13,826	0, 30
MV-LRS	İngilizce	Cümle	3,783	14,960 †	74,564	0-90 arasında değişir.
AVDigits	İngilizce	Rakam	53	10	795	0, 45, 90
		Söz	39		5,850	

BÖLÜM 3

OTOMATİK DUDAK OKUMA SİSTEMLERİ

Bu bölümde 2005 ve 2022 yılları arasında yayınlanan otomatik dudak okuma sistemleri üzerine yapılan araştırmalar incelenmektedir. Şekil 1.3, 2005 ve 2022 yıllar arasında bu alanda yıl bazlı yayınlanan toplam makale sayılarını göstermektedir. Bu zaman diliminde çalışmaların hızla arttığı görülmektedir. 2022 yılı henüz bitmediği için 2022'ye ait sayılar tam olarak verilememektir. İlerleyen bölümlerde görülebileceği üzere derin öğrenme mimarilerinin artan gelişimi gözlemlenmektedir. Derin öğrenme mimarilerinin bu alanda sıkça kullanılmasıyla, geniş ölçekli veri setlerinin yayınlanması aynı döneme denk gelmektedir. Bunun sebebi de derin öğrenme mimarilerinin çokça veriye ihtiyaç duymasıdır. Şekil 3.1, Şekil 3.2, Şekil 3.3, Şekil 3.4'teki grafikler, dudak okuma sistemlerinin ana özelliklerini özetlemektedir. Bu tablolarda ortaya çıkan bir diğer sonuç ise 2015 yılından itibaren dudak okuma sistemlerinin ezici bir çoğunlukla derin öğrenme mimarilerine kaymasıdır. Bu sebeple de derin öğrenmeden önceki (geleneksel olarak adlandırılan) yaklaşımlar ve derin öğrenme mimarilerini kullanan çalışmalar ayrı alt bölümlerde değerlendirilecektir.

Çalışmalar incelendiğinde çok fazla dudak okuma dışında teknik detaylar bulunmaktadır. Çünkü dudak okuma işlemi yapay zekâ, görüntü işleme, sınıflandırma gibi birçok alanla sıkı bir ilişki içindedir. Tez kapsamında geçmişte yapılan çalışmalar incelenirken dudak okumaya özgü yönlere odaklanılmıştır. Örneğin yüz analizi veya görüntü işlemede ön işleme aşaması gibi daha detay sayılabilecek basamaklar göz ardı edilmiştir. Tipik bir otomatik dudak okuma sistemi 3 aşamadan oluşmaktadır.

- Dudak konum tespiti.
- Görsel özelliklerin çıkarılması.
- Görüntü dizilerine (video) göre sınıflandırma.

İlk aşama olan dudak konum tespitine literatürde yapılan çalışmalar ve yöntemler kıyaslanırken çok fazla değinilmeyecektir. Bunun sebebi de yüz tespiti (face detection) konusu çok klasik ve görüntü işlemeye doğrudan bağlı bir konudur. Tez kapsamında geliştirilen yöntemdeyse bu konu dahil diğer konular incelenmiş olup diğer çalışmalarla benzerlikler ve aralarındaki farklar da listelenmiştir. İkinci aşama olan özellik çıkarmanın amacı, belirli bir anda veya karede görsel veriyi parametrelere döküp vektörel hale getirmektir. Üçüncü aşama olan sınıflandırma, ikinci aşamada kodu çözülen ve sayısal verilere dökülen değerlere dudak okuma gibi zaman serisi veriler için zamansal kısıtlamaları dahil ederek görsel özellikler ile konuşulanları eşleştirmeyi amaçlar. Sınıflandırma aşaması oldukça önemlidir. Çünkü sınıflandırmanın yanında yaptığı bazı ekstra işler de vardır. Örneğin videolardan gelen görüntülerde yer alan gürültülere karşı modelin yanılmamasını sağlamalıdır. Ayrıca "kemal" ve "keman" gibi görsel olarak benzer konuşma birimleri veya söz öbekleri arasındaki ayrımı yapmaya yardımcı olmalıdır. Otomatik dudak sistemlerinin incelendiği bu bölümün geri kalanında 2. ve 3. aşama incelenmektedir: "Özellik Çıkarma" ve "Sınıflandırma".

Dudak okumada ana adımlardan olan 2. adımda özellik çıkarımı genelde 4 farklı yöntemle yapılmaktadır [21, 201].

- Görüntü Tabanlı (Image Based)
- Hareket Tabanlı (Motion Based)
- Geometrik Özellik Tabanlı (Geometric Feature Based)
- Model Tabanlı (Model Based)

Örneğin görüntü tabanlı yaklaşımlarda gri seviyede olan bir görüntü ya doğrudan ya da PCA ve DCT gibi bazı görüntü dönüşümlerinden sonra bir özellik vektörü olarak kullanılır. Dudağın hareketine dayalı sistemlerde derin öğrenme yöntemleri veya optik akışa (optic flow) yönelik sistemler bulunur. Geometrik özelliklere göre oluşturulan sistemlerdeki yaklaşımlar genişlik, yükseklik, alan ve en boy oranı gibi ağzın belirli geometrik özelliklerine dayanmaktadır. Model tabanlı yaklaşımlar için ağzın şeklini ve dokusunu birlikte ele alarak karakterize eden Active Appearance Model gibi klasik

modeller veya derin öğrenmede CNN tabanlı sistemler oldukça sık kullanılmaktadır. Bölüm 3.1 ve 3.2'te bu konulara değinilmiştir.

Geleneksel otomatik dudak okuma sistemleri Bölüm 3.1'de ve derin öğrenme tabanlı olanlar da Bölüm 3.2'de incelenmektedir. Her iki bölümde de geliştirilen sistemleri, hedefledikleri görev (örneğin harflerin, rakamların, kelimelerin veya cümlelerin tanınması) açısından sistematikleştirerek en sık kullanılan veri setlerinde yayınladıkları performansları karşılaştırıp sayısal verilere dayalı bir analiz yapılmaktadır. Yapılan çalışmalarda kullanılan veri seti de en az model kadar önemli olduğu için yöntemlere ait sonuçların farklı veri setlerinde, değişken sayıda konuşmacı, kelime havuzu, dil vb. ile farklı tanıma görevleri için yayınlandıkları göz önüne alındığında sağlıklı bir karşılaştırma için bu yöntem seçilmiştir. Özellikle Bölüm 3.2'de otomatik dudak okuma sistemleri için en popüler derin öğrenme mimarilerini temel alan metotları ve onların çeşitli varyasyonları karşılaştırılmaktadır. Ek olarak, bazı çalışmalardaysa dudak okuma sistemlerinde temel derin öğrenme mimarilerine yine derin öğrenme tabanlı olmak üzere klasik modellerden farklı alternatif sistemler de önerilmiştir. Bu tarz çalışmalar da incelenmektedir.

3.1. GELENEKSEL OTOMATİK DUDAK OKUMA SİSTEMLERİ

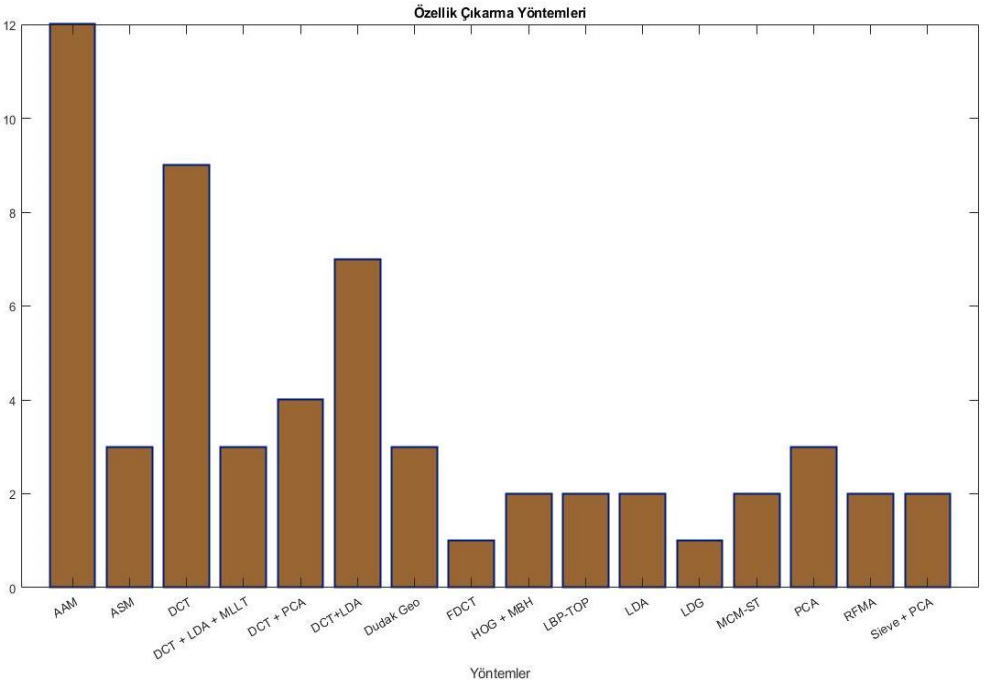
Geleneksel otomatik dudak okuma sistemleri, yüzün algılanması, ağız ve çevresini oluşturan bölgenin tespit edilip segmente edilmesiyle başlar. Bu ön işleme adımları bir kenara bırakılarak, konuşmacıların dudaklarının yeri belirlendikten sonra özellik çıkarma yöntemleri uygulanır. Bununla birlikte dudak okuma açısından hangisinin en iyi özellik çıkarma tekniği olduğu konusunda yapılan araştırmalarda henüz fikir birliği bulunmamaktadır. Ayrıca örneğin dudakların anlık konumunda mı yoksa hareketlerinde mi daha fazla ayırt edici veri olup olmadığı konusunda belirsizlikler vardır [119,120,121]. Bu yüzden birçok araştırmacı, görüntü dönüşümlerine (örneğin Discrete Cosine Transfrom), dudakların hareketine (optik akış), geometrisine (ağzın genişlik, yükseklik gibi parametreleri) ve istatistiksel modellere (Active appearance model) dayalı farklı görsel özelliklere sahip otomatik dudak okuma sistemleri önermiştir [11,22,122,123,124,125,126,127]. Buna karşılık, çoğu geleneksel dudak

okuma sistemi, özellik vektörlerini, söz veya konuşma birimlerine ayırmak/sınıflandırmak için Hidden Markov Model'i kullanır.

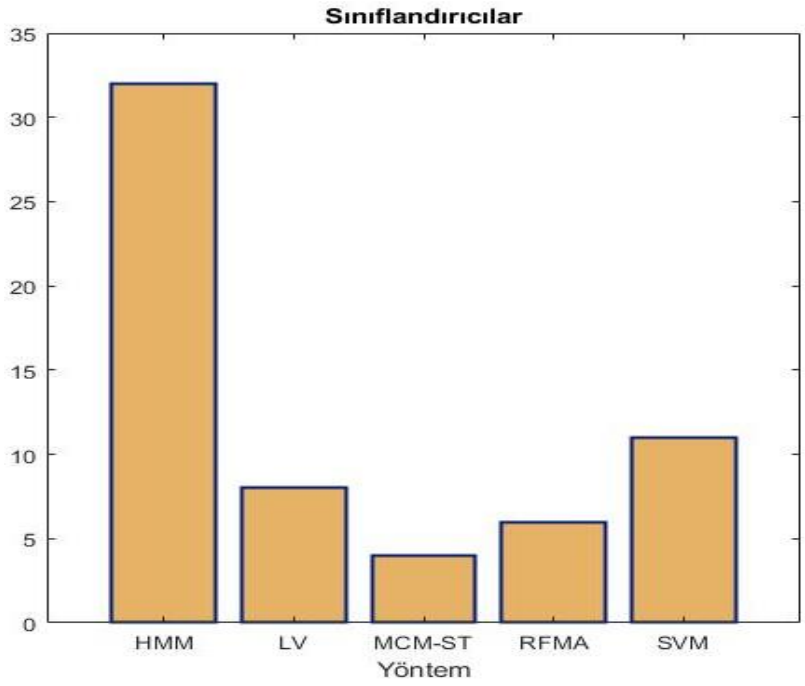
3.1.1. Harf ve Rakam Tanıma Yöntemleri

2000'li yıllardan günümüze kadar yapılan çalışmalarda rakam ve harf tanıma için geliştirilmiş 25 civarı sistem bulunmaktadır. Şekil 3.1, 3.2, 3.3 ve 3.4'te verilen grafikler incelendiğinde çoğu geleneksel sistemin görüntü dönüşümlerine [9,128,129,91,130] veya şekil tabanlı modellere [139,131,132,133,7] dayalı özellik çıkarma tekniklerini kullandığı görülmektedir. Şekil 3.1'de verilen grafikte, her bir özellik çıkarma tekniğinin rakam veya harf tanımayı amaçlayan otomatik dudak okuma sistemlerine kaç defa entegre edildiği gösterilmektedir. Aynı şekilde Şekil 3.2'de özellik çıkarma tekniklerinin de hangi oranlarda kullanıldığı gösterilmektedir. Bu bölümdeki yöntemlerin incelenmesi 3 aşamada yapılır. Birinci aşamada geliştirilen yöntemlerin hangi veri setleri üzerinde kıyaslanacağının belirlenmesidir. İkinci aşamada mimarinin temelleri ve basamakları incelenir. Üçüncü aşamada da yöntemler, performans açısından bölüm 3.2.4'te kıyaslanmaktadır.

Şekil 3.1'de, en çok kullanılan özellik çıkarım yöntemlerinden AAM, 2D-DCT, DCT veya DCT'nin LDA veya PCA gibi diğer dönüşümlerle birlikte kullanıldığı kombinasyonlarının olduğu görülebilir. Geleneksel yöntemlerde özellikle AMM özellik çıkarma için sık kullanılan yöntemlerin başında gelir. AAM modelinin böylesine sık kullanılmasının sebebi, ağız şekli ve hareketlerini ayrı ayrı kodlayabilmesidir. Bu sebeple de dudak okuma sistemlerinin doğruluğunu arttırdığı tespit edilmiştir [100,143]. AAM modeline ait çalışma prensipleri ve diğer detaylar ilgili çalışmada anlatılmaktadır [159]. Öte yandan, Şekil 3.2'de tüm rakam veya harf tanıma için en çok kullanılan sınıflandırma yöntemin, Hidden Markov Model olduğu görülmektedir. Destek Vektör Makineleri veya 1977 yılında Canonical Correlation Analysis yöntemine alternatif olması için yayınlanmış istatistiğe dayalı bir yöntem olan Redundancy Analysis gibi yöntemler de az da olsa kullanılmıştır.



Şekil 3.1. Harf ve rakam tanımadaki çalışma sayılarında özellik çıkarım yöntemlerine ait çalışma sayısı (2005 – 2002).



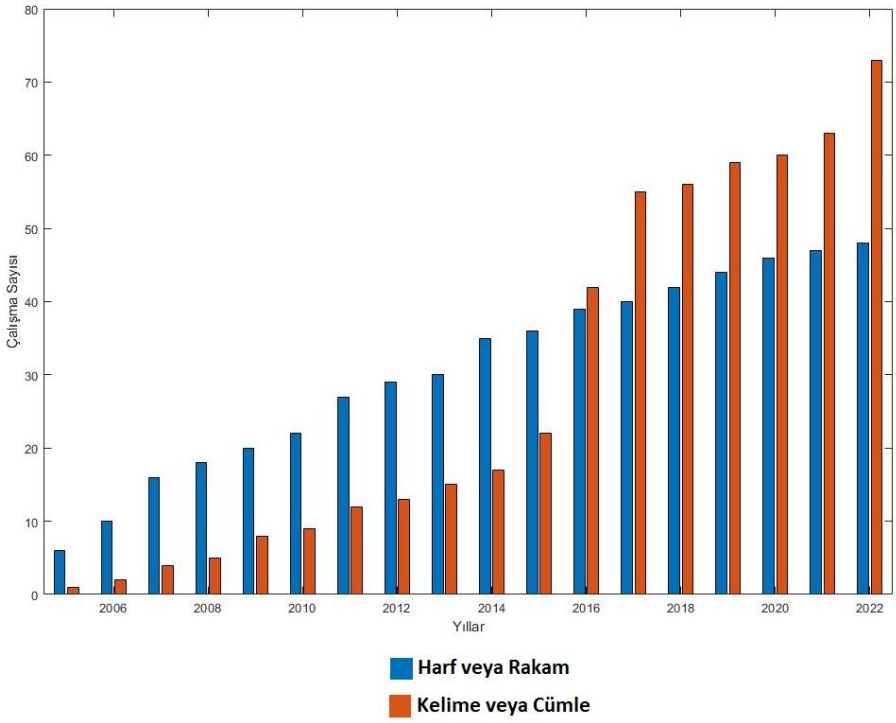
Şekil 3.2. Harf ve rakam tanımadaki çalışma sayılarında sınıflandırıcılara ait çalışma sayısı (2005 – 2002).

Çalışma sayılarına ait grafikler incelenirken [55]'te olduğu gibi bir çalışmada birden fazla sistem ve yöntem önerilebildiğini de unutmamak gerekir. Rakam veya harf tanımaya yönelik yöntemlerin çeşitliliği göz önüne alındığında, bunları performans açısından karşılaştırmak için çalışmaların detaylarına da bakmak gerekir. Karşılaştırmayı yapabilmek için öncelikle aynı veri setini kullanan yöntemler ve çalışmaların değerlendirilmesi sağlıklı olacaktır. Bu sebeple de rakam ve harf tanımada en yaygın kullanılan XM2VTS, CUAVE veya AVLetters2 gibi veri setleriyle çalışan yöntemler karşılaştırılmıştır. Bazı çalışmalarda da geliştirilen model birden fazla veri seti üzerinde çalıştırılmıştır. Örneğin [140]'teki model OuluVS2, CUAVE, AVLetters ve AVLetters2 veri setlerinin tamamında çalıştırılmıştır. [129,131,132,134,135,136,137]'de sunulan yöntemler CUAVE veri seti üzerinde test edilmiştir. Bu yöntemlerin doğruluk oranı, %50 ve %80 arasında değişen değerlere sahiptir. Özellik çıkarım yöntemi olarak da 5 tanesi HMM kullanırken, bir tanesi DCT [134] ve bir tanesi de LDA [129] yöntemini kullanmıştır. DCT yöntemiyle çalışan sistemde başarı oranı %53.12 iken LDA yöntemiyle çalışan sistemin başarı oranı %60 olarak verilmiştir. Benzer şekilde, Estellers tarafından sunulan modelde de DCT özellik çıkarım yöntemi kullanılmış olup %60.4 doğruluk oranı elde edilmiştir [137]. Buna karşılık, Papandreu ve arkadaşları tarafından sunulan her iki mimaride de AMM modelini kullanılmış ve sırasıyla %83 ve %75.7 WRR doğruluk oranına ulaşılmıştır [131,132]. %83 ile elde edilen sonuç bu veri setinde bildirilen geleneksel modeller arasında ulaşılmış en iyi sonuçtur.

Bununla birlikte Pachoud ve arkadaşları tarafından önerilen otomatik dudak okuma sisteminde, keypointlerin tespiti için kullanılan SIFT tanımlayıcıları (SIFT descriptor) ve yerel yer değiştirmeler algoritması olan MCM-ST kullanılmıştır. Her iki yöntemde de yaklaşık olarak %80 WRR elde edilmiştir. Yöntemde SIFT tanımlayıcılarıyla her kare üzerinde dudaklara ait keypointler tespit edilmiştir. Daha sonra bu keypointlerin sonraki karelerdeki yer değişimlerine bakarak bir özellik çıkarımı yapılmıştır. Bu veri seti için son olarak 2016 yılında Rekik ve arkadaşları tarafından önerilen yöntemde Histogram of Oriented Gradients (HOG) ve Motion Boundary Histograms (MBH) özellik çıkarım yöntemlerinin kombinasyonuyla özellik vektörü elde edilmiştir. Daha sonra da elde edilen özellikler SVM sınıflandırıcısıyla sınıflandırıp %70.1 WRR sonucu elde edilmiştir.

XM2VTS veri seti için Seymour ve arkadaşları geliştirdikleri sistemde sınıflandırıcı için HMM'i baz alarak bununla farklı görüntü dönüşümlerini birleştirmiştir. HMM sınıflandırıcısı ile sırasıyla DCT, PCA, LDA ve FDCT yöntemlerini birleştirmiştir. %85.36 ve %87.89 arasında doğruluk oranları elde edilmiştir. Öte yandan Stewart ve arkadaşları tarafından sunulan otomatik dudak okuma sisteminde, özellik çıkarımı için DCT ve sınıflandırma için HMM kullanılarak %70 WRR sonucu elde edilmiştir [141]. XM2VTS veri seti için en iyi performansı gösteren mimari, DCT yöntemiyle özellikleri elde ettikten sonra HMM ile sınıflandırılarak %87.89 WRR sonucu elde edilmiştir [9].

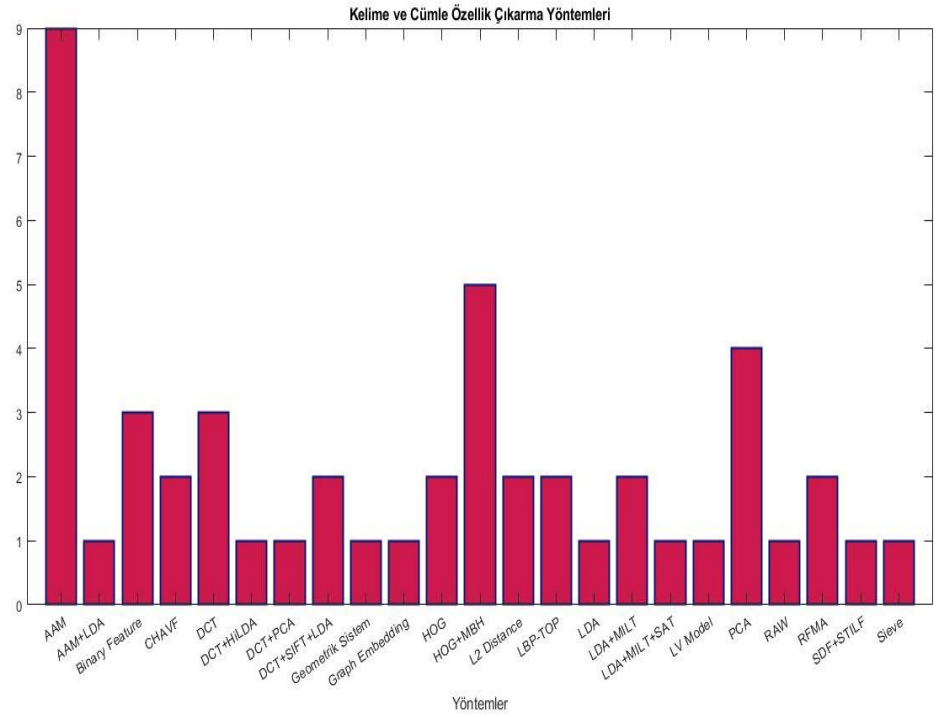
Son olarak harf veya alfabe tanıma için AVLetters2 en çok kullanılan veri setlerinden biridir. AVLetters2 için en fazla %91.8 WRR sonucuna ulaşabilen bazı geleneksel yöntemler önerilmiştir [7,138,139]. HMM tabanlı sistemler için Sieve filtresiyle birleştirilmiş PCA [139] ve AAM [139,7] gibi özellik çıkarma teknikleri kullanılmıştır. En iyi sonuç olarak %91.8 WRR değerini elde eden Random Forest Manifold Alignment yöntemine dayalı bir sistemi Pet ve arkadaşları yaptıkları çalışmayla duyurmuşlardır [138]. Hilder ve arkadaşlarıysa AVLetters2 veri seti için %75.24 WRR sonucunu elde etmişlerdir. Çalışmalar gösteriyor ki DCT özellik çıkarma yöntemi, geleneksel dudak okuma sistemlerini geliştirirken en çok uygulanan yöntemlerden biri olmasına rağmen AAM özellik çıkarımı yönteminin HMM sınıflandırıcısıyla birlikte kullanımı en yüksek WRR değerini vermiştir. Şekil 3.3'te 2005-2022 yılları arasında yapılan çalışma sayılarına ait grafik verilmiştir.



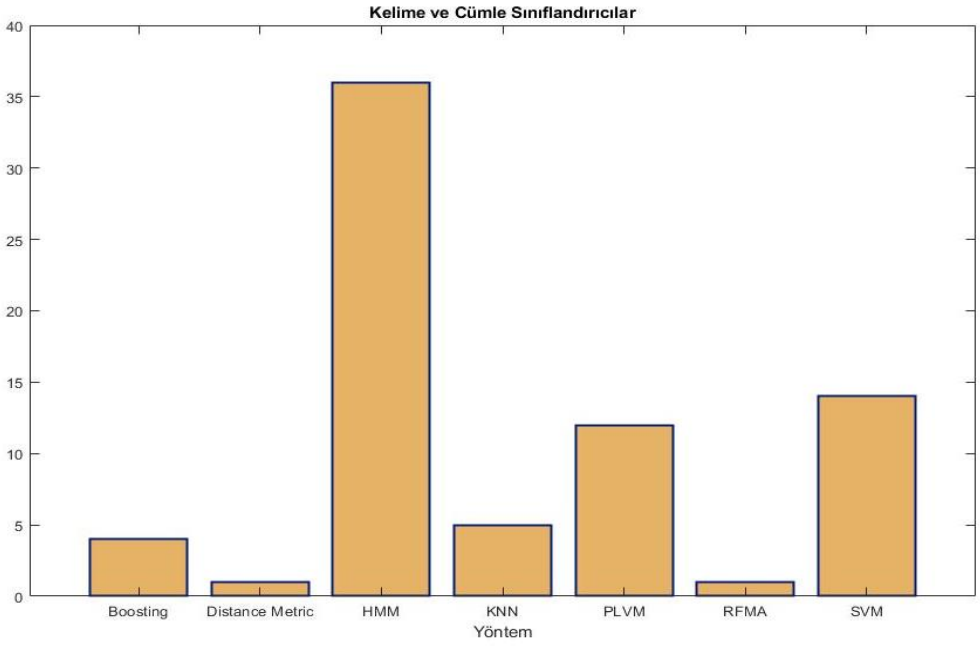
Şekil 3.3. 2005 – 2022 yılları arasındaki çalışmaların kümülatif toplamı.

3.1.2. Kelime ve Cümle Tanıma Yöntemleri

Rakam ve harf tanıma, dudak okuma çalışmalarının başladığı ilk yıllarda oldukça popüler olmasına rağmen bu amaçla ortaya çıkan modeller kelime veya cümle tanıma gibi daha karmaşık görevler için yetersiz kalmıştır. Bu nedenle de harf ve rakam tanıma için geliştirilen sistemlerin güncel problemlere uygulanabilirliği çok kısıtlıdır. Şekil 3.3'te 2005'ten 2022'ye kadar rakam veya alfabe ve kelime veya cümle tanımayı hedefleyen otomatik dudak okuma sistemlerinin sayısı belirtilmiştir. Şekilde belirtilen sayısal veriler, kontrollü (daha önceden bilinen ve kısıtlanmış) veri setleri üzerinden daha kolay tanıma görevlerini yapmaya çalışan ilk ve ilkel sayılabilecek sistemlerden, kelime veya cümle tanıma gibi daha karmaşık görevleri yerine getirmekle uğraşan sistemlere doğru belirgin bir eğilim gözlemlenmektedir. Şekil 3.4 ve 3.5'te kelime ve cümle tanıma için özellik çıkarım yöntemlerinin ve sınıflandırıcıların kullanılma sayıları verilmiştir.



Şekil 3.4. Kelime ve cümle tanıma için sık kullanılan özellik çıkarım yöntemleri ve sayıları (2005 – 2002).



Şekil 3.5. Kelime ve cümle tanıma için sık kullanılan sınıflandırıcı yöntemler ve sayıları (2005 – 2002).

Bu bölümde Çizelge 3.1'de verilen dudak okumada kelime ve cümle tanımayı hedefleyen geleneksel yöntemler derlenip kıyaslanmaktadır. Bu işlem, Bölüm

3.1.1'dekine benzer şekilde yapılacaktır. Önce sistemlerin temel basamakları detaylandırılır. Sonra da performanslarına göre kıyaslanır. Şekil 3.4 ve 3.5'te sırasıyla her bir özellik çıkarma yönteminin veya sınıflandırma tekniğinin kelime ve cümle tanımayı hedefleyen dudak okuma sistemlerine kaç çalışmada entegre edildiği gösterilmektedir. Şekil 3.4'te en çok kullanılan görsel özelliklerin PCA, DCT ve AAM gibi rakam ve harf tanımada kullanılanlara benzer yöntemler olduğu rahatlıkla görülebilmektedir. Burada asıl dikkat edilmesi gereken nokta, aslında bu yöntemler tek başına kullanıldığı gibi birbirleriyle birleştirilerek özellik çıkarma aşamasında da sıklıkla kullanıldığıdır. Örneğin PCA ve DCT gibi yöntemler tek başlarına kullanıldığı gibi birlikte kullanımı da mümkündür. Rakam veya harf tanıma sistemlerinde kullanılan özellik çıkarım yöntemlerindeki çeşitlilik havuzu, kelime ve cümle tanımada oldukça artmaktadır. Örneğin şekillerin farkına dayalı özellik tespiti yapan "Shape Difference Feature (SDF)" yöntemi, "Three Orthogonal Planes - Üç Ortogonal Düzlem- (LBP-TOP)" yöntemi, "Spatio-Temporal Lip Feature (STLF)" yöntemi bunlardan bazılarıdır. Şekil 3.5'te belirtildiği üzere sınıflandırıcılar açısından HMM yine en çok kullanılan sınıflandırıcıdır. Özellikle SVM sınıflandırıcısı başta olmak üzere HMM'ye alternatif sınıflandırıcıların kullanımında artış olmasına rağmen rakam veya harf tanımaya benzer bir eğilim devam etmektedir.

Performans değerlendirmesi açısından kelime veya cümle tanıma hedefi için en çok kullanılan veri setleri GRID, OuluVS, OuluVS2 ve RM-3000 olmuştur. GRID veri seti için Lan ve arkadaşları yaptıkları çalışmada tüm veri setini kullanmak yerine 15 konuşmacıdan oluşan bir alt grubu kullanmıştır. Çalışmalarının merkezine başta AAM olmak üzere DCT, Sieve, PCA gibi diğer özellik çıkarım yöntemlerinin kıyaslanmasını koymuştur. Kelime tanımayı amaçlayan bu sistemde özellik çıkarımını yaptıktan sonra sınıflandırma aşamasında her kelime başına bir HMM, toplamdaysa 52 HMM (51 kelime + konuşmama durumu) kullanmışlardır. En başarılı özellik çıkarım yöntemi olarak AAM yöntemi belirlenmiş olup %65 WRR sonucu elde edilmiştir. Diğer yöntemlerdeyse %40-60 WRR arasında sonuçlar alınmıştır [143]. Buna karşılık, Kolossa ve arkadaşları yaptıkları çalışmada yine GRID veri setini kullanmıştır fakat Lan'in yaptığı çalışmadan farklı olarak tüm veri setini kullanmışlardır. Geliştirilen sistem ise birbirlerine bazı detaylar haricinde oldukça benzerdir. Kolossa da tıpkı Lan'in çalışmasında olduğu gibi özellik çıkarımı için DCT'yi kullanmıştır.

Sınıflandırma için de yine kelime başına HMM'den oluşan benzer bir model önerip %57 WRR sonucuna ulaşılmıştır [144]. Biraz daha illeri yıllarda Wand ve arkadaşları sınıflandırıcı olarak SVM'yi kullanarak özellik çıkarımında 2 yöntemi (PCA ve HOG) karşılaştırmıştır. Wand yaptığı çalışmada yine tüm veri setini kullanmak yerine 20 konuşmacıdan oluşan bir alt grup kullanmıştır. Konuşmacıya bağlı bu çalışmada PCA yöntemi için %69.5 ve HOG için ise %71.2 WRR sonucu elde edilmiştir. Konuşmacıya bağlı çalışmada, sınıflandırıcılar için eğitim ve test verileri her zaman aynı konuşmacıdan alınır ve sonuçlar tüm konuşmacılar üzerinden ortalama alınarak hesaplanır.

OuluVS veri seti için literatür incelendiğinde 9 farklı mimari sunulmuştur [99,136,138,145,146,147,148,149,150]. Bu veri setinde değerlendirilen otomatik dudak okuma sistemleri için çeşitli özellik çıkarım yöntemleri kullanılmıştır. Sınıflandırıcı olarak ise çoğu çalışmada SVM kullanılmıştır. Rekik ve arkadaşları yaptıkları çalışmada özellik çıkarımı için MBH ve Spatio-Temporal HOG yöntemlerinin kombinasyonunu kullanarak elde ettikleri özellikleri SVM ile sınıflandırıp %68.3 WRR sonucunu elde etmiştir [136]. Sui ve arkadaşlarıysa Cascade Hybrid Appearance Visual Feature (CHAVF) adını verdikleri yeni bir özellik çıkarım yöntemi önermişlerdir. Bu yöntemin altında ufak değişikliklerle LBP-TOP ve DCT yöntemlerinin birlikte kullanılması yatar. Konuşmacıya bağlı testlerde SVM ile sınıflandırma sonucunda %68.9 WRR değeri elde edilmiştir [150].

OuluVS veri seti için yapılan 2 çalışma ilginç sonuçlar göstermiştir. Barnard ve arkadaşları LBP-TOP yönteminde elde ettiği özellikleri SVM ile sınıflandırmıştır. Sonuç olarak %62.4 WRR elde edilmiştir [99]. Buna karşılık olarak Zhou ve arkadaşları da LBP-TOP ile özellikleri elde edip SVM ile sınıflandırmıştır. Bu da demek oluyor ki 2 aynı yöntem kullanılmıştır [146]. Fakat Zhou, yaptıkları çalışmada Barnard'ın yaptığı çalışmadakiyle aynı yöntemleri ve aynı veri setini kullanmış olmasına rağmen %81.3 WRR bildirmiştir. Aradaki bu büyük %20'nin üzerindeki farkın sebebi, Zhou'nun yaptığı çalışmada girişte verilen orijinal videoyu "Curve Matching" adı verilen yöntemle bir eğri üzerine eşleyerek video sinyallerini belirli aralıkta normalize etmesidir. Daha sonra girişte verilen orijinal videodakiyle aynı sayıda kareye sahip video dizileri üretmek için yeniden örnekleyen bir işlem

gerçekleştirmiştir. Bu yöntemle de %20'lik bir farka sebep olmuştur. Daha sonra Ong ve arkadaşları 2 farklı çalışma yaparak 2 ayrı sistem önermişlerdir [147,148]. Birinde TGD-Boosting (Temporal Gradient Descend Boosting) yöntemini [147], diğerinde de SP-Boosting (Sequential Pattern Boosting) yöntemini [148] kullanarak sistemler önermişlerdir. Sırasıyla da %65.6 ve %86.2 WRR sonucunu bildirmişlerdir.

Pei ve arkadaşları OuluVS veri setinde RFMA'ya dayalı uçtan uca bir sistem önermiş olup bu veri setinde şimdiye kadar geleneksel yöntemlerle elde edilmiş en yüksek performans olan %87.7 WRR sonucuna ulaşmışlardır [138]. Yaptıkları çalışmada sadece OuluVS veri setini kullanmamışlardır. Bunun yanında KinectVS, AVLetters, AVLetters2, CUAVE gibi başka veri setlerini de kullanıp yöntemlerini test etmişlerdir. Sınıflandırıcı olarak da Unsupervised Random Forest yöntemini kullanmışlardır.

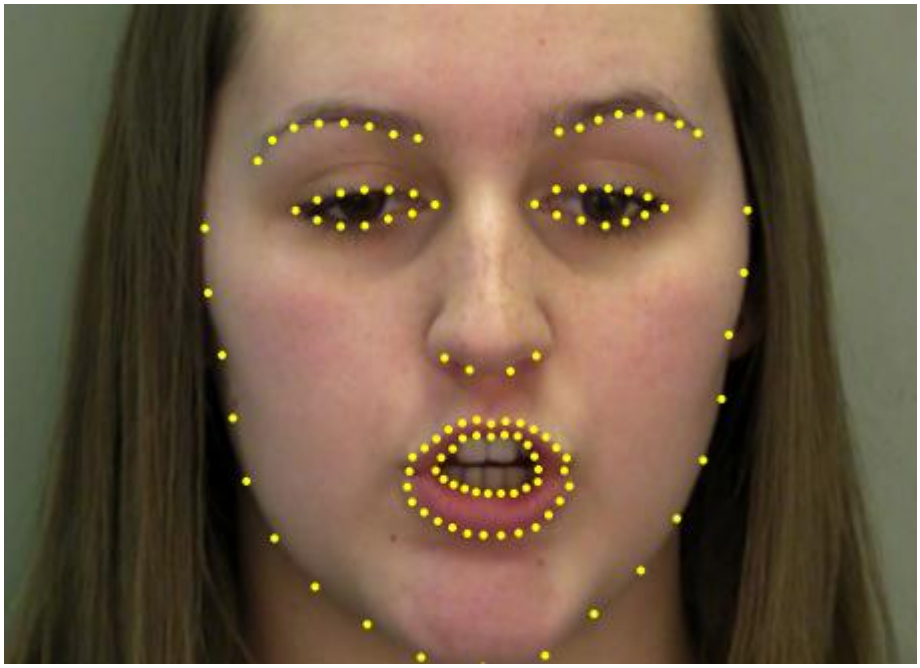
OuluVS veri seti için bir başka çalışmada da Zhou ve arkadaşları tarafından 2 sistem önerilmiştir [145,149]. Yaptıkları ilk çalışmada video dinamiklerini ve özelliklerini yakalayabilmek için "Graph Embedding" adını verdikleri bir yöntem önermişlerdir. Bu yöntemde konuşmacının söylediği aynı ifadelerin video kareleri arasındaki zamansal bağlantıları karakterize etmek için, kareler üzerinde sürenin (zamanın) yanına yeni bir uzaklık metriği tanımlanır. Sonrasında kareler arasındaki uzaklığa dayalı video dinamiklerini temsil etmek için grafikler oluşturulmuştur. Fakat burada diğer çalışmalardan farklı olarak bu mesafelerin hesaplanmasına yardımcı olması için videolardaki ses bilgisi de kullanılmıştır [145]. İkinci çalışmalarında da videolardaki görüntülerin dizisini oluşturmak için Latent Variable (LV) adını verdikleri bir model önerilmiştir [149]. İlk çalışmalarında [145] söylenen bir ifadeyi dışarda tutarak (leave-one-utterance-out yöntemi) çapraz doğrulama (cross validation) yapıp %90.6 WRR elde edilmiştir. İkinci çalışmalarında da [149] bu sefer bir konuşmacının söylediği tüm ifadeleri dışarda tutup çapraz doğrulama yaparak %74 WRR sonucunu elde etmişlerdir. Böyle yapılmasının amacı sistemin daha önce görmediği veriler üzerindeki performansının sağlıklı bir şekilde ölçülebilmesini sağlamaktır. Zhou ve arkadaşlarının yaptığı bu her 2 çalışmada da veri setinin tamamı kullanılmamıştır. Bu yüzden aldıkları sonuçlar her ne kadar yüksek olsa da en yüksek sonuçlar olarak değerlendirilmemiştir. OuluVS veri setine ait çalışmalar genel olarak incelendiğinde

değişik bir durum söz konusudur. Diğer çalışmaların ve veri setlerinin birçoğunda sınıflandırıcı olarak kullanılan HMM, bu veri setinde pek fazla tercih edilmemiştir.

Kullanılan veri setlerden birisi de OuluVS2 veri setidir. Wu ve arkadaşları [153] konuşulanları çözümleyebilmek için SDF (Spatiotemporal Deep Feature) ve STLF yöntemleriyle elde ettiği özellikleri SVM ile sınıflandırmıştır. Bu teknikle %55 WRR elde edilmiştir. Buna karşılık Lee ve arkadaşları yaptıkları çalışmada [154] 3 farklı sistem önermişlerdir. Geliştirdikleri sistemleri de veri seti imkân verdiği için 5 farklı açıdan (tam karşıdan, 30^o, 45^o, 60^o, profilden) ayrı ayrı test edip tüm sonuçlarla birlikte ortalama WRR değerini de vermişlerdir. Sınıflandırıcı olarak HMM tabanlı sistemlerin ilkinde DCT – PCA kombinasyonuyla özellik çıkarılmıştır. Sonuç olarak da sırasıyla her bir açıda %63, %62, %62, %63, %57 WRR değerlerini elde etmişlerdir. Önerdikleri ikinci sistemde de yine HMM sınıflandırıcısı temelinde DCT – HiLDA özellik çıkarım yöntemlerinin kombinasyonunu kullanarak %74, %72, %73, %73, %68 WRR değerlerini elde etmişlerdir. Üçüncü ve son sistemdeyse Latent Variable adlı modele, özellik olarak gelen görüntülerin ham şekildeki piksel değerleri verilip sınıflandırılmıştır. Burada da sırasıyla %73, %75, %76, %75, %70 WRR sonuçları elde edilmiştir. Çalışmaya ait tüm test sonuçları incelendiğinde profilden alınan görüntülerdeki başarı oranı tüm sistemlerde en düşük WRR değerine sahiptir. Bu çalışmada ayrıca derin öğrenme yöntemleriyle de önerilmiş sistemler bulunmaktadır. Derin öğrenme tabanlı sistemlerin incelendiği bölümde bu çalışmaya ait detaylara değinilmektedir.

1000 farklı kelimeden oluşan tek erkek konuşmacılı RM-3000 veri seti için, Thangthai ve arkadaşları geliştirdikleri AAM + HMM tabanlı ilk sistemde görüntüdeki iç ve dış dudak şekillerini içeren AAM özelliklerini çıkarmışlardır. AAM özelliklerinin toplam vektör boyutu 1x23 olarak verilmiştir. Bu vektörün ilk 12 elemanı dudağa ait şekil verisidir. Geri kalan 11 eleman ise görünüm bilgisidir. Bunun yanında yine sınıflandırıcı olarak HMM kullandıkları ikinci sistemlerinde özellik çıkarım yöntemi olarak HiLDA kullanılmıştır. AAM kullandıkları ilk sistemde %77.49 WRR sonucunu elde edilmişken HiLDA ile oluşturulan ikinci sistemdeyse %84.67 WRR sonucunu elde edilmiştir. Newman 2011 yılında yazdığı tezinde o yıllarda teknolojik sayılabilecek “Visual Phone” adı verilen bir cihazda telefonla dudak okuma işlemini

gerçekleştirmeye çalışmıştır. United Nations 1 ve United Nations 2 veri setleri üzerinden aynı videolar için 3 farklı çözünürlük tipi belirlemiştir (1920 x 1080, 1080 x 720 ve 640 x 360). Geliştirdikleri sistemden ziyade sonuç kısmında elde ettikleri doğruluk oranında çözünürlüğe bağlı anlamlı bir fark oluşmadığına ulaşılmıştır [156]. Bu sebeple de Howell ve arkadaşları [155] da Newman'in tezinden [156] vardıkları sonuca istinaden yaptıkları çalışmada, özellik çıkarma ve modelleme süreçlerinin verimliliğini artırmak için veri setindeki tüm görüntüleri 360 x 640 çözünürlüğe düşürmüşlerdir. Elle etiketleme yapmak için her bir kayıt oturumundan 20 - 30 adet kare seçilmiştir. Amaçları, mümkün olduğunca fazla şekil ve görünüm olasılıklarını yakalamak amacıyla ağız hareketlerinin uç noktalarını tanımlayan kareler bulmaktır. Özellik çıkarımı için AAM kullanılmış olsa da diğer sistemlerden farklı olarak "Inverse Compositional Project Out AAM" yöntemini kullanmışlardır. Bu yöntemle birlikte konuşmacıya yüz bölgesinde Şekil 3.6'ta da gösterildiği üzere 111 tane nokta oluşturulmuştur. Bu noktalar yüzün sınırlarını, gözü, kaşları, burnun deliklerini ve dudakların sınırlarını tanımlar. Daha sonra bu noktalar üzerinde ilk olarak sadece HMM ile sınıflandırma yapmışlardır. Burada %75.58 WRR sonucuyla 20500 tane kelimeyi başarıyla tespit etmiştir. Daha sonra geliştirilen yöntemde karar alma aşamasında özellikle konuşma tanımada, sinyal işlemede oldukça sık kullanılan "Weighted Version of the Finite State Transducer (WFST)" eklenmiştir. WFST'lerin önemli bir özelliği videoda olduğu gibi bir giriş etiketi dizisini bir çıkış etiketi dizisine eşleyebilmesidir. WFST, sonlu durum dönüştürücüsünün (FST – Final State Transducer) ağırlıklandırılmış versiyonudur. HMM + WFST sınıflandırıcısı le %76.14 WRR sonucuna ulaşılmıştır.



Şekil 3.6. AAM ile elde edilen 111 nokta.

Lan ve arkadaşları [156] oldukça eski bir yöntem olan temeli 1968 yılında atılmış ve G. Fisher'ın soyadını verdiği "Fisher Phoneme to Viseme Mapping" yöntemini [157] kullanmıştır. Buradaki temel yöntem, AMM ile elde edilen özelliklere LDA yönteminin uygulanmasıdır. Ardından HMM ile sınıflandırılmıştır. Fakat Lan'in yaptığı çalışmanın diğer çalışmalardan farklı bir yönü vardır. Çalışmasında dudak okuma amacıyla geliştirilen sistemlerin gerçek dudak okuma uzmanlarına göre kıyaslandığında nasıl bir performans sağlayacağı konusundaki belirsizliklere değinmeye çalışmıştır. Bu probleme pek değinilememesinin 2 sebebi bulunmaktadır. İlki dudak okuyucuların sayıca azlığı ve ikincisi de çoğu dudak okuma sisteminin yalnızca önemsiz ve dolayısıyla insan konuşmasını çok da kapsamayan verileri işleyip tanımaya çalışmasıdır. Çalışmada hem RM-3000 veri seti hem de kendi oluşturdukları çoklu görünümlü veri setini kullanmışlardır. 6 dudak okuma uzmanı 2 aşamadan oluşan teste katılmıştır. Dudak okuma uzmanları tüm test aşamalarında tek başına ve diğer herkesten bağımsız yer almaktadır. Bu testte RM-3000 veri setindeki sesin olmadığı sadece görüntünün yer aldığı videolar kullanılmıştır. İlk aşamada dudak okuma uzmanlarına bir kişiye ait 10 adet video ve videolarda da ses olmadığı için bu videolarda söylenen ifadeler bir metin olarak verilmiştir. Bu videolarla ile ilgili yapmak istedikleri tüm çalışmaları yapmalarına izin verilmiştir. Böylece dudak okuma

uzmanlarının, konuşmacının tarzını öğrenebilmesi sağlanmıştır. Daha sonra o dudak okuma uzmanına aynı konuşmacıya ait ilk verdiklerinden farklı 10 sessiz video daha verilip bunlarda nelerin söylendiğinin tespit edilmesi istenmiştir. İlk aşamada yapılan bu testin amacı dudak okuma uzmanlarının temel seviyedeki yeteneklerinin ölçülmesidir. Testin sonucu uzmanlara söylenmeden ikinci aşamaya geçilmiştir. Testin ikinci aşamasındaysa dudak okuma uzmanlarına tıpkı bir yapay zekâ sistemini eğitir gibi daha önceki test aşamasında verdikleri 10 video haricindeki 1000 cümleden ve 971 kelimeden oluşan tüm veri setini metinlerle beraber vermişlerdir. Daha sonra da 1. test aşamasındaki 10 videoyu tekrardan çevirmeleri istenmiştir. 6 dudak okuyucudan 4'ü testlerin her 2 aşamasını da sağlıklı bir şekilde tamamlayabilmiştir. Çalışma sonuçları dudak okuyucuları için incelendiğinde tüm dudak okuma uzmanları 2. aşamda tüm veri setini görmesiyle beraber başarı oranını arttırmıştır. Örneğin uzmanlardan biri 1. aşamada %18.31 WRR oranına sahipken ikinci aşamaya geçildiğinde %40.85 WRR oranına sahip olmuştur. Çalışmada geliştirilen sistem %41.08 başarı göstermiştir. Bu oran dudak okuyucuların 4'ünden daha yüksektir. Çalışmalarında nihai sonuç olarak geliştirilen sistemin dudak okuyuculardan daha iyi performans gösterdiğini ve ilerdeyse çok daha iyi sistemlerle aradaki farkın açılacağı belirtilmiştir.

LILiR veri seti için, Bowden ve arkadaşları [11] önerdikleri sistem, Bölüm 2.1.3'te verildiği üzere çok açılı veri setlerinde 30^o en iyi sonucu vermiştir. Fakat yayınlanan çoğu veri setinin tam karşıdan çekilmiş görüntülerden oluşmasından dolayı geliştirilen birçok sistem karşıdan çekilmiş görüntüler üzerine odaklanmıştır. Yaptıkları çalışmada sistem performansında sadece küçük bir kayıpla diğer kamera açılarını optimal olarak eşleştirmek için bir teknik geliştirilmiştir. Kısaca amaç sistemi tüm açılara göre eğitip daha karmaşık hale getirmek yerine sistemi en iyi açı olan 30^o üzerinden eğiterek diğer açılara göre bir bakıma haritalama fonksiyonu gibi çalışacak bir yöntemle eğitilen sistemin diğer açılardaki görüntülere de uyarlanmasını sağlamaktır. Bu yöntemde sistem için optimal açı konusunda yeterli eğitim yapıldığında, daha sonra diğer açılarda kaydedilen videolar için de kullanılması amaçlanmaktadır. LILiR veri seti ilgili bölümde de anlatıldığı üzere 2 adet HD (0^o ve 90^o) 3 adet SD (30^o, 45^o, 60^o) kamera bulunmaktadır. Geliştirdikleri sistemlerinde SD kamera açıları olarak 30^o, 45^o, 60^o açılarının diğer iki açıya (0^o ve 90^o) kıyasla yüksek

karıştırılabilme olasılığı bulunmasından dolayı açı tespitini yapacak yöntemin geliştirilmesi ve test edilmesi için seçilmiştir. Kamera açısının tespitine ait problemi iki ayrı yaklaşımla çözmeye çalışmışlardır. Birinde görüntüler üzerindeki keypointleri AAM, diğerinde de LGO (Local Gradient Orientation) yöntemiyle izlemişlerdir. Hem AAM tabanlı kamera açısının tespitinde hem de LGO tabanlı sistemde %99'un üzerinde bir başarı gösterilmiştir. Geliştirdikleri sistemi tüm veri setine uyguladıklarında ise başarı oldukça düşerek %30.2 WRR değerine ulaşılmıştır.

Almajai [115] tıpkı Lan'in [156] çalışmasında olduğu gibi Phoneme to Viseme Mapping adı verilen eşleştirme yöntemini [157] kullanmıştır. Yaptıkları çalışmanın diğer çalışmadan bazı farkları bulunur. Sistemlerini eğitirken Context-Dependent HMM (İçeriğe Bağlı HMM- CD-HMM) ve Context-Independent HMM (İçeriğe Bağımlı HMM - CI-HMM) yöntemini kullanmıştır. Buradaki temel amaç modelleri eğitirken doğrudan kelime modelleri oluşturmak yerine, fonem modellerini sisteme tanıtmaktır. Daha sonra, kelime modellerini tanımlayabilmek adına her kelimeye karşılık gelen ses birimlerini birleştirmişlerdir. Bu durumda kelime modeli "modelin modeli" olmaktadır. Önerdikleri sistem spesifik olarak birinci ve ikinci dereceden türev özellikleri kullanan tek sesli (monophone) ve monoviseme modellerine dayalı bir CI-HMM önerilmiştir. Ayrıca bunun yanında LDA, LDA+MLLT ve LDA + MLLT + SAT yöntemleriyle gelen özelliklere sahip önceki sistemden farklı olarak tek sesli değil de üç sesli (trifon-triphone) ve triviseme modellerine dayalı CD-HMM önermişlerdir. Bu durumda CI-HMM tek sesli sistemlere dayalı olduğu için içerikten bağımsız olarak çalışmaktadır. CD-HMM ise 3 ses üzerinden tanıdığı için içeriğe bağlı bir şekilde çalışır. Yapılan test sonuçlarında sistemler viseme'e dayalı modelleri kullanmak yerine foneme bağlı modellerin kullanılması durumunda %8'e kadar başarının arttığını ve tüm veri seti için %43 WRR sonucuna ulaşıldığı görülmüştür. Fakat İspanyolcada AV@CAR için yapılmış olan bir çalışma [158] tam tersini iddia etmektedir. İspanyolca için yapılan bu çalışmada fonem–viseme eşleştirmesi çok uzun olmayan kelimeler için en yüksek WRR değerini sağlamıştır. Bu gibi çalışmalardaki çelişkiler gösteriyor ki otomatik dudak okuma sistemlerinin "viseme" üzerine kurulması konusu hala kesinlik kazanmamıştır.

Geliştirilen yöntemlerde geniş ve çok farklı kelime çeşitliliğine sahip büyük veri setleri için 2 farklı modelleme önerilmiştir. Birinde kelimelerin modele gösterilip eğitilmesi sonucunda benzer yapıların öğrenilmesidir. İkinci yaklaşımdaysa kelimeleri eğitmenin o kadar faydalı olmayacağı düşünülerek kelimeleri eğitmek yerine bir fonem veya arda da dizilmiş üçlü fonem yapısıyla eğitmek daha faydalı görünmektedir. Buna gerekçe olarak da bu şekilde yapıldığında veri seti çok büyük olmasa dahi sınıf başına eğitim için mevcut örnek sayısının artırılmasını göstermektedirler.

Sözcük veya cümle tanımayı hedefleyen sistemler özetlenecek olursa her bir veri seti için hem özellikler hem de sınıflandırıcılar açısından farklı mimarilerin önerildiği görülmektedir. Rakam ve harf tanıma sistemlerinin aksine, her bir veri setinde değerlendirilen özelliklerin farklılığı, hangisinin en iyi performans gösterdiği sonucuna varmayı zorlaştırmaktadır. Sınıflandırıcılar açısından da benzer bir durum söz konusudur. GRID veri seti için HMM, OuluVS veri seti için SVM ve OuluVS2 veri seti için LV modelleri en iyi performansları göstermiştir. Ancak, OuluVS veri setinde HMM veya LV modellerine dayalı hiçbir sistem test edilmemiştir. Genel olarak çalışmaların temelinde SVM sınıflandırıcısı bulunmaktadır. OuluVS için değil de OuluVS2 için bazı HMM tabanlı sistemler önerilmiş olsa da performans açısından en iyi sonucu gösteren sistemler bunlar olmamıştır. Bu nedenle, farklı özellik çıkarım yöntemlerinin ve sınıflandırıcıların çalışmalardaki kullanım sıklığını da göz önünde bulundurarak performans açısından adil bir karşılaştırma yapılabilmesi zordur.

Özellikle geleneksel yöntemler için söz konusu olan bir diğer husus ise veri setleri, dil gibi değişken parametreler özellik çıkarım yöntemi ve sınıflandırıcılar üzerinde ciddi performans değişikliklerine sebep olmaktadır. Bu yüzden geleneksel yöntemler için en iyi özellik çıkarım yöntemi veya en performanslı sınıflandırıcı gibi seçimler yapmak güçtür. Aynı şekilde bu sistemler üzerinde kesin genellemeler yapmak da bu sebeple yanlış olacaktır. Bunun yanında geliştirilen sistemler aynı veri seti üzerinde çalışılmış gibi görünse de bazı çalışmalar kullandığı veri setinin tamamını kullanmayıp bir alt veri seti oluşturarak karşılaştırmayı iyice zorlaştırmaktadır. Çizelge 3.1'de gelenek dudak okuma sistemleri listelenmiştir.

Çizelge 3.1. Geleneksel yöntemlerle oluşturulmuş çalışmalar (2005-2022).

Çalışma	Özellik Çıkarım	Sınıflandırıcı	Kullanılan Veri Seti	Amaç	WRR (%)
Shao [164]	DCT	HMM	GRID	Deyim	%58,4
Papandreou [131]	AAM	HMM	CUAVE	Rakam	%83
Ong [147]	Binary feature	TGD-Boosting	OuluVS	Deyim	%65,6
Ong [148]	Binary feature	SP-Boosting	OuluVS	Deyim	%86,2
Fu [162]	LDG	HMM	AVICAR	Rakam	%37,87
Kumar [114]	Mouth geometry	HMM	CMU AVPFV	Kelime	%32,39
Lucey [128]	DCT+LDA	HMM	IBMSR	Rakam	%68,58
Marcheret [163]	DCT+LDA+MLLT	HMM	IBMIH	Rakam	%63
Cox [139]	Sieve+PCA	HMM	AVLetters2	Harf	%83
	AAM	HMM	AVLetters2	Harf	%85
Lucey [134]	DCT+LDA	HMM	CUAVE	Rakam	%53,12
Lucey [91]	DCT+PCA	HMM	IBMSR	Rakam	%66,21
Pachoud [135]	MCM-ST	Prob. Seq. Matching	CUAVE	Rakam	%80
Papandreou [132]	AAM	HMM	CUAVE	Rakam	%75,7
Seymour [9]	DCT	HMM	XM2VTS	Rakam	%87,89
	PCA	HMM	XM2VTS	Rakam	%86,57
	FDCT	HMM	XM2VTS	Rakam	%85,36
	LDA	HMM	XM2VTS	Rakam	%86,35
Wang [133]	ASM	RDA	Own data	Rakam	%88,32
	ASM	HMM	Own data	Rakam	%91,27
Gurban [129]	DCT+LDA	HMM	CUAVE	Rakam	%60
Hilder [7]	AAM	HMM	AVLetters2	Harf	%75,24
Kolossa [144]	DCT	HMM	GRID	Deyim	%57
Lan [143]	Sieve	HMM	GRID	Deyim	%40
	DCT	HMM	GRID	Deyim	%40
	Eigenlips	HMM	GRID	Deyim	%52
	AAM	HMM	GRID	Deyim	%65
Zhao [99]	LBP-TOP	SVM	AVLetters	Harf	%62,8
	LBP-TOP	SVM	OuluVS	Deyim	%62,4

Çizelge 3.1. (devam ediyor)

Çalışma	Özellik Çıkarım	Sınıflandırıcı	Kullanılan Veri Seti	Amaç	WRR (%)
Saitoh [165]	Keypointler arasında L2	HMM	Kendi Veri Seti	Kelime	%68,93
Zhou [145]	Graph Embedding	OuluVS	Deyim	%90,6	
Cappelletta [120]	Optical flow	HMM	VIDTIMIT	Cümle	%57
	PCA				%60,1
Navarathna [166]	DCT+PCA	HMM	AVICAR	Rakam	%25
Zhou [146]	LBP-TOP	SVM	OuluVS	Deyim	%81,3
Chi¸tu [168]	Mouth geometry	HMM	NDUTAVSC	Rakam	%84,24
Estellers [117]	DCT	HMM	CUAVE	Rakam	%60,4
Estellers [193]	DCT+LDA	HMM	Kendi Veri Seti	Rakam	%71
Lan [113]	AAM	HMM	LILiR	Cümle	%33
Lan [156]	AAM+LDA	HMM	LILiR	Cümle	%14.08
Bowden [11]	AAM	HMM	LILiR	Cümle	%30,2
Huang [130]	DCT+LDA	HMM	Kendi Veri Seti	Rakam	%35,2
	DCT+LDA	DBN	Kendi Veri Seti	Rakam	%35,7
Pei [138]	RFMA		AVLetters	Harf	%69,6
			AVLetters2	Harf	%91,8
			OuluVS	Deyim	%89,7
Zhou [149]	Latent variables	Cross correlation	OuluVS	Deyim	%74
Bear [170]	AAM	HMM	AVLetters2	Harf	%38
Bear [171]	AAM	HMM	LILiR	Cümle	%61,8
Biswas [172]	AAM	HMM	AVICAR	Cümle	%28,23
Bear [27]	AAM	HMM	AVLetters	Harf	%35
Noda [169]	CNN	MS-HMM	ATR	Kelime	%37
Stewart [141]	DCT	MS-HMM	XM2VTS	Rakam	%70
Pass [116]	DCT	HMM	QuLips	Rakam	%98
Ngiam [167]	ST-PCA	Autoencoder	AVLetters	Harf	%64,4
Puviarasan [62]	Manifold	SVM	AVLetters	Harf	%62,34
			OuluVS	Deyim	%65,26

3.2. DERİN ÖĞRENME TABANLI OTOMATİK DUDAK OKUMA SİSTEMLERİ

Geleneksel yöntemlerin uygulandığı dudak okuma sistemlerine ait detaylara Bölüm 3.1'de yer verilmiştir. Fakat derin sinir ağlarındaki gelişmelerle ve büyük ölçekli veri setlerinin yayınlanmaya başlamasıyla birlikte son yıllardaki otomatik dudak okuma sistemlerinde derin öğrenme modellerinin kullanımında önemli bir artış ve performanslarda da önemli gelişmeler meydana gelmiştir.

Derin öğrenme modellerinin gelişimiyle ses tabanlı ve video tabanlı konuşma tanıma sistemlerinde kullanılma şekli arasında güçlü bir paralellik bulunmaktadır. Aslında bu alanda yapılan ilk çalışmalar tamamen derin öğrenme tabanlı olmamıştır. İlk olarak geleneksel yöntemlerin derin öğrenme modelleriyle birlikte kullanımıyla hibrit sistemler ortaya çıkmıştır. Bu çalışmaların birçoğunda da sınıflandırıcı aşamasında hala HMM kullanılırken özellik çıkarımı aşamasındaysa derin öğrenme modelleri kullanılmıştır. Bunun sebebi de o dönemlerde henüz zamana bağlı veya arka arkaya birbirleriyle bağlantılı verilerin sınıflandırılabilmesine dair derin öğrenme tabanlı ciddi bir çalışmanın olmamasıdır. Daha sonrasında derin öğrenme dünyasına "Yinelemeli sinir ağı (Recurrent neural network)" kavramı girmiştir. Bu ağlara örnek olarak Uzun-Kısa Süreli Bellek (Long-Short Term Memory- LSTM) gösterilebilir. LSTM dudak okuma sistemlerinde HMM'ye ciddi bir alternatif oluşturmuştur. 2014 yıllarında özellikle RNN'lerin kullanılmasıyla beraber, uçtan uca derin öğrenme ağları, ses üzerinden otomatik konuşma tanıma (ASR) sistemleri tamamen değişmiştir. Değiştirmenin yanında geleneksel sistemlerden önemli ölçüde daha yüksek performans göstermiştir [176-178].

Derin öğrenme ağları ilk başta otomatik konuşma tanıma veya konuşmadan yazıya (speech to text) dönüştüren sistemlere uyarlanmıştır. Başarılı sonuçlar alındığı görülünce bu sefer video tabanlı sistemler için kullanımında da benzer bir gelişme görülmüştür. Çizelge 3.2'de ilk olarak 2011'de önerilen hibrit dudak okuma sistemlerinin yapısına bakıldığında tıpkı konuşma tanıma sistemlerinde olduğu gibi birinci aşamada özellik çıkarımı için kullanılıp geleneksel sınıflandırıcıların kombinasyonundan oluşmaktadır [176,178,181,182,183]. Daha sonrasındaysa dudak

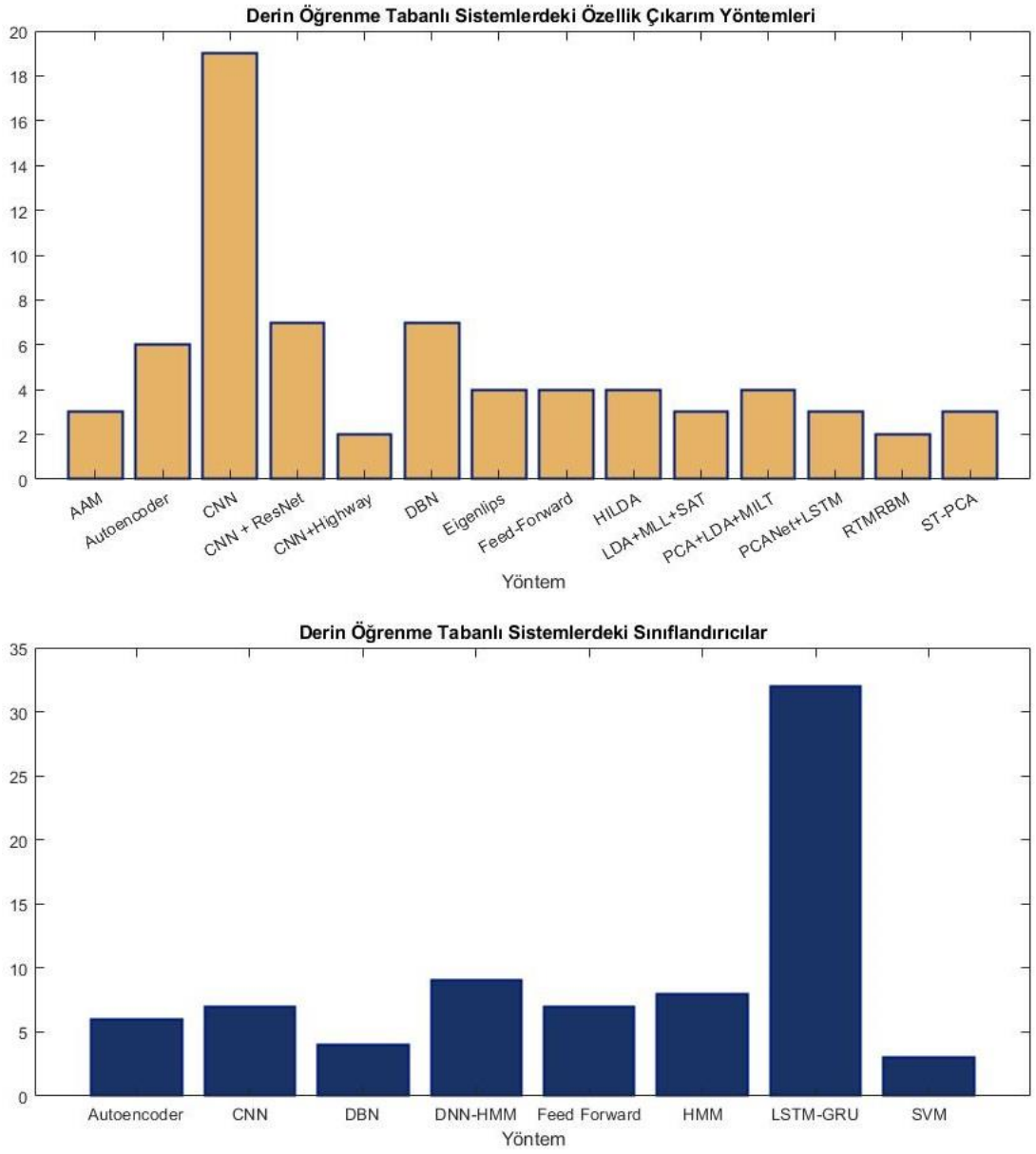
okuma sistemlerinin genelinde uçtan uca derin öğrenme ağlarına dayalı mimariler kullanılmıştır. Bu bölümde kısmen Çizelge 3.2'deki derin öğrenme tabanlı sistemler analiz edilmektedir. Bölüm 3.1' de olduğu gibi önce mimarilerin adımları açıklanmaktadır. Sonrasında sistemler performans açısından karşılaştırılmaktadır.

3.2.1. Derin Öğrenme Mimarilerin Adımları

Uçtan uca derin öğrenme ağlarına dayalı dudak okuma sistemleri, geleneksel sistemlere benzer adımlar dizisinden oluşur. Bir önceki bölümde (Bölüm 3.1) belirtildiği üzere;

- Dudakların kırpılması
- Özniteliklerin çıkarılması
- Sınıflandırma

Aşamalarından oluşmaktadır. Sistemler ise özniteliklerin çıkarımı ve sınıflandırma aşamalarına göre kıyaslanacaktır. Şekil 3.7'de verilen üstteki grafikte özellik çıkarım yöntemlere ait alttaki grafikte de sınıflandırıcılara ait kullanım sayıları verilmektedir.



Şekil 3.7. Derin öğrenme tabanlı sistemlerdeki özellik çıkarma yöntemleri.

Çizelge 3.2'de ve Şekil 3.7'de farklı derin öğrenme ağlarının özellik çıkarımı aşamasında veya sınıflandırma aşamasında ne sıklıkla kullanıldığı grafik olarak verilmiştir. Şekil 3.7 'de öznitelikleri çıkarmak için en çok kullanılan ağların Evrişim Sinir Ağları (CNN) olduğu görülmektedir. Bununla birlikte öznitelik çıkarımında CNN'lerin yerine İleri Besleme Ağları (Feedforward Networks) veya Derin İnanç Ağları (Deep Belief Networks) gibi ağlar da kullanılmıştır. Sınıflandırıcılar açısından, her ne kadar bazı çalışmalarda CNN'ler, İleri Beslemeli Derin Öğrenme Ağları ve Derin İnanç Ağları kullanılmış olsa da Şekil 3.7'de LSTM – BiLSTM sınıflandırıcılarının baskınlığı görülmektedir.

Çizelge 3.2'ye bakıldığında 13 tanesi CNN'ler ve RNN'lerin (GRU veya LSTM) kombinasyonlarından oluşan 45 tane uçtan uca derin öğrenme mimarisi ve çalışma olduğu görülmektedir. Dikkat edilmesi gereken durum, sadece bu derin öğrenme mimarileri bulunmuyor. Bunların dışında başka çalışmalarda başka mimariler kullanılmış olabilir. Tabloda bahsedilen durum sık kullanıma göredir. Dolayısıyla bu kombinasyonlar dudak okuma sistemleri için en çok kullanılan derin öğrenme mimarileri olarak öne çıkmaktadır. Bu yüzden sadece listelenen mimariler detaylı bir şekilde incelenecektir.

Tipik bir CNN – LSTM kombinasyonundan oluşan bir sistemde bir dizi video karelerinin evrişimli ağ ve ardından tekrarlayan ağ (recurrent network) tarafından işlenmektedir. CNN'ler tez kapsamında geliştirilen yöntem içinde de kullanıldığından ötürü detayları Bölüm 4.2.2'de verilmektedir. Fakat bu aşamada bir önbilgi verilmesi uygun görülmüştür. CNN'ler görüntü tanıma ve sınıflandırma aşamalarında [184,185] görsel özellikleri çıkarmak için güçlü bir model olarak kurulmuştur. CNN'lerin çok farklı yapısı olmasına rağmen genelde önemli 2 katmanları vardır. Bunlar evrişim katmanları ve havuzlama katmanlarıdır. Evrişim katmanı, CNN modellerine giriş olarak verilen görüntüyü işleyen ilk katmandır. Görüntüler aslında bünyelerinde belirli piksel değerlerini barındıran bir çeşit matrisler veya sinyallerdir. İlk katman olan evrişim katmanı, görüntünün gerçek boyutlarından daha küçük boyutlara sahip bir filtreyle görüntünün üzerinden pencere mantığıyla geçer. Bu filtrenin uygulanmasıyla (iç çarpımlarla) belirli özellikler elde edilir. Belirli adımlardan sonra doğrusal olmayan bir aktivasyon fonksiyonu çalışır. Örneğin tanh, sigmoid, ReLu bunlardan bazılarıdır.

Öte yandan LSTM veya BiLSTM, geçici bir bellek (cell) oluşturan ve döngüsel bağlantıları sayesinde art arda gelen veri dizilerini modellemek için kullanılan bir çeşit tekrarlayan sinir ağlarıdır (RNN). LSTM'ler tipik RNN'lerde görülen "vanishing and exploding gradient" (kaybolan ve patlayan gradyan problemi) sorununu [186] çözdükleri için yaygın olarak kullanılmaktadır. Bu problem aslında 2 alt problemden oluşmaktadır. Bunlar vanishing gradient ve exploding gradient problemleridir. Vanishing gradient problemi (kaybolan gradyan), derin öğrenmede kullanılan ağların eğitilmesinin zorluğu ve veri ihtiyacının fazla olmasının ardında yatan asıl problemdir.

Bu problemin oluşmasının asıl nedeni, aktivasyon fonksiyonlarının tüm girişlere karşılık ürettiği değerlerin, -1 ve 1 arasında oldukları müddetçe geri yayılım algoritmasının kullanılması durumunda değerin aşamalar boyunca çarpıla çarpıla 0'a yakınsamasıdır. Bu problemin önüne geçmek için RNN gibi yeni yöntemlere ihtiyaç duyulmaktadır. Ayrıca çok katmanlı derin sinir ağları için relu fonksiyonunun aktivasyon fonksiyonu olarak seçilmesiyle de bu problem bazı durumlarda aşılabilmektedir. Patlayan gradyan problemi ise yöntem içinde kullanılan aktivasyon fonksiyonlarının ürettiği sonuçlarının geriye yayılım algoritmasında kullanılırken stabil olmayan sonuçlar üretmesine dayanmaktadır. Aşamalar boyunca gradyan çarpanlarının büyüyerek anlamsız ve kullanışsız hale gelmesiyle oluşur. Çözümler arasında yine LSTM kullanımı vardır. Ayrıca yine aktivasyon fonksiyonu olarak relu'nun kullanılmasıyla çok katmanlı ileri beslemeli mimarilerde, öğrenme oranı düzgün verilmezse gradyan patlama probleminin oluşma ihtimali vardır. RNN ve LSTM hakkında daha detaylı bilgiler Bölüm 4.2.3 ve 4.2.4'te verilmektedir.

Çizelge 3.2. Derin öğrenme tabanlı yöntemler.

Çalışma	Özellik Çıkarım	Sınıflandırıcı	Kullanılan Veri Seti	Amaç	WRR (%)
Moon [173]	DBN		AVLetters	Harfler	%55,3
Mroueh [108]	Scattering coeffs+LDA	Feed-Forward	IBM AV-ASR	Cümleler	%30,64
Ninomiya [174]	DBN	MS-HMM	CENSREC-1-AV	Rakamlar	%39,3
Noda [183]	CNN	MS-HMM	ATR	Kelimeler	%22,5
Sui [182]	DBM+DCT+LDA	HMM	AusTalk	Rakamlar	%69,1
Thangthai [36]	AAM	CI-HMM	RM-3000	Cümleler	%33,32
	AAM	CD-HMM	RM-3000	Cümleler	%47,48
	AAM	Feed-Forward	RM-3000	Cümleler	%77,49
	HiLDA	Feed-Forward	RM-3000	Cümleler	%84,67
Almajai [115]	LDA	HMM	LILiR	Cümleler	%23
	LDA+MLLT	HMM	LILiR	Cümleler	%25
	LDA+MLLT+SAT	HMM	LILiR	Cümleler	%43
	LDA+MLLT+SAT	Feed-Forward	LILiR	Sözler	%53
Assael [188]	3D-CNN	Bi-GRU	GRID	Sözler	%93,4
Bear [151]	AAM	HMM-bigramnet	LILiR	Cümleler	%23
Chung [187]	VGG-M	LSTM	OuluVS2	Sözler	%31,9
	SyncNet	LSTM	OuluVS2	Sözler	%94,1
Chung [55]	CNN		LRW	Kelimeler	%61,1
	CNN		OuluVS	Sözler	%91,4
	CNN		OuluVS2	Sözler	%93,2
Howell [155]	AAM	CD-HMM	RM-3000	Cümleler	%75,58
Hu [207]	RTMRBM	SVM	AVLetters	Harfler	%64,63
	RTMRBM	SVM	AVLetters2	Harfler	%31,21
Lee [154]	DCT+PCA	HMM	OuluVS2	Sözler	%63,0
	RAW	PLVM	OuluVS2	Sözler	%73,0
	DCT+HiLDA	HMM	OuluVS2	Sözler	%74,0
	CNN	LSTM	OuluVS2	Sözler	%81,1
Petridis [181]	DBNF+DCT	LSTM	AVLetters	Harfler	%58,1
	DBNF+DCT	LSTM	OuluVS	Sözler	%81,8
Rekik [136]	HOG+MBH	SVM	CUAVE	Rakamlar	%70,1
	HOG+MBH	K-NN	MIRACL-VC	Sözler	%58,1
	HOG+MBH	SVM	OuluVS	Sözler	%68,3
	HOG+MBH	HMM	MIRACL-VC	Sözler	%69,6
	HOG+MBH	SVM	MIRACL-VC	Sözler	%79,2
Fernandez [104]	DCT+SIFT+LDA	HMM	VLRF	Cümleler	%20
Fernandez [158]	DCT+SIFT+LDA	HMM	AV@CAR	Cümleler	%23

Çizelge 3.2. (devam ediyor).

Çalışma	Özellik Çıkarım	Sınıflandırıcı	Kullanılan Veri Seti	Amaç	WRR (%)
Saitoh [56]	CFI+NIN		OuluVS2	Sözler	%81,1
	CFI+AlexNet		OuluVS2	Sözler	%82,8
	CFI+GoogLeNet		OuluVS2	Sözler	%85,6
Takashima [208]	CBN	HMM	ATR	Kelimeler	%51,0
Wand [142]	Eigenlips	SVM	GRID	Sözler	%69,5
	HOG	SVM	GRID	Sözler	%71,2
	Feed-Forward	LSTM	GRID	Sözler	%79,5
Wu [153]	SDF+STLF	SVM	OuluVS2	Sözler	%87,55
Zimmermann [209]	PCA_{NN}+LSTM	HMM	OuluVS2	Sözler	%73
Bear [210]	AMM	HMM	AVLetters2	Harfler	%36,53
	AMM	HMM	LILiR	Cümleler	%41,53
Chung [57]	CNN	LSTM+Attention	OuluVS2	Sözler	%91,1
	CNN	LSTM+Attention	MV-LRS	Cümleler	%43,6
Chung [112]	CNN	LSTM+Attention	LRW	Kelimeler	%76,2
	CNN	LSTM+Attention	GRID	Sözler	%97
	CNN	LSTM+Attention	LRS	Cümleler	%49,8
Petridis [202]	Autoencoder	LSTM	OuluVS2	Sözler	%84,5
Petridis [35]	Autoencoder	Bi-LSTM	OuluVS2	Sözler	%91,8
Petridis [203]	Autoencoder	Bi-LSTM	OuluVS2	Sözler	%94,7
Rahmani et al. [51]	ASM	HMM	CUAVE	Rakamlar	%56,3
	DBNF	HMM	CUAVE	Rakamlar	%63,4
	ASM	DNN-HMM	CUAVE	Rakamlar	%58,9
	DBNF	DNN-HMM	CUAVE	Rakamlar	%64,9
Stafylakis et al. [32]	3D-CNN+ResNet	Bi-LSTM	LRW	Kelimeler	%83
Sterpu [33]	DCT	HMM	TCD-TIMIT	Cümleler	%31,59
Sui [150]	CHAVF	SVM	OuluVS	İfadeler	%68,9
	CHAVF	HMM	AusTalk	Rakamlar	%69,18
Thangthai [34]	PCA+LDA+MLLT	DNN-HMM	TCD-TIMIT	Cümleler	%43,61
Thangthai [211]	Eigenlips	DNN-HMM	TCD-TIMIT	Cümleler	%42,97
Wand [194]	Feed-Forward	LSTM	GRID	İfadeler	%42,4
Afouras [55]	3D-CNN+ResNet	Bi-LSTM+LM	LRS	Cümleler	%37,8
		Depthwise CNN			%45
		Attention encoder+LM			%50
Fung [192]	3D-CNN	Bi-LSTM	OuluVS2	İfadeler	%87,6
Petridis [54]	3D-CNN+ResNet	Bi-GRU	LRW	Kelimeler	%82

Çizelge 3.2. (devam ediyor).

Çalışma	Özellik Çıkarım	Sınıflandırıcı	Kullanılan Veri Seti	Amaç	WRR (%)
Petridis [93]	Autoencoder	Bi-LSTM	AVDigits	İfadeler	%69,7
				Rakamlar	%68
Wand [53]	Feed-Forward	LSTM	GRID	İfadeler	%84,7
Xu [52]	3D-CNN+highway	Bi-GRU+Attention	GRID	İfadeler	%97,1
Mesbah [212]	HCNN + DA (SI)		AVLetters	Harfler	%59,23
			OuluVS2	Rakamlar	%93,72
			LRW	Kelimeler	%89,95
Huang [213]	Autoencoder + VGG16		GRID	Kelimeler	%45,81

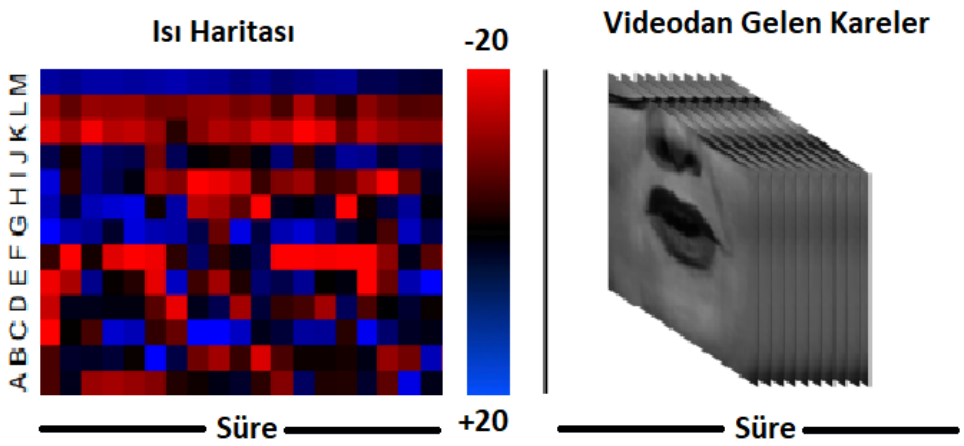
3.2.2. CNN ve LSTM Tabanlı Mimariler

Birçok araştırmacı Bölüm 3.2.1'de bahsedilen 3 aşamalı adımdan oluşan temel algoritmayı takip eden CNN-LSTM tabanları ağları önermiştir. Örneğin Oxford Üniversitesindeki bir çalışma grubundan Chung ve arkadaşları [187] cümle düzeyinde sınıflandırma gerçekleştiren bir ağ önermiştir. "Cümle düzeyinde sınıflandırmanın" sistemin çıktısının sınırlı sayıda cümleyle sınırlandırıldığı anlamına geldiği ve bir tür sınıflandırma probleminin oluşmasına sebep olduğu açıklanmıştır. Ayrıca bu çalışmanın önemli bir özelliği de ses ve görüntünün simultane kullanılmasıdır. Bu yüzden sistemleri 2 parçadan oluşur. Birinci parçasında ses akışıyla ilgili adımlar yer alır. İkinci parçasında da görsel verilerin akışı sağlanır. Geliştirdikleri yönteme giriş olarak videonun her bir karesini gri seviyesine çekilmiş haliyle yollamaktadırlar. Genel mimaride gri seviyedeki görüntüler, beş evrişim katmanından geçtikten sonra 2 tam bağlı katmandan oluşan SyncNet adlı evrişimli başka bir ağa girer. SyncNet adını verdikleri ağ hem ses hem de görüntü girdilerinin 0.2 saniyelik kısımlarını alır ve çalışmada söz edilen ses ve görüntünün simultane işlenmesini sağlayan alt yapıyı sunar. SyncNet ağına bir video geldiği için hem ses hem de görüntü verisi bulunmaktadır. BBC'nin 2013-2016 yılları arasındaki görüntülerinden oluşturdukları kendi veri setlerinde videolara ek olarak söylenen ifadeler veya ses-video arsındaki zaman farklı gibi ekstra veriler yayınlanmamıştır. Ancak çalışmalarında ses ve görüntünün senkronize olduğu varsayılmıştır. SyncNet ağını her birinin detayını vererek ses için ayrı video için ayrı 2 asimetrik akıştan oluşturmuşlardır. Giriş ses

verileri, Mel-Frekans Kepstral Katsayıları (Mel-Frequency Cepstral Coefficients) değerleridir. Mel frekans ölçeği, ses frekanslardaki değişimin insan kulaklarındaki algılama seviyesini gösteren bir ölçek türüdür. Videoda görüntüler ve sesler ayrıldıktan sonra bu seslere ait sinyalin kısa zamanlı güç spektrumunun doğrusal olmayan bir mel ölçeğindeki konumu veya ifadesi bu çalışmada giriş olarak kullanılan MFCC değerine karşılık gelmektedir. Bu değer Eşitlik 3.1'e göre hesaplanır.

$$M = 1125\, x \ln\left(1 + \frac{f}{700}\right) \tag{3.1}$$

Chung yaptıkları çalışmada 13 mel frekans bandını kullanmıştır. Sesten elde edilecek özellikler, 0.2 saniyelik bir giriş sinyali için 20 zaman adımı (time step) verilerek 100 Hz'lik bir örnekleme hızında hesaplanmıştır. Elde edilen değerler en nihayetinde sayısal veriler olduğu için, her bir zaman adımı ve bir mel frekans bandı için MFCC değerlerini temsil eden bir ısı haritası oluşturulmuştur. Şekil 3.8'te bu mimari sunulmuştur. Bu mimaride sol kısımda ses için ısı haritaları olarak zamansal temsilleri bulunur. Haritanın solunda yer alan A'dan M'ye kadar olan 13 satır, farklı frekans bölgelerindeki güçleri temsil eden 13 MFCC özelliğinin ayrı ayrı kodlarıdır. Sağda ise sisteme gelen gri seviyedeki görseller yer alır.

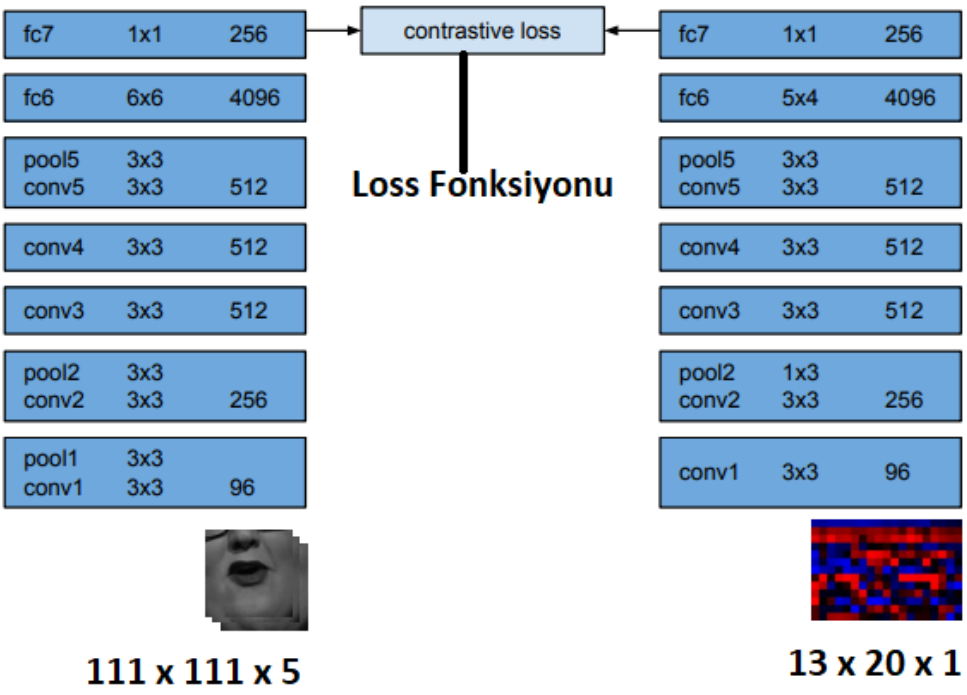


Şekil 3.8. Isı haritası mimarisi.

Ses akışında ısı haritası oluşturulduktan sonra sınıflandırma aşaması bulunur. Bu aşamada da normalde görüntü tanıma için tasarlanmış olan VGG-M modeli

kullanılmıştır. Sadece filtre boyutlarında ufak değişiklikler yapılmıştır. VGG-M'e 224 x 224 piksel boyutunda bir kare görüntü gelirken, girdi boyutu zaman yönünden 20 piksel (zaman adım sayısı) ve diğer yönden de sadece 13 pikseldir. Böylece VGG-M girişi 13x20 piksel boyutunda olmaktadır.

Chung'ın çalışmasında ikinci aşamada görsel akışa ait yapı sunulmuştur. Bu yapının ses akışından çok bir farkı yoktur. Görsel ağa Şekil 3.8 ve 3.9'da belirtildiği üzere sadece ağız bölgesinin yer aldığı gri seviyede görseller gelmektedir. Giriş boyutları 25 Hz kare hızında 0.2 saniyeden gelen 5 kare için 111 x 111 x 5 şeklindedir. Şekil 3.9'da genel mimari verilmiştir.

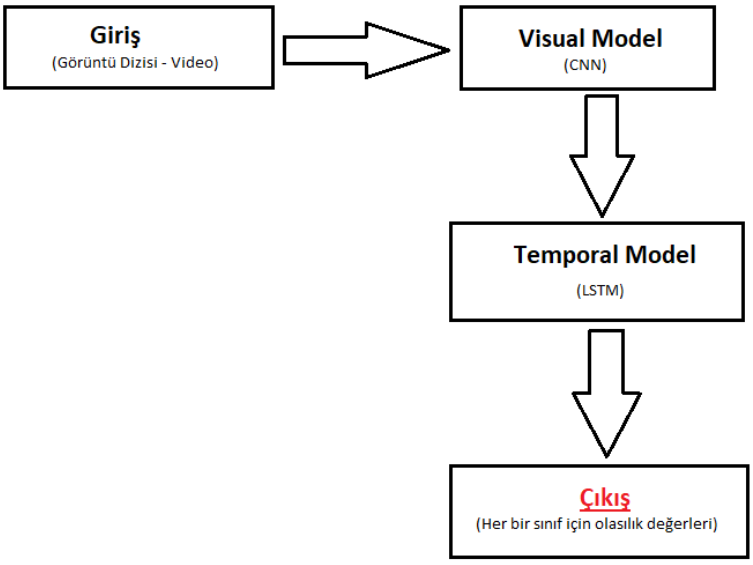


Şekil 3.9. Simultane çalışan Chung sistem mimarisi.

Chung ve arkadaşları yine aynı çalışmada CNN'i VGG-M olarak bilinen önceden eğitilmiş bir ağ (pre-trained network) ile performans açısından kıyaslamıştır. VGG-M adı verilen derin öğrenme ağı, yine hazır bir veri seti olan ImageNet [179] veri setinde önceden eğitilmiş 5 evrişim katmanını takip eden tam bağlantılı 3 katmandan oluşur. VGG-M çıkışı, tıpkı SyncNet'e benzer şekilde katman dizisinin sonunda sınıflandırmayı gerçekleştiren LSTM katmanı için bir giriştir. Performans karşılaştırmalarının yapıldığı Bölüm 3.2.4'te görülebileceği üzere ek bir tam bağlantılı

katmana sahip olmasına rağmen önceden eğitilmiş VGG-M, SyncNet gibi eğitiminin dudak okuma görevine çok daha özel bir şekilde yapıldığı bir ağ kadar iyi performans gösteremediği görülmektedir.

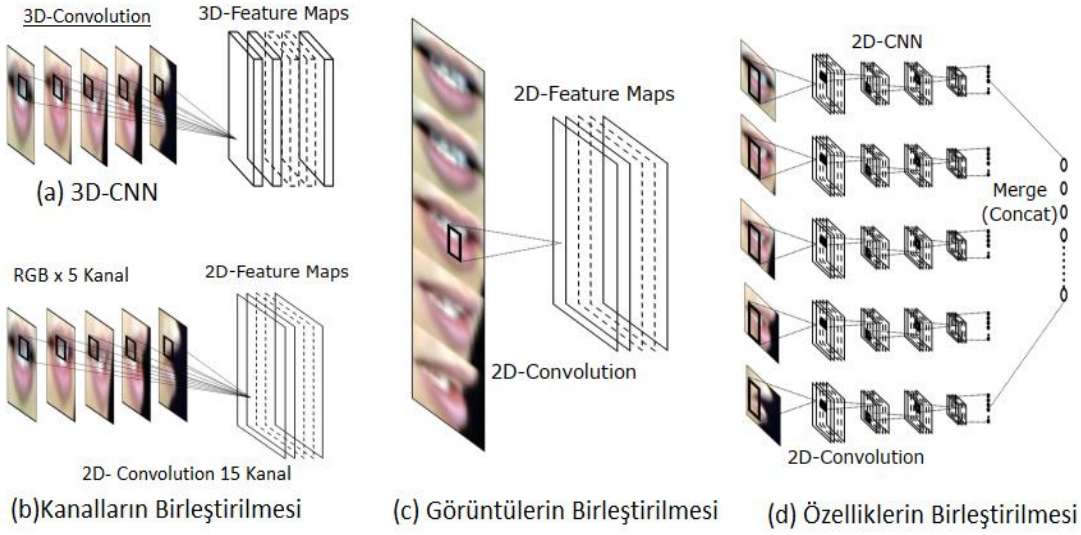
Lee ve arkadaşları [154] cümle seviyesinde tanıma ve sınıflandırma yapabilmek için DNN mimarisini önermişlerdir. Sistemlerine 2 evrişimli katman ve bir tam bağlı bir katmandan oluşan bir CNN tarafından işlenen RGB görüntüler, giriş olarak gönderilir. Bu işlemden sonra CNN tarafından elde edilen özellikleri alan ve dizinin (videodaki kareler) sonuna kadar her bir karenin çıkıştaki anlama katkısını toparlayan 2 LSTM katmanına dayalı bir zamansal model tanımlanmıştır. Böylece tüm dizi, sınıflandırılmasını ve bir cümleye döndürülmesini sağlayacak olan tam bağlantılı bir katman tarafından işlenir. Asıl sınıflandırma işlemini bu tam bağlantılı katman işlemi yapar. Şekil 3.10'da Lee'nin oluşturduğu mimariye ait yapılar belirtilmiştir.



Şekil 3.10. Lee mimarisi [154].

Lee, oluşturdukları sistemi OuluVS2 veri seti üzerinde denemiştir. Bu veri seti, Bölüm 2.1.3'te belirtildiği üzere 5 farklı açıdan (frontal, 30°, 45°, 60°, profil) çekilmiş görüntüleri barındırmaktadır. Aynı çalışmada 2D-CNN + LSTM, 3D-CNN (MV-3D) mimarilerini biraz değiştirerek "Kanalların Birleştirilmesi (Merge Channels MV-MC)", "Görüntülerin Birleştirilmesi (Merge Images MV-MI)", "Özelliklerin Birleştirilmesi (Merge Features – MV-MF)" teknikleri kullanılmıştır. Bu tekniklere ait

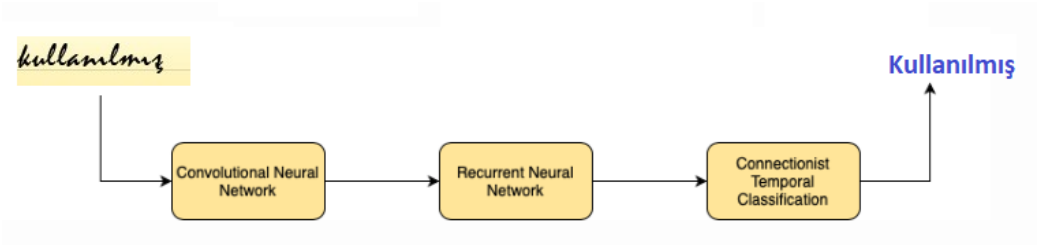
yapı Şekil 3.11'de gösterilmiştir. Kanalları birleştirirken üç (Red, Green ve Blue) kanallı 5 farklı açıdan alınmış görüntülerden oluşan toplamda 15 kanallı bir görüntü, giriş olarak gönderilir. Görüntülerin birleştirilmesi tekniğindeyse modele giriş için her bir konuşmacının söylediği belirli bir ifadenin 5 farklı açıdan çekilmiş görüntüleri birleştirilerek tek bir görüntü olarak verilir. Bundaki amaç aynı ifadenin farklı açılardaki görünümlerini tek seferde 2D-CNN mimarisine öğretmektir. Özelliklerin birleştirilmesi tekniğinde, farklı açılardaki görüntüler ayrı ayrı 2D-CNN mimarisine gönderildikten sonra elde edilen özellikleri tek seferde birleştirilir.



Şekil 3.11. Lee sistem teknikleri ve adımları [154].

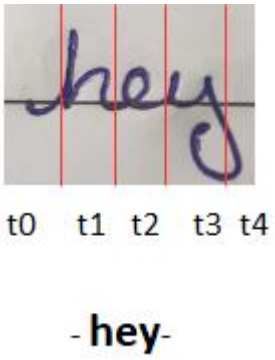
Assael ve arkadaşları [188] cümle düzeyinde sınıflandırma gerçekleştiren ve dudak okuma literatüründe oldukça önemli bir model olan uçtan uca bir derin öğrenme mimarisi LipNet'i önermiştir. Modelin girdisi, tümü aynı ve sabit boyutlarda RGB görüntü dizileridir. Girişe verilen görüntüler, üç tane spatiotemporal (uzamsal ve zamansal - hem zaman hem de konum ile ilişkili) evrişim katmanı tarafından işlenir. Genelde bu tarz yapılara "Spatiotemporal Convolutional Neural Network (STCNN)" denilmektedir. STCNN'lerin en önemli özelliği, video verilerini zamansal ve uzamsal (mekânsal - spatial) boyutları birbiriyle ilişkilendirerek evrişim işlemini uygulayabilmesidir. STCNN yapısında elde edilen özellik değerleri 2 tane Bidirectional Gated Recurrent Network (BiGRU) yapısına gönderilir. Bu katmanlardan sonra her zaman adımında veya her bir video karesi için doğrusal

dönüşüm (Linear Transformation) ve softmax uygulanır. Yani çıkarılan özellikler 2 Bi-GRU tarafından işlenir ve işlenen BiGRU çıktıları doğrusal katman (linear layer) ve bir softmax tarafından işlenir. Bu uçtan uca model, veri setindeki sınıf sayısına ek olarak bir de boş karakter için ek bir birim içeren softmax çıktı katmanına sahip Connectionist Temporal Classification (CTC) [176] ağı ile eğitilmiştir. CTC, el yazısı tanıma, konuşma tanıma veya dudak okuma gibi diziyle gelen ve ayrıca zamanlamayla ilişki veriler barındıran sorunların çözümü için oluşturulan bir çeşit Sinir Ağı Çıktısıdır. CTC'nin diğer yapılardan farkını oluşturan en önemli nokta hizalanmış veya sıraya konulmuş bir veri kümesine ihtiyaç duyulmamasını sağlamasıdır. Bu da eğitim sürecinde karmaşıklığı azaltır. CTC'nin el yazısını tanıma üzerine basit bir sistem diyagramı Şekil 3.12'de belirtilmiştir.

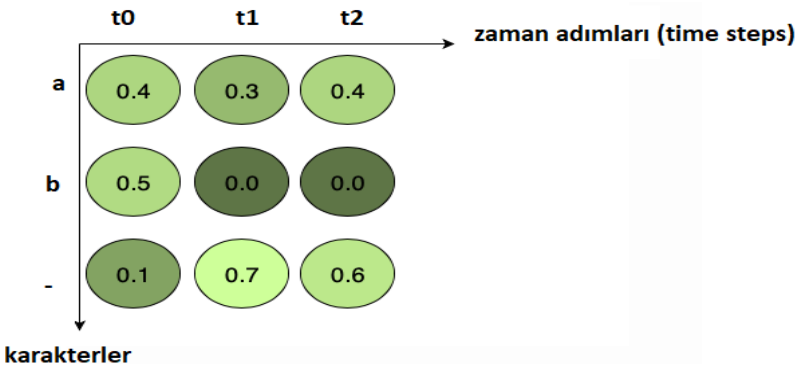


Şekil 3.12. CTC ile el yazısı tanıma sistemi.

Örneğin karakter tanıma sisteminde CTC, Şekil 3.13'te gösterildiği gibi her bir zaman adımı için Şekil 3.14'te gösterildiği gibi bir matris ile temsil edilen karakter puanı hesaplar.

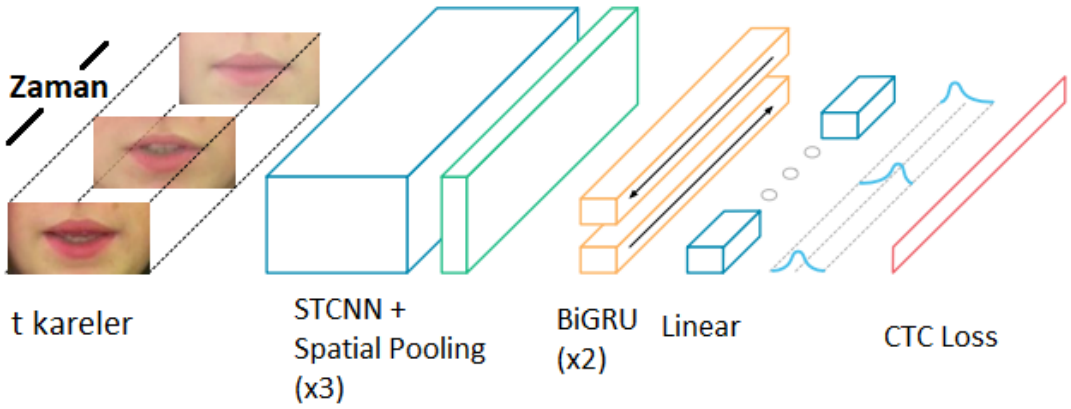


Şekil 3.13. CTC zaman adımları.



Şekil 3.14. Matris hesaplama.

Oluşturulan matris üzerinden t0, t1 ve t2 için aday olasılık değerleri satırdaki değerlerin çarpımıyla elde edilir. Örneğin "a--" olasılık değeri için 0.4 * 0.7 * 0.6 = 0.168 olarak hesaplanır. Benzer şekilde "aaa" için 0.4 * 0.3 * 0.4 = 0.048 olarak hesaplanır. Bu şekilde tüm olasılıklar ve yollar hesaplanır. Hesaplanan değerler üzerinden en iyi aday bulunur. Bu işlem dudak okumada da benzer şekilde yapılır. Örneğin dizi uzunluğu 3'e sabitlenirse, CTC "el" dizisinin olasılığını p(eel) + p(ell) + p(_el) +p(e_l) + p(el_) olarak tanımlar. Böylece model kare etiketlerini tahmin eder ve tahminler ile çıktı dizisi (önceden tanımlanmış veri setindeki sınıflar içinde olan bir cümle) arasındaki en uygun çıktıyı bulur. CTC aynı mantıkla LipNet içinde de kullanılmıştır. LipNet yapısı için mimari özetle Şekil 3.9'da gösterilmektedir.



Şekil 3.15. LipNet mimarisi.

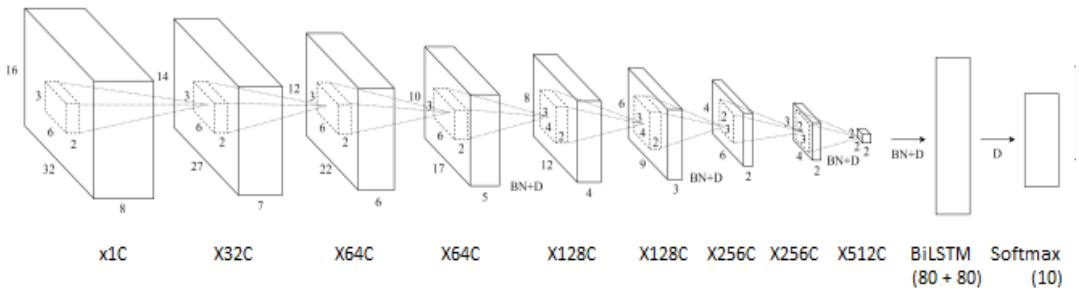
LipNet mimarisinin oluşturulduğu çalışmada GRID ve LRW veri setleri kullanılmıştır. Ön işleme aşamasında girişe 3 saniye ve 25 FPS olan videolar verilmiştir. Bu videolar,

“DLib” yüz dedektörü ve “Online Kalman” filtresi ile birleştirilmiş 68 landmarktan oluşan “iBug Face Landmark” adlı dedektör tarafından taranmıştır. Elde edilen yer işaretleri (landmark) kullanılarak tüm karelerde 100x50 piksel boyutunda ağız merkezli bir kırpma yapmak amacıyla Affine Transformation uygulanmıştır. Ayrıca overfitting problemini azaltmak için de veri artırımı (data augmentation) yapılmıştır ve videolara uygulanan basit dönüşümlerle veri seti zenginleştirilmiştir. Veri artırımının ilk aşamasında hem normal hem de yatay olarak “mirroring (aynalama veya yansıtılma)” işlemi uygulanmıştır. İkinci olarak veri setinde kelime başlangıç ve bitiş zamanları bilinen cümle videoları için cümle düzeyinde eğitim verilerini, ek eğitim örnekleri olması için cümledeki tüm kelimelerin videolarıyla zenginleştirilmiştir. Üçüncü olarak karelerin silinmesi ve çoğaltılması yoluyla yine veriler artırılmıştır. Böylece değişen video hızlarına uyum sağlanmıştır. Bu işlem kare başına %5 olasılıkla rastgele gerçekleştirilmiştir. Her bir kare için 1-100 arasında rastgele bir değer üretilir. Eğer bu değer 5’ten küçükse o kare silinir.

Öte yandan Stafylakis [51] yaptığı çalışmada kelime düzeyinde sınıflandırma gerçekleştiren bir sistem önermiştir. Modellerine giriş olarak 1 saniyelik sabit uzunlukta gri seviyeye dönüştürülmüş videolar vermiştir. Önerilen mimari, “spatia temporal evrişim katmanı” ve onun ardından gelen residual network (ResNet) [191] ağından oluşmaktadır. ResNet’in bu modelde kullanılma sebebi her zaman adımında ResNet’in oluşturacağı çıktıyı tek boyutlu bir vektör haline getirene kadar maksimum havuzlama ile uzamsal boyutu azaltan 34 katmandan (evrişim, havuzlama ve tam bağlı katmanlar dahil) oluşmasıdır. Daha detaylı bilgiler Bölüm 4.2.2.7’de verilmektedir. Böylece çok boyutlu bir matris yerine ResNet sayesinde1 boyutlu vektör elde edilmiş olur. Daha sonra bu vektörler sınıflandırma için birleştirilen 2 tane Bidirectional LSTM (BiLSTM) [189] için girdi olarak kullanılır. Diğer çalışmalardan farklı olarak, tüm video dizisi LSTM tarafından kodlandıktan sonra LSTM çıktısının son zaman adımında sınıflandırma yapılmaz. Bunun yerine her zaman adımı için softmax uygulanır. Bu nedenle toplam kayıp tüm zaman adımlarında ayrı ayrı hesaplanan kayıpların toplamı olarak tanımlanır.

Bu son iki sistemin [51,188] yalnız tek bir yönde çalışan standart LSTM’lerin aksine hem geçmiş hem de gelecek bağlamlarda koşullandırılmış çıktılar üretebildikleri için

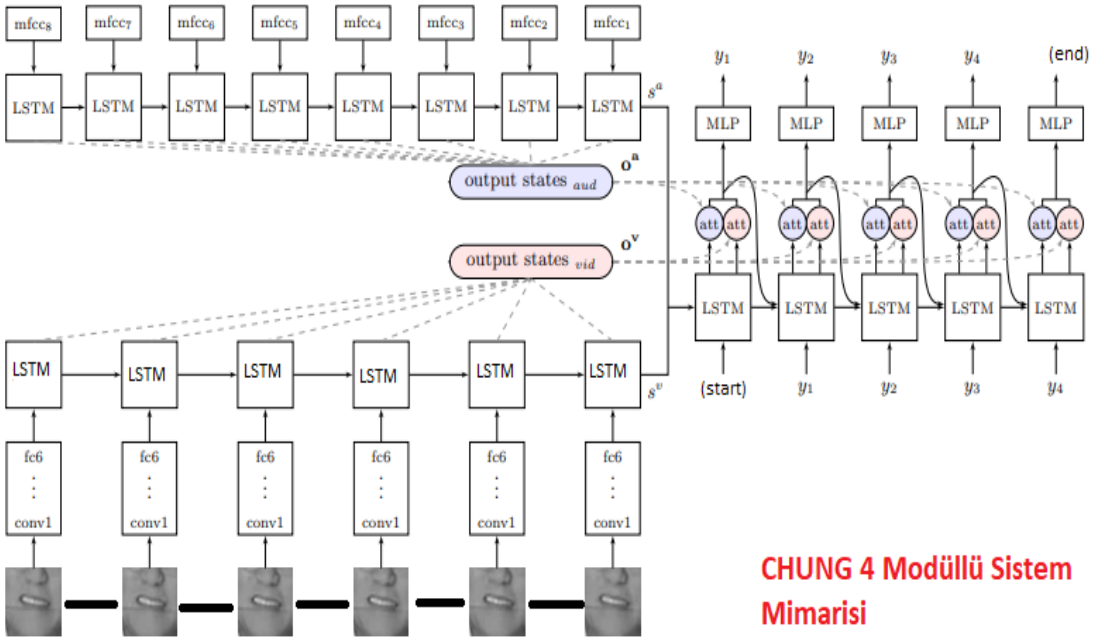
BiLSTM veya BiGRU'ları kullandıklarına dikkat edilmelidir. Zaten modellerin başlarındaki "Bi" ifadesi "Bidirectional" yani "çift yönlü" anlamına gelir. Bu araştırmalar gibi diğer araştırmalardan sonra dudak okumayla ilgili yeni çalışmalar iki yönlü ağların kullanımını ve geliştirilmesini araştırmıştır. BiLSTM ile ilgili detaylı bilgiler Bölüm 4.2.4'te verilmiştir. Petridis ve arkadaşları [54], [51]'e çok benzer bir model önermiştir. Burada her iki dudak okuma mimarisi arasındaki en önemli fark, [54]'ün [51]'de kullanılan BiLSTM ağları yerine daha fazla sayıda gizli birim (hidden units) barındıran 2 katmanlı BiGRU ağlarını kullanmasıdır. Çıktı katmanında her bir kare için sınıf ataması yapan bir softmax katmanı kullanılmıştır. Bir video için sınıf ataması, en yüksek ortalama olasılık değerine göre yapılır. Bu çalışmada OuluVS veri seti kullanılmıştır. Oluşturulan sistem dudak okuma açısından sert koşulların bulunduğu gerçek ortamlarda kelime tanıma için bilimsel araştırmalara açık en büyük veri setlerinden biri olan OuluVS'ye ilişkin sonuçlar paylaşılmıştır. Geliştirdikleri uçtan uca görsel-işitsel model, temiz koşullar ve düşük gürültü seviyeleri altında standart MFCC tabanlı sistemlerden biraz daha iyi performans göstermiştir. Öte yandan Fung ve arkadaşları da [192] cümle düzeyinde sınıflandırma için Bi-LSTM'leri kullanmıştır. Genel mimarisine bakıldığında da sistem diğer birçok sistemde olduğu gibi 2 kısımdan oluşuyor. İlk kısım 8 katmanlı evrişim katmanı içerir. İkinci kısım tek katmanlı BiLSTM içerir. Evrişim katmanlarının hepsi, sıfırla doldurma (zero-padding), kaydırma adımı (stride) içermeyen 3 boyutlu spatialtemporal evrişim katmanlarıdır. Spatialtemporal katmanlarından hemen sonra herhangi bir havuzlama katmanı olmadan ReLU birimi veya maxout birimi, aktivasyon fonksiyonu olarak yer almaktadır. Oluşturulan bu 8 katmanlı 3D CNN mimarisi ardından BiLSTM katmanı gelmektedir. Son olarak modelin üreteceği çıktı, son zaman adımında softmax katmanında elde edilir. Bu katmanların yer aldığı model mimarisi Şekil 3.16'da verilmiştir.



Şekil 3.16. Fung mimarisi.

Chung ve arkadaşları AV-ASR veri seti [112] için bir sistem ve ALR veri seti [57] için başka bir sistem önermiştir. Diğer çalışmalardan farklı modüler yapılı bir sistem önerilmiştir. AV-ASR sistemi için konuşulan cümlelerde karakterleri tahmin etmeye çalışan "Watch (İzle)", "Listen (Dinle)", "Attend (Katıl)" ve "Spell (Hecele)" olmak üzere dört ana modüle dayalı uçtan uca bir derin öğrenme ağı önerilmiştir. Watch modülü tıpkı bir insanın videodaki görseli izleyip anlamlandırması gibi çalışmaktadır ve görüntü kodlayıcı (image encoder) olarak çalışır. Watch modülü, videoyu girdi olarak alır ve 5 adet 3 boyutlu evrişim katmanından, 1 adet tam bağlı katmandan ve farklı abstraction seviyelerini yakalayabilmek için arka arkaya dizilmiş 3 tane LSTM katmanından oluşur. Girişteki her karenin boyutu 120 * 120 olarak sabitlenmiştir. İkinci modül olan "Listen" modülüyse videodaki sesi dinleyip anlamlandırabilmesi için oluşturulmuştur. Sesi işlemek için Listen modülünde de Watch modülüne benzer bir ağ kullanılmıştır. Listen modülünün Watch modülünden farkı evrişim katmanlarını içermemesidir. Listen, ses kodlayıcı (audio encoder) olarak çalışır. 3. modül olan Spell ise karakter kodlayıcı (character encoder) olarak çalışır. Spell modülü 3 adet LSTM, 2 adet dikkat (attention) mekanizması (Watch ve Listen modülü tarafından sağlanan işitsel ve görsel veriler) ve bir tane multi-layer perceptron (MLP) yapılarından oluşmaktadır. Bahsedilen 2 adet dikkat mekanizmasının amacı Watch modülüyle oluşturulan görsel hafızayı ve Listen modülüyle oluşturulan işitsel hafızayı takip etmektir. Bu nedenle Spell modülünde kullanılan LSTM'ler 3 veriyi kullanır: (1) bir önceki karakter, (2) önceki LSTM durumu, (3) son zaman adımında yer alan Watch-Listen modüllerindeki LSTM durumlarının birleşimi. Ardından Attend modülüne gelindiğinde işitsel ve görsel verilere dayalı 2 tane özellik vektörü hesaplanır. Bu özellik vektörleri, dikkat mekanizmaları tarafından her zaman adımında hesaplanır.

Oluşturulan dikkat mekanizmaları hesaplamalarının yapabilmesi için Watch ve Listen modüllerindeki LSTM'ler tarafından her zaman adımında veya videodan gelen her kare için üretilen çıktıyı ve o zaman adımındaki Spell modülünün LSTM çıktısını kullanır. Son olarak, çıktı karakterinin olasılık dağılımı, çıktı üzerinde bir softmax katmanı ile MLP tarafından gerçekleştirilir. Bu işlemi Şekil 3.17'de sağdaki bölümde gerçekleştirmektedir.

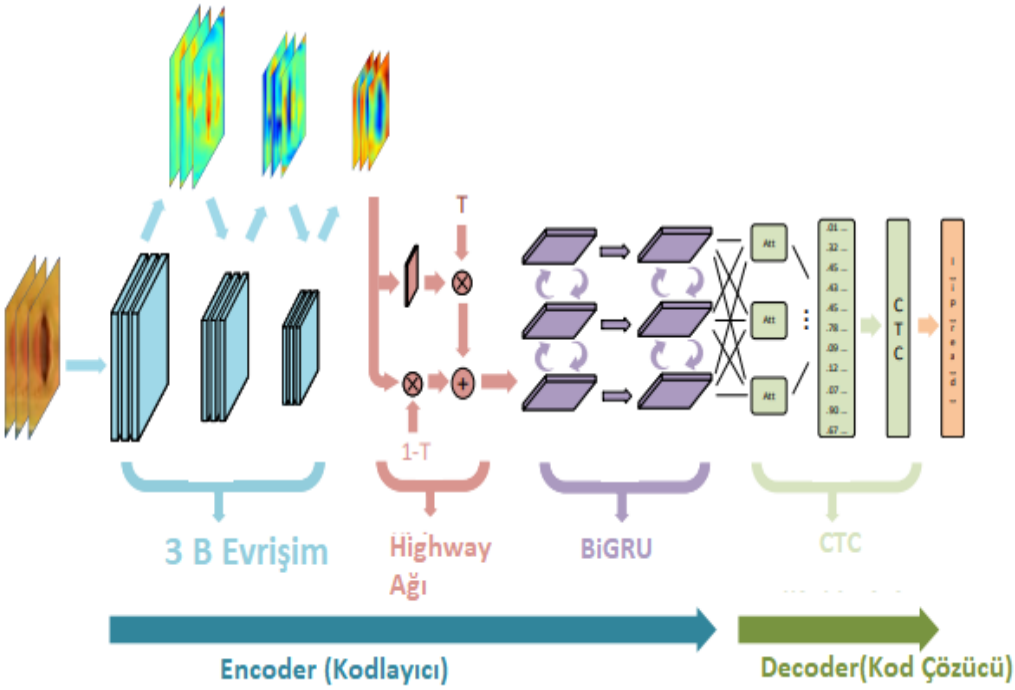


Şekil 3.17. Chung sistem mimarisi.

Chung geliştirdiği sistemde görüntüleri, diğer çalışmalardan farklı olarak normal zaman akışının tersinde ve gri ölçekli olarak göndermiştir. Yani videonun ilk karesi son giriş olacak şekilde yollanıp bunun da başarıyı arttırdığı vurgulanmıştır. Ayrıca kurdukları modüler sistemde dikkat (attention) mekanizmasının kritik öneme sahip olduğu belirtilmiştir. Çünkü o olmazsa temelde 2 problemin meydana geleceği ön görülmektedir. İlk problem dikkat mekanizması olmadığında tüm model giriş sinyalini unutur. Yani ona bağlı bir özellik sonraki durumlara aktarılamaz. İkinci problem ise birinci veya ikinci kelimenin (kodlayıcılar tarafından çözümlenecek son kelime olacağı için model bunu kendi içinde düzeltebilir) uzağındaki girişlerle bağlantılı olmayan bir çıkış dizisi üretir. Bu da oldukça ciddi bir soruna yol açar. Ek olarak Watch ve Listen modüllerindeki tek yönlü kodlayıcılar değiştirilerek çift yönlü kodlayıcılar haline getirilip karşılaştırma yapılmıştır. Fakat çift yönlü kodlayıcılı

ağların eğitilmesi önemli ölçüde daha uzun sürmüştür. Fakat eğitim süresinin artışına rağmen performansta belirgin bir artış görülmemiştir. [25]' te önerilen, ses verisinin bulunmadığı ALR sistemi için de üstte bahsedilen mimarinin neredeyse aynısı önerilmiştir. Sadece ALR içinde ses verisi olmadığından dolayı Listen modülü ve buna bağlı olarak da dikkat mekanizmasındaki Listen modülünden gelen verilerin takibini yapan ilgili yapı çıkarılmıştır. Geri kalan tüm sistem aynıdır.

Klasik CNN-LSTM mimarisinin son örneklerinden biri olarak Xu ve arkadaşları [52] LCANet adını verdikleri karakter seviyesinde tanıma işlemi gerçekleştiren bir modeli sunmuşlardır. LCANet, giriş olarak verilen video için 3 bileşenden oluşan bir video kodlayıcısına sahiptir: (1) 3 boyutlu evrişim katmanları, (2) klasik yapay sinir ağlarından çok daha derin olan ve yüzlerce katman barındıran ilk derin ileri beslemeli sinir ağı olan Otoyol Ağı (Highway Network) ve (3) BiGRU Ağları. LCANet hem görsel hem de kısa süreli zamansal bilgileri tanımlayabilmek için veri yapılarındaki yığın mantığıyla 3 boyutlu bir evrişimli sinir ağına videodaki ardışık 3 kareyi yollar. Yığın yapısında 3 boyutlu evrişim katmanının üstüne 2 katmanlı otoyol ağı eklenmiştir. Otoyol ağı modülü, bir çift dönüşüm kapısına sahiptir. Ayrıca derin sinir ağının bazı giriş verilerini doğrudan çıkışa aktarılmasına izin veren bir kapı da bulunmaktadır. Bu ağların, çok daha zengin, derin anlamsal özellikleri çözümleyebilmesi sağlanmıştır. Video kodlama işleminin sonunda Bi-GRU ağlarına gönderilmesi gereken veriler, uzun vadeli zamansal verileri çözümleyebilmek için otoyol ağlarından sonra gönderilir. Daha uzun bir kontekstten anlamlı veriler yakalayabilmek için, LCANet'te son aşamada çalışan ve uzay-zamansal özellikleri kademeli çözen daha önce detaylarından bahsedilmiş olan "Connectionist Temporal Classification Loss" kod çözücüsü eklenmiştir. CTC'nin eklenmesi dudak okuma probleminde modelleme yeteneğini geliştirir ve görsel olarak benzer kelime ve söz öbekleri için daha iyi tahminler üretebilir. Modele ait yapı Şekil 3.18'de verilmiştir.



Şekil 3.18. LCANet Mimarisi.

3.2.3. Diğer Derin Öğrenme Tabanlı Mimariler

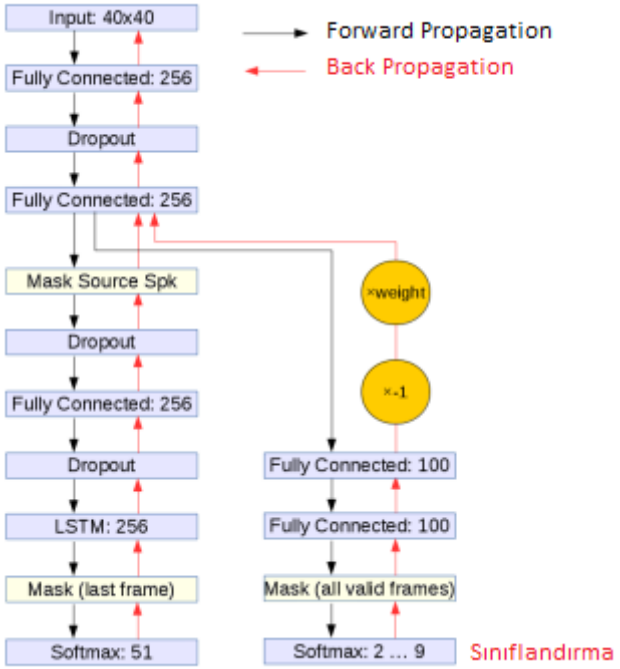
Bazı araştırmacılar, Bölüm 3'ün başında belirtilen ve 3 aşamadan oluşan klasik yöntemin temel çizgilerini takip etmeyen uçtan uca mimariler önermişlerdir. Örneğin Wand ve arkadaşları kelime düzeyinde sınıflandırma gerçekleştiren üç DNN mimarisi [53,142,194] önermiştir. Çalışmada [142] önerilen sistemde ilk aşamada bir ileri besleme katmanı yerleştirilmiştir. İleri besleme katmanından sonra 2 LSTM ve sınıflandırmayı sonuçlandırmak için softmax katmanı eklenmiştir. Çalışmada veri seti olarak GRID veri seti (34 konuşmacı, 1000 cümle, 28 saat) kullanılmıştır. Bölüm 3.2' de belirtildiği üzere GRID veri setinde "Normal" ve "Yüksek" kalitede videolar vardır. Çalışma kapsamında tüm veri setini kullanmak yerine 360 x 288 piksel çözünürlüğe sahip gri seviyeye çekilmiş "Normal" kalitede videolar kullanılmıştır. Ayrıca GRID veri setindeki 8 numaralı konuşmacı için kendilerince okunamayan videolar da test aşamasında kullanılmamıştır. Veri setiyle birlikte verilen kare düzeyinde sıralama verileri kullanılarak, videoda geçen cümlelerde kelime düzeyinde segmentasyon yapılmıştır. Böylece videolardaki cümleler, kelimelere bölünebilmiştir. Bu şekilde bir işlemden sonra konuşmacı başına 6 x 1000 = 6000 kelime elde edilmiştir. 6 sayısı

Bölüm 3.2'de belirtildiği üzere GRID veri setindeki her bir cümle 6 kelimelik bir yapıdan oluştuğu için denkleme eklenmiştir. Her konuşmacı 1000 adet cümle söylediği için 1000 ile çarpılmıştır. GRID veri setinin sesle ilgili herhangi bir parçası kullanılmamıştır. Her video karesinden ağız bölgesi 40 x 40 piksellik bir kareyle kırpılmıştır. Çalışma kapsamında klasik yöntemlerle LSTM kıyaslanmıştır. Mimaride her ikisi de SVM sınıflandırıcısı ile birleştirilmiş 2 özellik çıkarımı (Eigenlips ve HOG) ve LSTM performans açısından kıyaslanmıştır. Oluşturdukları LSTM, (1) bir ileri besleme katmanından ve (2) sonrasında gelen 2 tekrarlayan LSTM (her biri 128 hücreden oluşan) katmanından ve (3) kelime sınıflandırmasını yapan 51 birimli softmax katmanından oluşur. Eigenlips, HOG ve LSTM için sırasıyla %6814, %71.1 ve %79.4 WRR değerleri elde edilmiştir. Sonuçlar incelendiğinde LSTM klasik yöntemlerden daha iyi performans göstermiştir. Çalışmada ayrıca tam bağlantılı ileri besleme katmanlar, CNN ile değiştirilmiştir. Ancak sonuçlarda ciddi bir değişiklik olmadığı bildirilmiştir. Bu duruma olası nedenlerden biri olarak 40 x 40 piksellik alanın zaten sınıflandırma için yeterli bilgiyi içermesi olarak gösterilmiştir. Ayrıca daha büyük boyutlardaki görüntülerle çalışan CNN tabanlı çalışmaların yanı sıra konuşmacıdan bağımsız bir CNN-LSTM mimarisinin daha uygun olacağı belirtilmiştir.

Benzer şekilde aynı gruba ait [142] çalışmasından sonra yapılan [194]'te önerilen sistem konuşmacıdan bağımsız dudak okuma için geliştirilmiştir. Bu çalışmanın amacı [142]'de geliştirilen sistemin bilinmeyen konuşmacılarda ciddi bir performans düşüşü göstermesidir. Eğitim aşamasında modelin eğitildiği konuşmacılarla (bilinen konuşmacılar) bilinmeyen test konuşmacıları arasındaki uyuşmazlığı kısmen gidermek için Ganin ve Lempitsky [195] tarafından önerilen "domain adversarial training" yöntemi kullanılmıştır. Bu yöntemde kısaca hedef konuşmacıdan gelen metne dökülmemiş veriler, ağın etki alanından bağımsız (Domain-Agnostic teknik anlamda bir sistemin verinin türüne ve tipine bağımlı kalmadan herhangi bir performans kaybı yaşamadan, hata almadan çalışmasıdır) olan, başka bir deyişle giriş verilerinin bir kaynaktan gelip gelmediğine bağlı olmayan bir ara veri kümesi olur. Modelin bu verileri öğrenmesi sinir ağına veya derin öğrenme modeline ek eğitim veri girişi olarak kullanılabilmesini sağlar. Spesifik olarak modeldeki ikinci ileri besleme katmanına konuşmacının sınıflandırılmasını gerçekleştirmek için 2 ileri besleme katmanından ve

bir softmax katmanından oluşan ek bir ağ entegre edilmiştir. Eklenen bu ağ Şekil 3.19'un sağ altında yer almaktadır. Modele eklenen bu ağın temel amacı konuşmacıları sınıflandırmaktır. Söylenen ifadeyle ilgilenmezler. Modele bu şekilde ek bir ağ eklenmesinin faydalı olduğu düşünülmektedir. Çünkü bu model, ters gradyanla ve geri beslemeyle ana modeli besleyerek tüm sistemin spesifik olarak sadece o konuşmacıya bağlı özellikleri öğrenmesini veya ezberlemesini engeller.

Çalışmada önceki çalışmayla kıyaslanabilmesi açısından yine GRID veri seti kullanılmıştır. Veri setinde sadece s1-s19 arasındaki konuşmacılar üzerinde çalışılmıştır. s1-s9 arasındaki konuşmacılar, modelin optimal parametrelerini belirlemek için geliştirme aşamasında kullanılmıştır. s10-s19 arasındaki konuşmacılar da sistemin nihai değerlendirilmesinde kullanılmıştır. Her bir konuşmacıdan gelen veriler rastgele olarak eğitim, doğrulama ve test veri setlerine bölünmüştür. Bunlardan doğrulama ve test veri setleri her kelimenin 5 örneğini (kelime sayısı) 51 * 5 = 255 (örnek) içerir. Sonuç olarak eğitim verileri oldukça dengesizdir. GRID veri setindeki cümleler, kalıplar üzerinden oluşturulduğu için örneğin "a" harfinden "z" harfine kadar olan her harf 30 kez kullanılırken her renk 240 defa kullanılmıştır. Buna ilişkin diğer detaylar Bölüm 2.1.2'de verilmiştir. Geliştirilen mimari incelendiğinde üç ileri besleme katmanından, onun ardından gelen bir LSTM katmanından ve son olarak da kelimeleri sınıflandıracak olan softmax katmanından oluşur. Kelimeleri sınıflandıracak olan katman Şekil 3.19'da sol alt kısımdaki bölümdür.



Şekil 3.19. Wand mimarisi [194].

Wand ve çalışma grubunun önerdiği son sistem [53] bir LSTM katmanın ve ona bağlı üç ileri besleme katmanından ve veri dizisinin sonunda kelime sınıflandırmasını gerçekleştiren softmax katmanından oluşur. Bu mimaride, LSTM katmanı dahil tüm katmanlar aynı sayıda nörona sahiptir.

Chung ve arkadaşları sessiz bir videoda cümle düzeyinde sınıflandırma yapmak için bir çalışma yayınlamıştır [55]. Bu amaçla 3 derin sinir ağı modeli önerilmiştir. Modeller BiLSTM, Temporal Convolution ve Kodlayıcı-Kod Çözücü tabanlı yapılardan oluşmaktadır. Geliştirdikleri tüm modeller 2 modülden (veya alt ağlardan) oluşur. Bunlar (1) genel hatlarıyla kırpılmış dudak bölgelerinin bir dizi görüntüsünü giriş olarak aldıktan sonra bir özellik vektörü üreten bir bakıma giriş modülü (visual frontend – Şekil 3.20'de en soldaki bölüm) ve (2) (1)'den gelen özellik vektörlerini alıp sırayla işledikten sonra karakter olarak çıktı üretip sonrasında bir cümleyi karakter tabanlı kodlayarak sınıflandıran ve Şekil 3.20'de a,b ve c bölümlerinde gösterilen çıkış modülüdür.

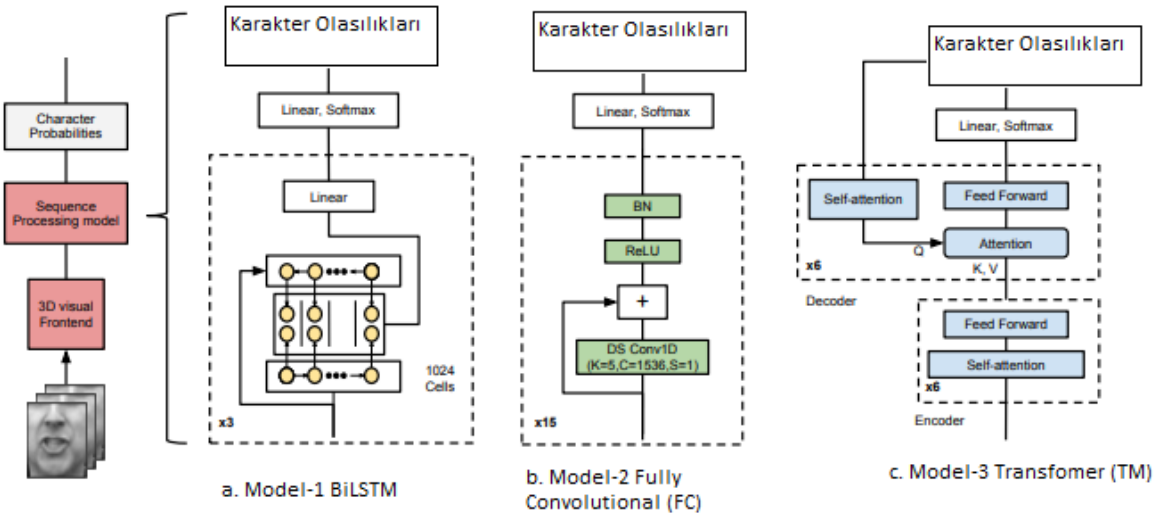
Visual frontend modülü önerilen 3 modelin tamamında bulunur ve genel mimari ResNet mimarisine oldukça benzemektedir. BiLSTM tabanlı ilk sistemde "visual frontend" modülünden gelen özellik vektörlerini alınır ve son BiLSTM katmanı her bir girişe verilen kare için karakter olasılığını hesaplar. BiLSTM'lerin her biri 1024 hücreye sahiptir. BiLSTM ağının uygulanması LipNet [188] modelindeki kullanıma benzerdir. Çünkü bu ağ da CTC ile eğitilmiştir. Bu şekilde de çıktı karakterleri, CTC'nin ekstra üretebildiği boşluk karakteriyle düzenlenir. Böylece sistem, tahmin edilen karakterlerden bir cümle oluşturur. "Fully Convolutional" adını verdikleri ikinci model, bir dizi zamansal evrişim katmanından oluşur. Bir video için temelde 2 boyut bulunmaktadır. Birinci boyut o görüntünün kendisi ve ikinci boyut da zamandır. İlk olarak evrişim işlemi her kanal için zaman boyutuna ayrı ayrı uygulanır. Bunu yapan zamansal evrişim katmanıdır. Sonrasında benzer şekilde görüntülerin tüm kanallarına (kırmızı, yeşil, mavi) filtre genişliği 1 olan konum bazında bir evrişim işlemi ayrı ayrı uygulanır.

Her evrişimden sonra bir kısayol bağlantısı eklenmiştir. Bu kısayol bağlantısının eklenmesi, bir ağın residual network mimarisine uygun olması için gereklidir. Kısayol bağlantıları modelde gradyan için eklenir. Bu şekilde ilk birkaç katman ağırlık güncellemeleri gibi değişiklikleri çok hızlı alır ve kaybolan gradyan sorunu ortadan kalkar. Kaybolan gradyan problemi çözülünce de rahatlıkla çok katmanlı derin mimariler oluşturulabilir. Kısayol bağlantısının sınıflandırma veya nesne tespiti gibi konularda doğrudan bir etkisi yoktur. Eklenmesi tamamen ağın derinleştirilmesiyle ve katman sayısının sağlıklı bir şekilde arttırılmasıyla ilgisi bulunmaktadır. Daha detaylı bilgiler Bölüm 4.2.2.7'de verilmiştir.

Kısayol bağlantısından sonra yığın normalleştirme (batch normalization) ve ReLU eklenmiştir. Yığın normalleştirme, yapay sinir ağlarında özellikle de katman sayısı fazla olan derin sinir ağlarında bulunan gizli katmanlar arasında hesaplanan kovaryansın değerini düşürmek amacıyla kullanılan bir tür normalizasyon yöntemidir. Çeşitli kullanım amaçları vardır. Örneğin Yapay sinir ağları için ağırlık güncellemesi geriye yayılım yöntemiyle yapılarak öğrenme gerçekleşir. Fakat ağlardaki geriye yayılım işlemi esnasında ağın eğitimi için gerekli süreyi arttıran "Internal Covariate Shift" adı verilen bir durum oluşur. Bu durum katmanların hatasının kendi içlerinde

düzeltilmesiyle ilgilidir. Eğitim aşamasında ağdaki katmanlar hatalarını düzeltmek için çalışmalarına rağmen bunu tüm katmanlar ayrı ayrı yapmaya çalışır. Örneğin 3. katmanın ürettiği çıkış vektörünün normalize edilmesiyle beraber sonrasında gelen 4. katman beslenir. Bu amaçla da katmanın ürettiği ilgili çıkış vektörüne 2 işlem uygulanır. İlk olarak vektör ve hesaplanan ortalama değerinin farkı alınır. Daha sonra da bu fark standart sapma değerine bölünür. Bölünmesiyle beraber çıkış vektörü normalize edilmiş olur. Katmanda oluşan normalize edilmiş vektör hemen sonraki katmana girdi olarak gönderilir ve bu işlemler bütün katmanlar arasında aynı mantıkla sürer. Her katman kendisinden önce gelen katman tarafından beslendiği için bir katmanda oluşacak herhangi bir değişiklik sonraki katmanlarda da değişime sebep olur. Özellikle giriş katmanlarında oluşabilecek bu tarz değişiklikler ve kaymalar sebebiyle hesaplamaların boyutu ve süresi zaman geçtikçe artar. Bunun sebebi de her katmanın öğrenmesinin gerçekleşmesi için bir önceki katmanın öğrenmesinin bitmesi gerekir. İşte bu kaymaları azaltmak için yığın normalizasyonu işlemi uygulanır. Normalizasyon işlemindeki temel amaç bütün girişlerin standart sapmasını 1 ve dağılımının ortalamasını 0 değerine ulaştıracak duruma getirmektir. Bu yüzden değerler -1 ve +1 arasına indirgenir. Normalizasyon işlemi sinir ağındaki girişlere uygulanabilir. Fakat bu normalizasyon işleminden sonraki katman olan 2. katman faydalanamaz duruma gelir. Bu yüzden de normalizasyon işlemi, katmanlar arasında yapılır. Yığın normalizasyonu sayesinde ağdaki tüm katmanlar, kendisinden önce gelen katmanların öğrenmesini beklemekten kurtulmuş olur. Bu şart ortadan kalkınca da tüm katmanlar açısından eş zamanlı eğitime olanak sağlanmış olur. Eğitimin de hızlanması sağlanır. Yığın normalizasyonda dikkat edilmesi gereken bir diğer husus da bu tür bir normalizasyon kullanmadan yüksek bir öğrenme oranı kullanılırsa gradyanların kaybolması problemi ortaya çıkar. Fakat yığın normalizasyonda yukarda bahsedildiği gibi bir katmanda oluşabilecek güncelleme veya değişiklik öncesinde, sonrasında yer alan hiçbir katmana yayılmayacağından dolayı daha yüksek öğrenme oranları kullanılabilir. Böylece yığın normalleştirme ile ağ daha düzenli ve kararlı bir yapıya kavuşmuş olur [196,197,198].

Üçüncü modelde de 6 kodlayıcı ve 6 kod çözücü katmanından oluşan ileri beslemeli bir yapı sunulmuştur. 3 mimarinin de sınıflandırma aşamasında Linear ve Softmax fonksiyonları kullanılır. Sırasıyla 3 mimari Şekil 3.20'de verilmiştir.



Şekil 3.20. Chung çalışmasındaki 3 sisteme ait mimariler.

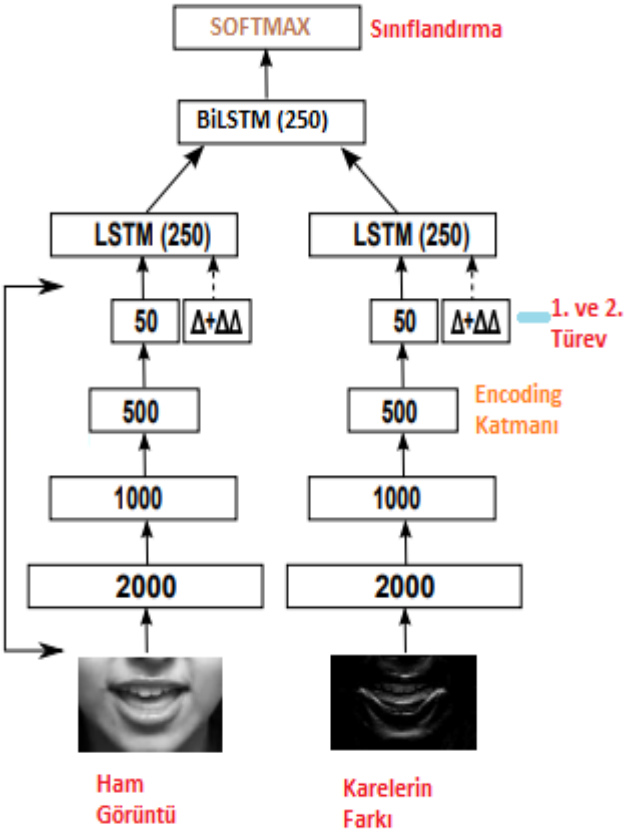
Saitoh ve arkadaşları [56] birçok açıdan dudak okumaya yönelik değerlendirmeler yaptıkları bir çalışma yayınlamışlardır. Yöntemde diziyi kare tabanlı işlemek yerine "Birleştirilmiş Kare Görüntüsü" (Concatenated Frame Image) yöntemini kullanarak tüm videonun kare dizisini makro bir görüntüye indirgeyerek cümle tanıma ve rakam tanımaya çalışılmıştır. Bu durum aslında diğer modellerde kullanılan "bir cümleyi karakter veya hece bazlı tanımlamaktansa bir bütün olarak ele alarak tanıma" mantığına benzer bir mantık olup farklı bir yöntemdir. Ayrıca CFI için veri artırımı da uygulanmıştır. CFI önceden eğitilmiş 3 CNN ağı ile test edilmiştir. Çalışmalarında OuluVS2 veri seti kullanılmıştır. Modellerini konuşmacıdan bağımsız bir sistem üzerine kurmaya çalışmışlardır. Test için OuluVS2'den 12 konuşmacı (s06, s08, s09, s15, s26, s30, s34, s43, s44, s49, s51 ve s52, 10 erkek ve 2 kadın) ve eğitim için de kalan 40 konuşmacı kullanılmıştır. Veri artırımı işleminde $\alpha = 0$ için 4 adet "γ" değeri (0.6, 0.8, 1.2 ve 1.4) ve 4 adet "α" değeri (-2, -1, 1, 2) için γ=1.0 değerleri kullanılmıştır. Bu şekilde 8 farklı tip CFI üretilmiştir. Sonuç olarak test verileri veri artırımından önce 360 (10 cümle x 12 konuşmacı x 3 örnek) CFI içermekteydi fakat veri arttırımından sonra 10800'e (10 cümle x 40 konuşmacı x 3 örnek x 9) yükselmiştir.

Eğitim aşamasında NIN [199], AlexNet [179] ve GoogleNet [184] olmak üzere 3 adet iyi bilenen CNN modeli kullanılmıştır. CNN modellerini oluşturmak ve eğitmek için "Chainer 2" adı verilen framework kullanılmıştır. Tüm modelleri 0.9 momentum değerine sahip stokastik gradyan inişi, yığın boyut değeri olarak 32, öğrenme hızı 0.01 olarak ayarlanmıştır. Bölüm 2.1.3'te OuluVS2 incelendiğinde 6 kameradan görüntüler alındığı ve çoklu görünümde bir veri seti olduğu belirtilmiştir. Bu çalışmada sadece HD kayıt yapan HD1, HD2, HD3, HD4 ve HD5 isimli 5 kameradan görüntüler alınmıştır. Her ne kadar veri seti çoklu görünüm özelliğine sahip olsa da bu çalışma tek yönlü dudak okuma gerçekleştirmiştir. Bu yüzden de her kameradan alınan görüntüler için ayrı ayrı tanıma test ve eğitim işlemleri gerçekleştirilmiştir.

Sonuçları da veri artırımıyla birlikte ve veri artırımı olmadan ki halleriyle ayrı ayrı verdikleri için, bu işlemin ne kadar etkili olduğu da gösterilmiştir. Örneğin GoogleNet veri seti için veri artırımı olmadan sırasıyla her bir kamera için 12.2, 8.8, 8.9, 10.3, 28.3 ve ortalama olarak da 13.7 sonucu elde edilmiştir. Veri artırımı yapıldığında 86.9, 90.6, 85.3, 85.3, 85.6 ve ortalama da 86.7'e yükselmiştir. Sonuçların bu kadar yüksek çıkmasının bir diğer sebebi de test aşamasında modelin daha önce karşılaşmadığı bir konuşmacı olmamasıdır. Ayrıca modellerin katman sayısı üzerinden de bir kıyaslama yapılmıştır. Örneğin NIN modelinin sırasıyla NIN (49), NIN (64) ve NIN (81) türleri için ortalama 37.5, 39.6 ve 42.2 değerleri elde edilmiştir. Bu sonuçlar da gösteriyor ki modeldeki katman sayısı bir seviyeye kadar arttırılması performansı yükseltmektedir. Ayrıca GoogleNet, NIN ve AlexNet arasında performans açısından genel olarak %5'i aşmayan çok ciddi derecede olmayan farklılıklar görülmüştür. Başka bir değerlendirme de kameraların çekim açısının performansa etkisi üzerinden yapılabilmektedir. Çünkü tüm kameralardan aldıkları veri ayrı ayrı değerlendirilmiştir. Bunun için örneğin rakamlar için 0^o, 30^o, 45^o, 60^o ve 90^o açıları için sırasıyla 89.4, 92.5, 90.8, 91.7 ve 87.5 sonuçları elde edilmiştir. Sonuçlar diğer birçok çalışmada olduğu gibi düşünülenin aksine tam karşıdan çekilmiş 0^o'lik görüntüler değil de 30^o'lik açılarla alınmış görüntülerde en iyi sonucu vermektedir. Fakat bu durum cümle tanıma için ufak bir farkla geçerli olmamaktadır. Cümle tanımada 85.6, 82.5, 82.5, 83.3 ve 80.3 değerleri elde edilmiştir. Bu sonuçlara göre de 0^o'lik görüntüler daha iyi sonuç vermektedir. Elde edilen bu çelişki sayılabilecek durum tekrardan hangi açıdan çekilen

görüntüler en iyi performansı gösterir probleminin henüz kesin bir cevabının olmadığını tekrar göstermektedir.

Petridis [35,93,202,203] cümle düzeyinde sınıflandırma amacıyla 4 farklı çalışma yayınlamıştır. İlk olarak [202]'de yaptıkları çalışma OuluVS2'nin çıkış yıllarına yakındır. Bu yüzden OuluVS2 ve ikinci veri seti olarak da CUAVE kullanılmıştır. O güne kadar ki OuluVS2 için alınan en iyi sonuç, klasik makine öğrenmesi yöntemlerini kullanarak DCT özellik çıkarımı + HMM sınıflandırmasıyla beraber %74.8 sonucu olarak elde edilmiştir. Çalışmalarında birbirinden bağımsız çalışan 2 farklı akışa dayalı sistem önermişlerdir. İlk akışta özellikler doğrudan o kare üzerinden elde edilirken ikinci akıştaysa birbiri ardına gelen karelerin farkını alarak özellikler elde edilmiştir. İkinci akıştaki amaç video esnasında zamanla dudak oluşan harekete bağlı lokal hareketlilikleri (dinamizmleri) yakalamaktır. Sonrasında görüntülerin 1. türevi ve 2. türevi üzerinden de hesaplamalar yapılarak özellikler hesaplanır ve darboğaz katmanına yollanır. Darboğaz yapısı genelde evrişim işlemlerinde 1x1 boyutlarına sahip evrişim işleminin uygulanmasından ibarettir. Bu da boyut azaltımını sağladığından dolayı hesaplama maliyetini azaltır.



Şekil 3.21. Petridis'in mimarisi.

Sadece bu mimaride değil genel olarak darboğaz katmanı bir önceki katmanlara göre daha az düğüm içeren bir katmandır. Bu katmanın amacı genelde kendisine gönderilen giriş üzerinde bir bakıma boyut indirgeme yaparak hesaplama miktarını azaltır. Darboğaz katmanının klasik kullanım amaçlarından biri "Doğrusal Olmayan Boyutsal Küçültme" işlemi için darboğaz katmanlarına sahip otomatik kodlayıcıların kullanılmasıdır. Daha detaylı bilgiler Bölüm 4.2.2.8'de verilmektedir. Petridis'in mimarisinde darboğaz katmanında ek olarak kodlama katmanları çalışır. Bu kodlama katmanları "Kısıtlı Boltzmann Makineleri" kullanılarak eğitilmiştir. Sistem videoyu sırayla kare bazlı işlediği için oluşan çıktı aslında bir zamansal dinamik olmakta ve elde edilen değer hemen sonrasında çalışan LSTM katmanında modellenmektedir. Daha sonra her 2 akıştan gelen özelliklerin birleştirilmesi için BiLSTM kullanılmıştır. O da Softmax fonksiyonuna bağlanarak sınıflandırma gerçekleştirilir. Çalışma sonuçları incelendiğinde OuluVS2 veri setindeki başarı oranı CUAVE veri setinden daha yüksektir. Ayrıca her bir akışın sınıflandırma başarısı da verilmiştir. Sadece ham görüntünün kullanıldığı akışla yapılan bir sınıflandırma OuluVS2 veri seti için %78

iken, fark görüntüsü üzerinden yapılan akışta %75 ve her ikisinin Şekil 3.21'de olduğu gibi birlikte kullanımıyla başarı oranı %84.5'te çıkmıştır. Benzer sonuçlar CUAVE veri seti için de geçerlidir. Akışların birleştirilmesiyle sonuçlarda iyileşme gözlemlenmiştir.

Petridis'in yaptığı bir diğer çalışmada [35] önceki çalışmasına [202] benzer bir model kullanmıştır. Ek olarak bir de videodaki sesi işleme almıştır. Özellikle [202]'deki yapılan çalışmadaki ikinci akışta yer alan karelerin farkı yerine sesten elde edilen özellikler kullanılmıştır. Ayrıca her akışın sonundaki LSTM katmanlarını BiLSTM katmanlarıyla değiştirmiştir.

Önerdikleri üçüncü sistem olan [203]'te cümle düzeyinde çoklu görünümlü OuluVS2 veri setini kullanarak bir dudak okuma sistemi geliştirilmiştir. Özellik vektörü önceki sistemlere benzer 3 akış mekanizmasından elde edilir. Her akışta o ifadenin farklı açılardan çekilmiş görüntüleri ayrı ayrı ele alınmaktadır. Bu akışlar bir kodlayıcı ve bir BiLSTM olmak üzere temelde 2 bölümden oluşmaktadır. Modelde ilk olarak yüksek boyutlara sahip bir giriş görüntüsünü darboğaz katmanında özelliklerini fazla kaybetmeden daha düşük boyutlu yapıya sıkıştıran darboğaz mimarisi ve ondan sonra da kodlayıcı katmanı gelir. Sırasıyla 2000, 1000 ve 500 boyutlarında 3 tamamen bağlı gizli katman ve ardından bir doğrusal darboğaz katmanından oluşan önceki çalışmalarına benzer bir mimari kullanılmıştır. Birinci türevler ve ikinci türevlere ait özellikler de darboğaz özelliklerine dayalı olarak hesaplanır ve darboğaz katmanına gönderilir. Bu şekilde eğitim aşamasında kodlama katmanlarını, yapılmak istenen görev için ayırt edici olan özelliklerle ve boyut indirgenerek elde edilmiş özelliklerle öğrenmeye zorlar. İkinci bölümde de BiLSTM ve softmax fonksiyonu bulunur. Bunlar, elde edilen özellik vektörlerine göre sınıflandırma yapar. Geliştirdikleri sistemi yine her akış için ayrı ayrı değerlendirip sonuçlarını üretmiştir. Ayrıca tek görünümlü test çalışmaları yapıp her açı için yine ayrı ayrı sonuçlar üretmiştir. Sonrasında çoklu görünümlü test çalışmalarına ait sonuçlar da paylaşılmıştır. Fakat çoklu görünümdeki tüm açılardan elde edilen görüntüler kullanılmamıştır. 3 tane açının kombinasyonu kullanılmıştır. Örneğin $0^0 + 30^0 + 90^0$ gibi 4 farklı kombinasyon kullanılmıştır. Genel sonuçlar incelendiğinde çıkarılacak sonuçlardan bir diğeri de çoklu görünümde, oluşturulan kombinasyona önden görünüme ek olarak profil

görünümünün eklenmesi performans açısından iyileştirmeler göstermiştir. Ayrıca kombinasyonların çoğunda 45^0 kombinasyonun eklenmesi ufak bir artışa yol açmıştır. Bu yüzden en iyi sonuca 4 farklı kombinasyon arasından 0^0 + 45^0 + 90^0 kombinasyonuyla ulaşılmıştır.

Son olarak geliştirdikleri dördüncü sistemleri [93], [35]'in bazı modifikasyonlardan geçirilmiş halidir. 2 sistem arasındaki temel fark, [93]'te tanıtılan sistemde 2 akış yerine tek bir akış mekanizmasının kullanılmasıdır. Akışta kullanılan ağ yapısı aynı olmakla birlikte katman sayılarda bazı değişiklikler bulunmaktadır. Akışa verilen giriş, karenin orijinal halidir. Fark görüntüsü bu modelde kullanılmamıştır. Bu çalışmada temelde odaklandıkları durum normal konuşma, fısıltıyla konuşma ve sessiz konuşma üzerine bir çalışmadır. Özellikle sessiz konuşmacı tanımayı amaçlayan sistemleri eğitmek için sesli ve normal konuşma setiyle eğitmenin en iyi performansı göstermediği belirtilmiştir. Çünkü normal konuşma üzerinde eğitilmiş bir modelin performansı, fısıldayarak konuşma üzerinde test edildiğinde %8.5 düşmüştür. Benzer şekilde fısıltılı konuşma üzerine eğitilmiş bir model normal konuşmalı veri kümesi üzerinde test edildiğinde performansta %5.7 düşüş yaşanmıştır. Yine modelin hangi veriler üzerinde eğitildiğine bakılmaksızın sessiz konuşma performansı tüm durumlarda en düşük değere sahiptir. Bunun en önemli sebebiyse işitsel geri bildirim eksikliğinden kaynaklanmaktadır. Fakat bu duruma yönelik de teorik de bir çalışma mevcuttur. Janke [204] yaptığı çalışmada EMG sinyallerinin kullanılması, sessiz konuşmada işitsel geribildirim eksikliği somatosensorik geribildirimle daha güçlü bir odaklanmayla giderilebilir. Hatta daha iyi sonuçlar da verebilir. Bunun için ses çıkmasa bile o sesleri daha güçlü ifade ediyor gibi konuşması gerekir. Şu an bu işlem teorik olsa bile ilerde buna yönelik çalışmalarla pratikte bazı sonuçlar elde edilebilir.

Moon harf tanıma işlemini gerçekleştirmek amacıyla transfer öğrenimi tabanlı bir derin öğrenme modeli sunmuştur [173]. Sistem, birden fazla Kısıtlı Boltzmann Makinesiyle, standart bir derin inanç ağı kullanarak verilerin soyut özelliklerini (anlamlarını) öğrenmek için birbirinden bağımsız olarak videodan elde ettiği işitsel ve görsel verileri kullanır. Bu yapı, kaynak ve hedef modüller arasında anlamsal düzeyde bilgi/özellik aktarımını sağlar. Çünkü bir derin inanç ağının her bir ara katmanındaki çıktı değerleri, modüller arasında takip edilebilir bir bilgi aktarımına olanak sağlayarak

giriş özelliklerine ait vektörü verir. Bu amaçla da aynı sayıda ara katmana sahip ses ve video verileri için ayrı 2 derin inanç ağı oluşturulmuştur. Çalışmada Stanford ve AVLetters olmak üzere 2 veri seti kullanılmıştır.

Bilgisayarlı görü alanında normal bir insandan daha sağlıklı sonuçlara ulaşan çalışmalar varken zaman ilerledikçe daha efektif ve verimli uygulanabilen çözümler ile performansta geliştirmeler sürdürülmektedir. Şu ana kadarki çalışmalar incelendiğinde nesnelerin sınıflandırılması ve karmaşık nesnelerin tespitine yönelik çalışmalar; yüksek başarı, gerçek zamanlı çalışma, düşük çözünürlüğe sahip görüntülerde bile ufak boyutlu nesne tahminleri gibi daha fazla yeteneğe kavuşmuştur. Fakat her şeye rağmen yine de nesne algılama modelleri eğitim, test ve sınırlı parametrelerle iyi bir şekilde genelleştirilebilen birleşik bir mimari açısından klasik sınıflandırma modellerinin sahip olduğu basitlik ve sadeliğe sahip değildir. Bu sorunu gidermek amacıyla yürütülen çalışmalardan biri de Ağustos 2020'de Facebook bünyesinde çalışan bir çalışma grubu tarafından yayınlanan Transformerslarla uçtan uca nesne tespiti (End-to-End Object Detection with Transformers), kendilerinin verdiği başka bir isimle "DEtection TRansformer (DETR)" adlı çalışmadır. DETR bu probleme yönelik getirdiği ilk çözüm, nesne algılamada yeni sayılabilecek modellerle karşılaştırıldığında ayarlanması gereken hiper parametre sayısı ciddi ölçüde daha az olmasıdır. Örneğin nesne algılamada oluşturulan ağ eğitilmeden önce veri setindeki nesneleri tahmin / tespit etmek amacıyla öncesinden belirlenmiş bir genişlik ve yüksekliğe sahip "Anchor Box" adı verilen kutular bulunmaktadır. Anchor Box isimli kutuların boyutu, geliştirilen modelin veri setindeki performansını arttırmak için ayarlanması gereken yüksek öneme sahip parametrelerden bir tanesidir. DETR yapısındaysa Anchor Box sayısı, kutuların en boy oranı, varsayılan koordinatları gibi birçok parametrenin ayarlanmasına gerek yoktur. Tüm bu ayarlamalar DETR bünyesinde bulunan ve daha sonra açıklanacak olan 3 modül aracılığıyla düzenlenir. DETR açık kaynak kodlu olup Github üzerinden tüm kaynak kodlarına erişilebilmektedir [214].

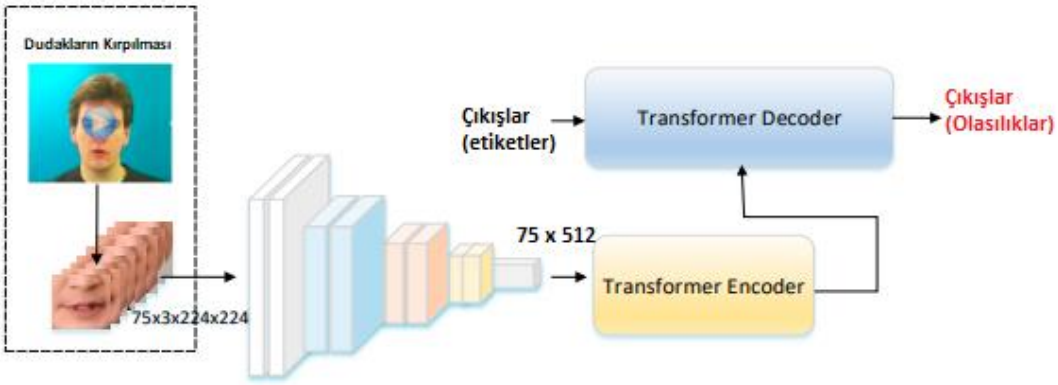
Transformer ağları son zamanlarda doğal dil işleme, dudak okuma gibi derin öğrenme mimarilerinin sık kullanıldığı çalışmaların temel yapıtaşını oluşturmaya başlamıştır. Transformer ağları dikkat mekanizmalarına dayalı standart bir ağ mimarisi sayılabilir.

Yapay zekâ modellerine verilen girişlerin belli başlı olanları seçip odaklanmasını, böylelikle daha etkin ve hızlı bir öğrenme hedeflenmiştir. Ayrıca pekiştirmeli öğrenme, sembolik matematik ve konuşma tanıma gibi çeşitli görevlerde de transformer ağları kullanılmaya başlanmıştır. Aslında Google ilk olarak transformer ağlarını kendi çeviri sistemlerinde kullanmıştır. Elde ettikleri başarılı sonuçlardan sonra [215]'teki çalışmalarında ilk defa bilgisayarlı görü tekniklerini kullanarak nesne tanıma, nesne tespiti gibi sorunları çözmek için kullanmıştır.

DETR daha önce yapılmış nesne tanıma sistemlerine ve algılama sistemlerine kıyasla tümüyle farklı bir mimariyle iş akışının merkezine transformer yapılarını koyan ilk çalışma özelliği taşımaktadır. DETR, bünyesinde birçok parametre ayarlayan ve düzenlemeler yapan 3 yeni modülden meydana gelmektedir. Birincisi gelen veriye göre daha kompakt bir özellik çıkarılması için yeni bir CNN modeli önerilmektedir. İkincisi, graf yapılarında da benzeri kullanılan "Bipartite Matching (2 Parçalı Eşleştirme) Loss" modülüdür. Bu modülün temel amacı bire bir eşleştirme yaparak, düşük kaliteye sahip tahminlerin elenmesi veya sayısal olarak önemli ölçüde azaltılmasıdır. Bipartite matching loss, macar algoritmasına göre tasarlanmıştır. Üçüncü modül ise "Tansformer Kodlayıcı/Kod Çözücü" modülüdür. Transformer kod çözücü normal çözücülerden farklıdır. Çünkü normal çözücüler N adet giriş bile olsa, tek seferde sadece bir öğenin kodunu çözer. Sabit boyutlu N sınırlayıcı kutu seti tahmin edildiğinde N genellikle ilgili görselde odaklanılan nesnelerin gerçek sayısından çok daha fazladır. Bununla birlikte kutuda hiçbir nesnenin bulanamadığını göstermek amacıyla ek bir özel sınıf etiketi ∅ tanımlıdır. Bu sınıf, klasik nesne algılama çalışmalarındaki "arka plan" sınıfına benzer bir mantığa sahiptir. Fakat DETR veya transformer sistemindeyse paralel kod çözme tekniğiyle N giriş için N çıkışı paralel olarak çözer. Sonuçta üretilecek tahmin ise İleri Besleme Ağı (FFN) tarafından hesaplanır. FFN normalize edilmiş haliyle merkez koordinatlarını, genişlik ve yükseklik değerlerini tahmin eder. FNN, ReLU x boyutunda (değişken boyutta) gizli katman ve 3 tane doğrusal katmandan oluşan bir algılayıcıdır. Sonrasında gelen doğrusal katmanda da softmax fonksiyonuyla sınıf tahmin edilir [215].

2022 yılında yayınlanan çalışmada [213], transformer network tabanlı literatürde ilk sayılabilecek çok basit bir sistem önerilmiştir. Sistemde VGG16 tabanlı CNN modelle

özellikler çıkarılıp transformer kodlayıcıya ve daha sonrasında kod çözücüsüne yollanır. Burada da sınıflandırma için olasılıklar hesaplanır. Şekil 3.22'de Huang'ın sunduğu transformer tabanlı mimari verilmektedir.



Şekil 3.22. Huang transformer mimarisi.

3.2.4. Performans Kıyaslaması

Bu bölümde hem hibrit sistemler (geleneksel yöntemlerin ve derin öğrenme yöntemlerin birlikte kullanıldığı) hem de uçtan uca derin öğrenme ağı tabanlı mimariler performans açısından kıyaslanmaktadır. Çizelge 3.2'de verilen genelde sık kullanılan veri setleri (AVLetters, GRID, LRW, OuluVS2 gibi) üzerinde çalışmış mimariler değerlendirilmektedir. Çünkü çalışmaların performansını en çok etkileyen parametrelerden biri de kullanılan veri setidir. Ayrıca değerlendirme aşamasında modellere ait bilgiler özet şeklinde sunulmaktadır

Harf tanıma amacıyla sık kullanılan veri setlerinden olan AVLetters veri setini kullanan [167,173,181,207,212] olmak üzere 5 tane çalışma sunulmuştur. İlki Ngiam ve arkadaşları tarafından sunulmuştur [167] ve derin bir otomatik kodlayıcı ve PCA ile özellikler elde edilmiştir. Bu çalışmada %64.4 WRR sınıflandırma başarısı gösterilmiştir. Buna karşılık Moon ve arkadaşları standart bir derin inanç ağı kullanarak ham görüntülerin özelliklerini ve sınıfını elde etmek için önerdikleri yöntemde görsel verileri ses verilerinden aktarılan ek bilgilerle besleyen bir model kurgulamışlardır. %55.3 WRR elde edilmiştir. Petridis [181] geliştirdiği sistemde ilk olarak girişe verilen yüksek boyutlu bir karenin verilerini, düşük boyutlu bir

temsile(vektöre) indirgemek için darboğaz modeline ait özellikleri de kullanarak derin bir otomatik kodlayıcı eğitmeyi önermiştir. Daha sonra darboğazdan çıktı olarak üretilen vektörler, DCT yöntemiyle özellik vektörü haline getirilir. DCT'nin ürettiği özellik vektörü, zamansal dinamikleri modellemek için bir LSTM ağına gönderilir ve sınıflandırılır. Bu modelle %58.1 WRR elde edilmiştir. Hu ve arkadaşlarının çalışmasında [207] çok duyulu verilerden (video için ses ve görüntü) anlamlı bilgiler çıkarma ve görsel-işitsel modaliteler (görme, duyma gibi algıların ana yollarından her biri) arasında ortak bir öğrenme yeteneğine sahip olan "Tekrarlayan Zamansal Çok Modlu Kısıtlı Boltzmann Makineleri" adlı çok modlu RBM'lere dayalı bir sistem önerilmiştir. Bu sistem ile %64.63 WRR bildirilmiştir. Mesbah [212] çalışmasında 2 boyutlu Hahn CNN tabanlı bir sistem önermiştir. Özellik çıkarım ve sınıflandırma tamamen bu yapıyla yapılmıştır. Sadece ekstradan veri artırımı gerçekleşmiştir. Bu çalışmayla %59.23 WRR sonucuna ulaşılmıştır. İlginç bir şekilde bu sonuçlar çoğu geleneksel sistemlerle elde edilenlerin bile altındadır. Örneğin [138]'de sunulan RFMA tabanlı sistem %69.6 WRR elde etmiştir. Bu nedenle, AVLetters gibi veri setlerinde harf tanıma için geleneksel veya hibrit sistemler hala derin öğrenme tabanlı sistemlerden daha iyi performans gösterir. Bunun nedeni, sağlam derin öğrenme sistemlerini eğitmek için çok yüksek sayıda veri gerekmektedir. AVLetters sahip olduğu veri boyutuyla bunu karşılamıyor olabilir.

Kelime veya cümle tanıma için en çok kullanılan veri setleri GRID, LRW ve OuluVS2 olmuştur. GRID veri seti için 7 farklı mimari tespit edilmiştir. [213]'te Transformer, kodlayıcı, kod çözücü sistemiyle oluşturulan mimariyle kelime sınıflandırmada %45.81 WRR elde edilmiştir. Wand GRID veri seti için 3 model sunmuştur. Birincisi [142] bir ileri besleme katmanı, ardından 2 adet tekrarlayan LSTM katmanından oluşur. Birinci sistemiyle %79.5 WRR sonucu elde edilmiştir. İkinci ve üçüncü sistemlerinde [53,194] üç tane ileri besleme katmanı bir LSTM katmanına bağlanır. Konuşmacıya bağlı testlerde %83.3 ve %84.7 WRR performansı göstermiştir. [194]'te oluşturulan sistem, sistemin daha önce verisini görmediği konuşmacılarla da test edilmiş olup %42.4 WRR bildirilmiştir. Buna karşılık, Assael ve arkadaşları tarafından sunulan çalışmada [188] Bi-LSTM'lerle birlikte spatio-temporal CNN'in kullanıldığı bir sistem önerilmiştir. Yapılan konuşmacıya bağlı test sonucunda %93.4 ile daha yüksek bir tanıma oranı elde edilmiştir. Chung ve arkadaşları dikkat mekanizmalarıyla

CNN ve LSTM ağlarına dayalı bir sistemle konuşmacıya bağlı bir test için %97 WRR elde etmişlerdir. Son olarak Xu ve arkadaşları da [165] 3D-CNN'leri, otoyol ağlarını, Bi-GRU'ları ve dikkat mekanizmalarını birleştiren bir sistemle önceki yöntemlerin çoğundan daha iyi bir performans göstererek [112]'den %97.1 WRR ile daha yüksek performans göstermiştir. Konuşmacıdan bağımsız en yüksek doğruluk oranlarından biri de [144] tarafından bildirilen %57 WRR sonucudur. Bu sistemle birlikte geleneksel sisteme göre performansta önemli gelişmeler gösterilmiştir.

LRW veri seti için Chung ve arkadaşları [55] CNN'lere dayalı uçtan uca bir mimari sunmuş ve sonuç olarak %61.1 WRR bildirilmiştir. Stafylakis [32] 3D-CNN, ResNet ve BiLSTM'lere dayalı bir sistem sunmuştur. Stafylakis, [55]'te yapılan çalışmaya göre %20'den fazla gelişme göstererek %83 WRR bildirmiştir. Benzer şekilde, Petridis ve arkadaşları [163] 3D-CNN, ResNet ve Bi-GRU ağlarına dayalı bir sistem sunmuş ve %82 WRR bildirilmiştir. Yine başka bir katkı da Chung ve arkadaşları tarafından sağlanmıştır [112]. Bu çalışmada dikkat mekanizmalarıyla birleştirilmiş CNN ve LSTM ağlarına dayalı bir sistem önerilmiştir. Şu ana kadar LRW için elde edilen en yüksek ikinci sonuç olan %84.5 WRR elde edilmiştir. Mesbah ve arkadaşları da [212] yaptıkları çalışmada HCNN + Veri Artırımı (SI) mimarisini kullanmıştır. LRW için en yüksek performans olan %89.95 WRR değeri elde edilmiştir.

OuluVS2 için 13 mimari sunulmuştur. Saitoh [56] ve Chung [55] temelinde CNN bulunan birkaç uçtan uca sistem sunmuştur. Saitoh tarafından önerilen üç sistem çeşitli test türlerinde, %81.1 ile %86.5 WRR arasında tanıma oranları bildirirken Chung ise %94.1 WRR oranına ulaşmıştır. Bu 2 çalışma arasındaki temel fark, [56]'daki ağların girdi olarak CFI'ları kullanması, [55] ise doğrudan videodan gelen kareyi kullanmasıdır. Ayrıca Saitoh CNN'lere dayanan, literatürde de sıkça kullanılan NIN [199], AlexNet [179] ve GoogleNet [184] kullanmıştır. Chung ise geliştirdiği ağı, dudak okuma gibi bir görev için özel olarak eğitmiştir. Sınıflandırıcılar olarak da LSTM'ler veya BiLSTM'ler ile çeşitli mimariler önerilmiştir. Sınıflandırıcıda LSTM, BiLSTM kullanan mimarilerde özellik çıkarma için çeşitli yöntemler uygulanmıştır. [57,154] çalışmalarında CNN'ler, [187] çalışmasında VGG-M ve SyncNet, [35,202,203] çalışmalarında otomatik kodlayıcılar, [192] çalışmasında 3D-CNN, [209]'da ise PCA-NN kullanılmıştır. [209]'da sınıflandırıcıya ek olarak, zamansal

dinamikleri modellemek ve sınıflandırmak için HMM'ler kullanılmıştır. Bu mimariler için bildirilen tanıma oranları %31.9 ve %94.7 WRR arasındadır. En düşük tanıma oranı VGG-M [187] kullanılan sistemde gerçekleşmiştir. Bu düşük doğruluk oranı, VGG-M'in nesne tanıma ve sınıflandırma görevleri için geniş bir veri seti olan ImageNet üzerinde önceden eğitilmiş olması, ancak dudak okumaya özgü olmaması nedeniyle açıklanabilir. Buna karşılık Petridis ve arkadaşları [203], %94.7 WRR ile en yüksek performansı bildiren kodlanmış özelliklere dayalı bir sistem sunmuştur. [203]'e çok yakın başarı oranına sahip bir başka çalışma olan [19]'da %94.1 WRR elde edilmiştir. Geleneksel yöntemlerde elde edilmiş en yüksek başarı oranının %74 ile [154] çalışmasında olduğuna göre derin öğrenme mimarileriyle beraber geleneksel yöntemlere kıyasla başarı oranlarında %20'den fazla bir gelişme gösterilmiştir.

Yukardaki paragraflardan, DNN'lerin, kelime veya cümle sınıflandırma görevlerine odaklanan GRID ve OuluVS2 gibi veri setlerindeki dudak okuma sistemlerine önemli performans iyileştirmeleri kazandırdığı görülebilmektedir. Bu gelişmeler, araştırmacıları daha gerçekçi ayarlamalara yönelmeye ve sürekli dudak okumayı hedefleyen sistemler üzerine çalışmalarına sebep olmuştur. Sürekli dudak okuma sistemlerinin genel amacı, normal sistemlerdeki tek bir girdiye karşılık sınıf üretmekten farklıdır. Çünkü normal sistemler eğitilip test aşamasına geldiğinde ayrı ayrı süreçlerde veriler gönderilir. Örneğin 2 cümle varsa bunlar teker teker yollanır. Sürekli dudak okuma sistemlerindeyse amaçlanan durum, bu 2 cümlenin birlikte yollanmasıdır. Bu tür ayarlamalar, kelime veya cümle sınıflandırma görevlerinde bulunanlardan çok daha zordur, çünkü her cümle bilinmeyen bir yapıya sahiptir ve zaman sınırları önceden bilinmeyen yani kelimenin-cümlenin başlayıp bittiği zaman aralıklarının bilinmediği rastgele sayıda kelime içerebilir. Bu nedenle sürekli dudak okumayı hedeflerken, minimum ayırt edilebilir fonemlere (dil birimlerine) yönelik tahmin mekanizmaları geliştirmek çok daha uygun olur. Başka yollar denendiğinde kelime veya cümle yapılarının varyasyonları çok daha karmaşık ve sayıca fazla olacaktır. Örneğin sadece bir kelimeye gelecek ek sayısı bile sistemi oldukça karmaşık hale getirir. Fakat sistem, hece gibi çok ufak ses birimleri üzerinden tanıma işlemi gerçekleştirilebilirse çok daha sağlıklı çalışır. Uçtan uca derin öğrenme mimarilerindeki son gelişmeler, gerçekten de tam sözcükler veya önceden tanımlanmış cümleler yerine fonemleri [108,169,183,208] veya karakterleri

[52,57,112,55] tahmin etmeye çalışan otomatik dudak okuma sistemlerine odaklanmıştır. Örneğin Mroueh ve arkadaşları [108], büyük ölçekli fakat bilimsel araştırmalara açık olmayan bir görsel işitsel veri seti olan IBM AV-ASR veri setini kullanarak fonemleri tahmin etmek için ileri beslemeli DNN'leri önermiştir. Noda [169,183] ve Takasima [208] ise yaptığı çalışmalarda CNN'leri ve HMM'leri kullanan başka mimariler sunmuştur. [216] çalışmasında "ATR Japanese" korpusunu kullanarak Japon ses birimlerini tanımaya çalışmışlardır. Geliştirdikleri 3 sistemde sırasıyla %22.5 WRR, %37 WRR ve %51 WRR elde edilmiştir. Çok kullanılan GRID veri setini kullanan başka bir mimari, yakın zamanda Xu ve arkadaşları tarafından karakter tabanlı sınıflandırma için sunulmuştur [52]. Bu çok derin ve kompleks ağ, 3D-CNN'leri, otoyol ağlarını, Bi-GRU'ları ve dikkat mekanizmalarını birleştirmiştir. Sistemin elde ettiği performans ise %97.1 WRR olarak bildirilmiştir.

Chung ve arkadaşları [57,112] dikkat mekanizmalarıyla birleştirilmiş CNN ve LSTM ağlarına dayalı bir mimari sunmuştur. Sistemlerini de kısmen yeni sayılabilecek MV-LRS ve LRS gibi büyük ölçekli veri setlerinde test etmişlerdir. Karakter tabanlı tanıma için her bir veri setinde sırasıyla %43.6 WRR ve %49.8 WRR elde etmişlerdir. Daha yakın zamanlarda, Afouras ve arkadaşları [55], LRS veri setinde test edilen karakter tabanlı tanıma görevine odaklanan üç mimarinin performans kıyaslamasını yapmışlardır. Mimarilerde kullanılan özellik çıkarım yöntemi neredeyse aynıdır. Sadece sınıflandırma aşamasında ciddi farklılıklar göstermiştir. Sınıflandırma adımında Bi_LSTM'leri kullanan model için %37.8 WRR, her giriş kanalı için tek bir evrişim filtresinin uygulandığı evrişim türü olan "Depthwise Convolution" katmanlarını kullanan model için %45 WRR ve bir dikkat mekanizmasından birkaç kez paralel olarak geçen dikkat mekanizmaları içeren modül olan "Multi-Head Attention (Çok Başlı Dikkat Mekanizması)" katmanları ve kodlayıcı-kod çözücü yapılarını içeren mimariyle de %50 WRR elde etmişlerdir.

Bu nedenle, en yeni derin öğrenme ağlarını kullanan çalışmalarda farklı deneysel ayarlamalara rağmen geleneksel sistemler tarafından bildirilen performansın neredeyse 2 katı olan WRR bildirilmiştir. WRR değerleri açısından iyileştirme ortalama %20'dir [104,156,158]. Bu iyileştirmeler, dudak okumada geleceğe yönelik atılmış büyük bir adım olsa da elde edilen sonuçların hala görsel konuşmayı tamamen

çözebilecek bir sistemden uzak olduğunu belirtmekte fayda vardır. Gündelik hayattaki gerçek dünya senaryolarında en iyi performans gösteren dudak okuma sistemleri şu anda %50'lik WRR değerlerine yaklaşıyor. Bu da konuşmada söylenilen ifadelerin yarısını anlayamayacağımız anlamına gelmektedir. Bu nedenle DNN tabanlı sistemler ve büyük ölçekli veri setleri dudak okuma alanını önemli ölçüde ilerletmiş ve iyileştirmiştir. Ancak sürekli otomatik dudak okuma problemi ise tam anlamıyla çözülmemiş bir problem olmaya devam etmektedir.

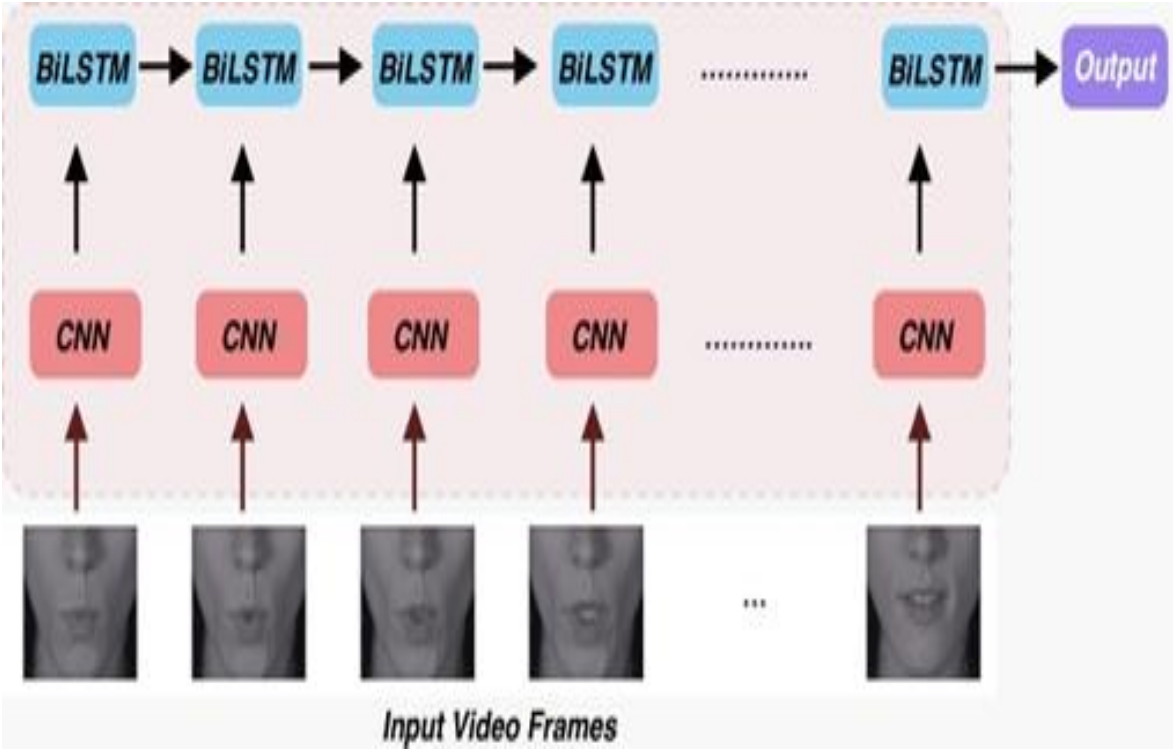
BÖLÜM 4

DENEYSEL ÇALIŞMALAR

Video gibi zaman serisi verileri, belirli ve eşit zaman dilimlerinde kesikli bir şekilde kronolojik sıraya göre saklanan verilerdir. Bu tip verilerde her görüntü bir önceki ve bir sonraki kareyle ilişkili olduğundan aralarındaki korelasyonun iyi belirlenmesi gerekir. Önemli olan bir diğer husus ise işlenmesi gereken karenin zaman düzleminde hangi noktada bulunduğudur. Bulunduğu zaman sırasına veya konumuna göre veri kümesindeki anlamı ve temsil ettiği değer farklılaşır. Literatürde Türkçe dudak okumaya dair herhangi bir veri seti yoktur. Böyle bir veri seti bulunmadığından dolayı otomatik dudak okumaya dair Türkçe için önerilmiş bir model de doğal olarak bulunmamaktadır. Bu sebeple ilk olarak literatüre "Türkçe Dudak Okuma Veri Seti" kazandırılmıştır. Sonraki bölümde oluşturulan veri setinin test edileceği örnek bir model geliştirilmiştir. Model için temel adımlar şu şekildedir;

- Dudak tespiti ve kırpılması
- Özellik çıkarımı
- Sınıflandırma
- Etiketleme

Tez kapsamında önerilen modelde dudak okuma için gerçekleştirilen temel adımlar Şekil 4.1'de özetlenmektedir.



Şekil 4.1. Geliştirilen Yöntemin Temel Aşamaları.

İlk aşamada giriş videosunda dudak tespiti yapılır. Tespit edilen bölge kareden kırpılır. İşlemlere kırpılan dudak bölgesine ait görüntüyle devam edilir. Bu kırpılan görüntüden CNN tabanlı bazı modeller kullanılarak özellikler çıkarılır ve BiLSTM ile sınıflandırma gerçekleştirilir. Bu aşamalar ilerleyen bölümlerde detaylandırılmıştır.

4.1. TÜRKÇE DUDAK OKUMA VERİ SETİ

Dudak okuma alanında hazırlanmış veri setleri hazırlanış amacı (kelime, cümle, harf, renk, aylar vs.), kullanılan dil, kişi sayısı, telaffuz sayısı, veri sayısı, FPS, çözünürlük gibi özelliklerine göre birbirinden ayrılır. Literatürde şu anki haliyle bilimsel araştırmalara açık olacak şekilde yayınlanmış Türkçe dilinde herhangi bir veri seti olmadığından dolayı tez kapsamında 111 kelimelik ve 113 cümlelik bir veri seti oluşturulmuştur. Veri seti oluşturulurken literatürdeki setlerin iyi incelenmesi gerekmektedir. Çünkü literatürdeki veri setlerinin hangi kurallara göre oluşturulduğu veya hangi ihtiyacı gidermek amacıyla yayınlandığı gibi hususlar tespit edilebilmektedir.

Veri setinde yer alan video görüntülerinin tamamı aynı ortam ve ışık şartlarında elde edilmiştir. Videolarda konuşmacı ve kamera arasında 1.5 metre mesafe yer almaktadır. Görüntülerin çekimi esnasında iPhone 11 marka mobil cihazın 12 megapiksel geniş açılı kamerası kullanılmıştır. Çalışma kapsamında oluşturulan veri setleri otomatik dudak okuma sistemlerinden iki alan olan kelime ve cümle tanıma üzerine oluşturulmuştur. Literatürde istenen kriterlere uygun olarak hazırlanan dudak okuma veri setlerine ait özellikler "Kelime" veri seti için Çizelge 4.1'de ve "Cümle" veri seti için çizelge 4.2'de belirtilmektedir.

Çizelge 4.1. Kelime veri setine ait özellikler.

Veri Seti Özellik	Değer
Kişi	18 Kadın + 6 Erkek
Tekrar Adeti	15
Sınıf	111 farklı kelime
Çözünürlük	1920 x 1080 piksel
FPS	30
Toplam Video	39960
Tür	Kelime
Konuşmacı Sayısı	Tekli Konuşmacı
Mesafe	1,5 Metre
Açı	90° / Frontal View
Cihaz	İphone 11 Mobil
Kamera	12 MP Kamera
Görüntü	Mevcut
Ses	Mevcut
ROI	Mevcut Değil
Ortam	Kontrollü Ortam

Çizelge 4.2. Cümle veri setine ait özellikler.

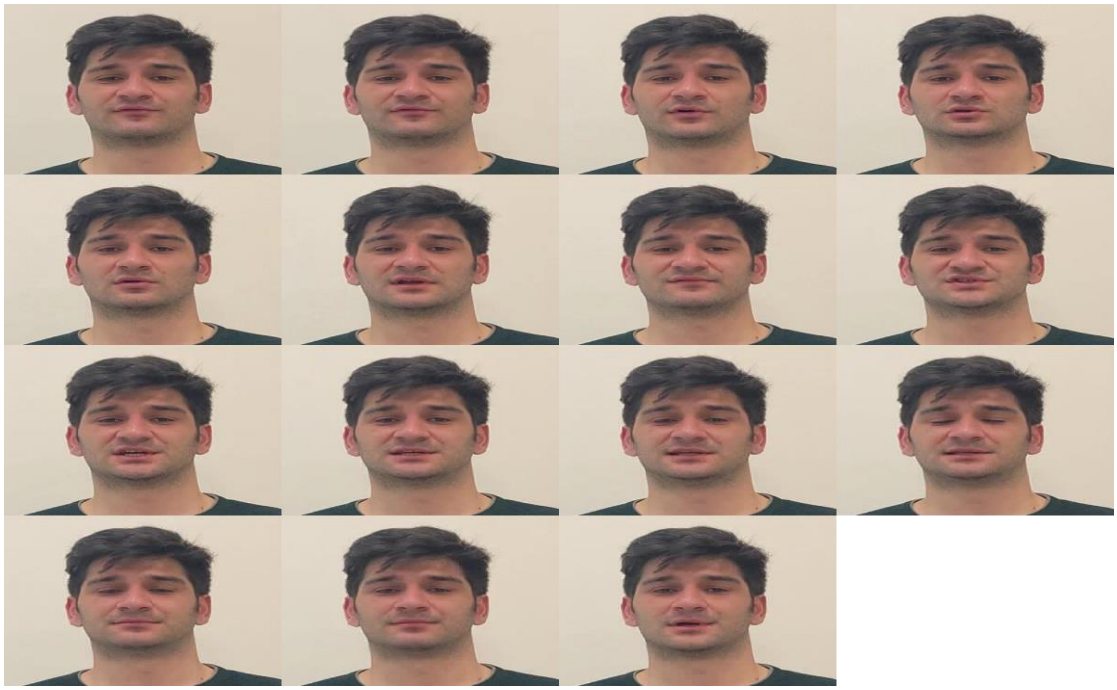
Veri Seti Özellik	Değer
Kişi	18 Kadın + 6 Erkek
Tekrar Adeti	10
Sınıf	113 farklı cümle
Çözünürlük	1920 x 1080 piksel
FPS	60
Toplam Video	27120
Tür	Cümle
Konuşmacı Sayısı	Tekli Konuşmacı
Mesafe	1,5 Metre
Açı	90° / Frontal View
Cihaz	İphone 11 Mobil
Kamera	12 MP Kamera
Görüntü	Mevcut
Ses	Mevcut
ROI	Mevcut Değil
Ortam	Kontrollü Ortam

Bu veri setlerin oluşturulması konusunda dikkat edilen ilk husus literatürdeki çalışma sıklığı ve sayısı olduğundan dolayı daha çok kişinin çalışacağı bir veri seti üzerinde

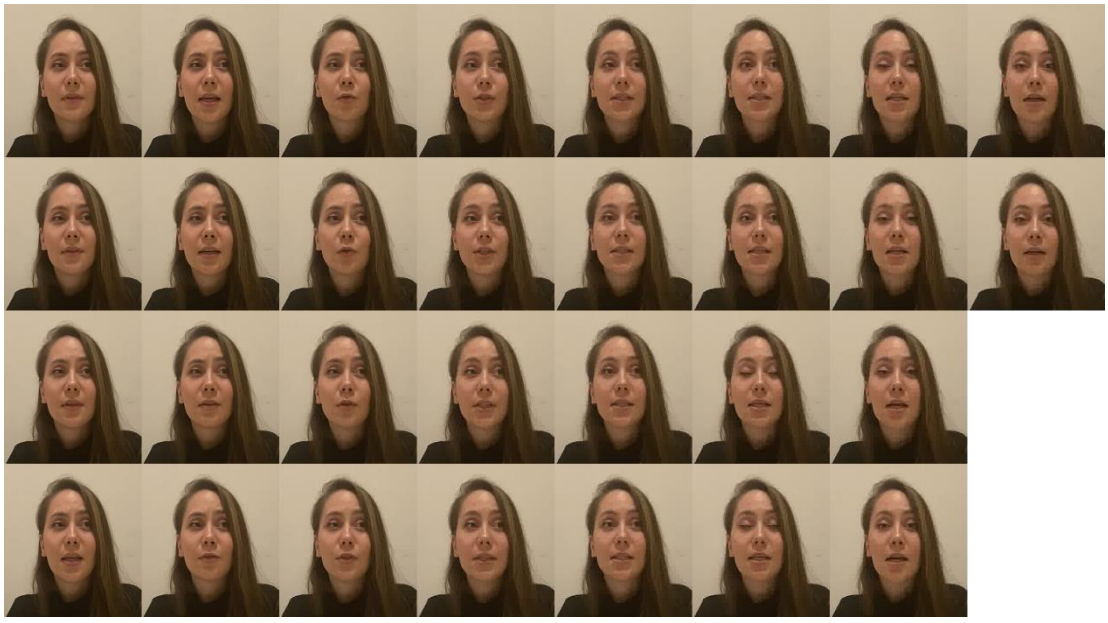
durulmuştur. Bu amaçla ilk olarak Bölüm 2.1'de veri setlerinin detaylandırıldığı bölümde belirtildiği üzere kelime ve cümle tanıma görevleri diğer alanlara göre daha popüler olması göz önünde bulundurulmuştur. Diğer çalışmalardan farklı olarak ise çözünürlük, FPS, kontrollü ortam, konuşmacı sayısı, tekrar sayısı gibi özellikler gösterilebilir. Diğer çalışmaların çoğu 1920 x 1080 piksel çözünürlükten daha düşük çözünürlüklere sahiptir. Kelime veri seti 30 FPS olmasına karşın cümle veri seti 60 FPS oranına sahiptir. Çizelge 2.1 incelendiğinde 60 FPS oranına sahip veri seti sayısının 2 tane olduğu görülmektedir. Bir konuşmacının bir ifadeyi tekrar sayısı da kelime için 10 ve cümle için 15 olması literatürdeki veri setlerinin çoğundan daha yüksektir. Tekrar sayısının yüksek olması konuşmacıya bağımlı veya konuşmacıdan bağımsız test performansını arttırmaktadır.

Kelime veri setinin 30 FPS ve cümle veri setinin 60 FPS olacak şekilde farklı oluşturulmasının sebebi, literatürdeki birçok çalışma 30 FPS üzerinden oluşturulmuş olmasıdır. Fakat gelişen teknolojiyle birlikte telefonlarla bile yüksek çözünürlük ve FPS oranına sahip videolar çekilebildiğinden dolayı yenilik olması açısından 60 FPS / yüksek çözünürlük şeklinde cümle veri seti oluşturulmuştur. Kelime veri setiyse literatürdeki veri setlerinin çoğuyla benzer FPS oranına sahip olacak şekilde oluşturulmuştur. Böylece geliştirilen bir model diğer veri setleriyle birlikte oluşturmuş olduğumuz veri setiyle de test edilebilir ve performans karşılaştırması yapılabilir.

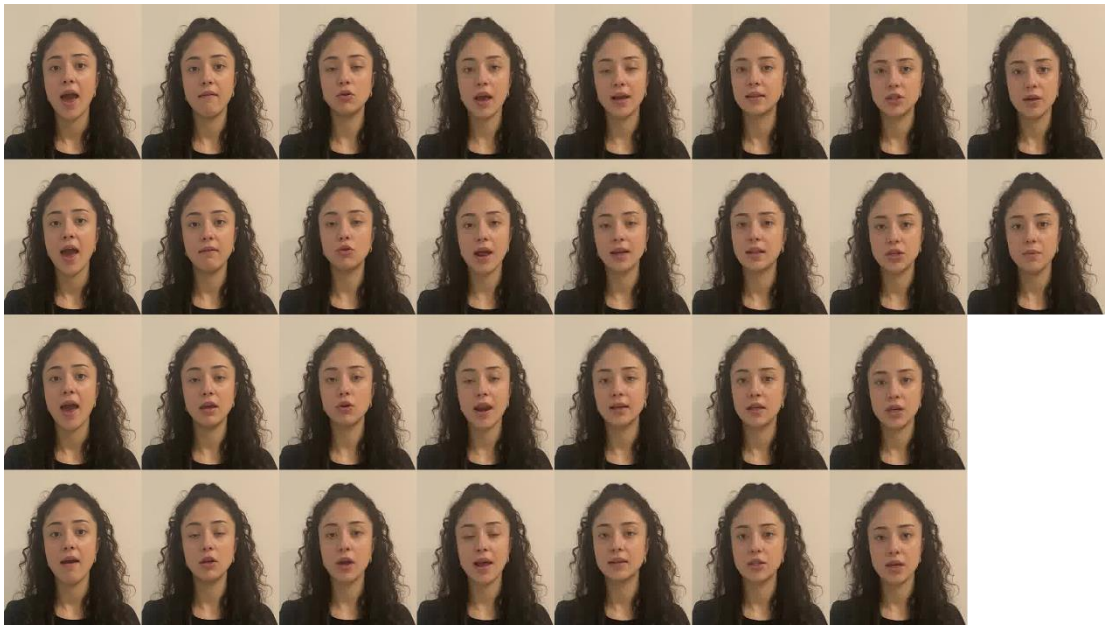
Tez kapsamında oluşturulan veri setlerine ait örnek görseller Şekil 4.2, Şekil 4.3 ve Şekil 4.4'te verilmiştir.



Şekil 4.2. Kişi-1 “Fotoğraf Çekinelim mi?” cümlesi.



Şekil 4.3. Kişi-2 “Teknoloji Hızla İlerliyor” cümlesi.

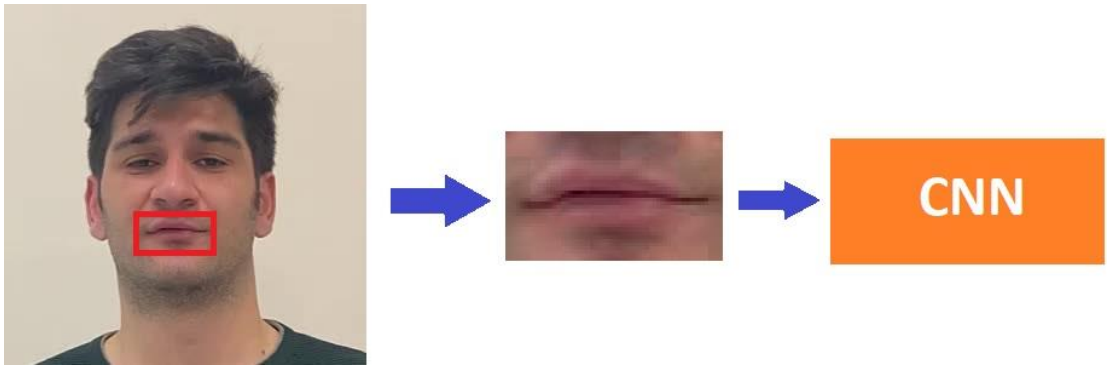


Şekil 4.4. Kişi-3 “Eve Yürüyerek Gideceğim” cümlesi.

Veri seti https://www.furkansabaz.com/turkish_lip_reading linkinde diğer araştırmacıların kendi çalışmalarında kullanabilmesi için paylaşılmaktadır.

4.2. GELİŞTİRİLEN YÖNTEM ve MİMARİ

Bu çalışmada Türkçe dudak okuma için önerilen CNN-BiLSTM tabanlı model Şekil 4.1’de gösterilmektedir. Önerilen modele bir videonun her karesi giriş olarak ayrı ayrı yollanır. Modelde ilk aşamada dudak bölgesinin tespiti yapılıp ilgili alan görüntüden kırpılmaktadır. Dudakların tespiti ve kırpılmasına ait detaylar Bölüm 4.2.1’de verilecektir. Videodan kırpılan dudak bölgesi, ilgili CNN ağının giriş boyutuna göre yeniden ölçeklendirilir. Videodaki tüm kareler kırpıldıktan sonra ikinci aşamada CNN tabanlı ağa gönderilir. CNN ile her bir kare için özellik vektörü elde edilir. Bir videoya ait her karenin özellik vektörü oluşturulduktan sonra bu vektörler eşit boyutlu olacağından dolayı tek bir sayısal matris içerisinde birleştirilerek tek başına o videonun özellik vektörünü temsil eder hale gelir. Böylece bir kelime veya cümlenin söylenişinin sınıflandırılması için gereken özellik vektörleri tek parça halinde elde edilmiş olur. n adet kareye sahip videonun özellik matris boyutu Eşitlik 4.1’e göre hesaplanır.



Şekil 4.5. Dudağın kırpılması ve CNN modeline gönderilmesi.

Ayrıştırılan görseller, özellik çıkarmak amacıyla CNN modeline gönderildiğinde modelin yapısına göre elde edilecek özellik vektörünün boyutu ve yapısı değişmektedir. Fakat tüm karelere aynı CNN modeli uygulandığı için tüm karelerden elde edilen özellik vektörünün boyutu hep sabit kalır. Farklı CNN modelleri için farklı özellik vektörleri elde edileceğinden dolayı hangisinin daha iyi sonuç verdiği sonuçlar bölümünde incelenmektedir. Çünkü modeldeki BiLSTM katmanı tüm modellerde değişmeden kalmaktadır. Modelin sonucuna direkt olarak etki eden en önemli husus, CNN katmanında kullanılan ağdır.

$$FV = k\ x\ n \tag{4.1}$$

Veri setinde yer alan bir videoya ait toplam kare sayısı k ve CNN modelinin tek bir görüntü için çıkardığı özellik vektörünün boyutu 1xn ise CNN modeli kullanarak elde edilen özellik vektörünün boyutu k x n olur. Bu yöntemle art arda gelen karelerin özellik vektörü tek bir matriste birleştirilir. Söz konusu özellik matrisi sırayla BiLSTM hücrelerine yollanarak sınıflandırma yapılır.

4.2.1. Dudakların Kırpılması

Geliştirilen modelin ilk adımı olan dudak kırpma işleminin gerçekleştirilebilmesi için genelde yapılan ilk işlem yüzün tespitidir. Bu yüzden aslında problemin adı literatürde sıkça rastlanılan yüz tespiti (face detection) olmaktadır. Yüz tespiti için literatürde her ne kadar çok fazla yöntem olsa da temelinde 2 tip yöntem bulunmaktadır. Bunlardan ilki algoritmaya dayalı yöntemlerdir. Algoritmaya dayalı bu yöntemler genelde konum

bazlı, görüntü işleme tabanlı çalışmaktadır. İlk yöntemde bazı durumlarda makine öğrenmesi de kullanılabilmektedir. İkinci yöntem ise derin öğrenme tabanlı sistemlerden oluşmaktadır.

Tez kapsamında hem klasik yöntemlerden hem de makine öğrenmesi yöntemlerinden birer tanesi kullanılmıştır. 2 ayrı yöntemin modeldeki performansları, ilgili bölümlerde değerlendirilmiştir.

4.2.1.1. Klasik Yöntem- Haar Cascade

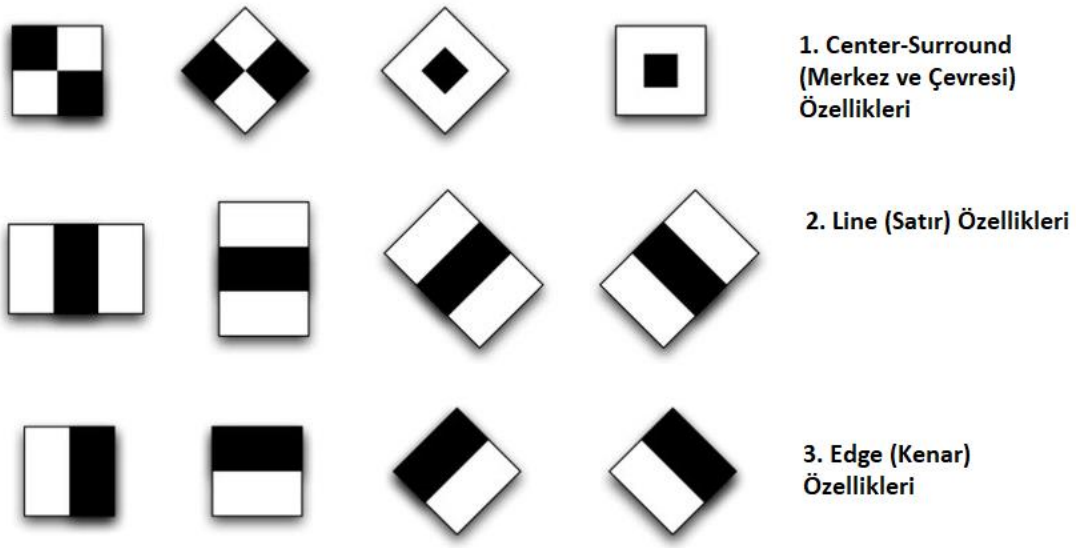
Veri setindeki video karelerinde kişinin öncelikle dudaklarının tespit edilmesi ve görüntüden kırpılması gerekmektedir. Yüzün tespiti aşamasında geleneksel yöntemlerde temelde 2 alt yöntem bulunmaktadır. Bunlar analitik yaklaşımlar ve bütünsel yaklaşımlardır. Bütünsel yaklaşımlarda insan yüzü bir bütün olarak ele alınmaktadır. Huang ve arkadaşlarının [219] geliştirdikleri "Point Distribution Model (Nokta Dağıtım Modeli- PDM)" ve Pantic'in [220] önerdikleri modelde, görüntülerde yer alan yüzü hem yan profilden hem de önden işleyen bu sistemler bütünsel yaklaşıma örnek gösterilebilir. Analitik yaklaşımdaysa profil veya diğer açılardan yüzle ilgilenmek yerine sadece ön yüze ait karakteristik yüz bölgelerini veya elemanlarını beraber ele alır [221]. Hara ve arkadaşlarının önerdikleri "iris localization (iris lokalizasyonu)" yöntemi [222], Rowley'in [223] temelinde yapay sinir ağlarının bulunduğu yöntem ve [224]'te yer alan yöntem gibi yaklaşımlar analitik yaklaşıma örnek verilebilir.

Yüz tespiti gibi konularda geleneksel yöntemler arasında oldukça sık kullanılan "Haar Cascade" algoritması ve videoda dudakların takibi için de Viola Jones algoritması kullanılmıştır. Haar özellikleri vasıtasıyla nesne saptama yaklaşımı, Haar dalgacık dönüşümünü kullanarak Viola ve Jones tarafından geliştirilmiştir [217]. Algoritma temelde 4 adımdan oluşur.

- Haar özelliklerinin hesaplanması.
- İntegral görüntülerinin oluşturulması.

- Adaboost öğrenme algoritmasının kullanılması.
- Cascading sınıflandırıcısının entegre edilmesi.

İlk aşamada haar özellikleri hesaplanır. Bu aşama 2. adım olan integral görüntülerinin oluşturulmasıyla bağlı çalışır. Haar özelliklerinin hesaplanması, integral görüntüleri üzerinden olduğu için bunlar ayrı olarak tanımlanmış olsalar bile birlikte çalışan iki adımdır. Özelliklerin hesaplanma amacı yüz tespitini yaparken Viola Jones algoritmasının eğitilmesi için bir veri seti gereksiniminden dolayıdır. Bu algoritma, diğer makine öğrenmesi modellerine benzer şekilde sınıflandırıcıyı eğitmek için insan yüzlerinin çok sayıda pozitif görüntüye (insan yüzüne ait görseller) ve negatif görüntülere (yüz olmayan) ihtiyaç duymaktadır. Pozitif görüntü içindeki nesneyi tespit etmek için Şekil 4.6'da verildiği gibi oluşturulan alt pencereler, tüm görüntü üzerinde kaydırılarak nesne taraması yapar. Bir haar özelliği, esas olarak bir pencerede belirli bir konumdaki komşu veya bitişik dikdörtgen bölgeler üzerinde gerçekleştirilen hesaplamalarla oluşur. Hesaplama, her bölgedeki piksel yoğunluklarının toplanmasını ve her bölgede toplanan değerler arasındaki farkların hesaplanmasından ibarettir. Şekil 4.6'da verilen pencereler de haar özelliklerini temsil etmektedir.



Şekil 4.6. Haar-Like özellikleri.

Bu özelliklerin büyük bir görüntü için belirlenmesi zor olabilir. İntegral görüntülerin kullanılmasıyla beraber işlem sayısı da azaldığından dolayı ikinci aşamada integral

görüntülerinin oluşturulması devreye girer. İntegral görüntülerin temelde yaptığı en önemli şey haar özelliklerinin hesaplanmasını hızlandırmaktır. Görüntüdeki her piksel için hesaplama yapmaktansa bunun yerine alt dikdörtgenler veya bölgeler oluşturulur. Bu alt dikdörtgenlerin her biri için referanslar oluşturulur. Referanslar da daha sonra haar özelliklerini hesaplamak için kullanılır. İntegral görüntüler, orijinal görüntüyle aynı boyutlara sahip bir matris olacak şekilde 2 boyutlu arama tabloları olarak tanımlanabilir. Bu görüntüler Eşitlik 4.2, 4.3 ve 4.4'e göre hesaplanır.

$$ii(x,y) = \sum_{x=0,y=0}^{n} i(x',y') \tag{4.2}$$

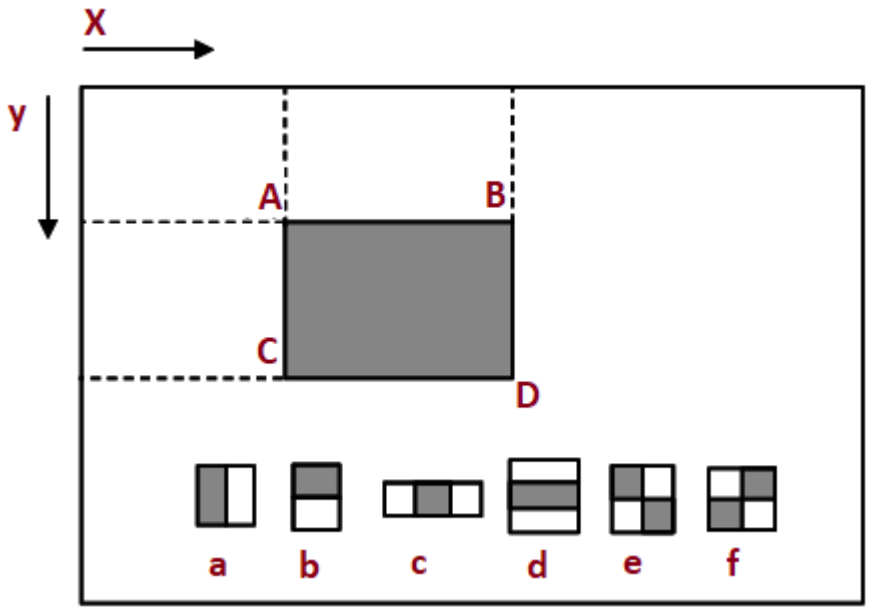
$$\boldsymbol{x' \leq x} \tag{4.3}$$

$$\boldsymbol{y' \leq y} \tag{4.4}$$

Eşitlik 4.2'deki i görüntüyü temsil etmektedir. İntegral görüntünün her öğesi tüm (x,y) ikilisi için, orijinal görüntünün sol üst bölgesinde (pencerenin konumuna göre) bulunan tüm piksellerin toplamını içerir. Bu işlem, yalnızca 4 noktayı kullanarak herhangi bir konumda veya ölçekte yer alan görüntüdeki dikdörtgen alanların Eşitlik 4.4'e göre toplamını hesaplamayı sağlar.

$$\boldsymbol{sum = I(C) + I(A) - I(B) - I(D)} \tag{4.4}$$

Eşitlik 4.5'te verilen A, B, C ve D noktaları Şekil 4.7.'de gösterildiği gibi integral görüntüsü I'ye aittir.

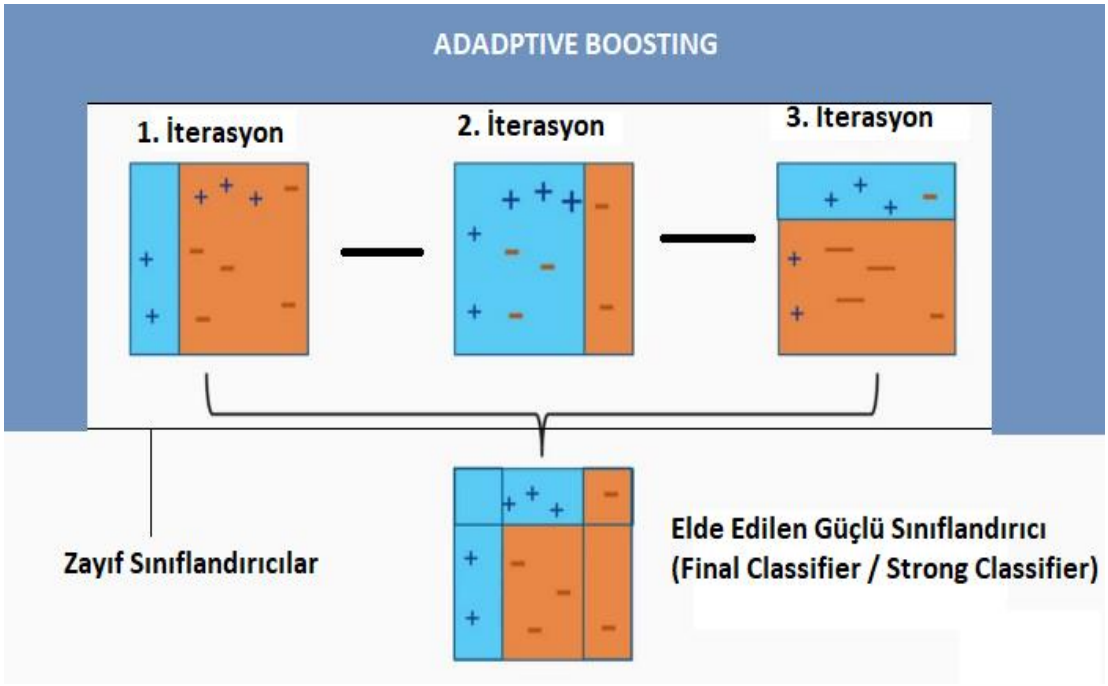


Şekil 4.7. İntegral görüntüsüne ait noktalar.

Haar algoritmasında nesne algılama yapılırken hesaplanan tüm haar özelliklerinin birbirinden alakasız olacağına dikkat etmek önemlidir. Önemli olan sadece nesnenin özellikleridir. Fakat neredeyse yüz binleri bulan haar özelliğinden o nesneyi temsil eden iyi özellikleri belirlemek ayrı bir sorun teşkil etmektedir. Çünkü tüm görüntü üzerinde Şekil 4.6'da verilen küçük alt pencerelerin gezdirilmesiyle beraber negatif ve pozitif alanların tespit işlemi fazlasıyla zaman alıcı olmaktadır. Adaboost ise bu kısımda devreye girmektedir.

Adaboost'un yaptığı şey temelde en iyi özellikleri seçmek ve sınıflandırıcıları da seçtiği özellikleri kullanmaları için eğitmektir. "Adaptive boosting" veya "boosting" işlemi Haar algoritmasına özel değildir. Tüm makine öğrenmesi yöntemlerine uygulanabilir. Algoritmanın nesneleri algılamak için kullanabileceği güçlü bir sınıflandırıcı (strong classifier) oluşturabilmesi için birkaç zayıf sınıflandırıcının (weak classifier) oluşturduğu bir kombinasyon kullanılır. Zayıf sınıflandırıcılar, bir pencereyi giriş görüntüsü üzerinde gezdirerek görüntünün pencerenin olduğu yerdeki her alt bölümü için haar özelliklerinin hesaplanmasıyla oluşturulur. Bu değerler, nesne olmayanları nesnelerden ayıran öğrenilmiş bir eşik değeri ile karşılaştırılır. Bunlar zayıf sınıflandırıcılar olduğu için yüksek doğruluklu güçlü bir sınıflandırıcı

oluşturulabilmesi için çok sayıda haar özelliğine ihtiyaç duyulmaktadır. Şekil 4.8' de Adaptive Boosting için 3 iterasyonlu bir yapı gösterilmektedir.



Şekil 4.8. Boosting çalışma prensibi.

Şekil 4.8'de belirtildiği üzere 1. iterasyonda yani başlangıçta basit bir sınıflandırıcı, sınıflandırılacak veriyi 2 farklı sınıfla sınıflandıracak şekilde yerleştirilir. Sınıfı ne olursa olsun eğer değer doğru olarak sınıflandırılmışsa sonraki iterasyonlarda ona daha az ağırlık verilir. İterasyonlar ilerledikçe ağırlıklar, yanlış sınıflandırılan sınıflara göre değiştirilir daha doğrusu kenar olarak büyür. Örneğin 3. iterasyondaki "-" işaretlerinin boyutuna ve durumuna bakıldığında bu durum daha net anlaşılır. Tüm iterasyonlar tamamlandıktan sonra her bir iterasyonda her sınıflandırıcı için ayrı ayrı hesaplanan ağırlıklar ve hata değerleri üzerinden zayıf sınıflandırıcılar kullanılarak nihai güçlü sınıflandırıcı elde edilir.

Eşitlik 4.5 incelendiğinde ilk olarak her bir veri için başlangıç ağırlık değerleri verilir.

$$W_i = 1/N \tag{4.5}$$

N veri setindeki veri sayısıdır. Hata değerleri de Eşitlik 4.6 üzerinden hesaplanır. Hata değerleri bir sonraki aşamada hesaplanacak ağırlık değerleri için dolaylı olarak kullanılır.

$$err_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i} \tag{4.6}$$

Hata değeri hesaplandıktan sonra Eşitlik 4.7 ve 4.8 ile m değeri üzerinden alfa değeri hesaplanır.

$$a_m = \log\left(\frac{1 - err_m}{err_m}\right) \tag{4.7}$$

$$w_i < -w_i * \exp\left[a_m * I\left(y_i \neq G_m(X_i)\right)\right], i = 1, 2, \ldots, N \tag{4.8}$$

Çıktı değeriyse Eşitlik 4.9 üzerinden hesaplanır.

$$output = G(x) - sign\left[\sum_{m=1}^{M} a_m G_m(x)\right] \tag{4.9}$$

Her bir veri için ağırlıklar Eşitlik 4.9'a göre güncellenir. Bu şekilde yanlış sınıflandırılmış veriler için ağırlıklar yükselir. Doğru sınıflandırılmış olanlar içinse ağırlık düşürülür. Böylece AdaBoost öğrenme algoritması kullanılarak hızlı bir şekilde hem yüz binlerce haar özelliği arasından en iyi olanlar seçilmiş olur hem de elde edilen özellikleri kullanan sınıflandırıcıların eğitimi yapılmış olur.

Son aşamadaysa sınıflandırıcı çalışır. Haar Cascade sınıflandırıcısının karar verme yöntemine ait eşitlikler, Eşitlik 4.10 ve 4.11'de verilmektedir.

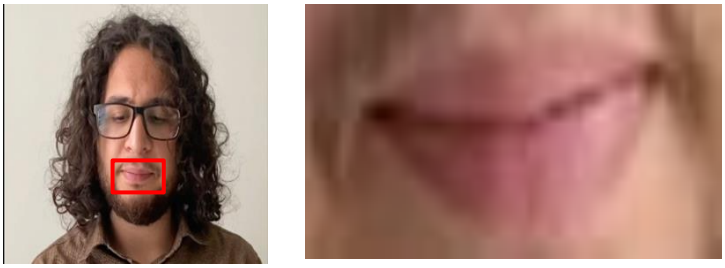
$$C_m = \begin{cases} 1, \sum_{i=0}^{i_m - 1} F_{m,i} > \theta_m \\ 0, diğer \end{cases} \tag{4.10}$$

$$F_{m,i} = \begin{cases} \alpha_{m,i}, eğer\ f_{m,i} > t_{m,i} \\ \beta_{m,i}, diğer \end{cases} \quad (4.11)$$

Eşitlik 4.10'da verilen $\boldsymbol{F_{m,i}}$, 2 boyutlu integrallerin ağırlıklı toplamına eşittir. $\boldsymbol{F_{m,i}}$ değeri ayrıca "i. öznitelik çıkarıcının" karar eşik değeri olarak da kullanılır. $\boldsymbol{\alpha_{m,i}}$ ve $\boldsymbol{\beta_{m,i}}$ i. öznitelik çıkarıcıyla ilişkili sabit değerleri temsil eder. $\boldsymbol{\theta_m}$ değeri de m. sınıflandırıcının karar eşik değeridir.

Vioala Jones yöntemi ilk çıkarıldığı dönemlerde çalışma hızı olarak maksimum 15 fps değerine kadar çıkabilmiştir. Daha sonra yapılan çalışmalarda CPU ve GPU işlem birimlerinin beraber kullanılmasıyla algoritmanın hızının arttırıldığı ve 30 fps videolar üzerinde yüzün tespit edilip takip edilebildiği görülmüştür [225,226,227].

Özetlenecek olursa Haar Cascade algoritması temelde herhangi bir nesnenin alt parçacıklarının farklı renk yoğunluğu, dağılımı, parlaklık vb. değerlere sahip olduğu bilgisi kullanılarak görüntü veya görüntüde tespit edilen bir nesne farklı alt bölümlere ayrılır. Bu ayrılan bölümlere ait özellikler farklı özellik kümeleri ile temsil edilerek nesnenin tamamı tespit edilir. Örneğin, kişilerin yüzünde çoğu zaman gözlerin bulunduğu bölge yanak bölgesine göre daha koyu olur. Yanak ve göz bölgesinin her birine, birbirine komşu iki dikdörtgen koyarak dikdörtgenlerin altındaki bölgelerde hesaplanan yoğunlukları farkı, yüzün tespit edilmesinde kullanılır. Yüz üzerinde buna benzer başka özelliklerin de kullanılmasıyla beraber kişinin yüzünü diğer nesnelerden ayırabilecek özellikler kümesi oluşturulur. Elde edilen özellik kümesi farklı eğitim kümeleri kullanılarak bir makine öğrenmesi yöntemi ile nihai nesne tanımlayıcıyı elde etmek üzere eğitilir. Sonuç olarak eğitilen nesne tanımlayıcı, aynı tip nesneleri örneğin insan yüzü veya insan yüzündeki özel bir alan olan dudakların tespit edilmesi için kullanılabilir [218].



Şekil 4.9. Orijinal resim karesi ve tespit edilen dudak kesiti.

Viola Jones algoritmasının 30 fps sınırının olması ve cümle veri setinin 60 fps olması bir uyuşmazlık problemini meydana getirmiştir. Bu hem çalışmanın hızına (eğitim ve test) ve hem de performansına etki etmiştir. Özellikle cümle veri setinde modelin türüne göre değişmekle birlikte %15-20 arasında düşüş göstermiştir. Çalışma hızı açısından Media Pipe'a göre %30 civarında daha yavaş çalışmaktadır.

4.2.1.2. MediaPipe

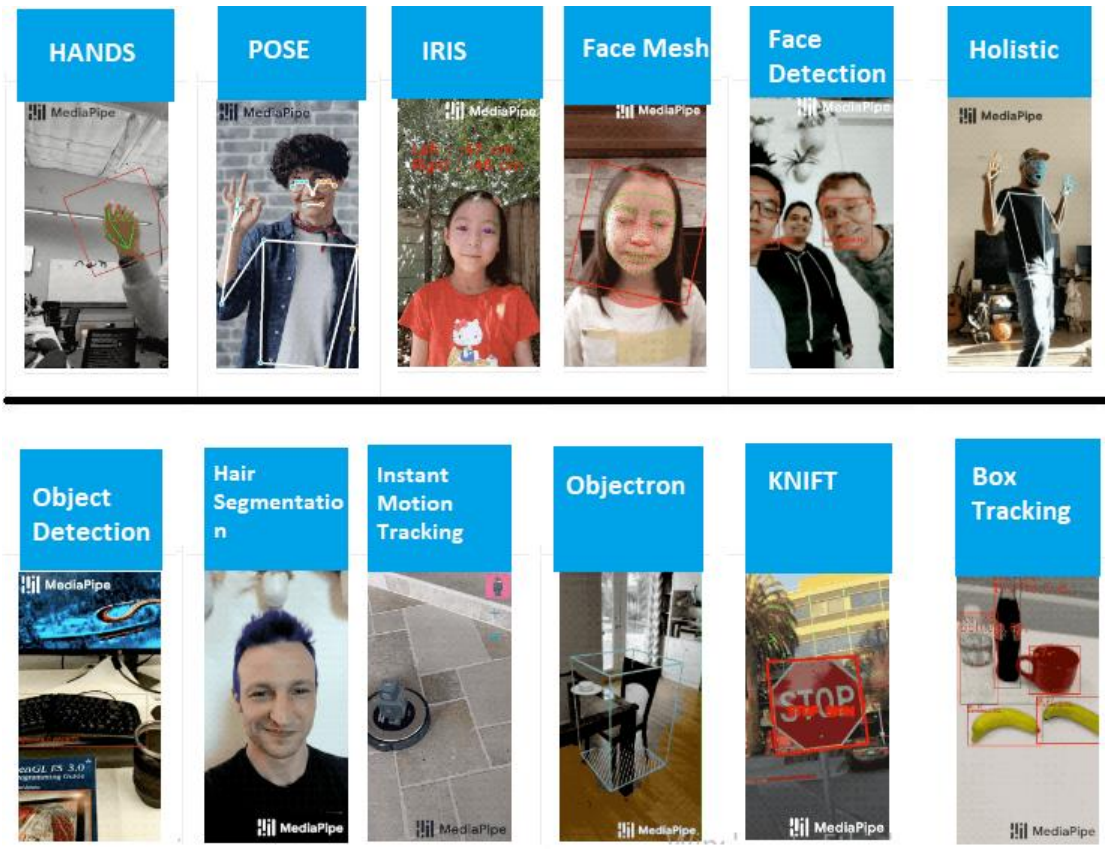
MediaPipe, Google tarafından geliştirilen makine öğrenimi çözümleri ve servisleri oluşturmak için kullanılan, kaynak kodları Github üzerinden paylaşılan bir servistir (framework). MediaPipe sahip olduğu modüler yapısıyla beraber yazılımcılara hızlı ve kolay geliştirilebilir bir mimari sunmaktadır. Sadece masaüstü uygulamalarda değil birçok platformda kullanılması büyük bir avantaj sağlamaktadır.

MediaPipe kütüphanesi aslında yazılım dünyasına çığır açacak derecede bir yenilik getirmemiştir. Sadece hali hazırda var olan fakat uygulanması ve gerçekleştirilmesi çok fazla zaman, çaba, çalışma isteyen birçok teorik araştırmayı, çalışmayı büyük bir modül haline getirip son kullanıcıya sunmaktadır. Son kullanıcıya bu kadar kolay sunulması önemlidir çünkü yazılım geliştirmeye yeni adım atmış bir kullanıcı açısından makine öğrenmesi ve görüntü işleme tabanlı projeler yapabilmesi eskisine göre çok daha kolay hale gelmiştir. MediaPipe kütüphanesinin sunduğu bazı servisler aşağıda listelenmiştir.

- Yüz Tespiti- Face Detection
- Yüz Noktalarını Oluşturma- Face Mesh

- Video Kesme- AutoFlip: Automatic video cropping pipeline
- Çoklu El Takibi- Multi-hand Tracking
- Poz Sınıflandırma- Pose
- Saç Segmentasyonu- Hair Segmentation
- Mobil Cihazlarda Takip, Yüz Tespiti vb. - Holistic
- Nesne Takibi- Box Tracking
- Nesne Tespiti - Object Detection
- 3 Boyutlu Nesne Tespiti- Objectron
- Nesne Eşleştirme- KNIFT

Mediapipe'ın sunmuş olduğu hizmetlerin birçoğuna ait görseller Şekil 4.10'da örnekleriyle verilmiştir.



Şekil 4.10. MediPipe modülleri.

MediaPipe birçok farklı dilde çalıştırılabildiği için farklı platformlara geliştirme yapılabilmektedir. Android, iOS, Java Script, Python ve C++ Mediapipe'ın çalıştırılabildiği platform ve dillerden bazılarıdır. Fakat 2022 itibariyle tüm modüller tüm platformlarda yer almamaktadır. Örneğin "Face Detection" tüm platformlarda ve dillerde kullanılabilirken "Hair Segmentation" sadece Android ve C++ için geçerlidir.

Çalışma kapsamında yüz tespiti ve dudakların konumlarını bulduktan sonra kırpılması için MediaPipe'ın Face Mesh çözümü kullanılmıştır. Face Mesh, mobil cihazlarda bile 468 tane 3 boyutlu yüz noktalarını (3d face landmark) gerçek zamanlı olarak tahmin edebilecek kapasiteye sahip bir çözümdür. Tahmin etmek ifadesi burada daha uygun olur. Bunun sebebi de direkt konum bazlı bir tespitten ziyade makine öğrenmesi tabanlı bir tahmin yapılmaktadır. Tahminin yapılabilmesi amacıyla yüzün 3 boyutlu yüzeyini çıkarmak için makine öğrenimi kullanılır. Face Mesh'in burada fark yarattığı nokta, 3 boyutlu yüzey oluşturmak için özel bir derinlik algısına sahip kameraya ihtiyaç duymadan yalnızca tek bir kameradan aldığı görüntülerle bunu yapabilmesidir. İşlem hattı boyunca GPU hızlandırmayla birlikte kullanımı çok fazla kaynak ve zaman gerektirmeyen hafif bir modelden faydalanarak gerçek zamanlı deneyimler için performanslı bir çözüm sunar [230].

Face Mesh aslında tek parça bir modül değildir. Altında birden fazla farklı teknoloji ve modül çalıştırır. Örneğin bunlardan biri de insan yüzünün üstünde tespit edilen noktalar ile gerçek zamanlı arttırılmış gerçeklik uygulamaları arasındaki boşluğu kapatan ve haberleşmeyi sağlayan Face Transform modülüdür. Face Transform, yazılımsal 3 boyutlu metrik bir uzay oluşturur ve bu uzayı kullanarak tespit ettiği yüze ait noktaları ekranda gösterebilmek amacıyla ekranda olması gereken konumlarını hesaplar. Bu şekilde "yüz dönüşüm verilerini" oluşturur. Böylece biri Face Mesh ile ekrana baktığında ilk olarak noktalar tespit edilir ve daha sonra da dönüşüm yapılarak bu noktaların ekranda ilgili konumda işaretlenerek çizilmesi sağlanır. Yüz dönüşüm verileri, yüz pozunu dönüştürme matrisi (kişi kameraya hangi açıyla bakıyorsa bunun dönüşümünü yapacak matris) ve bir tane de üçgensel yüzde ızgara çizilmesini sağlayacak 3 boyutlu temel öğelerden oluşur. Bu işlemin altında yatan yöntemde sağlam, performanslı ve taşınabilir cihazda da çalıştırılabilmesini sağlamak için şekiller kümesinin dağılımını analiz etmek amacıyla kullanılmış istatistiksel analiz

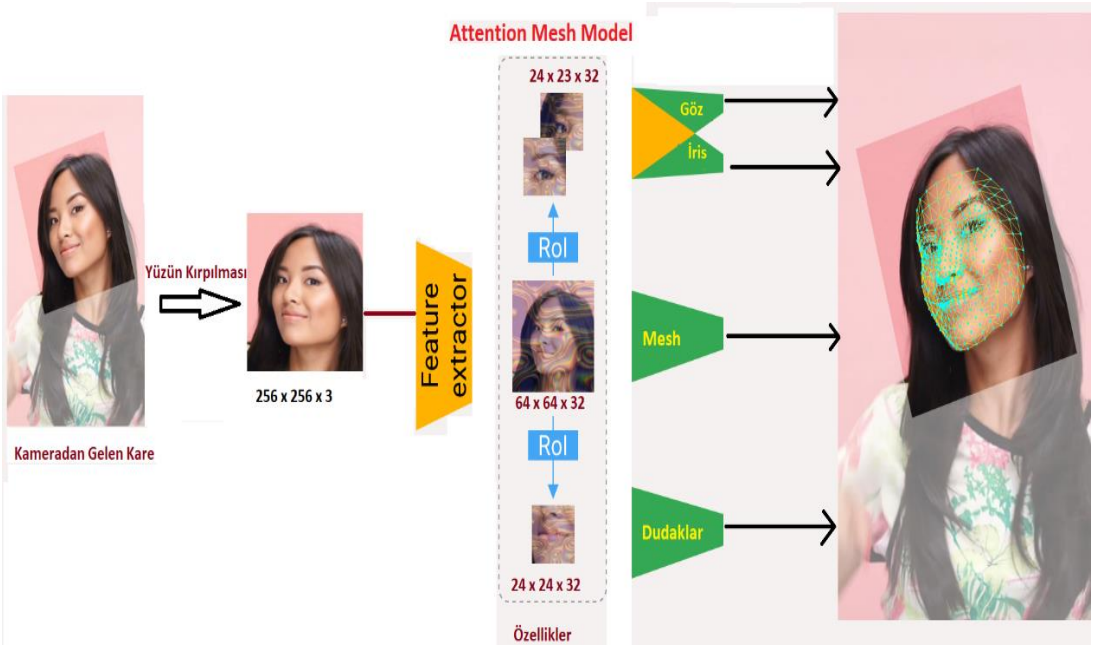
yöntemi olan “Procrustes Analizi” kullanılır. Analiz, cihazın işlemcisi üzerinde çalışır ve makine öğrenmesi model çıkarımının üzerinde minimum hız/bellek kullanımına sahiptir.

Face Mesh çözümündeki işlem hattı, birlikte çalışan 2 gerçek zamanlı derin sinir ağı modelinden oluşmaktadır. Birinci sinir ağı tüm görüntü üzerinde işlem yaparak yüzü tespit eder ve yüzün ilgili bölgelerindeki 3 boyutlu noktaları işaretler. İkinci sinir ağıysa regresyon yoluyla yaklaşık olarak 3 boyutlu yüzeyi tahmin eder. Yüzün doğru bir şekilde tespit edilip kırpılması döndürme, öteleme ve yeniden ölçeklendirme benzeri değişikliklerden oluşan “affine dönüşümü” gibi özellikle derin öğrenme modellerinde yaygın ihtiyaç duyulan veri artırma bağımlılığını büyük ölçüde azaltmaktadır. Böylece ağ kapasitesinin çoğu, koordinat tahmininin doğruluğunu arttırmak için ayrılmıştır. Ek olarak işlem hattındaki sonuçlar, videoda bir önceki karede tanımlanan yüz noktalarına dayalı olarak da oluşturulabilir ve yalnızca eğer landmark modeli ekranda herhangi bir yüz tespit edemezse veya tanımlayamazsa yüzü yeniden konumlandırmak için yüz dedektörü servisi çağrılır. Bu işlemin çok benzeri MediaPipe Hands çözümü için de geçerlidir. Çünkü bu çözümde de Face Mesh benzeri bir mantık çalışır. Hands çözümünde yüzdekine benzer olarak bu sefer avuç içi tespit edilmeye çalışılmaktadır. Face Mesh’te “face detector” yüzü bulmaya çalışırken hands çözümündeyse “palm detector” avuç içini bulmaya çalışır. Bu dedektörler bir kere çağrılırlar ve herhangi bir karede algılanamama durumu oluşmadığı sürece bir daha çağrılmazlar ve bir önceki kareden gelen değerler kullanılır. Bu da performansı ciddi oranda artırır [42,229].

Yüz Algılama, 6 nokta ve çoklu yüz desteği ile gelen (bir karede aynı anda birden fazla yüzün tespit edilip takip edilebilmesi) ultra hızlı bir yüz algılama çözümüdür. Face Mesh çözümünde yer alan yüz dedektörü (face detector), MediaPipe’ın yüz algılama çözümünde kullandığı detektörle aynıdır. İkisinde de mobil GPU üzerinden sonuç elde edilmesi için özel olarak tasarlanmış hafif ve iyi performans gösteren bir yüz dedektörü olan “BlazeFace” modeli kullanılır. Dedektörün gerçek zamanlı performansı 3 boyutlu yüz landmark tahmini (MediaPipe Face Mesh) başka işlemlerin de yapılabilmesine olanak sağlar. Örneğin yüz özellikleri, yüzdeki ifadenin sınıflandırılması ve yüz segmentasyonu gibi başka problemlerin çözülmesi amacıyla geliştirilen özel modeller

için girdi olarak doğru bir yüz ROI'si gerekir. Bu tarz modellere de uygulanabilmesini sağlar. BlazeFace, MobileNet V1 ve V2'den ilham alınarak geliştirilmiştir. Ancak onlardan temelde bazı farkları bulunur. (1) Hafif bir özellik çıkarma ağı, (2) "Single Shot Multibox Detektörünün (SSD)" değiştirilmesiyle elde edilen GPU dostu bir bağlantı şeması ve (3) görüntü üzerinde tespit edilen hedef veya nesneler için oluşturulan sınırlayıcı kutular arasında güven değeri en fazla olan kutunun ekrana çizilmesini sağlayan "Non-max Suppression" algoritmasına alternatif başka bir yaklaşıma sahip olması bu farklardan bazılarıdır [228].

3 boyutlu yüz dönüm noktaları için transfer öğrenimi kullanılmıştır ve çok çeşitli amaçlara sahip bir ağ eğitilmiştir. Eğitilen ağ, sentetik işlenmiş veriler üzerindeki 3 boyutlu dönüm noktası koordinatlarını ve ne olduğunun tamamen bilindiği gerçek dünya verileri üzerindeki 2 boyutlu anlamsal konturları (2 nesnenin ayrıldığı sınırları temsil eder) paralel olarak eş zamanlı tahmin eder. Böylece ortaya çıkan ağ, sadece sentetik veriler üzerinde değil, aynı zamanda gerçek dünya verileri üzerinde de makul sayılabilecek 3 boyutlu dönüm noktalarının tahminini gerçekleştirebilir [229].



Şekil 4.11. Face Mesh mimari yapısı.

MediaPipe içerisinde Face Landmark Modeline ek olarak tespit edilmesi gereken yüz bölgelerine odaklanan ve bu nedenle dudaklar, gözler ve iris çevresindeki odak

noktalarını daha fazla hesaplama pahasına yüksek doğrulukla tahmin eden başka bir model daha sunulmuştur. Bu tarz modellere ait detaylar tez kapsamı dışında olduğu için daha fazla detay verilmeyecektir.

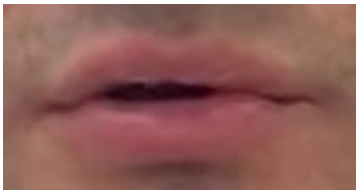
Face Mesh videoda veya bir görüntüde 468 tane yüz noktası vermektedir. Buna ait görsel Şekil 4.12'de verilmiştir.

Şekil 4.12. Face Mesh'ten gelen 468 nokta.

Elde edilen 468 nokta arasından dudakların çevresini de içine alacak biçimde Şekil 4.13'te verildiği gibi 18, 57, 164 ve 287 numaralı noktalar alınıp işaretlenmiştir. Bu noktaların seçilme sebebi sadece dudakları değil biraz daha geniş olacak şekilde dudağın etrafını alan noktalar olmasıdır.

Şekil 4.13. Dudakları içine alacak şekilde alanın çizilmesi.

Şekil 4.13'teki gibi alan belirlendikten sonra sadece ilgili bölge kalacak şekilde dudak orijinal görüntüden kırpılmıştır. Kırpılma sonucu oluşan ve CNN modeline giriş olarak gönderilecek olan görüntü Şekil 4.14'te verilmiştir.



Şekil 4.14. Kırpılmış dudak bölgesi.

Konuşmacıların dudak yapıları ve boyutları farklı olduğundan ve veri örneklerin alınması sırasında kameranın mesafesi biraz değişse bile kırpılacak dudağın boyutu değişeceğinden dolayı tüm dudak bölgeleri 60 x 35 piksel olarak yeniden boyutlandırılmıştır. Zaten görüntüler eğer hazır CNN modelleri için kullanılacaksa bu adıma da gerek yoktur. Çünkü bu modellerin sabit bir giriş boyutu bulunur. Her durumda o boyuta yeniden ölçeklemek gerekir. Elde edilen görüntüler RGB renk uzayındadır. Face Mesh Python ile kodlanmıştır.

Face Mesh çözümünün Haar Cascade ve Viola Jones ikilisinden hız açısından oldukça daha hızlı olduğu görülmüştür. Ayrıca Face Mesh çözümünün FPS ile ilgili bir sorunu bulunmamaktadır. WRR değeri olarak da %20'e kadar MediaPipe lehine fark oluşturmuştur.

4.2.2. CNN

Derin Öğrenmeden bahsedilince akla ilk gelen kavramlardan biri de CNN veya Türkçe karşılığıyla "Evrişim Sinir Ağları" olmaktadır. Derin öğrenmenin son dönemlerde oldukça popüler hale gelmesinin birçok sebebi bulunmaktadır. Fakat bunların arasından en önemlisi günümüzde deyim yerindeyse gerçekleşen veri patlamaları sonucu oluşan Büyük Veri (Big Data) kavramının, kendisiyle beraber ciddi hesaplama iş yükü ve karmaşıklığını getirmesidir. Burada tek problem verinin büyük olması ve hesaplamaların daha karmaşık hale gelmesi değil. Bunun yanında artık araştırmacılar bu kadar büyük veriler üzerinden çok daha fazla anlam çıkarmak istemektedirler ve büyük veriden eskiye göre daha fazla şey beklemektedirler. Gelinen noktada, bazı veri setlerinde milyarlar ile ifade edilecek seviyeye ulaşan parametreleri hesaplamak için veriden farklı bilgilerin elde edilmesi gerektiği anlaşılmıştır. Nesne takibi, görüntülerin sınıflandırılması, nesnelerin tespiti, stil transferi, doğal dil işleme gibi pek

çok problemlerin çözülmesi için gereken işlemler bütünü yine yapay sinir ağı merkezlidir.

Verinin aşırı büyümesi ve verilerden daha anlamlı bilgileri elde etmek, öznitelik çıkarımıyla ilgili performans optimizasyonu yapmayı zorunlu hale getirmektedir. Bunun sebebi de klasik bir yapay sinir ağında katmanlar ve nöronlar arasındaki bağlantılar ve öğrenilen parametreler çok fazla hesaplama problemlerine sebep olmaktadır. Bu problemin çözümü için Yann LeCun "Lenet" adını verdiği en temel evrişimli sinir ağlarını ortaya atmıştır [231].

Evrişimli sinir ağı (ConvNet/CNN), bir girdi görüntüsünü alıp görüntüdeki çeşitli nesnelere/yönlere önem (öğrenilebilir ağırlıklar ve bias) atayan, birini diğerinden ayırt edebilen bir tür derin öğrenme algoritmasıdır. CNN'de gereken ön işleme, diğer sınıflandırma algoritmalarına kıyasla çok daha azdır. Çünkü ilkel sayılabilecek diğer yöntemlerde filtreler manuel tasarlanırken, yeterli eğitim yapıldığında CNN bu filtreleri/özellikleri öğrenme yeteneğine sahiptir. Evrişimli sinir ağları temelde 5 katmandan oluşmaktadır.

- Evrişim Katmanı
- Doğrusal Olmayan Katman –Aktivasyon Katmanı
- Havuzlama Katmanı
- Düzleştirme Katmanı
- Tam Bağlantı Katmanı

4.2.2.1. Evrişim Katmanı

Bu amaçla yapılacak ilk adımda 2 boyutlu evrişim işlemi uygulanmaktadır. Evrişim katmanının asıl amacı giriş görüntüsünden özelliklerin elde edilmesini sağlamaktır. Evrişim temelde 2 basit fonksiyonun matematiksel işleminden ibarettir. CNN yapısında evrişim işlemi basitçe filtre adı verilen çekirdek fonksiyonunun giriş görüntüsü üzerinde kaydırılarak gezdirilmesidir. Görüntüye uygulanacak olan filtre genelde Eşitlik 4.12'de verildiği şekilde hesaplanır. Filtreler belirli bir genişlik,

yükseklik ve derinlik değerine sahip olur. Örneğin 6 x 10 x 4 boyutlarına sahip filtrede 6 genişlik, 10 yükseklik ve 4 filtrenin derinliğini temsil eder.

$$f1 = w\ x\ h\ x\ d \tag{4.12}$$

Ayrıca 2 boyutlu bilgiye veya görüntüye uygulanması gereken filtrenin x–y eksenlerine göre simetrisi alınmaktadır. Her pencere kaydırma işlemi ve a giriş resmi, h çekirdek filtresi olmak üzere evrişim işlemi Eşitlik 4.13'te belirtildiği şekilde giriş görüntüsünün ilgili konumundaki değerlerle filtredeki değerler çarpılıp toplanır.

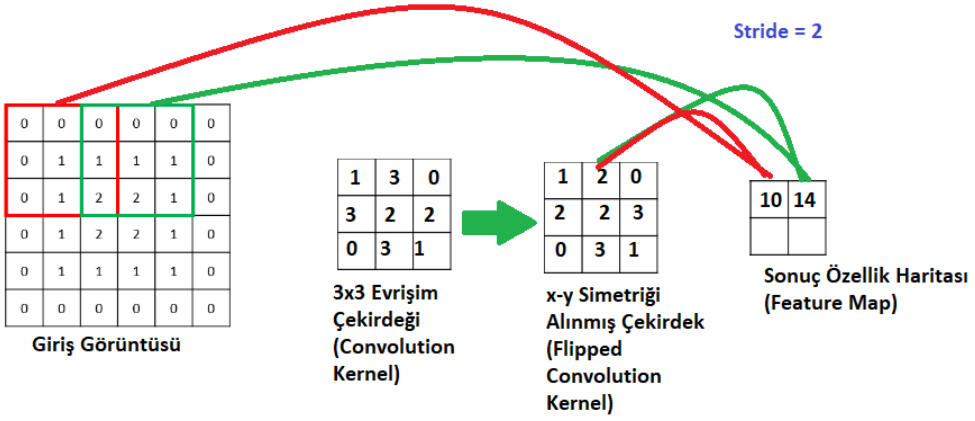
$$a\ x\ h = \sum_{k}\sum_{l} a(k,l)\ x\ h(i-k, j-l) \tag{4.13}$$

3x3'lük bir matrise 3x3'lük bir filtrenin uygulanmasına ait işlemler Eşitlik 4.14 ve 4.15'te verilmiştir.

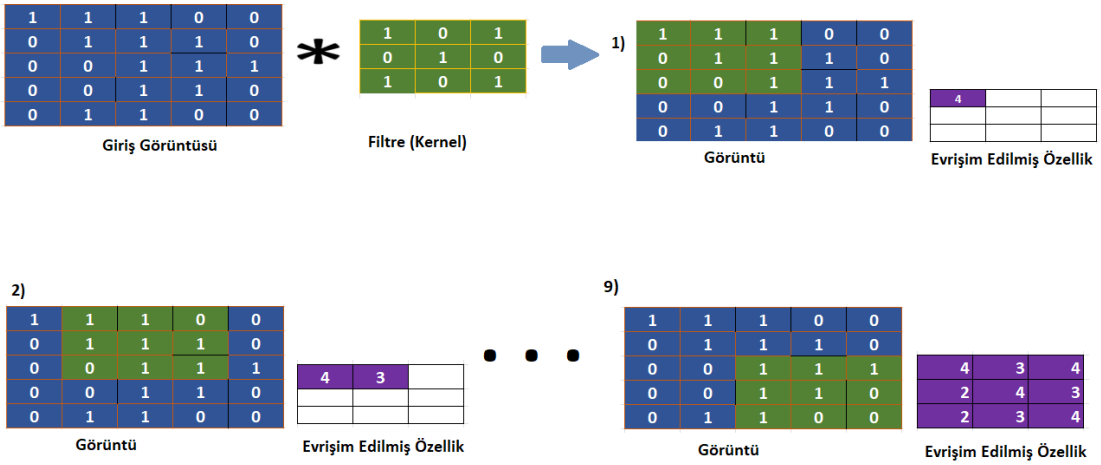
$$a\begin{bmatrix} a1 & a2 & a3 \\ a4 & a5 & a6 \\ a7 & a8 & a9 \end{bmatrix}\ h = \begin{bmatrix} h1 & h2 & h3 \\ h4 & h5 & h6 \\ h7 & h8 & h9 \end{bmatrix} \tag{4.14}$$

$$\begin{aligned} a\ x\ h = {} & a1xh9 + a2xh8 + a3xh7 + a4xh6 + a5xh5 + a6xh4 \\ & + a7xh3 + a8xh2 + a9xh1 \end{aligned} \tag{4.15}$$

Elde edilen toplam değeri de sonuç matrisinin ilgili konumuna yazılır. Tüm görüntü üzerinde pencerenin kaydırılmasıyla Şekil 4.15 ve 4.16'da gösterildiği üzere özellik haritası adı verilen evrişim işleminin çıktı değerleri elde edilir.



Şekil 4.15. Evrişim işleminin genel uygulanışı.



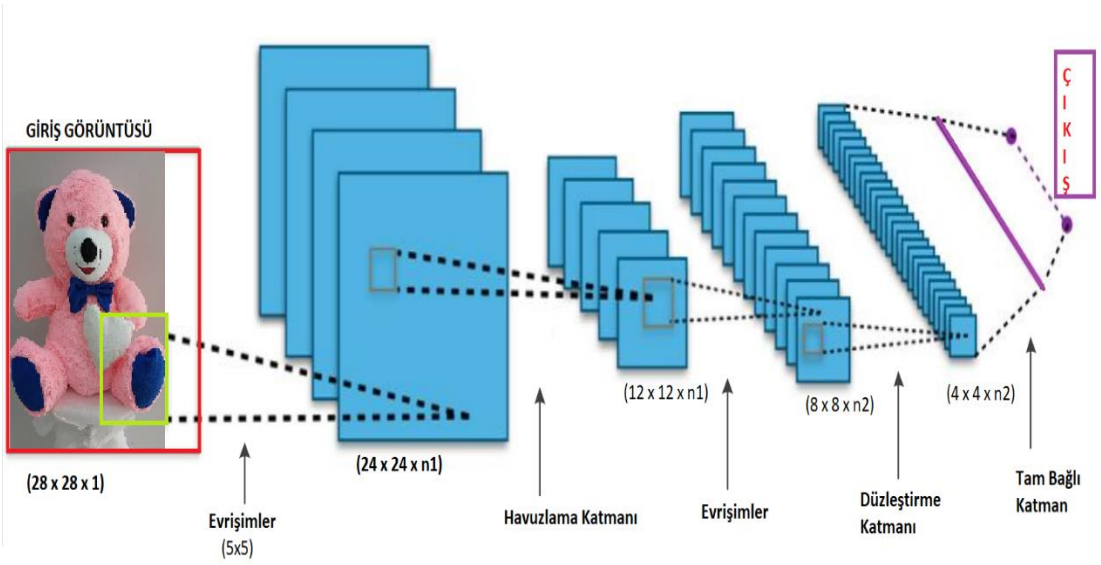
Şekil 4.16. Evrişim işlemine ait hesaplamalar.

Şimdiye kadar ifade edilenler hep tek bir filtre üzerinedir. Bir filtre uygulanmasıyla sadece 1 özellik elde edilebilir. Genellikle, birden fazla özelliğin elde edilmesi isteniyorsa yine birden fazla filtrenin kullanılması gerekir. Bu durumda da CNN yapısında birden fazla evrişim işlemi uygulanır.

Klasik yapay sinir ağında bir nöron girdiler, çıktılar ve eğim değerinden meydana gelir. Çıkış fonksiyonu sigmoid veya ReLU gibi herhangi bir aktivasyon olabilir. Eşitlik 4.16'da bir nörondaki hesaplama verilmiştir.

$$f = \sum_i w_i * x_i + b \tag{4.16}$$

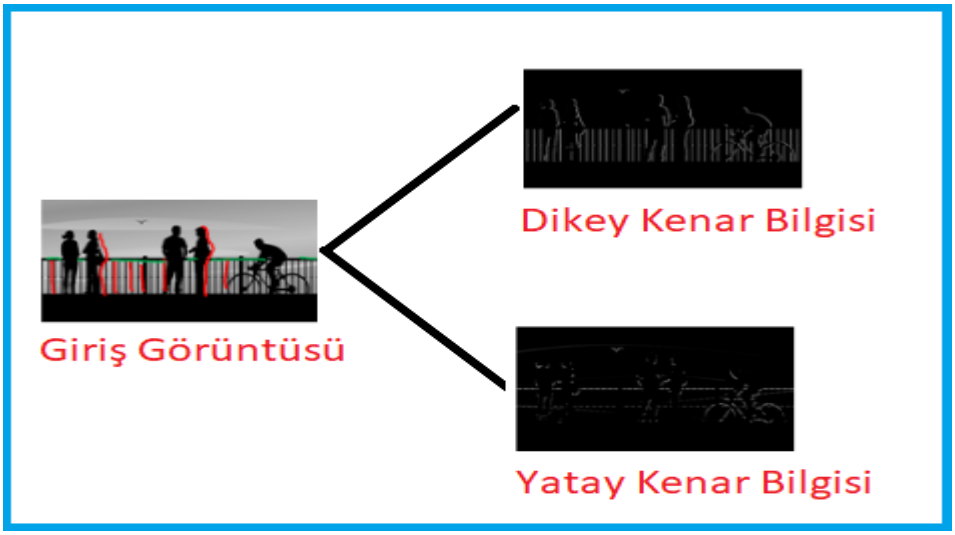
Eşitlik 4.16’da belirtilen f fonksiyonu biraz daha açılırsa CNN, temelde bir ya da daha fazla evrişim katmanı, havuzlama katmanı ve onun ardından klasik çok katmanlı bir sinir ağı gibi bir ya da daha fazla tam bağlı katmandan oluştuğu düşünülebilir. Çünkü Eşitlik 4.16’da verilen hesaplamalar klasik bir yapay sinir ağındaki katman olarak düşünülürse girdi görüntüsü ve uygulanan filtreler aslında devamlı bir şekilde geri yayılım ile güncellenen ağırlık değerleridir (matrisidir). Aktivasyon fonksiyonunun uygulandığı sonuç matrisine son adımda sabit bir bias değeri eklenmektedir.



Şekil 4.17. CNN’in en temel mimarisi.

4.2.2.1.1. Kenar Bulma

Kenar bulma, kenar bilgileri, CNN için görüntüden elde edilmesi en çok istenen özelliklerin başında gelir. Kenarlar genelde bir görüntüde yüksek frekanslı bölgeleri temsil etmektedir. Kenarlara ait özelliklerin elde edilebilmesi için yatay ve dikey olmak üzere 2 ayrı filtre uygulanır. 2 ayrı filtrenin uygulanma sebebi hem yataydaki hem de dikeydeki kenar bilgilerinin elde edilmesidir. Klasik yöntemlerde (Gabor, Sobel gibi filtreler) filtreyle görüntüye evrişim işlemi uygulanır. Böylece elde edilen görüntü kenar bilgilerini temsil eder.



Şekil 4.18. Hem yatay hem dikey kenarların filtreyle bulunması.

Şekil 4.18'de verilen kenar filtreleri oldukça basit filtrelerdir. Çeşitli kenar bulma filtrelerini kullanarak aydınlıktan karanlığa geçiş durumları, karanlıktan aydınlığa geçişler, açısal kenarlar gibi birbirinden ayrı birer özellik olarak değerlendirilip hesaplanır. Kenarlar genelde evrişimli bir ağ mimarisinde ilk katmanlarında hesaplanır. Tabi bütün bu hesaplamalar yapıldığında Eşitlik 4.17'de hesaplanan şekliyle giriş ve çıkış arasında boyut farklılıkları oluşur.

$$O = (n - f + 1)\, x\, (n - f + 1) \tag{4.17}$$

Örneğin girdi görüntüsün (n) boyutu 7x7, kenar bulma filtresinin boyutu da (f) 4x4 olduğu durumda evrişim işlemi sonucunda oluşacak çıkış görüntüsünün boyutu 4x4 olur. Eğer bu şekilde giriş ve çıkış görüntüleri arasında boyut farklılığının oluşması istenmiyorsa bunu engellemek için piksel ekleme (padding) işlemi uygulanabilir. Özellikle kaydırma işlemi varsa çıkış matrisinin boyut hesaplanması da Eşitlik 4.17'den biraz farklı olarak Eşitlik 4.18'e göre yapılmaktadır. S değeri kaydırma miktarını temsil etmektedir.

$$O = ((n - f) / S + 1)\, x\, ((n - f) / S + 1) \tag{4.18}$$

4.2.2.1.2. Doldurma İşlemi

Evrişim işleminin ardından giriş ve çıkış matrisleri arasında meydana gelen boyut farkına müdahale edilebilir. Eğer giriş ve çıkış matrislerinin eşit boyutlara sahip olması isteniyorsa giriş matrisine piksel ekleme işleminin uygulanması gerekmektedir. Doldurma veya piksel ekleme işleminde standart olarak 2 yöntem bulunmaktadır. Birinci yöntemde filtre matrisi, görüntünün üzerinde kaydırılırken görüntü matrisinin dışına taştığı kısımların tamamı Şekil 4.19'da gösterildiği gibi 0 değeriyle doldurulur.

0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	2	5	6	3	6	7	3	0	0
0	0	2	3	4	6	7	5	1	8	4	0	0
0	0	8	7	6	5	7	6	3	3	4	0	0
0	0	2	3	5	6	7	8	2	7	3	0	0
0	0	4	5	3	2	1	6	8	7	2	0	0
0	0	1	4	5	3	2	6	7	8	1	0	0
0	0	2	3	4	5	6	8	9	2	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0

Şekil 4.19. Sıfır ile doldurma işlemi.

İkinci yöntemde de görüntü matrisinin dışına taşan bölümlere Şekil 4.20'de görüldüğü gibi komşu elemanlardaki piksellerle doldurulur.

2	2	2	3	4	6	7	5	1	8	4	4	4
1	1	1	1	2	5	6	3	6	7	3	3	3
1	1	1	1	2	5	6	3	6	7	3	3	7
3	2	2	3	4	6	7	5	1	8	4	4	8
7	8	8	7	6	5	7	6	3	3	4	4	3
3	2	2	3	5	6	7	8	2	7	3	3	7
5	4	4	5	3	2	1	6	8	7	2	2	7
4	1	1	4	5	3	2	6	7	8	1	1	8
3	2	2	3	4	5	6	8	9	2	1	1	2
3	2	2	3	4	5	6	8	9	2	1	1	2
3	1	1	4	5	3	2	6	7	8	1	1	2

Şekil 4.20. Komşu elemanlarla doldurma işlemi.

Doldurma işlemi yapıldıktan sonra oluşacak görüntünün boyut hesaplaması Eşitlik 4.19 ve 4.20'ye göre yapılır.

$$p = (f-1) / 2 \tag{4.19}$$

$$O = (n + 2p - f + 1) \times (n + 2p - f + 1) \tag{4.20}$$

f, filtre boyutu olmak üzere p değeri, giriş görüntüsüne eklenmesi gereken piksel boyutunu temsil eder. Örneğin 7x7 boyutlara sahip bir görüntüde çıkış görüntüsünün oluşturulması için kullanılacak filtrenin boyutları da 5x5 olduğunda çıkış matrisinin boyutu Eşitlik 4.19'a göre 3x3 olur. Eşitlik 4.20'ye göre ise (7 + 2 x 2 – 5 + 1) x (7 + 2 x 2 – 5 + 1) = 7 x 7 olur ve böylece giriş–çıkış görüntüleri arasındaki boyut farkı ortadan kalkmıştır.

Değinildiği üzere 2 doldurma yöntemi bulunmaktadır. Bunlardan hangisinin daha iyi sonuç vereceği projenin amacına göre değişebilir. Ancak örneğin bir evrişim işlemi yapıldığı zaman piksel doldurma yöntemlerinden 0 ile doldurma işlemini uygularsak, filtre matrisinin görüntü üstünde kaydırılması sırasında sonuç görüntüsünün pikselleri hesaplandığında 0 değerlerinin eklendiği kısımlarda hesaplanan değerler düşer. Bu sebeple çıkış görüntüsünde birbirinden hayli uzak değerler meydana gelecektir. Çoğu zaman bu durum istenmemektedir.

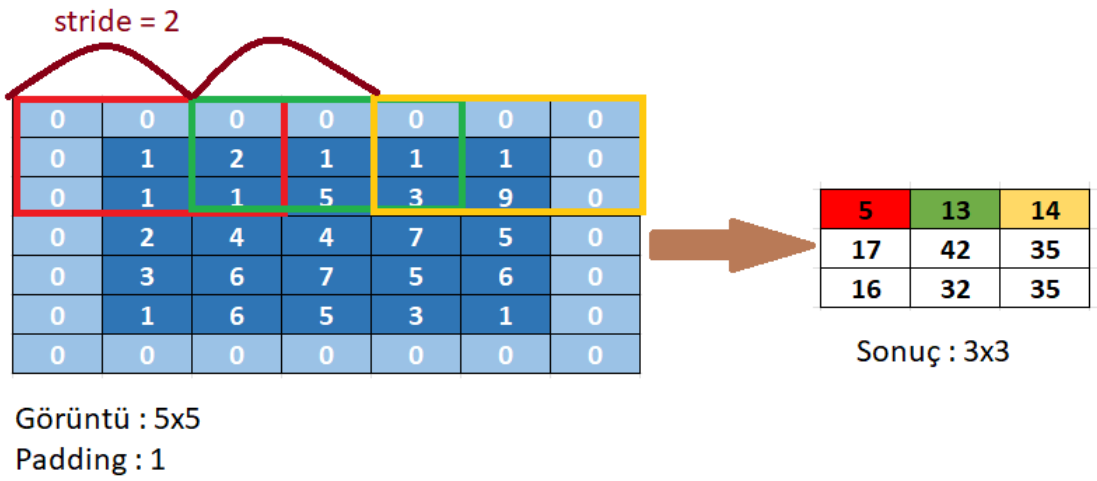
Diğer yöntemde aynalama mantığıyla görüntü matrisinin dışında kalan pikseller yine kendisine ait sınır değerleriyle doldurulduğunda eklenen bu değerler görüntüye aitmiş gibi davranacağından dolayı özellikle çıkış görüntüsü hesaplanacağı zaman piksel değerleri arasında uçurum olmaz. Böylece uygulanan evrişim katmanı çıkışında 0 ile doldurmaya kıyasla sonuç görüntüsünde orijinal görüntüye yaklaşan çok daha fazla bilgiler ortaya çıkar. Bu yöntemin de 0 ile doldurma yöntemine göre dezavantajı işlem hacminin fazla olmasıdır.

4.2.2.1.3. Kaydırma (Stride) İşlemi

Evrişim katmanında işlemlerden bir diğeri de "Kaydırma Adımı" değerine göre kaydırma işlemidir. Kaydırma değeri evrişim katmanında ağırlık matrisi görevi gören filtre matrisinin görüntü üzerinde gezdirilirken uygulanacak adım sayısını ayarlar. Filtre, birer piksellik adımlarla mı yoksa daha büyük adımlarla mı görüntü üzerinde gezdirileceğinin bilgisini vermektedir. Bu sebeple de aslında sonuç görüntüsünün boyutunu direkt olarak etkileyen bir parametre halini alır. nxn boyutlarına sahip giriş görüntüsüne fxf boyutlarındaki filtreyi p padding ve s kaydırma değerlerine sahip olacak şekilde uygulanırsa sonuç görüntüsünün boyutu Eşitlik 4.21'ye göre hesaplanır.

$$o = \left[\frac{n + 2 * p - f}{s} + 1\right] * \left[\frac{n + 2 * p - f}{s} + 1\right] \tag{4.21}$$

Uygulanışı da Şekil 4.21'de gösterilmiştir.



Şekil 4.21. Kaydırma (stride) işleminin uygulanışı.

4.2.2.1.4. Aktivasyon Fonksiyonu- Doğrusal Olmayan (Non-Linearity) Katmanı

Aktivasyon fonksiyonu barındırmayan yapay sinir ağı, standart bir lineer regresyon modelinden hiçbir farkı bulunmamaktadır. Video, ses, görüntü ve diğer karmaşık gerçek dünya verilerinin yapay sinir ağı tarafından öğrenebilmesini sağlamak için aktivasyon fonksiyonları kullanılır.

Aktivasyon fonksiyonu, sonuç görüntüsündeki doğrusal olmama durumunu arttırmak için evrişim katmanının son bileşenidir. Bir diğer deyişle aktivasyon işlemi, giriş verisi veya sinyali üzerinden yapılan doğrusal olmayan bir dönüşüm yöntemidir. Nörondan çıkan bu dönüştürülmüş sonuç verisi bir sonraki katmanda yer alan nörona giriş olarak gönderilir. Tüm evrişim katmanlarından sonra genelde doğrusal olmayan katman gelir. Bu katmanın eklenme sebebiyse görüntüdeki doğrusallık problemidir. Problem, bütün katmanlar doğrusal bir fonksiyon olabileceğinden dolayı bütün sinir ağının tek bir perceptron mantığıyla davranmasından kaynaklanır. Böylece sonuç değeri, çıktıların doğrusal kombinasyonları olarak hesaplanır.

Genel olarak aktivasyon fonksiyonu olarak ReLU, sigmoid, tanh gibi fonksiyonlar kullanılmaktadır. ReLU hız konusunda daha iyi sonuçlar verdiği için aktif olarak diğer fonksiyonlara göre daha sık kullanılmaktadır. Bu durum evrişim katmanı için geçerlidir. Fakat çıkış katmanında kullanılacak aktivasyon fonksiyonu, amaca göre farklılık göstermektedir. Sigmoid fonksiyonu için Eşitlik 4.22'de, tanh için 4.23'te, maxout için 4.24'te, ELU için 4.25'te, Leaky ReLU için 4.26'da ve son olarak ReLU için de 4.27'de verilmiştir.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{4.22}$$

$$\mathrm{o(x)} = \tanh(x) \tag{4.23}$$

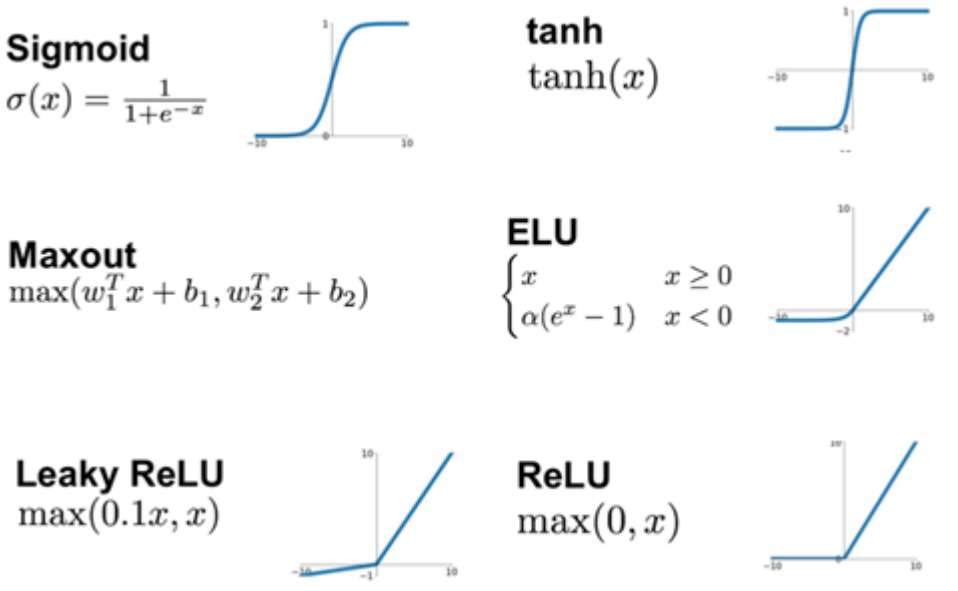
$$\mathrm{o(x)} = \max(w_1^T \mathrm{x} + b_1, w_2^T \mathrm{x} + b_2) \tag{4.24}$$

$$\mathrm{f(x)} = \begin{cases} \mathrm{x}, & \mathrm{x} \geq 0 \\ a(e^x - 1), & \mathrm{x} < 0 \end{cases} \tag{4.25}$$

$$\mathrm{o(x)} = \max(0.1\mathrm{x}, \mathrm{x}) \tag{4.26}$$

$$\mathrm{o(x)} = \max(0, x) \tag{4.27}$$

Fonksiyonlara ait grafikler Şekil 4.22'de verilmiştir.



Şekil 4.22. Aktivasyon fonksiyonlara ait grafikler.

ReLU fonksiyonu bir özellik haritasına uygulandığında Şekil 4.23'teki gibi bir sonuç elde edilir.

Şekil 4.23. Özellik haritasına ReLU uygulanması.

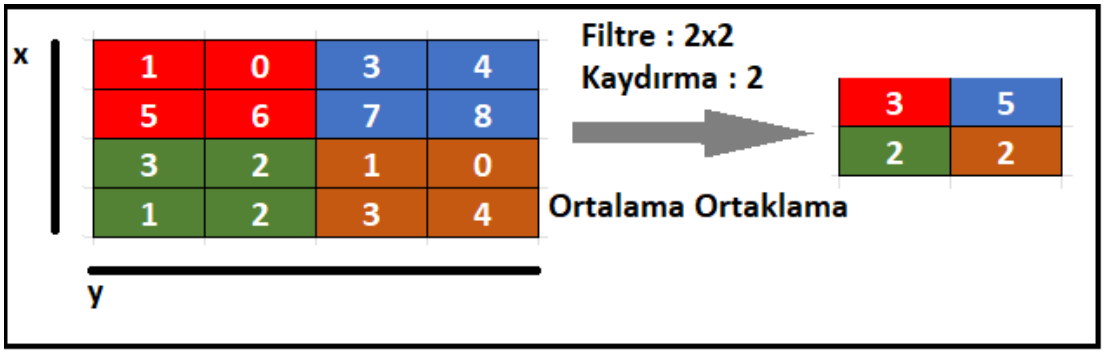
Özellik haritasında bulunan siyah bölgeler piksel değeri olarak negatif değerlere sahip olduğundan ve ReLU fonksiyonu da negatif değerleri 0'a çektiğinden dolayı fonksiyon özellik haritasına uygulandıktan sonra siyah değerler silinip yerine sıfır değeri yerleştirilir.

4.2.2.2. Ortaklama Katmanı

Ortaklama katmanı, kullanımı zorunlu olmamakla birlikte ardışık evrişim katmanları arasında sıklıkla kullanılan bir katmandır. Literatürde ortaklama veya havuzlama gibi isimlerle yer alan bu katmanda boyut indirgeme yapılır. Hedeflenen işlem, giriş görüntüsünün kanal sayısını değiştirmeden genişlik ve yüksekliklerinde indirgemeye gitmektir. Bu sayede modellerin en büyük problemlerinden biri olan hesaplama karmaşıklığı ciddi oranda azalmış olur. Fakat Hinton'ın kapsül ağ [232] teorisine göre bu boyut indirgeme sırasında ciddi oranda bilgilerin kaybolması söz konusu olduğunu söyleyerek başarı oranından fedakarlıklar yapıldığını iddia etmektedir. Hinton kapsül ağıyla beraber bu problemi çözdüğünü belirtmektedir. Fakat kapsül ağlarda da yüksek miktarda veriyle çalışmada çalışma oranı ciddi derecede uzamaktadır. Bunun dışında havuzlama katmanında öğrenmeye dair herhangi bir işlem gerçekleştirilmez ve öğrenilen herhangi bir parametre de bulunmaz.

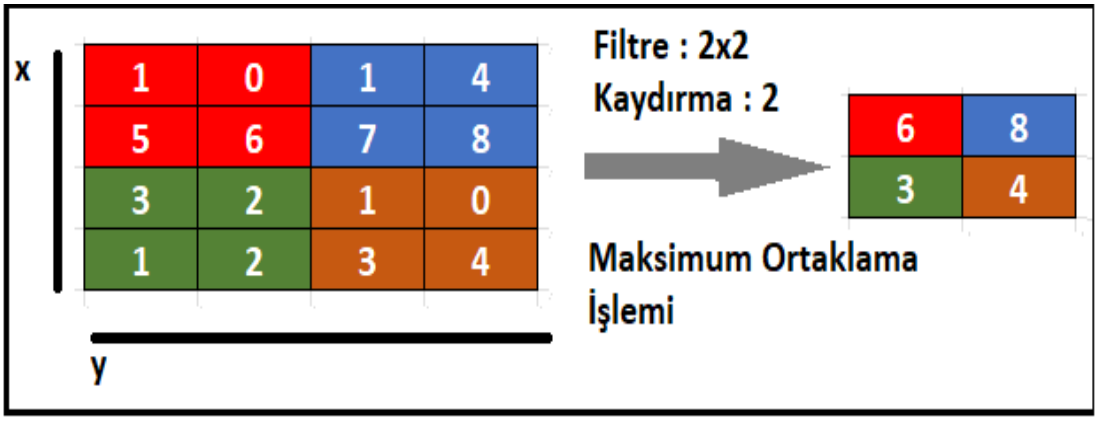
Genelde ortaklama işlemi için literatürde kullanılan 2 yöntem bulunmaktadır: (1) Ortalama Ortaklama ve (2) Maksimum Ortaklama şeklindedir.

Evrişim katmanından sonra kullanılan ortaklama katmanında ortalama ortaklama yöntemi kullanılırsa filtre belirtilen kaydırma miktarına uyarak görüntü üzerinde gezdirilir. Görüntü ve filtrenin kesiştiği alandaki piksel değerlerinin ortalaması alınır. Bu işlem tüm görüntü üzerinde gezdirilerek tekrarlanır ve boyutu orijinal görüntüye göre azalmış sonuç matrisi elde edilir. Şekil 4.24'te belirtilen 4x4 boyutlarına sahip matris için ortalama ortaklama yöntemi uygulanırsa sonuç matrisinin boyutları 2x2 olmaktadır. Kırmızı gösterilen 4 tane değerin toplamı 12 ve ortalaması alındığında 3 değeri sonuç matrisinin ilgili konumuna yazılmıştır.



Şekil 4.24. Ortalama ortaklama işlemi.

Ortaklama katmanında ortalama maksimum ortaklama yöntemi kullanılırsa filtre belirtilen kaydırma miktarına uyarak görüntü üzerinde gezdirilir. Görüntünün filtre ile kesiştiği alandaki piksel değerleri arasından en büyük olanı alınır. Bu işlem tüm görüntü üzerinde gezdirilerek tekrarlanır ve boyutu orijinal görüntüye göre azalmış sonuç matrisi elde edilir. Şekil 4.25'te belirtilen 4x4 boyutlarına sahip matris için ortalama ortaklama yöntemi uygulanırsa sonuç matrisinin boyutları 2x2 olmaktadır. Kırmızı ile gösterilen 4 tane değer arasında (1, 5, 6 ve 0) en büyük değer olan 6, sonuç matrisinin ilgili konumuna yazılmıştır.

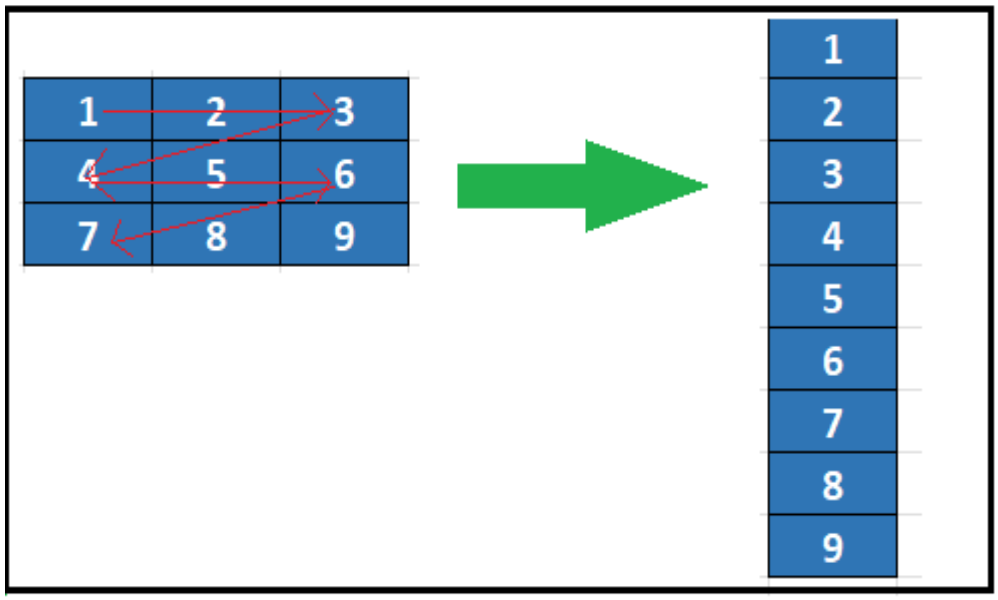


Şekil 4.25. Maksimum ortaklama işlemi.

4.2.2.3. Düzleştirme Katmanı

Düzleştirme katmanına kadar yapılan tüm işlemler matrisler üzerinden gerçekleştirilmiştir. Fakat yapılan işlemlerin bir sonraki katmana gönderilip işlenebilir hale getirilebilmesi için yapay sinir ağlarının beklediği format olan tek boyutlu bir

vektör haline getirilmesi gerekmektedir. Düzleştirme katmanında kendisine gelen matris, tek boyutlu bir vektöre dönüştürülür. Bu görevin yerine getirilme sebebiyse ağdaki en önemli ve son katman olan "Tam Bağlı Katman'ın1 girişine verilerin hazırlanmasıdır.



Şekil 4.26. Düzleştirme işlemi.

4.2.2.4. Tam Bağlantı Katmanı

Tam bağlantı katmanı ConvNet'in son ve en önemli katmanıdır. Tam bağlantı katmanında düzleştirme katmanında tek boyutlu vektör haline getirilen veriler alınıp yapay sinir ağlarına giriş olarak gönderilir. Böylece ilgili öğrenme süreci için işlem başlamış olur.

4.2.2.5. Toplu Normalleştirme Katmanı

Normalleştirme işlemi veri kümesinde bulunan değerlerin belirli aralıklara çekilerek standartlaştırmak amacıyla kullanılan bir çeşit ön işleme yöntemidir. Toplu normalleştirme işlemiyse sinir ağında yer alan katmanlar arasında gerçekleştirilen normalleştirme işlemidir. Bütün veriye olduğu gibi değil de küçük gruplar halinde normalleştirme işlemi uygulanmaktadır. Eğitimin hızlandırılmasını sağlar. Bunun

yanında daha yüksek öğrenme oranlarının kullanılmasını sağlayarak öğrenme sürecini daha kolay hale getirir. Genelde evrişim katmanıyla aktivasyon katmanı arasında kullanılır.

4.2.2.6. Seyreltme Katmanı

Basitçe ifade edilecek olursa "seyreltme" ifadesi, rastgele seçilen bazı nöron gruplarının eğitim süreci sırasında yok sayılması veya göz ardı edilmesi anlamına gelmektedir. Bir başka deyişle seyreltme hiper parametresinin görevi bir katmandaki belirli bir düğümün Eşitlik 4.28'e göre eğitilme olasılığıdır. Varsayılan olarak 1.0 değeri seyreltme olmayacağı, 0.0 ise seyreltme uygulanarak herhangi bir çıktı üretemeyeceği anlamına gelir. Gizli katmanlarda seyreltme değeri olarak genelde 0.5 ve 0.8 arasında değerler kullanılır. Giriş katmanlarındaysa genelde gizli katmanlardan daha büyük seyreltme değerleri kullanılır. Bu katmanın eklenme sebebi çoğu zaman modelin aşırı öğrenme problemiyle karşılaşmış olmasıdır. Eğer geliştirilen ağ çok büyükse, veri setindeki veri sayısı az ise veya modelin eğitim süresi çok uzunsa aşırı öğrenme problemiyle karşılaşma olasılığı yükselmektedir. Seyreltme katmanı da devreye girip nöronları yok sayarak performansta artış sağlar [233].

$$w_j' = \begin{cases} w_j, & \mathrm{P(c)} \\ 0, & \text{diğer} \end{cases} \tag{4.28}$$

P(c), c satırının ağırlık matrisinde tutulma olasılığını temsil ederken, w_j seyreltme işleminden önceki ağırlık matrisinde yer alan gerçek satır değerleridir. Çünkü seyreltme işleminde ilgili satır ağırlık vektörlerinin yer aldığı matristen tamamen silinir.

4.2.2.7. ResNet (Residual Network)

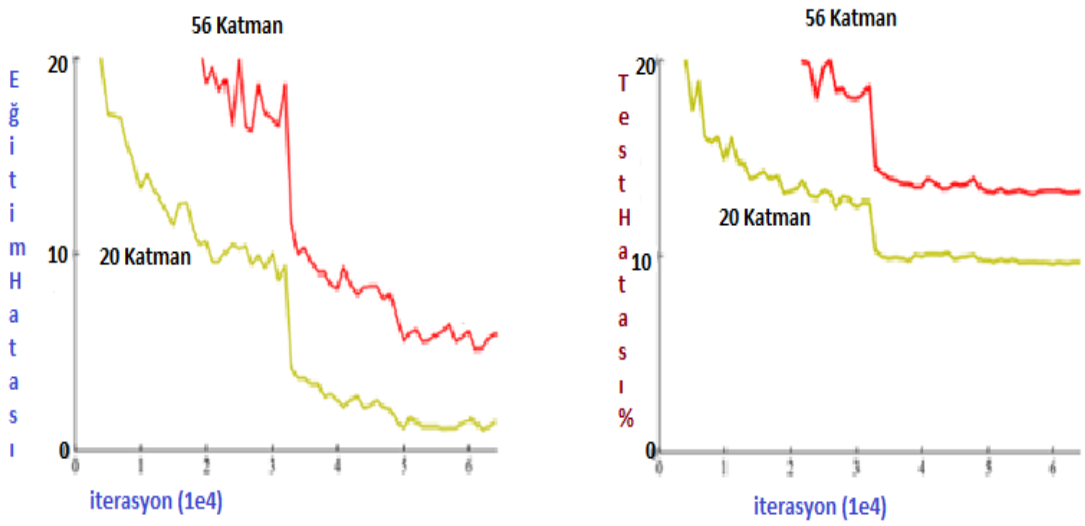
Sinir ağlarında gittikçe daha fazla gizli katman kullanılmaya başlanmıştır. Fakat modellerdeki gizli katman sayısı arttıkça yani çok fazla derinleşmeye başlandığında değerlerin hesaplanması ve öğrenilmesi süreci o ölçüde daha zor ve karmaşık hale gelir. ResNet, 2015 yılında Kaiming He ve arkadaşları tarafından o ana kadarki

ağlardan çok daha derin olan ağların eğitimini kolaylaştırmak amacıyla geliştirilen bir sinir ağı çeşididir.

ResNet modelinde VGG-19 mimarisi temelinde kurulmuştur. VGG mimarisinden daha az karmaşıklığa ve daha sade filtrelere sahip 34 katmanlı standart (düz) bir ağ mimarisi kullanılmaktadır. Bu düz ağa sonraki kısımlarda bahsedilecek olan kısayol (atlama) bağlantıları ve artık blokların eklenmesiyle mimari ResNet mimarisinin temellerine dönüştürülmüştür.

ResNet derin öğrenmeyle ilgili test veya yarışmalarda oldukça iyi performans göstermiştir. Örneğin ImageNet adı verilen çok geniş bir veri setinin kullanıldığı ILSVRC 2015 sınıflandırma testinde, VGG ağlarından daha derin ve karmaşık olmasına rağmen yine de daha düşük kompleksliğe sahip 152 katmanlı ResNet ağı %3.57 hata oranı ile 1. olmuştur. VGG-16 katmanlarının ResNet-101 ile değiştirilmesiyle Faster R-CNN'de COCO adında nesne tespiti için oluşturulan veri setinde ise önceki duruma göre %28 performans artışı görülmüştür [191].

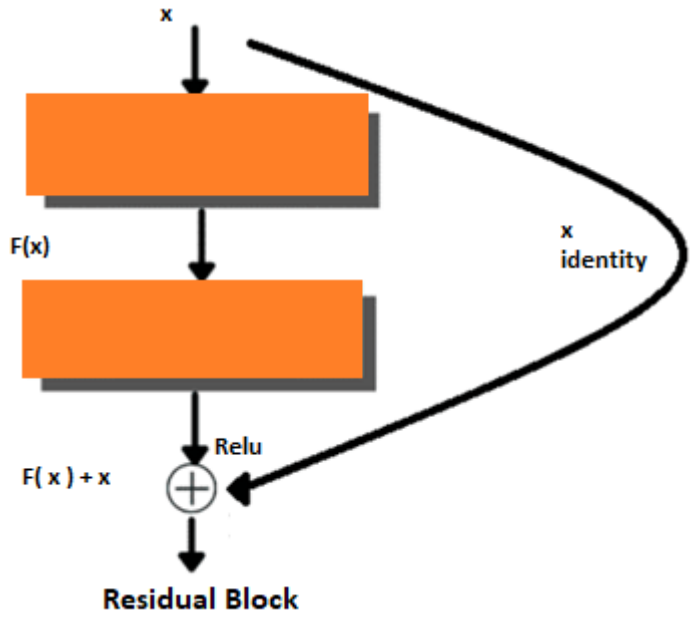
Derin CNN yapılarıyla birlikte görüntü sınıflandırması işlemi için bazı adımlar atılmıştır. Bu ağlarla birlikte görüntü (nesne) sınıflandırma ve görüntü (nesne) tanıma gibi görevlerde performans artmaktadır. Böylece yıllar geçtikçe derin öğrenme modellerinin giderek daha karmaşık problemleri çözebilmesini amaçlayarak modellere daha fazla katman eklenmiştir. Bu da modelleri git gide daha karmaşık hale getirmiştir. Bu durum aslında bir yere kadar nesne tanıma ve sınıflandırmaya dair işlemlerde performansı iyileştirmiş ve modelleri daha sağlam hale getirmiştir. Katman eklenmeye devam edilmesi durumunda farklı durumlar söz konusu olabilir. Benzer model yapılarına sahip biri 20 katmanlı diğeri 56 katmanlı 2 ağ için eğitim ve test süreçlerine dair grafik Şekil 4.27'de verilmiştir.



Şekil 4.27. Katmanlı modellerin kıyaslanması.

Şekil 4.27'de soldaki grafikte modellerin eğitim hatası ve sağdaki grafik ise test sırasında elde edilen hata oranları gösterilmektedir. Eğitim ve test aşamalarının her ikisinde de 20 katmanlı ağın 56 katmanlı ağdan daha düşük hata oranlarına sahip olduğu görülmektedir. Grafikler, modellere sürekli katman eklenmesinin çoğu zaman performansı azalttığına dair ipuçları vermektedir. Derinleşen sinir ağlarının eğitilmesi, kaybolan/patlayan gradyan problemiyle yakınsamayı engellediği için daha da zorlaşmaktadır. Teorik olarak ağı derinleştirdikçe eğitim hatasının düşmesi beklenir. Ancak pratikte doğruluk değeri bir yerde doyar ve sonrasında doğruluk değeri hızlı bir şekilde düşerek eğitim hatasını arttırır. Eğitim hatasının artması aşırı öğrenme sebebiyle değil de degredasyon/optimizasyon sorunundan kaynaklıdır. Bu sorunun oluşması, derinleşen ağların optimize edilmesinin kolay olmadığını göstermektedir. Tam da bundan hareketle ResNet modelinin oluşturulmasına karar verilmiştir.

Ağ mimarilerinin gittikçe daha derinleşmeye başladığı bir dönemde kendinden önceki mimarilerden farklı bir temele sahip ResNet; Şekil 4.28'de gösterildiği şekilde residual value (artık değer) ile residual block (artık blok) adı verilen sonraki katmanları besleyen blokların modele eklenmesiyle oluşturulmuştur. ResNet bu özelliğiyle standart mimarilerden ayrışmaktadır.



Şekil 4.28. Residual Block yapısı.

Şekil 4.28'de bulunan x katman girdisini sonraki bölümlerde yer alan toplama operatörüne dahil eden düz çizgi, kısayol bağlantısı veya artık bağlantı olarak adlandırılmaktadır. Kısayol bağlantılarıyla bir ya da daha fazla katman atlanmaktadır. Bu tekniğe "bağlantıları atlamak (skip connection)" tekniği denir. Bağlantılar atlama tekniği sayesinde modelde birkaç katman atlanır ve doğrudan çıktı katmanına ulaşılabilir. Residual bloklara gelen girdiler, katmanların arasında yer alan bağlantılarla birlikte daha hızlı yayılmaktadır. Bu bloklarda genelde aynı sayıda çıktı kanalı bulunduran 2 adet 3x3 evrişim katmanı yer alır. 2 tane olan evrişim katmanlarının her biri, toplu normalizasyon katmanı ve ReLU etkinleştirme katmanı barındırır. Sonrasında bu 2 evrişim katmanı atlanıp girdi doğrudan son katmanda yer alan ReLU aktivasyon fonksiyonuna eklenir.

Artık bağlantı kullanılmadan önce x girdisi, katmanın sahip olduğu ağırlık değerleriyle çarpılır. Çarpım değerine de bias değeri eklenir. Sonrasında "f" aktivasyon fonksiyonundan geçirilir. Bu şekilde yukarıda da bahsedildiği üzere patlayan/kaybolan gradyan probleminin önüne geçilir. Çıktı değeri de H(x) şeklinde Eşitlik 4.29'da belirtildiği şekilde elde edilir.

$$H(x) = f(wx + bias) \tag{4.29}$$

Veya

$$H(x) = f(x) \tag{4.30}$$

Eşitlik 4.30'da artık bağlantı değeri toplama işlemine dahil edildiğinde H(x) değeri Eşitlik 4.31'de belirtildiği gibi hesaplanır.

$$H(x) = f(x) + x \tag{4.31}$$

Ağın derinliklerine inildikçe, çok sayıda katman yer aldığından dolayı H(x) değerlerini hesaplamak zorlaşmaktadır. Bu sebeple zorlaşmasını önlemek için "bağlantılara atlama" yapılır. Çizelge 4.3'te ResNet için farklı katman sayılarına göre hesaplanan parametre sayıları verilmiştir.

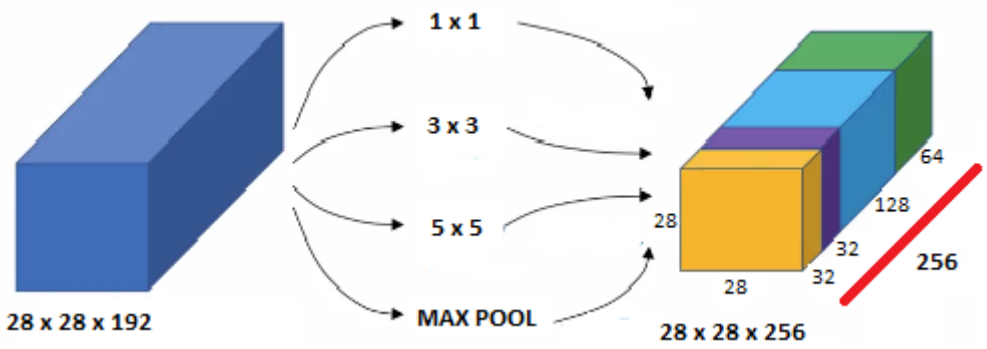
Çizelge 4.3. ResNet için katman sayılarına göre parametre sayıları.

Model	**Katman Sayısı**	**Parametre Sayısı**
ResNet	18	11.174M
ResNet	34	21.282M
ResNet	50	23.521M
ResNet	101	42.513M
ResNet	152	58.157 M

4.2.2.8. GoogleNet (Inception Ağlar)

Görüntü sınıflandırma problemlerinde sınıflandırmayı etkileyen özellikler, sınıflandırılacak görüntü özelinde büyük ölçüde değişiklik gösterebilir. Bu yüzden filtre boyutuna karar vermek zorlaşmaktadır. Görüntülerde geniş bölgelere yayılmış biraz daha genel özellikleri görüntüden elde etmek amacıyla filtre boyutu büyük olan filtreler tercih edilmektedir. Tam tersi filtre boyutu küçük olan yapılar görüntüdeki sadece alana özgü özniteliklerin algılanmasını sağlamaktadır. Sonuç olarak efektif bir şekilde özniteliklerin elde edilmesini sağlamak ve özelliklerin anlaşılması istendiğinde

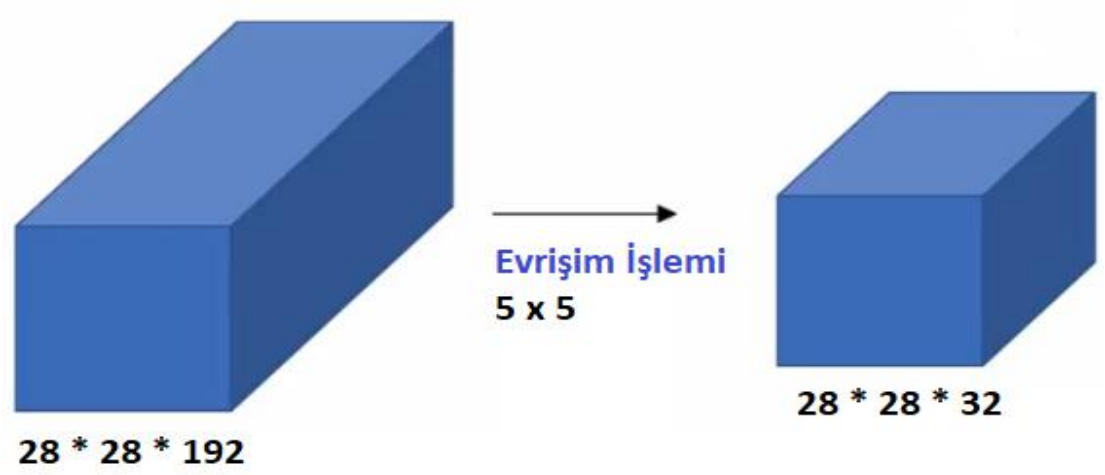
değişken boyutlarda çekirdeklere ihtiyaç duyulmaktadır. Başka bir yaklaşımda da katman sayılarının artırılmasıyla derine inmek yerine ağ daha da genişletilir. Inception mimari alt yapısı araştırmacılara bunu sunar. Şekil 4.29'da bir inception modül yapısı sunulmaktadır.



Şekil 4.29. Inception modül yapısı.

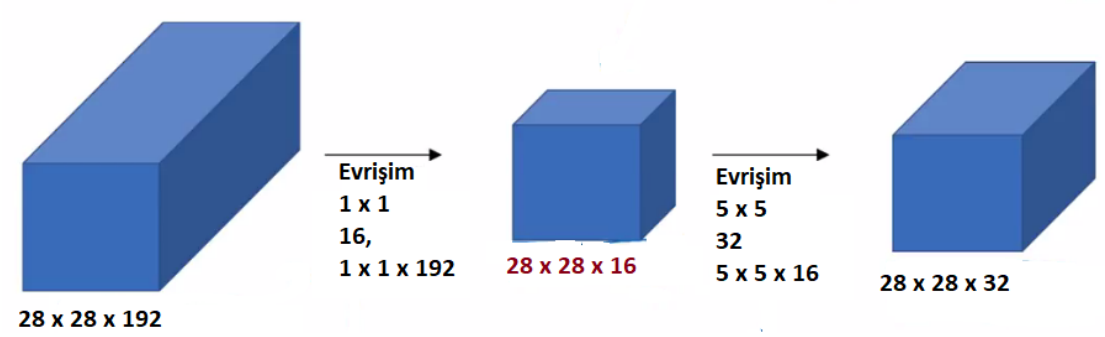
Modüllerden oluşan inception ağ modelinde yer alan her bir modül, farklı boyutlu evrişim ve maksimum havuzlama işlemlerinden oluşmaktadır. Şekil 4.29'da verilen görüntü, bir inception modülünü temsil etmektedir. Görülebildiği üzere her bir modülde sırasıyla 1x1, 3x3 ve 5x5 boyutlarında olmak üzere 3 ayrı evrişim ve bir maksimumum havuzlama katmanı sonucunda 28 x 28 x 256 boyutlarına sahip bir tensör oluşmuştur. İlk adımdaki 1x1 evrişim katmanı genel olarak birçok modelde kullanılan ve derinlik azaltmada kullanılan bir yapıdır. İlk adımda 1x1'lik bir evrişim işlemi değil de 5x5 boyutlarında yapılacak bir evrişim işlemi uygulanmış olsaydı hesaplanması daha kompleks ve zor bir adım olurdu. Tüm mimaride yer alan her bir modül farklı düzeylerde ayırt edici özellikleri ortaya çıkarabilir [188].

Şekil 4.30'da örneğin 5x5'lik evrişim sürecindeki parametre sayısı hesaplanıp olayın karmaşıklığı değerlendirilebilir.



Şekil 4.30. 5x5 Evrişim işleminin uygulanışı.

Şekil 4.30'da belirtilen evrişim işlemi için [5 x 5 x 192] x [28 x 28 x 192] sonucunda 120 milyon parametrenin hesaplanması gerekmektedir. Bu şekilde belirtilen diğer evrişim ve maksimum ortaklama katmanındaki işlemlerin yükü hesaplanabilir. Szegedy ve arkadaşlarının [184], Şekil 4.31'de gösterilen "Network in Network" [199] başlıklı çalışmaya atıfta bulunarak bütün evrişim katmanlarının öncesinde 1x1 evrişim katmanının kullanılmasıyla işlem yükü optimize etmişlerdir. Bu sayede karmaşıklaşan ağ modellerinin daha az hesap yapılarak daha hızlı bir şekilde tasarlanması mümkün hale gelmiştir.



Şekil 4.31. Network in Network.

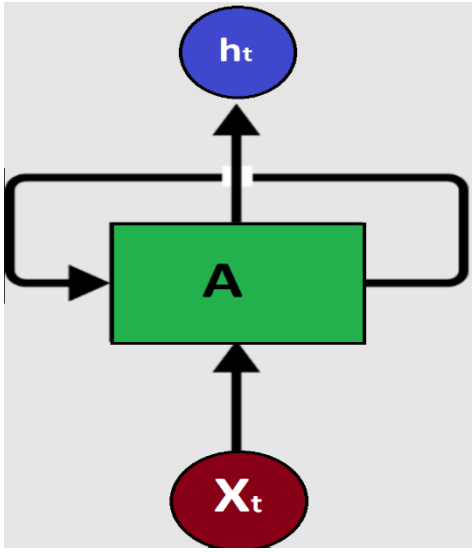
Bu durumda 1x1 evrişim işlemi sırasında (1x1x192) x (28x28x16) sonucunda 2,4 milyon adet parametre ve sonraki katmanda da 5x5 evrişim sırasında da (5x5x16) x (28x28x32) sonucunda 10 milyon adet parametre hesaplanır. Toplamdaysa 12,4 milyon adet parametrenin hesaplanması gerekir. Bu da ilk duruma göre yaklaşık 11

kat daha az parametrenin hesaplanması gerektiğini gösterir. Uygulanan 1x1'lik evrişim işlemine literatürde darboğaz tanımı yapılmaktadır.

Bu şekilde her bir modüle inception adı verilir. GoogleNet adı verilen ağ mimarisinde toplamda 9 inception bulunur. GoogleNet mimarisinde modelin kendisi de genişler. Inception ismi bir Hollywood yapımı olan 'inception' isimli filmden gelmektedir.

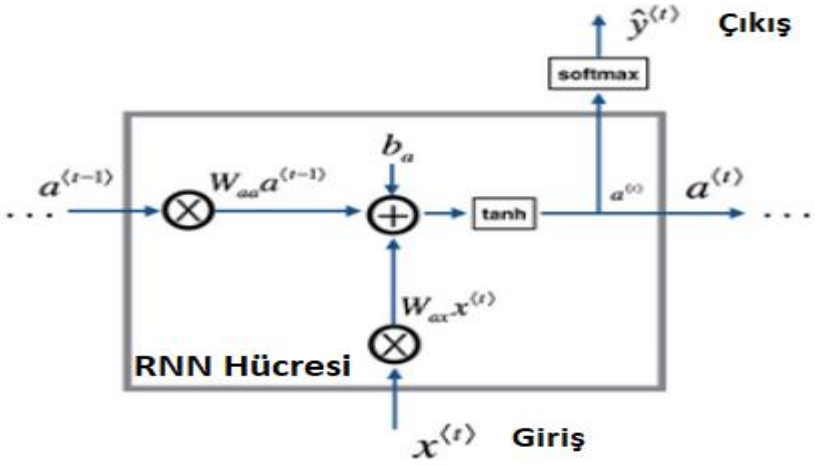
4.2.3. RNN (Tekrarlayan Sinir Ağları)

Geleneksel standart sinir ağlarında tüm girdiler ve çıktılar birbirinden bağımsızdır. Ancak bir cümlenin sonraki kelimesinin tahmin edilmesi gibi durumlarda, önceki kelimeler gereklidir ve bu nedenle sinir ağının önceki kelimeleri (sonraki kelime önceki girişe bağlı olacağından) hatırlama ihtiyacı doğmaktadır. Başka güncel bir örnek vermek gerekirse insanlar bir film izlediğinde filmin ortasında durup filmin sonunu tahmin etmek istediğinde yapacağı tahmin, o ana kadar filmin ne kadarlık bir kısmını izlediğine ve o zamana kadar beyninde oluşturulan bağlamların ortaya ne kadar çıktığına bağlı olacaktır. Benzer şekilde RNN (tekrarlayan sinir ağları) bu tarz durumlarda her şeyi hatırlaması için tasarlanmıştır. Geleneksel ağlarda böyle bir hafıza yapısı mümkün olmadığından yeni bir mimari sunularak problem çözülmeye çalışılmıştır. Çözümün en temel ayağında standart sinir ağlarına eklenen gizli bir katman bulunur. RNN, ayrıca yapısında döngüler barındırır. Bilginin aktarılması Şekil 4.32'de gösterilen bu döngüler sayesinde olur [205].



Şekil 4.32. RNN hücre yapısı.

Döngü incelendiğinde, X_t girişi ve çıkış değeri olarak da h_t değeri bulunmaktadır. Bu tarz döngüler bilginin bir adımdan diğer adıma geçmesini veya genel olarak sinir ağında bilginin aktarılmasını sağlar. Burada bir sinir ağı ünitesi olan A ünitesi, anlık t zamanında giriş olarak X_t girişini alıp h_t çıktısını oluşturmaktadır. Oluşturulan h_t çıktısı bir sonraki t+1 zamanında giriş olarak kullanılır. Bu yapı video, metin gibi zaman serisi şeklinde oluşturulan ve zaman adımları arasında ilişkilerin bulunduğu verilerin bir sinir ağı aracılığıyla işlenmesini mümkün hale getirmektedir. Şekil 4.32'deki hücre yapısı biraz daha detaylı incelendiğinde Şekil 4.33'teki gibi bir yapı görülür.



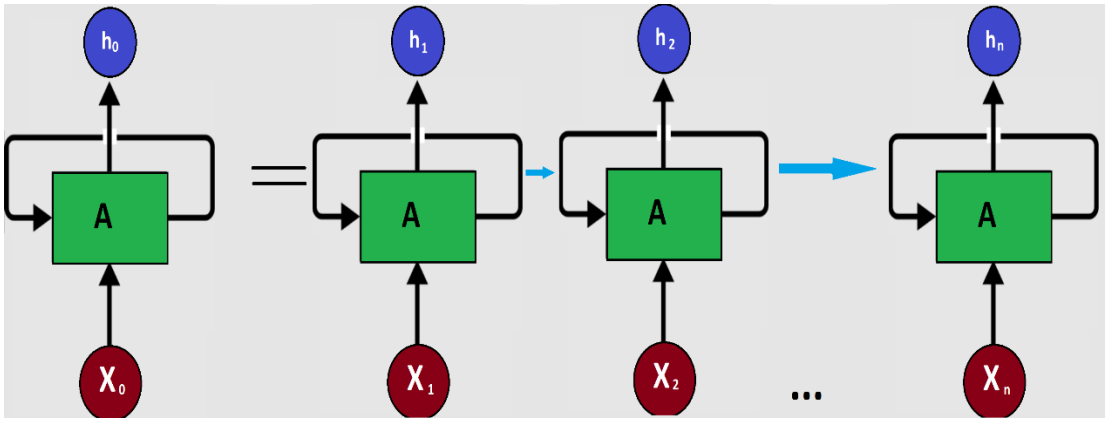
Şekil 4.33. RNN hücresinin detayları.

Bir RNN hücresinde çıkış ve sonraki hücreye gönderilecek değerler Eşitlik 4.32 ve 4.33'ee göre hesaplanır.

$$a^{(t)} = \tanh\left(W_{ax}X^{(t)} + W_{aa}a^{(t-1)} + b_a\right) \tag{4.32}$$

$$y^{(t)} = \text{softmax}\left(W_{ya}a^{(t)} + b_y\right) \tag{4.33}$$

Şekil 4.32 ve 4.33'te gösterilen RNN yapısı, standart sinir ağlarına oldukça benzemektedir. Çünkü RNN'ler birden fazla standart yapay sinir ağının zaman içinde tekrar etmesi şeklinde elde edilir ve her bir ünite (A) bir çıktı üretip sonrakine yollar. Buna ilişkin yapı Şekil 4.34'te gösterilmektedir.



Şekil 4.34. RNN döngülerinin birleştirilmesi.

Şekil 4.34 ve Eşitlik 4.32, 4.33'te belirtilen işlemler işlemlerin standart halini belirtir. Fakat döngülerin birleştirilmesi sırasında oluşan her bir çıktının adım adım nasıl oluştuğunu anlamak önemlidir. Bu işlemler temelde Eşitlik 4.34'e göre başlar.

$$h_t = f(h_{t-1} - X_t) \tag{4.34}$$

Çıkış değerinde aktivasyon fonksiyonu kullanılırsa Eşitlik 4.35 kullanılır.

$$h_t = tanh(W_{hh}h_{t-1} + W_{xh}X_t) \tag{4.35}$$

Çıktı katmanında hesaplamalar Eşitlik 4.36'da belirtildiği gibi yapılır.

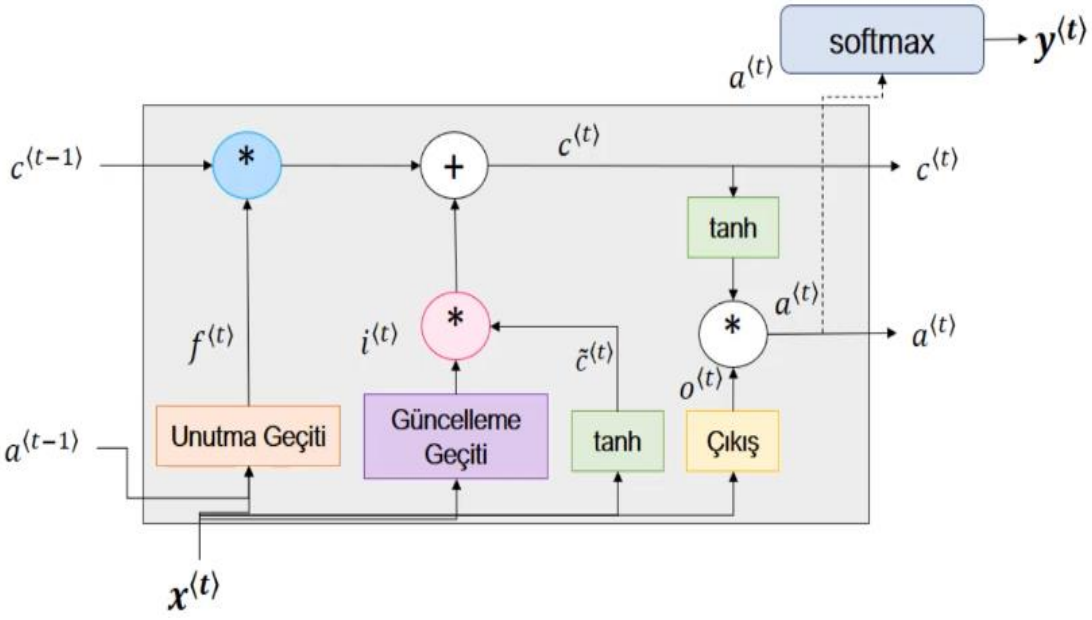
$$y_t = W_{hy} h_t \tag{4.36}$$

RNN'lerin öğrenmesi güzel bir gelişme olsa da bazı problemleri de bulunmaktadır. Örneğin kaybolan gradyan problemi bunlardan biridir. Çünkü aktivasyon fonksiyonları kullanıldığında üretilen sonuç değerleri belirli bir aralığa indirgenir. Bu aralık genelde 0 ve 1 ya da -1 ve 1 aralığında olmaktadır. Dar bir alana indirgenen bu değerler için girdide oluşan büyük değişimler aktivasyon fonksiyonu açısından o kadar büyük bir değişime sebep olmayabilir. Bu yüzden de türev değeri küçük olmaktadır. Türev değerinin çok küçük olması durumunda ilgili katman yeterli bir şekilde öğrenme işlemini gerçekleştiremez. Tekrarlayan ağlarda çok erken aşamalarda bile bu problem ortaya çıkabilir. Katmanlar yeterince öğrenemediğinden dolayı, tekrarlayan ağlar çok uzun metin veya videolarda gördüklerini unutmaya başlar. Bu yüzden de hafızası uzun değil kısa süreli olacağı için "uzun vadeli bağımlılık" problemi ortaya çıkar. Bu tarz problemler LSTM ile çözülmüştür.

4.2.4. LSTM- BiLSTM

RNN'in çözemediği uzun vadeli bağımlılık ve kaybolan gradyan problemlerini LSTM kapılar sayesinde çözer. Bu kapılar neyin hatırlanması neyin unutulması gerektiğini ayarlar. Eğer gönderilen girdinin önemsiz olduğuna karar verilirse unutulur, önemliyse bir sonraki duruma aktarılır. Bu işlemler kapı ve hücre durumları yardımıyla yapılır.

LSTM temelde, tekrarlayan ağlar içindeki normal bir adımın farklılaştırılmasıyla oluşturulur. Uygulanan farklılaşma tekniğinde ihtiyaç duyulan işlemlerin yapılıp düz bir şekilde ilerlendiği klasik sinir ağı yapısı yoktur. İlerleme sırasında ek olarak farklı matematiksel işlemlerin daha yapılıp yeni parametrelerin hesaplanması ve saklanması gibi süreçler bulunmaktadır. Yapılan ek hesaplamalar, kapılar ve hücre durumları sayesinde önceki adımlardaki girdi veya çıktı bilgilerine gerek duyulduğunda erişilebilir. Şekil 4.35'te klasik bir LSTM hücresi bulunur.



Şekil 4.35. BiLSTM hücre yapısı.

LSTM hücrelerinde genelde kapılar üzerinde hesaplama yapılır. Hesaplanan en önemli değerlerden bir tanesi ilgili LSTM hücresinin "hücre durum" değeridir. $\tilde{c}^{(t)}$: t anındaki girişten hesaplanan $c^{(t)}$ (hücre durum) için aday değeri olmak üzere Eşitlik 4.37'de verildiği şekilde hesaplanır. Bu eşitlikte t değeri 1'den T'ye kadar zaman değerini temsil etmektedir.

$$\tilde{c}^{(t)} = tanh\,(Wc\left[a^{(t-1)}, x^{t}\right] + b_c) \tag{4.37}$$

LSTM yapısındaki kapılardan Unutma Kapısı, adından da anlaşılabileceği üzere hangi bilgilerin unutulacağını hangi bilgilerin tutulması gerektiğine karar verir. Çalışma mantığı çoğu zaman oldukça basittir. Bir sayı ne kadar büyük bir değere sahip olursa olsun eğer 0 ile çarpılırsa elde edilecek değer 0 olur. Unutma kapısı da unutmak istediği değer için ilgili girdinin ağırlık değerine 0 verilir. Burada asıl karar verilmesi gereken durum ise unutulacak bilgilerin belirlenmesidir. Bunun için her hücre içinde kendisinden önceki gizli katmandan gelen verileri ve güncel verileri sigmoid fonksiyonuna yollar. Elde edilen değer 0'a ne kadar yakınsa unutulma oranı o derecede olur. 1'e yakınlık derecesi de hafızada tutulma derecesini belirler. İlgili hesaplamalar Eşitlik 4.38'de belirtildiği gibi yapılır.

$$\Gamma_f = \sigma\left(W_f\left[a^{(t-1)}, x^t\right] + b_r\right) \quad (4.38)$$

Güncelleme Kapısı, modelin unuttuğu hücre durumunu güncellemek için kullanılmaktadır. Unutma kapısında kullanılan sigmoid fonksiyonu bu kapıda da kullanılır. Klasik bir tekrarlayan ağ yapısında kapı, güncellenmesi gereken durumları Eşitlik 4.39'a göre yapmaktadır.

$$\Gamma_u = \sigma\left(W_u\left[a^{(t-1)}, x^t\right] + b_u\right) \quad (4.39)$$

Hücre durumunun (cell state) en kritik görevi bilginin taşınmasını sağlamasıdır. Taşınmasına ihtiyaç duyulan bilgileri alır. Aldığı bilgileri hücrenin sonuna aktarır. Oradan da diğer hücrelere taşınmasını sağlar. Bu şekilde ağın tamamında bilginin akışı "hücre durumu" sayesinde gerçekleştirilir. Şekil 4.35'te de belirtildiği üzere unutma geçidinden gelen değer $f^{(t)}$ ile bir önceki katmandan gelen durum verisi $c^{(t-1)}$ çarpılır. Çarpım değerine, tanh fonksiyonuna gönderilen ve güncelleme geçidinden gelen $c^{(t-1)}$ değerinin sonucu $\tilde{c}^{(t)}$ değeri eklenir.

Çıktı kapısı, bir sonraki katmana gönderilecek olan değerin belirlenmesi görevini üstlenmiştir. Bu değer önemlidir çünkü tahmin sırasında kullanılmaktadır. Şu anki girdi değeriyle önceki hücrenin durum değeri sigmoid fonksiyonuna yollanır. Hücre durumundan gelen değer tanh fonksiyonuna yollanır. Elde edilen 2 değer çarpılır. Çarpım sonucu sonraki katmana "Önceki Hücrenin Durum" değeri olarak gönderilir.

$$\Gamma_o = \sigma(W_o\left[a^{(t-1)}, x^t\right] + b_o) \quad (4.40)$$

LSTM hücresine ait nihai hücre durum değeri olan değeri Eşitlik 4.41'e göre hesaplanır.

$$c^{(t)} = \Gamma_u * \tilde{c}^{(t)} + (\Gamma_f) * c^{(t-1)} \quad (4.41)$$

$$a^{(t)} = \Gamma_o * c^{(t)} \quad (4.42)$$

Dudak okuma veya doğal dil işlemede olduğu gibi bir kelimeyi veya cümleyi anlamak için sadece önceki ifadelere değil aynı zamanda sonraki verilere de ihtiyaç duyulur. Örneğin doğal dil işlemede kullanılan bir kelime kullanıldığı yere göre anlam ve özellik kazanır. Aynı şey dudak okuma için de geçerlidir. Bir videonun anlamlandırılması için o karenin bulunduğu zaman, öncesinde ne özellikler elde edildiği, sonradan hangi karelerin geldiği gibi veriler, oluşturulacak anlamı değiştirir. Fakat klasik LSTM'de sadece ileri yönlü yayılım yapıldığı için sonraki karelerdeki durum değeri, geçmişi çok da fazla değiştirmemektedir [234].

BiLSTM yapısındaysa biri ilerdeki hücreler için ve diğeri de gerideki hücreler için olmak üzere 2 defa ileri-geri yayılım uygulanmaktadır. Bununla ilgili her iki aktivasyonda da t zamanında $\widehat{y_t}$ çıktı değeri Eşitlik 4.43'e göre hesaplanır.

$$\widehat{y_t} = g(W_y \left[\overrightarrow{a^{<t>}}, \overleftarrow{a^{<t>}} \right] + b_y \tag{4.43}$$

Böylece LSTM hücrelerinin 2 yönlü olarak çıktılarının güncellenmesi sağlanarak verilerin birbirleriyle olan ilişkisi güçlendirilmiştir.

4.2.5. Mimarinin Oluşturulması

İlk adımda, videodaki karelerin her birinden dudak bölgesi kırpılır. Daha sonra özellik vektörünün elde edilmesine yönelik işlemler başlamaktadır. Kırpılan dudak görüntülerinden özellikler, CNN tabanlı modeller aracılığıyla çıkarılmaktadır. Dudak algılama işlemiyle video karelerindeki ilgilenilen bölgeler kırpıldıktan sonra RGB renk uzayındaki görüntüler CNN modeline gönderilir. Öznitelik vektörleri, tez kapsamında birçok CNN modeli kullanılmış olmasına rağmen temelde kullanılan modellerden sırasıyla ResNet-18 modelinde "pool-5" katmanından ve GoogleNet modelinde "pool5-7x7_s1" katmanından elde edilmiştir. Görüntülerin öznitelik vektörleri elde edildikten sonra sınıflandırma aşamasında bir RNN modeli olan BiLSTM kullanılmıştır.

Videolardan alınan kareler, konuşmacıların dudak yapısındaki farklılıkları ve kameradan kaynaklanabilecek olası açısal değişiklikleri düzeltmek için saat yönünde

ve saat yönünün tersine 10 derece döndürülüp veri artırımı işlemi gerçekleştirilmiştir. Böylece veri setinin boyutu üç katına çıkmıştır. Bu, modelin eğitiminden hemen önce yapılır. Çalışmanın test bölümünde verilen eğitim video sayılarına, rotasyon işlemi ile üretilen veriler dahil değildir.

Şekil 4.36. Döndürülmüş kareler.

BiLSTM modelinde, "Gizli Durumdaki (Hidden State)" gizli birimlerin sayısı, video gibi zaman serisi verilerindeki zaman adımları arasında hatırlanan bilgi miktarına denk gelir. Gizli durum, zaman serisinin uzunluğundan bağımsız olarak önceki tüm zaman adımlarından bilgiler içerir. Fakat gizli birimlerin sayısı çok fazla ise modelin eğitim aşamasında Bölüm 4.2.3 ve 4.2.4'te belirtildiği üzere ciddi performans ve genelleştirme sorunu ile karşılaşılmıştır. Gizli birim sayısı az olduğunda eğitimin başarı oranı düşmektedir. Çalışma kapsamında sınıf sayısı ve veri miktarının çokluğu göz önüne alındığında sırasıyla 2000 2000 ve 1000 adet gizli birim kullanılmıştır.

Test çalışmalarımızda unutma kapılarının oranı yüksek ise eğitim başarı oranının oldukça düştüğü gözlemlenmiştir. Küçük değerler girilmesi durumunda model yeni veriler gördükçe test başarı oranı da düşmüştür. Bunun için bu parametrenin optimum değeri olarak 0.4 girilmiştir.

Bir videonun her karesi öznitelik vektörünün çıkarılması için aynı evrişim işlemlerinden geçtiği için vektörler eşit boyutlara sahip olmaktadır. Bu şekilde bir video için tüm karelerden ayrı ayrı elde edilen özellik vektörleri tek bir matriste birleştirilir. Birleştirme işlemi ile elde edilen vektör, tek başına o videonun öznitelik vektörünü temsil eder hale gelmektedir. Bu işlemler sonucunda bir kelimenin veya cümlenin telaffuzunun sınıflandırılması için gerekli olan öznitelik vektörü elde edilmiştir.

Karelerin CNN modeline özellik çıkarımı amacıyla gönderilmesiyle elde edilecek özellik vektörünün boyutları ve yapısı tamamen kullanılan CNN'in yapısına göre belirlenmektedir. Ancak tüm karelere aynı CNN modeli uygulandığı için tek bir kareden elde edilen özellik vektörünün boyutu sabit kalmaktadır. Farklı CNN modelleri için farklı özellik vektörleri elde edildiğinden hangisinin daha iyi sonuç verdiği test sonuçlarının değerlendirildiği Bölüm 5'te incelenmektedir.

Bölümün başında da belirtildiği üzere veri setindeki bir videonun kare sayısı ile CNN modelinin bir görüntüye ait çıktı için ürettiği vektör boyutu çarpıldığında o video için elde edilecek matrisin boyutu hesaplanmış olur. Bu işlem sayesinde ardışık karelerin öznitelik vektörü tek bir matriste birleştirilir. Oluşturulan öznitelik matrisinin değerleri sıralı olarak sınıflandırmanın yapıldığı BiLSTM modeline gönderilir.

Önerilen yaklaşımda, video karelerinden özellik çıkarmak için kullanılan CNN modeli için önceden eğitilmiş modeller tercih edilmiştir. Bu çalışmada, önceden eğitilmiş CNN modelleri olarak ResNet-18, Resnet50, Xception, ShuffleNet, Nasnetmobile, AlexNet, Vgg16, Darknet53, Darknet59 mimarileri seçilmiştir.

Çalışmada kullanılmak üzere oluşturulan veri setinde 113 cümle ve 111 kelime bulunmaktadır. Bu nedenle kelimeler için 111, cümleler için 113 ayrı sınıf tanımlanmıştır. Ayrıca her videonun uzunluğu farklı olabileceğinden kare sayısı da farklıdır.

Hem cümle hem de kelime için yapılan çalışmada eğitim aşaması için 24 konuşmacıdan rastgele 18 konuşmacı (%75) seçilir ve bunlara ait veriler modelin eğitiminde, kalan 6 konuşmacının verileri ise modelin testi için kullanılır. Cümle veri setinde, konuşmacılar 113 farklı cümlenin her birini 10 kez telaffuz etmiştir. Kelime veri setinde konuşmacılar 111 farklı kelimenin her birini 15 defa telaffuz etmiştir. Test ve eğitim için kullanılan video sayıları Çizelge 4.4'te verilmiştir.

Çizelge 4.4. Eğitim ve test aşamasında kullanılan video sayıları.

	Tip	Kişi Sayısı	Sınıf Sayısı	Telaffuz	Video Sayısı	Toplam Video Sayısı
Eğitim	Kelime	18	111	15	29970	**39960**
Test	Kelime	6	111	15	9990	
Eğitim	Cümle	18	113	10	20340	**27120**
Test	Cümle	6	113	10	6780	

Yukarıda da belirtildiği üzere bir videodan elde edilen öznitelik vektörünün boyutu kullanılan CNN modeline göre değişmektedir. Birden fazla model kullanıldığı için tüm model adına sabit bir öznitelik vektör boyutu verilememektedir. Örneğin, ResNet-18, her kare için 512 x Çerçeve Sayısı boyutunda bir vektör sağlar. 1 kelimelik videonun 20 kareden oluştuğu düşünüldüğünde, modelin her kareye uygulanması gerektiği için 512 x 20 matris oluşturulmuştur. Diğer modeller de farklı boyutlarda çıktı üretir.

Modelin başka araştırmacılar tarafından da denenebilmesi için tüm parametrelerin ayrıntılı bir şekilde paylaşılması önem arz etmektedir. Bu sebeple modele ait diğer parametreler Çizelge 4.5'te verilmiştir

Çizelge 4.5. Modele ait detaylı parametre listesi.

Toplam video sayısı	Cümle Eğitim: 20340 Cümle Test: 6780 Kelime Eğitim: 29970 Kelime Test: 9990
Sınıflar	111 Kelime ve 113 Cümle
Sınıf başına video sayısı	Kelime veri setinde her bir kelime 15 defa, cümle veri setindeyse 10 defa söylenmiştir.
Ozellik vektörlerinin boyutu	Özellik vektörünün boyutu CNN modeline göre değişir.
Eğitim veri setinin boyutu (m tane video için)	512 x kare sayısı x m
Bir videonun özellik vektörünün boyutu	Kare özellik vektör boyutu x kare sayısı ResNet-18: 512 x kare sayısı
Öğrenme oranı	1e-4
Normalizasyon Yöntemi	Min-Maks Normalizasyon
Bi-LSTM katman sayısı	2000 2000 1000

Çizelge 4.5. (devam ediyor).

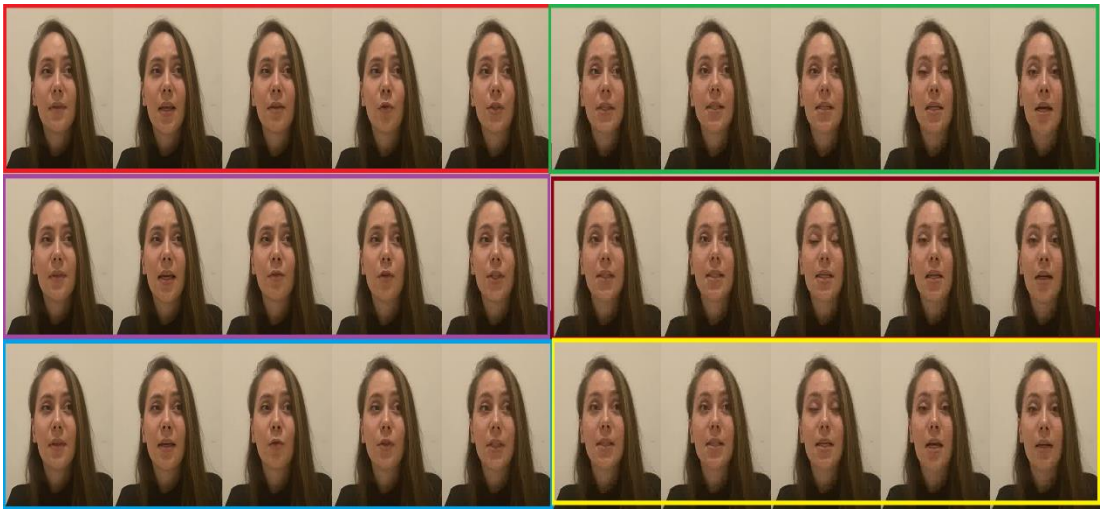
Çıkış sayısı	111-113
Yığın Boyutu	32
Karıştırma	Her bir epoch için.
Unutma oranı	0.4
Solver Algoritması	Adam
Gradyan Eşit Değeri	2
Çıkış sayısı	111-113
Yığın Boyutu	32

4.2.6. Gerçek Zamanlı Dudak Okuma Sistemi

Literatür incelendiğinde henüz dikkat çeken performansa ulaşmış gerçek zamanlı çalışan bir otomatik dudak okuma sisteminin geliştirilmediği fakat konuşma tanımada ciddi başarılar gösteren sistemlerin olduğu görülmektedir. Bunun sebebi olarak da dudaktaki görsel özelliklerin ses kadar ayırt edici özellikler sunmaması olarak düşünülmektedir.

Bu çalışma için cümle veri setinde test için kullanılan cümleler uç uca eklenerek tek parça bir video elde edilmiştir. Gerçek zamanlı test işlemi bu tek parçalı video üzerinden gerçekleştirilmiştir.

Gerçek zamanlı dudak okuma sistemi pencere tabanlı ve Bölüm 4.2.5'te belirtilen en iyi model olan ResNet-18–BiLSTM kombinasyonlu model kullanılarak geliştirilmiştir. Çeşitli pencere boyutları kullanılarak pencereler videonun üstünde Şekil 4.37'de gösterildiği şekilde gezdirilmektedir.

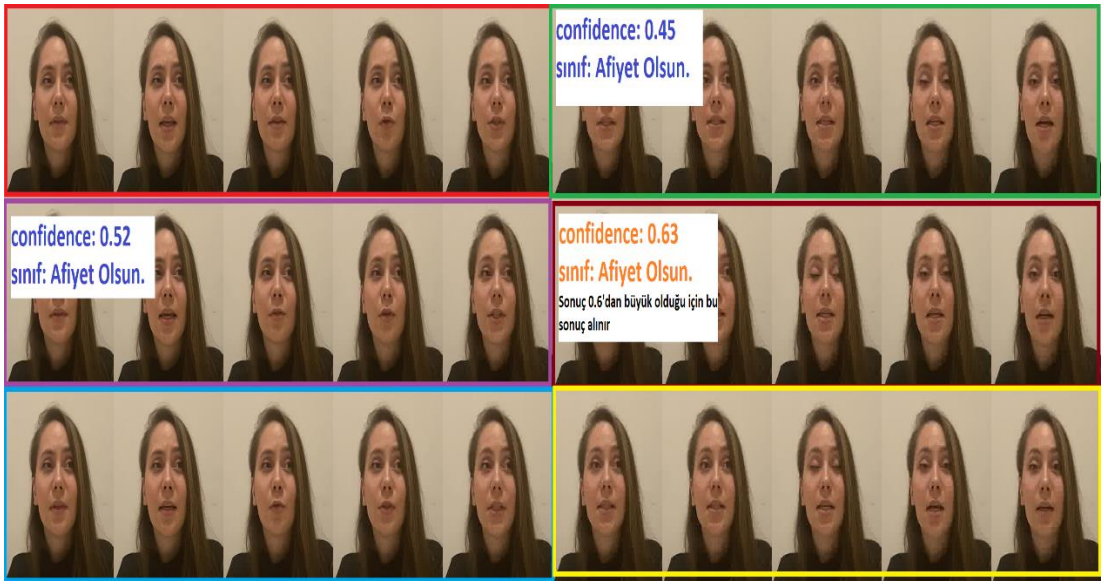


Şekil 4.37. Kare boyutu 5 olacak şekilde pencerenin yerleşimi.

Model geliştirilirken en büyük problem, bir kelime veya cümlenin nerde başlayıp nerde bittiğinin tespitidir. Ses işlerken bu durum ses sinyalinin örüntüsüne veya kelimeler/cümleler arasında oluşan çok az olsa durağanlığa vs. bakılarak yapılır. Fakat dudak okumada görseller için böyle bir durum oluşmamaktadır.

Problemin çözümü için ise 2 koşul parametresi eklenmiştir. Gelen kareler için toplu olarak özellikler elde edilip modele yollanır. Sınıflandırma için BiLSTM'e yollanır. Eğer confidence (güven) yani sınıflandırmanın güven değeri 0.6'dan yüksekse sınıflandırma sonucu kullanılır. Bu değerin minimum 0.4 olana kadar pencere sağa kaydırılır. Eğer 0.4'ten büyük ve 0.6'dan daha küçükse bu sonucun güvenilir olmadığı anlamına gelir. Bu yüzden de pencere tekrar kaydırılır ve o kareler tekrar CNN'e yollanıp özellikleri çıkarılır. Önceki özellik vektörlerinin yanına eklenir. Birleştirilmiş özellik vektörleriyle sınıflandırma işlemi tekrar yapılır. Pencerenin kaydırma işlemi, güven değeri 0.4 olarak bulunduktan sonra maksimum 3 defa yapılır. 1'er kaydırma daha yaparak toplamda 3 kaydırma yapılır. Her kaydırma işleminden sonra özellik vektörleri birleştirilir ve sınıflandırılır. Toplamda 4 farklı güven değeri arasından en yüksek değer sınıflandırma sonucu olarak işaretlenir. Sonraki sınıflandırma işlemi, pencerenin o güven sonucunun elde edildiği kareden itibaren sürüklenerek başlatılır. Örneğin pencere boyutu 5 olduğu düşünülürse 3. pencere kaydırmasında güven değeri 0.45 olarak geldiğinde 3*5=15 kareden oluşan bir özellik vektörü bulunur. Bununla birlikte artık 3 defa daha kaydırma şansı bulunur. Bir sonraki karede benzer işlemler

tekrarlanır. Bu işlem 3 kaydırma ile veya güven değerinin 0.6 üzerinde olmasıyla son bulunur.



Şekil 4.38. Pencerenin video üzerinde kaydırılması.

BÖLÜM 5

TEST AŞAMASI ve DEĞERLENDİRMELER

Tez kapsamında oluşturulan tüm modeller GPU desteği ile çalıştırılmaktadır. Tüm deneysel çalışmalar, Windows 10 işletim sistemi, Intel Xeon işlemci, çift NVIDIA rtx 3080 ti ekran kartı, 128 GB RAM bulunan bir cihaz üzerinde Matlab-2020B ortamında gerçekleştirilmiştir.

Dudak okuma ile ilgili çalışmalarda sistemin performansını belirleyen en önemli ölçüt, girdi olarak verilen kelimenin veya cümlenin doğru tanıyıp tanınmadığıdır. Dudak okuma ve derin öğrenme alanında çok önemli bir çalışmaya sahip olan Adriana Fernandez'in [40] 2017 yılında ve Hao ve arkadaşlarının [235] 2020 yılında yaptığı dudak okumada literatür taramasına yönelik çalışmalarında birçok algoritmayı karşılaştırmış ve diğer çalışmalarda olduğu gibi WRR (Kelime Tanıma Oranları) metriğini dudak okuma için performans kriteri olarak belirlemiştir. Dudak okumaya yönelik çalışmalarda genelde WRR veya doğruluk metriği şeklinde verilir [40,212,235]. Bu metriğin hem ses hem de görsel verilerin kullanıldığı ses-görsel yöntemlerinde WER (Kelime Hata Tanıma) gibi farklı bir parametresi vardır. Tez kapsamında sadece dudak okuma sistemleri incelendiğinden ötürü ses-görsel yöntemleri bu çalışmada çok fazla yer almamaktadır. WER parametresi sadece dudak okumanın yapıldığı ses işlemenin yapılmadığı çalışmalarda verilmezken ses sinyalinin de işlendiği ASR veya AV-ASR çalışmalarında verilmektedir [134,188]. Bu çalışmada model performansları WRR metriğine göre değerlendirilmiş olsa bile bazı test bölümlerinde WER değeri de verilmektedir. WER değeri Eşitlik 5.1'e göre hesaplanır.

$$WER = \frac{S + I + D}{N} \tag{5.1}$$

WRR veya recall değeri Eşitlik 5.2'ye göre hesaplanmaktadır.

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

Precision değeri Eşitlik 5.3'e göre hesaplanmaktadır.

$$Precision = \frac{TP}{TP + FP} \tag{5.3}$$

F1 skor değeri Eşitlik 5.4'e göre hesaplanmaktadır.

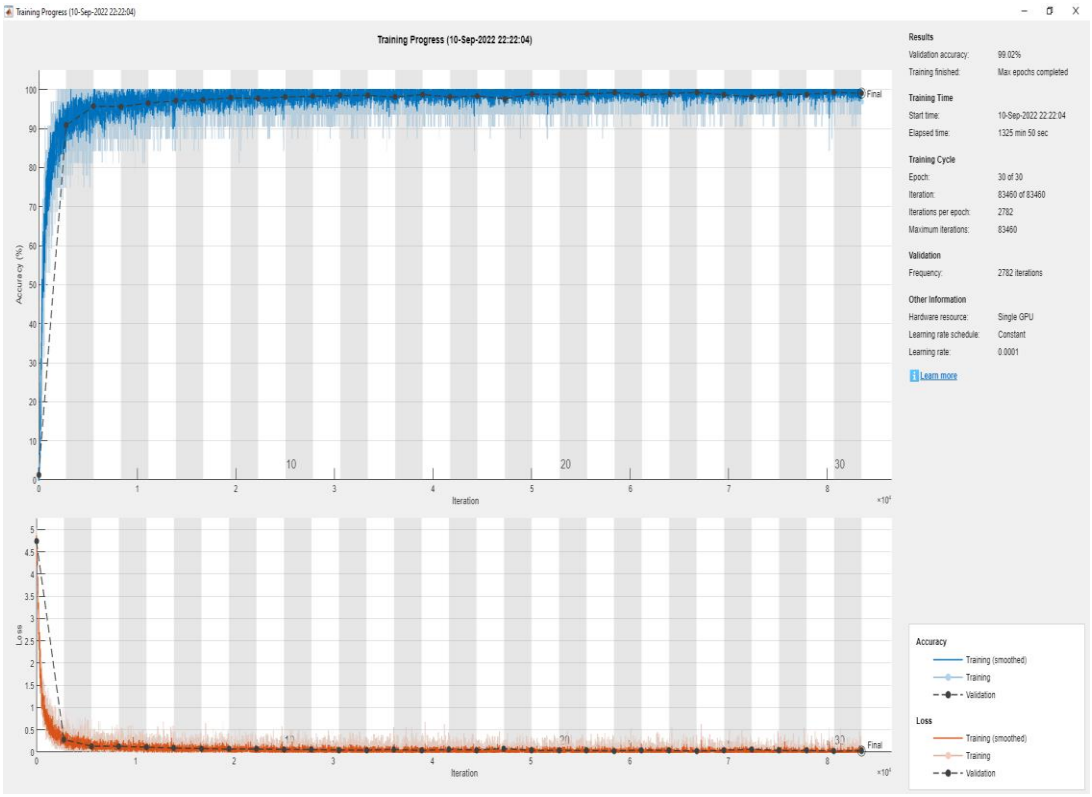
$$F1 = \frac{(2 * Recall * Precision)}{Recall + Precision} \tag{5.4}$$

Dudak okuma problemlerinde sadece "substitution" değeri hesaplanıp diğer değerler hesaplanamadığı için WER değeri aslında yanlış sınıflandırılan kelime oranına denk gelmektedir. Böylece WER değeri, 100-WRR veya 1-WRR değerine eşit olur.

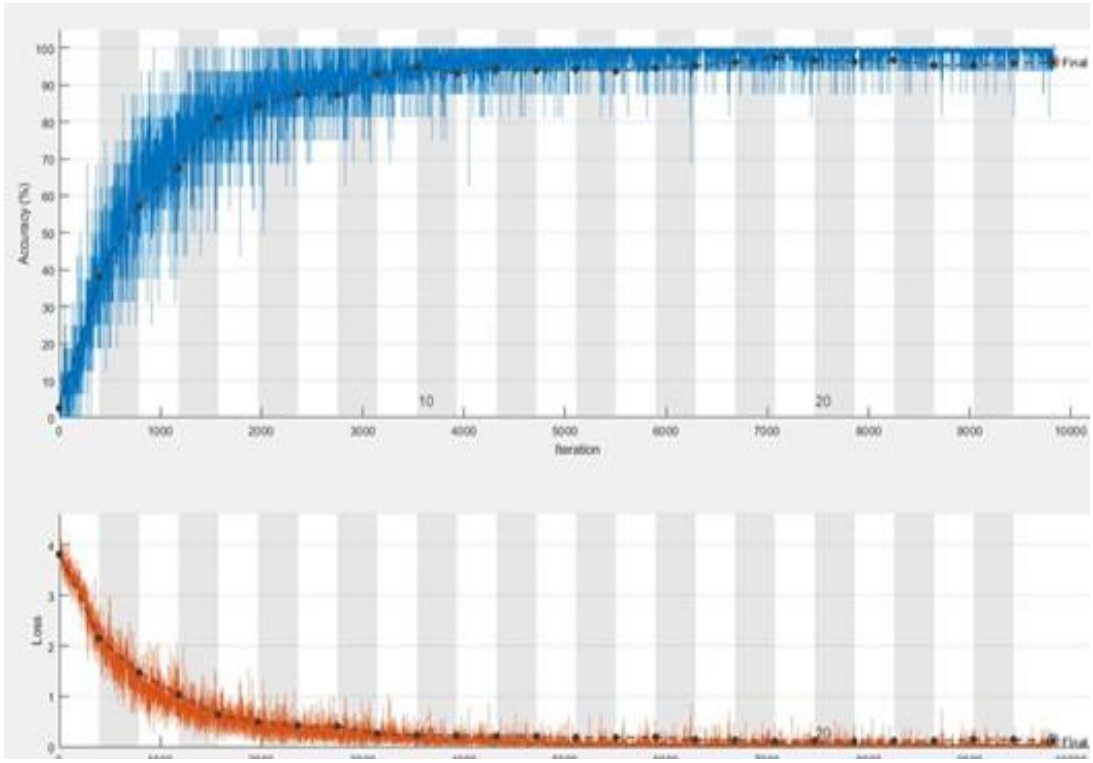
Dudak okuma alanındaki çalışmaların sonuçları diğer yöntemlerle karşılaştırıldığında, modelin hangi dil için geliştirildiği ve kullanılan veri setinin boyutu gibi birçok özellik de dikkate alınmaktadır [37,40,49]. Bu nedenle çalışmanın sonuçları diğer çalışmalarla karşılaştırılırken önemli olanın sadece doğruluk değeri olmadığı göz önünde bulundurulmalıdır.

Dudak okuma çalışmalarında dil çok önemli bir parametredir. Çünkü her dilin kendi telaffuzu ve lehçesi vardır. Çalışmalar genellikle kendi dillerinde yapılan çalışmalarla karşılaştırılmakta veya yapılan çalışmalarda dilin çok önemli bir performans ölçümü olduğu belirtilmektedir. [20,37,40,49]. Huyen, 2019 yılında yaptığı doktora tezinde

dudak okuma çalışmalarının genel odağının İngilizce dili üzerinde olduğunu ve kendisinin Almanca üzerine bir çalışma yaptığını belirtmiş. Çalışmasının sonuçlarını ve modelini, İngilizce yapılmış herhangi bir çalışma ile karşılaştırmamıştır [37]. Zhao et al. Çince bir dudak okuma sistemi geliştirdi, ancak bu sistemi diğer dillerle karşılaştırmamıştır [49]. Korece'deki başka bir çalışma da diğer dillerle karşılaştırılmamıştır [20]. Petridis ve arkadaşları çalışmalarında kullandıkları modeli İngilizce olarak 4 farklı veri seti üzerinde test etmişlerdir. Çalışma sonuçlarını karşılaştırırken, her veri seti için aynı veri setini kullanan diğer çalışma sonuçları kullanılmıştır [50]. Bazı çalışmalarda veri setinde görsel verilerle birlikte ses verileri de kullanılabilmektedir. Çalışmaların sonuçları karşılaştırılırken modelin uygulandığı dil, amaç, ölçek, karşılaştırılan dil, çalışmada kullanılan veri setinin telaffuz sayısı gibi birçok madde incelenir.



Şekil 5.1. Resnet-18 modeli için accuracy ve loss grafikleri.



Şekil 5.2. GoogleNet modeli için accuracy ve loss grafikleri.

Tez kapsamında oluşturulan veri seti ve geliştirilen model için 5 farklı test adımı ve gerçek zamanlı sistem için de 1 test adımı olmak üzere toplamda 6 farklı test süreci gerçekleştirilmiştir. Her bir test bölümü, kendi içinde detaylandırılmaktadır.

5.1. AYNI KİŞİLER (TEST-1)

Birinci bölümde tüm veri seti kullanılmamıştır. Burada amaç modelin daha ufak bir veri setindeki performansın ölçülmesidir. Bu sebeple rastgele seçilen 40 kelimeden ve 12 kişiden meydana gelen alt veri seti oluşturulmuştur. Oluşturulan bu alt veri setinden her bir konuşmacıdan elde edilen verilerden her bir sınıfın %80'i eğitim için kullanılırken geri kalan %20'si de test işlemi için kullanılmıştır. Veri setindeki her bir sınıf için kelime veri setinde 15 kelimeden 12 tanesi eğitim için 3 tanesi test için kullanılmıştır.

Bu bölümdeki amaç konuşmacıyla bağlı test işlemini gerçekleştirerek konuşmacıların farklı telaffuzlarını tespit etmek olduğundan dolayı model daha önce görmediği bir kişiyle test esnasında karşılaşmaz. Bu test ile veri setinin boyutu büyüdükçe eğitim

süresindeki değişim değerlendirilebilir. İlk bölümde en iyi test sonuçlarına sahip 2 modelin sonuçları Çizelge 5.1'te verilmiştir.

Çizelge 5.1. Test-1'e ait sonuçlar.

Model Adı	Sınıflandırıcı	Eğitimdeki Video S.	Eğitim Süresi(m)	Test Video S.	Başarılı Sınıf.	WRR
ResNet-18	**Bi-LSTM**	**5760**	**64**	**1440**	**1399**	**0.9715**
GoogleNet	Bi-LSTM	5760	51	1440	1376	0.9555

Çizelge 5.1'te verildiği üzere Resnet-18 modeli en iyi sonuçları vermektedir. Bu sonuçlar, konuşmacıların telaffuzlarıyla eğitilmiş bir modelin, aynı konuşmacı grubunun aynı kelime grupları için farklı telaffuzlarını tanımada oldukça başarılı olduğunu göstermektedir.

5.2. KELİME VERİ SETİ İÇİN FARKLI KİŞİLER (TEST-2)

İkinci test bölümünde veri setinin tamamı kullanılmamıştır. Kelime veri seti için rastgele 12 konuşmacı ve 40 farklı kelime seçilmiştir. Modeller 24 kişi arasından rastgele 12 konuşmacı seçilmiştir. 12 konuşmacı arasından da yine rastgele seçilmiş 9 konuşmacının verileri ile eğitilmiştir. Daha sonra eğitilen model, eğitim aşamasında hiç görmediği geriye kalan 3 farklı konuşmacıya ait bir veri seti ile test edilmiştir.

Bu testin amacı modelin daha önce hiç görmediği konuşmacı verileri üzerinde nasıl bir performans gösterdiğini belirlemektir. Sonuçlar Çizelge 5.2'de belirtilmiştir.

Çizelge 5.2. Test-2'ye ait model sonuçları.

Model	Sınıf	Konuşmacı	Eğitim	Test	Başarılı Sınıflandırma	WRR
GoogleNet	40	9 + 3	5400	1800	1385	0.7694
ResNet-101	40	9 + 3	5400	1800	1377	0.7650
ResNet-50	40	9 + 3	5400	1800	1215	0.6750
ResNet-18	**40**	9 + 3	**5400**	**1800**	**1537**	**0.8538**
Nasnet-Large	40	9 + 3	5400	1800	529	0.2938
Xception	40	9 + 3	5400	1800	827	0.4594
DarkNet53	40	9 + 3	5400	1800	965	0.5361
DarkNet19	40	9 + 3	5400	1800	922	0.5122
AlexNet	40	9 + 3	5400	1800	949	0.5272
Squeezenet	40	9 + 3	5400	1800	53	0.2940
DenseNet201	40	9 + 3	5400	1800	1322	0.7344

Çizelge 5.2'de görüldüğü gibi en yüksek sınıflandırma doğruluğu yine 0.8537 ile ResNet-18 modeline aittir.

5.3. CÜMLE VERİ SETİ İÇİN FARKLI KİŞİLER (TEST-3)

Üçüncü test bölümünde veri setinin tamamı kullanılmamıştır. Cümle veri seti için 24 kişi arasından rastgele 12 konuşmacı ve 40 farklı cümle seçilmiştir. 12 konuşmacı arasından yine rastgele seçilmiş 9 konuşmacının verileri ile eğitilmiştir. Daha sonra eğitilen model, eğitim aşamasında hiç görmediği geriye kalan 3 farklı konuşmacıya ait bir veri seti ile test edilmiştir.

Bu testin amacı modelin daha önce hiç görmediği konuşmacı verileri üzerinde nasıl bir performans gösterdiğini belirlemektir. Kelime ve cümlelerin ortalama kare sayısı ve videoların uzunluğu farklı olduğundan (cümleler daha uzun ve kare sayısı daha fazla) arada oluşan farklar gözlemlenmiştir. Sonuçlar Çizelge 5.3'te belirtilmiştir.

Çizelge 5.3. Test-3'e ait model sonuçları.

Model	Sınıf	Konuşmacı	Eğitim	Test	Başarılı Sınıflandırma	SRR
GoogleNet	40	9 + 3	3600	1200	929	0.7741
ResNet-101	40	9 + 3	3600	1200	889	0.7408
ResNet-50	40	9 + 3	3600	1200	835	0.6958

Çizelge 5.3. (devam ediyor).

Model	Sınıf	Konuşmacı	Eğitim	Test	Başarılı Sınıflandırma	SRR
ResNet-18	**40**	**9 + 3**	**3600**	**1200**	**1101**	**0.9175**
Nasnet-Large	40	9 + 3	3600	1200	354	0.2950
Xception	40	9 + 3	3600	1200	689	0.5741
DarkNet53	40	9 + 3	3600	1200	656	0.5466
DarkNet19	40	9 + 3	3600	1200	652	0.5433
Alexnet	40	9 + 3	3600	1200	701	0.5841
Squeezenet	40	9 + 3	3600	1200	104	0.0866
Densenet201	40	9 + 3	3600	1200	881	0.7341

Cümlelerdeki uzunluk ve dolayısıyla elde edilebilen özellik sayısının fazla olması nedeniyle neredeyse tüm modellerin kelime veri setine göre başarı oranının arttığı görülmektedir. Örneğin ResNet-18 modelindeki performans artışı %7 civarında gerçekleşmiştir.

5.4. GENEL TESTLER (TEST-4)

Test edilen cümle ve kelime veri setlerinin sonuçları karşılaştırıldığında, cümle veri seti kare sayısı açısından daha fazla veri sağladığından ResNet-18 modeli cümleyi sınıflandırmada daha iyi performans göstermiştir. Veri setinde başarı oranı olarak GoogleNet ikinci sırada yer almaktadır. Bu nedenle 111 kelimelik ve 113 cümlelik veri setinin tamamında ResNet-18 ve GoogleNet modelleri kullanılmıştır. Eğitim için 18, test için 6 konuşmacının verileri kullanılmıştır.

WRR değeriyle Recall değerinin hesaplanması aynı şekilde yapıldığı için tez kapsamında geçen WRR değeri aynı zamanda "recall" değeri olarak da düşünülebilir.

Çizelge 5.4. Genel test sonuçları.

Model	Veri Seti	Eğitim	Test	Başarılı Sınıf S.	WRR (Recall)	WER	Precision	F1 Skor
ResNet-18	**Kelime**	**29970**	**9990**	**8401**	**0.8409**	**0.1591**	**0.8674**	**0.8432**
	Cümle	**20340**	**6780**	**6004**	**0.8855**	**0.1145**	**0.9010**	**0.8931**
GoogleNet	Kelime	29970	9990	69.55	0.6961	0.3039	0.6810	0.6885
	Cümle	20340	6780	4904	0.7233	0.2767	0.7461	0.7345

Kelime veri setinde ResNet-18 modeli kullanıldığında 9990 test videosunun 8401 tanesi doğru sınıflandırılmıştır ResNet-18 modeli cümle veri setinde kullanıldığında 6780 test videosunun 6004'ünü doğru bir şekilde sınıflandırmıştır. Veri setindeki tüm verilerin kullanıldığı modelin detayları Çizelge 5.5'te verilmiştir. Çizelge 5.5'te bu çalışmada en iyi performansı gösteren ResNet-18 ve GoogleNet tabanlı modellerin eğitim süreleri, yanıt süreleri ve parametre sayıları verilmektedir.

Çizelge 5.5. Modele ait detaylı parametre listesi.

Model	**Dataset**	**Eğitim Süresi(h)**	**Cevap Süresi(s)**	**Parametre Sayısı (Milyon)**	**Giriş Boyutları**
ResNet-18	Kelime	21.55	97	11.511M	224x224x3
	Cümle	25.14	82	11.511M	224x224x3
GoogleNet	Kelime	14.38	66	6.9 M	224x224x3
	Cümle	19.52	75	6.9 M	224x224x3

Geliştirilen model test aşamasında her kelimeden 6 farklı konuşmacıdan gelen toplamda 90 tane test edilmektedir. Model bir kelimeyi %5'in üzerinde veya en az 5 defa hatalı olarak aynı sınıfa dâhil etmiş ise ilgili kelimenin model tarafından karıştırıldığı kabul edilmiş ve bu kelimenin hangi kelimeler ile karıştırıldığı bilgisi, yanlış sınıflandırma adetleri ve orijinal kelime ile tahmin edilen kelime arasındaki Levenshtein mesafesi Çizelge 5.6'da verilmiştir.

Çizelge 5.6. Test sonucunda karıştırılan kelimelerin listesi.

Orijinal	**Sınıflandırma Sonucu**	**Yanlış Sınıflandırma Adeti**	**Levenshtein Uzaklık Değeri**
Video	Veya	15	4
Jandarma	Akraba	9	6
Çanta	Anne	16	4
Gurbet	Doküman	15	7
Salıncak	Sekreter	9	7
	İnşaat	6	7
Masaüstü	Mikrofon	10	7
Bağlantı	Banka	7	5
Başlık	Pencere	7	7
	Perde	6	6
Müfettiş	Fanatik	8	6
	Beşiktaş	7	6
Türev	Holigan	7	7
	Tuzak	11	4
	Günaydın	13	7
Tüylenmek	Dolap	11	8
Şeffaf	Sekreter	11	7
Çimento	Hoparlör	7	8
Pilot	Masaüstü	7	7
Doküman	Kupa	21	6
Teknoloji	Kardeş	9	9
Çaydanlık	Sandalye	17	5
	Kardeş	9	7
Barfiks	Perde	8	6
	Pervane	7	6
Telefon	Fakat	6	7
	Anne	8	6
Sekreter	Sandalye	7	7
Holigan	Veya	6	6
Oldukça	Oyuncak	11	5

Çizelge 5.6. (devam ediyor).

Orijinal	Sınıflandırma Sonucu	Yanlış Sınıflandırma Adeti	Levenshtein Distance Değeri
Şınav	İnşaat	24	5
Meslek	Beşiktaş	10	7
	Pencere	13	5
Merdiven	Pırlanta	6	7
	Akraba	11	7
	Meyve	10	4
Perde	Merdiven	7	4
	Veya	10	4
	Banka	29	5
	Anne	6	4
Oyuncak	Oldukça	8	5
Kumanda	Banka	8	4
Televizyon	Şereflendirme	6	10
	Anne	12	9
Kaldırım	Anne	6	8
Asansör	Akraba	6	6
Öğrenci	Kardeş	20	6
Bezelye	Sandalye	9	5
Mikrofon	Merhaba	8	6
Pervane	Banka	6	6
Navigasyon	Anne	6	10
Fanatik	Sekreter	7	7
	Anne	22	6
Kuaför	Anne	6	6
Baba	Ama	7	3
Anne	Afiş	7	3
Kardeş	Anne	9	5
Meyve	Perde	35	3
Defter	Anne	33	5
İnşaat	Cinayet	11	5

Çizelge 5.6. (devam ediyor).

Orijinal	Sınıflandırma Sonucu	Yanlış Sınıflandırma Adeti	Levenshtein Distance Değeri
Bisküvi	Mikrofon	10	7
Dondurma	Toplumsal	7	6
Ameliyat	Damacana	8	7
Cinayet	Sekreter	8	7
	Anne	8	5
Futbol	Lunapark	7	7
Tümsek	Domates	8	5
Uçurtma	Dondurma	27	5
Günaydın	Üşengeçlik	14	9
	Babaanne	8	7
Stajyer	Çanta	8	6
	Sandalye	10	5
	Anne	12	6
Çekmece	Fenerbahçe	6	8
	Limonata	7	8
Kaplan	Ama	10	5
Köpek	Ama	19	5
	Hoparlör	6	7

Çizelge 5.6 üzerinden elde edilen verilere göre en çok karıştırılan 5 kelimenin listesi Çizelge 5.7'de verilmiştir. Çizelge 5.7'de sistem, ilk sütunda verilen kelimeleri ikinci sütundaki kelimeler olarak tahmin etmiştir. Tam tersi durum geçerli değildir. Örneğin model "şınav" kelimesini "inşaat" olarak tahmin etmiş olabilir fakat bu durum modelin "inşaat" kelimesini "şınav" olarak tahmin ettiği anlamına gelmez.

Çizelge 5.7. Test sonucunda en çok karıştırılan kelimelerin listesi.

Orijinal	Sınıflandırma Sonucu	Yanlış Sınıflandırma Adeti	Levenshtein Distance Değeri
Meyve	Perde	35	3
Defter	Anne	33	5
Perde	Banka	29	5
Uçurtma	Dondurma	27	5
Şınav	İnşaat	24	5

Çizelge 5.7'e göre en çok karıştırılan kelime "meyve" kelimesidir. 90 tane "meyve" kelimesi en çok "perde" kelimesiyle 35 defa karıştırılmıştır. Sınıflandırma sonuçları en iyi olan ilk 15 kelimenin listesi Çizelge 5.8'de verilmiştir.

Çizelge 5.8. Test sonucunda en iyi sınıflandırılan kelimelerin listesi.

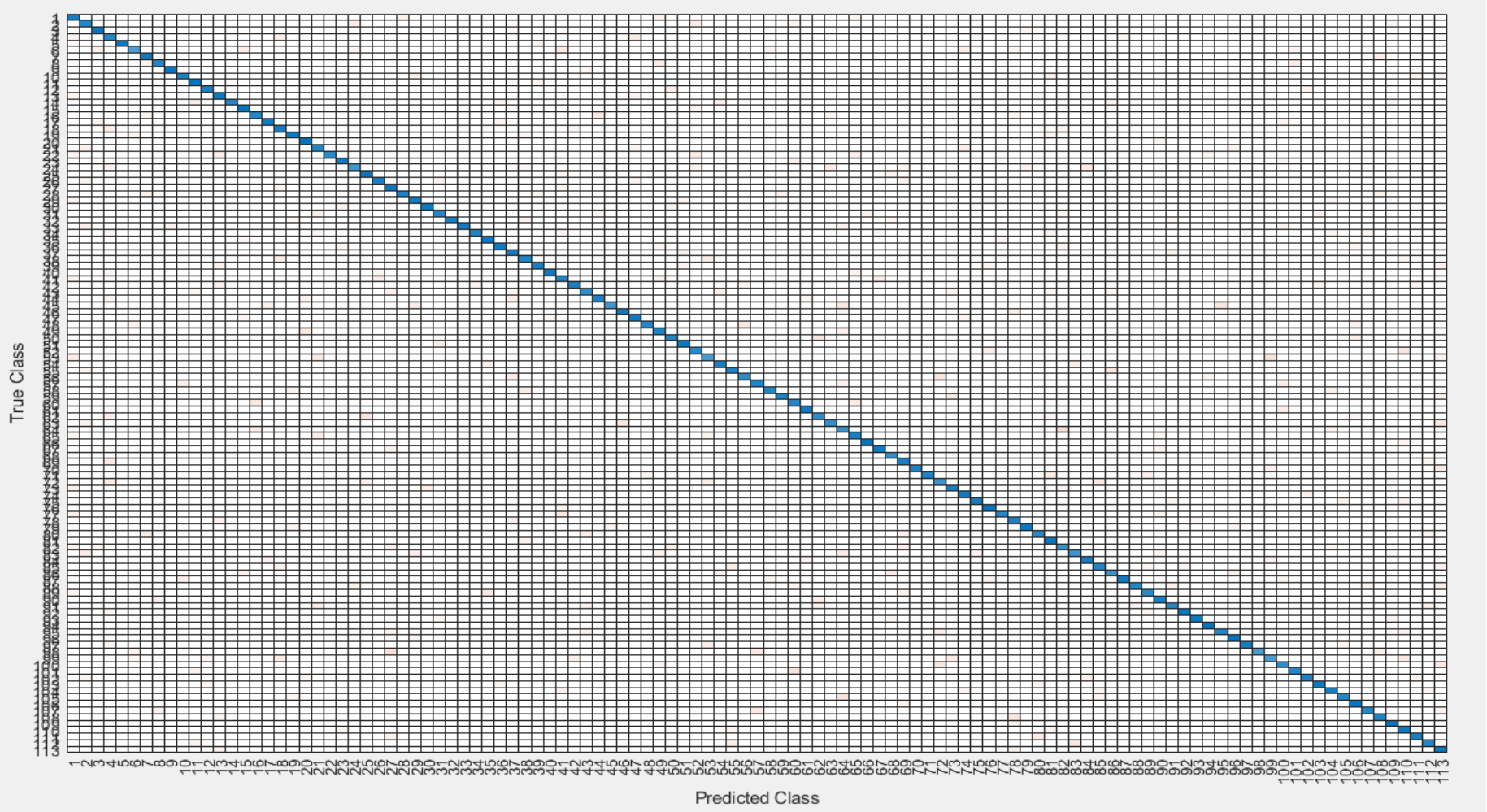
Kelime	Recall	Precision	F1 Score
Yazılımcı	%100	0.97	0.98
Merhaba	%100	0.927	0.96
Türkçeleştirmek	%100	0.96	0.87
Uzaklaşmak	%100	0.98	0.99
Ressam	%100	0.96	0.98
Münakaşa	%100	0.967	0.98
Buharlaştırmak	%100	0.98	0.99
Biçimlendirmek	%100	1	1
Fizyoterapist	%100	0.95	0.97
Ama	%100	0.70	0.82
Karşılamak	%100	0.95	0.97
Baklaçiçeği	%100	0.95	0.97
Yağmur	%100	0.95	0.97
Prefabrik	%98	1	0.99
Kaplumbağa	%98	0.98	0.98

En kötü sınıflandırma oranına sahip 15 kelimenin listesi Çizelge 5.9'da verilmiştir.

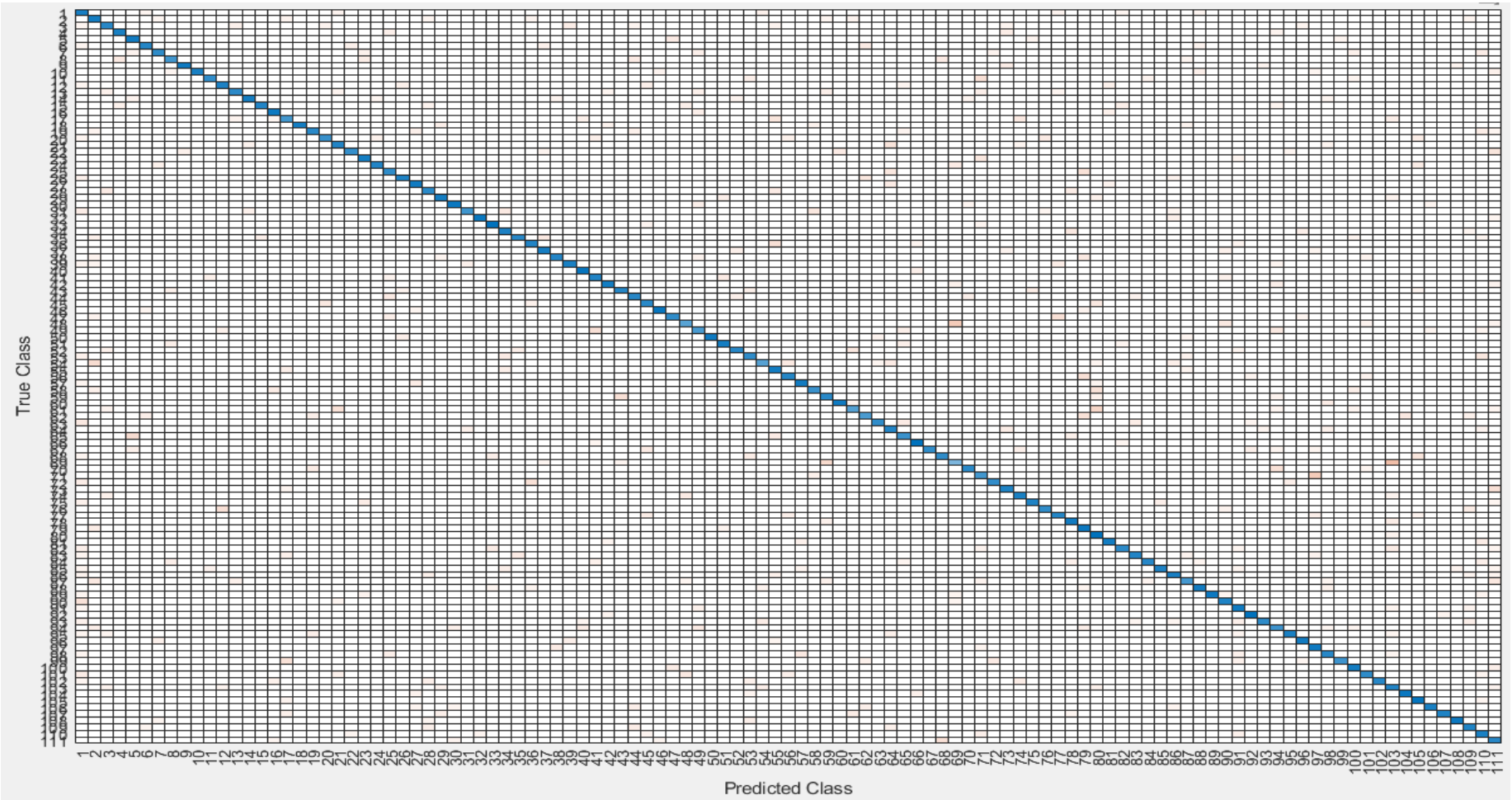
Çizelge 5.9. Test sonucunda en kötü sınıflandırılan kelimelerin listesi.

Kelime	**Recall**	**Precision**	**F1 Skor**
Perde	%35.5	0.34	0.34
Stajyer	%45.5	0.91	0.60
Fanatik	%50	0.76	0.60
Meyve	%51	0.7	0.59
Günaydın	%52.2	0.71	0.60
Fakat	%56.7	0.69	0.62
Defter	%56.7	0.81	0.66
Türev	%57.8	1	0.73
Merdiven	%60	0.81	0.68
Çaydanlık	%61.1	0.96	0.75
Telefon	%61.1	0.84	0.71
Şınav	%61.1	0.88	0.72
Döküman	%66.6	0.75	0.7
Uçurtma	%66.6	0.92	0.77
Meslek	%67.7	0.91	0.78

Tüm yapılan testlere yönelik detaylı analizler, değerlendirmeler ve yorumlar Bölüm 6'da verilmektedir.



Şekil 5.3. Cümle veri seti için ResnNet-18 confusion matrisi.



Şekil 5.4. Kelime veri seti için ResnNet-18 confusion matrisi.

5.5. GERÇEK ZAMANLI TEST(Test-5)

Gerçek zamanlı otomatik dudak okuma sistemi için cümle veri seti üzerinde bir test işlemi gerçekleştirilmiştir. Diğer test aşamalarında en iyi sonucu veren ResNet-18 BiLSTM kombinasyonuna ait model kullanılmıştır. Çeşitli pencere boyutları kullanılarak sonuçlar ayrı ayrı değerlendirilmiştir. Test 5'e ait sonuçlar Çizelge 5.10'da verilmiştir.

Çizelge 5.10. Gerçek zamanlı test sonuçları.

Model	Veri Seti	Sınıf Sayısı	Konuşmacı	Pencere Boyutu	SRR (Acc)
ResNet-18	Cümle	113	18+6	5	0.217
ResNet-18	Cümle	113	18+6	8	0.292
ResNet-18	Cümle	113	18+6	12	0.146

Gerçek zamanlı test sonuçları diğer test sonuçlarına göre oldukça düşük performans göstermiştir. Bunun sebebi de cümlelerin sınırlarının video içinde zaman adımı olarak belli olmamasıdır. Ayrıca pencere boyutları arasında da ciddi performans farklılığı oluşmuştur. Burada pencere boyutunun videonun FPS değeriyle ilişkili olması gerektiği unutulmamalıdır. Cümle veri setinin 60 FPS olduğu düşünülürse 8/60'lık oran pencere boyutu/FPS oranını olarak verebilir.

5.6. EKLERLE DEĞİŞTİRİLMİŞ KELİMELERLE CÜMLE TANIMA TESTİ (Test-6)

Bu testin 2 farklı amacı vardır. Bunlardan ilki geliştirilen modelin, test edilen kelimelerdeki değişimlere ne kadar duyarlı olduğunun ölçülmesidir. İkinci amaç ise cümleden kelime bazlı dudak okumanın gerçekleştirilmesidir. Fakat bu veri setinde videodaki kelimeleri birbirinden ayıran kırmızı bir kare eklenmiştir. Bu sayede kelimelerin başlangıç ve bitişi belli olmaktadır. Şekil 5.5'te veri setine ait bir kare verilmiştir.

Test-6 bölümünde veri setindeki kelimelerle anlamlı 20 adet birbirinden farklı cümle oluşturulmuştur. Cümlelerde kelime veri setindeki 111 kelimenin 58 tanesi (%52)

kullanılmaktadır. 20 cümlede toplamda 72 adet kelime olduğundan dolayı cümle başına 3.6 kelime düşmektedir. Test için modelin daha önce görmediği bir konuşmacıdan veri setindeki her bir cümleyi 5 defa söylemesi istenmiştir ve 360 (72x5) adet kelime oluşturulmuştur. Tekrarlarla birlikte cümleleri oluşturan 360 kelimeden 65 tanesi (12 farklı kelime) hiç ek almadan veri setindeki haliyle kullanılmakta olup geri kalan 295 tane kelime ise en az 1 harflik değişime uğramıştır. Bunlardan 45 tanesi 1 harflik ek, 200 tanesi 2 harflik ek, 45 tanesi 3 harflik ek, 5 tanesi de 4 harflik ek alarak değişime uğramıştır. Böylece veri setindeki kelimelerin %82'si orijinal halinden farklı halde cümle veri setinde kullanılmaktadır. Veri setine ait tüm bilgiler Çizelge 5.11'de verilmektedir.

Çizelge 5.11. Eklerle değiştirilmiş cümle veri seti bilgileri.

Toplam Cümle Videosu Sayısı	5x20=100
Toplam Kelime Videosu Sayısı	360
Sınıflar	58 Farklı Kelime
Cümle Başına Düşen Ortalama Kelime Sayısı	3.6
Farklı Cümle Sayısı	20
Tekrar Sayısı	5
Konuşmacı Sayısı	1
Veri Setindeki Haliyle Kullanılan Kelime Sayısı	65 Video – 12 Kelime
Veri Setindekinden Farklı Kelime Sayısı	295
1 Harf Değişimli Kelime Sayısı	45
2 Harf Değişimli Kelime Sayısı	200
3 Harf Değişimli Kelime Sayısı	45
4 Harf Değişimli Kelime Sayısı	5
Ek Almamış Kelimeler	Veya, Yazılımcı, İnşaat, Stajyer, Bozuk, Masaüstü, Barfiks, Türkiye, Münakaşa, Oldukça (2 ayrı cümlede), Şeffaf
Video Bilgileri	30 FPS, 1920x1080 Çözünürlük
Örnek Cümle 1	Türkiye Cumhuriyeti inşaat teknolojisi oldukça şeffaftır
Örnek Cümle 2	Asansörün mikrofonu bozuktur

Cümle veri setindeki en temel kısıt kelimelerin nerde başlayıp bittiğini belirlemenin çok zor olmasıdır. Bundan dolayı kelimeler arasına kırmızı kareler eklenmiştir. Bir kelimeden diğer kelime okunurken kırmızı kareye denk gelininceye kadar devam edilir.

Şekil 5.5. Cümle veri setinden örnek bir kare.

Veri setindeki kelimeler, aradaki kırmızı kareler sayesinde elde edildikten sonra "mediapipe" ile dudaklar kırpılmıştır. Kırpılan dudak bölgeleri diğer tüm test yöntemlerinde olduğu gibi önce Resnet-18 modeliyle özellik vektörleri elde edilip sonrasında BiLSTM ile sınıflandırılmıştır. Modele ait sonuçlar Çizelge 5.12'de verilmiştir.

Çizelge 5.12. Test-6'ya ait model sonuçları.

Model	**Sınıf**	**Kon.**	**Toplam Video**	**Başarılı**	**Başarısız**	**Recall**	**WER**	**Levenshtein Uzaklık**
Resnet-18	111	1	360	294	66	0.8166	0.1833	2.7694

Model bir kelimeyi en az 2 defa hatalı olarak aynı sınıfa dâhil etmiş ise ilgili kelimenin model tarafından karıştırıldığı kabul edilmiş ve bu kelimenin hangi kelimeler ile karıştırıldığı bilgisi, yanlış sınıflandırma adetleri ve orijinal kelime ile tahmin edilen kelime arasındaki Levenshtein mesafesi Çizelge 5.13'te verilmiştir.

Çizelge 5.13. Test sonucunda karıştırılan kelimelerin listesi.

Orijinal	Cümledeki Kullanımı	Sınıflandırma Sonucu	Levenshtein Uzaklık	Yanlış Sınıflandırma Adeti
Baba	Baban	Babaanne	3	4
Balkon	Balkonda	Merhaba	7	2
Cumhuriyet	Cumhuriyeti	Toplumsal	10	2
Damacana	Damacanada	Limonata	6	2
Dondurma	Dondurmalı	Toplumsal	6	4
Türkiye	Türkiye'yi	Teknoloji	7	3
Pencere	Pencerede	Beşiktaş	8	2
Doküman	Dokümanı	Toplumsal	7	4
Pencere	Penceresi	Başlık	9	2
Limonata	Limonatayı	Temizlikçi	9	3
Karşılamak	Karşılamadı	Yazılımcı	6	5
Fotoğraf	Fotoğrafı	Toplumsal	8	2

Model sonucunda elde edilen WER değeri, Test-4 bölümündeki WER değeriyle kıyaslandığında 0.0242'lik bir yükseliş göstermiştir. Performanstaki bu düşüşün en temel sebebi kelimelerin orijinal hallerinden farklı kullanılmasıyla birlikte gelen dudak hareketindeki değişimlerdir. Örneğin "baba" kelimesi normalde model tarafından "babaanne" kelimesiyle karıştırılmazken "baban" kelimesi haline geldiğinde levenshtein uzaklığı 4'ten 3'e düşerek "babaanne" kelimesine yaklaştığı için bu sınıflandırma hatası meydana gelmiştir. Aynı şekilde Türkiye kelimesi de normalde "teknoloji" kelimesiyle karıştırılmazken "türkiyeyi" haline geldiğinde levenshtein uzaklığı 8'den 7'e düşerek "teknoloji" kelimesiyle karıştırılmıştır. Bunun yanında levenshtein uzaklığı daha az olan kelimelerde Çizelge 5.13'te belirtildiği üzere daha fazla hata yapılmıştır. Diğer kelimelerin neden karıştırıldığına dair detaylı analizler Bölüm 6'da yapılmıştır. Cümle veri setindeki kelimelerin %82'sinin orijinalinden farklı olduğu düşünüldüğünde modelin başarısındaki bu değişim normal seviyede olduğu değerlendirilebilir.

Levenshtein uzaklık değeri de model sonucunda 2.76 değeri olarak elde edilmiştir. Bu değerin bu şekilde gelme sebeplerinin başında kelimelerin doğru tahmin edilse bile eğer kelime veri setindeki orijinal halinden farklı kullanıldıysa arada bir uzaklık değeri oluşmasıdır. Örneğin "karşılamadı" kelimesi doğru tahmin edilse bile orijinal

kelimenin veri setindeki hali “karşılamak” olduğu için arada 2 birimlik bir uzaklık meydana gelmektedir. Kelimelerin %82’sinin eklerle değiştirilmiş ve eklerle değiştirilen kelimelerin de %84’ü en az 2 harfli ekle değiştirilmiş olduğu göz önüne alındığında levenshtein uzaklık değerinin yükselmesinde bu durumun önemli bir etkiye sahip olduğu değerlendirilebilir. Etki eden bir diğer sebep ise kelime veri setindeki kelimelerin birbirine uzaklık metriği açısından yakın olmamasıdır. Örneğin “türkiyeyi” kelimesi “teknoloji” kelimesi olarak sınıflandırıldığında 7 birimlik bir uzaklık metriği oluşur. Halbuki veri setinde “türkiyeyi” gibi ekstra bir veri olsaydı veya buna yakın bir kelime olsaydı modelin o şekilde sınıflandırması durumunda böyle bir uzaklık oluşmayacaktı.

BÖLÜM 6

ANALİZ ve DEĞERLENDİRMELER

Tez kapsamında yapılan tüm testler bu bölümde değerlendirilerek yorumlanmaktadır. Değerlendirmeler aşağıda belirtilen maddeler üzerinden yapılmaktadır.

- Performans metrikleri
- Kelimedeki harfler
- Kelimelerin benzerliği
- Kelimelerin uzunluğu
- Sakal, bıyık gibi fiziksel etkenler
- Karıştırılan kelimeler
- CNN modellerinin performansı

6.1. PERFORMANS METRİKLERİ

Bu bölümde sadece metrikler değerlendirilmekte olup değerlendirmelere etki eden sebepler sonraki bölümlerde tartışılmaktadır. Literatürdeki dudak okuma sistemlerinin değerlendirilmesi için kullanılan tek parametre recall (WRR) parametresi olsa bile bu metrik tek başına bir sistemin değerlendirilmesi için yetersizdir. Çünkü recall değeri sadece bir kelimenin veya tüm sistem için verildiyse o sistemin doğru sınıflandırma başarısını vermektedir. Halbuki sınıflandırma yapan sistemlerde doğru sınıflandırma maliyeti kadar yanlış sınıflandırma maliyeti de bulunur ve hesaplanmalıdır. Bu yüzden de recall değeri literatürde yapıldığı gibi tek başına bir dudak okuma sisteminin performansını ölçmek için yeterli ölçüde bilgi vermemektedir. Tez kapsamında yapılan testlerde recall, WER, precision, F1 skoru metrikleri de hesaplanmıştır. Ayrıca Test-6 bölümünde cümle veri seti için yapılan test işleminde bunlara ek olarak kelimeler arasındaki benzerliği veren “levenshtein

uzaklık” değeri hesaplanmıştır. Bu değer diğer testlerde de kullanılmakta olup sistemin değerlendirilmesi sırasında da kullanılmıştır.

Recall değeri, dudak okuma sistemlerinde belirtilen bir kelimenin doğru tahmin edilme oranıdır. Örneğin veri setindeki test için her kelimeden 90 adet örnek kullanılmaktadır. Recall, her bir kelime için 90 tanesinden kaçının doğru sınıflandırıldığını veren bir orandır. 0.8409’luk recall oranı literatürdeki diğer yöntemlerle kıyaslandığında kabul edilebilir bir orandır. Çünkü yapılan çalışmada ses verisi kullanılmamıştır. Başka dil ve veri seti üzerinde çalışmış bile olsa Lipnet gibi %90’ın üzerinde başarı üreten neredeyse tamamında ses verisi de kullanılmıştır.

WER değeriyse recall değerinin tersine kelimelerdeki farklılıklar üzerine hesaplanan bir metriktir. WER değerinin hesaplanmasında “değiştirme”, “ekleme” ve “silme” gibi 3 ayrı parametre bulunmaktadır. Dudak okuma sistemlerindeyse veri setindeki kelimeler sabit olduğu için ekleme, silme yer almayıp sadece “değiştirme” değeri üzerinden hesaplanır. “Değiştirme” değeri toplam kelime sayısına bölündüğünde WER değeri hesaplanır. Recall ve WER değeri birbirini tamamlar. Çizelge 5.8’de en iyi sınıflandırılan ve recall değeri en yüksek olan kelimeler verilmiştir. Çizelge 5.7 ve 5.9’da karışan ve en kötü sınıflandırılan kelimelerin listesi verilmiştir. Bu kelimelerdeki performans düşüşleri sonraki bölümlerde tartışılmaktadır.

Dudak okuma sistemlerinde en önemli parametrelerden biri de precision değeridir. Bu değer, modelin sınıflar arasındaki sınıflandırma performansını ve olası başarısızlıklarını gözler önüne serdiği için oldukça değerli bir parametredir. Sistemin genel ortalama precision değeri 0.8674 veya %86.74’tür. Kelime bazlı incelendiğinde örneğin en kötü sınıflandırılan kelimelerin listelendiği Çizelge 5.9’da verilen kelimeler incelendiğinde “uçurtma”, “meslek”, “çaydanlık”, “stajyer” gibi kelimeler recall değeri olarak sistemin en kötülerinden olsalar bile precision değerleri ortalamanın üstündedir. Bu da aslında her ne kadar kötü sınıflandırılmış olsalar bile sistem başka kelimeleri, belirtilen bu kelimelerle karıştırmamıştır. Hatta “Türev” kelimesinin precision değeri 1 olduğu için başka hiçbir kelime bu sınıfa dahil edilmemiştir. Fakat bu durumun tam tersi en iyi sınıflandırılan kelimelerin listelendiği Çizelge 5.8’de görülebilir. Bu listedeki 15 kelimeden 1 tanesinin (“ama”) precision değeri

ortalamanın altındadır. Bu kelime %100 başarıyla sınıflandırılmıştır. Yani 90 tane "ama" kelimesi test edildiğinde 90 tanesi de doğru sınıflandırılmıştır. Fakat precision değeri 0.7'dir. Bu da başka kelimelerin de sık sayılabilecek derecede "ama" olarak sınıflandırdığını göstermektedir. Detaylarına bakıldığında "köpek" 19, "kaplan" 10, "baba" 7 defa "ama" kelimesiyle karıştırılmıştır. Diğer yandan karıştırılan sınıf sayısının çok fazla olmaması sadece 3 sınıf olması olumlu sayılabilir. Çizelge 6.1'de precision değeri en düşük kelimelerin listesi verilmektedir. Çizelge 6.1'de verilen listede son sütundaki kelimeler, sistem tarafından ilk sütundaki kelimeler olarak tahmin edilmiştir.

Çizelge 6.1. En kötü precision değerine sahip kelimeler.

Tahmin Edilen Kelime	**Recall**	**Precision**	**F1 Score**	**Orijinal Kelimeler**
Anne	%88.9	0.26	0.41	Fanatik (22), Fakat (20), Defter (33), Afiş (10)
Perde	%35.5	0.34	0.35	Meyve (35), Barfiks (8), Başlık (6)
Banka	%91.1	0.55	0.7	Perde (29), Kumanda (8), Bağlantı (7), Pervane (6)
Sandalye	%91.1	0.57	0.7	Çaydanlık (17), Sekreter (7), Stajyer (10), Bezelye (9)
Sekreter	%92.4	0.58	0.71	Cinayet (8), Fanatik (7), Salıncak (9),Şeffaf (11)

Çizelge 6.1'de son sütunda verilen kelimeler, ilk sütunda verilen kelimeler olarak sınıflandırılmıştır. Örneğin "defter" kelimesi 33 defa "anne" kelimesi olarak sınıflandırılmıştır. En az 5 defa karıştırılan kelimeler listelendiği için bu kelimelerin dışında da karıştırılan kelimeler bulunabilir. Precision değerinin seviyesi o kelimeye ait karıştırma durumunu vermektedir. Ayrıca Çizelge 6.1'de verilen kelimelerin, Çizelge 5.7'de yer alan kelimelerle direkt bir bağlantısı yoktur. Çünkü Çizelge 5.7'de 1. sütunda verilen kelimeler 2. sütunda verilen kelimelerin sınıfına dahil edilmiştir. Aslında 2. sütunda verilen kelimelerin recall değerini değiştirmeyip, precision değerini düşürmektedir. Örneğin Çizelge 5.7'de en çok karıştırılan kelime çiftlerinden biri "şınav-inşaat" çiftidir. "Şınav" kelimesi 24 defa "inşaat" olarak sınıflandırılarak recall

değerini düşürmüştür. Fakat "inşaat" sınıfına 24 defa "şınav" kelimesi ve 6 defa "salıncak" kelimesi dahil edildiği için precision değeri düşerek %74 olmuştur. Çizelge 6.1'de ki en düşük precision değerine sahip kelimeler ise hem fazla sayıda kelimeyle hem de yüksek miktarda karışmıştır. Çizelge 5.7'de verilen "perde" kelimesi aynı zamanda Çizelge 6.1'de de yer almıştır. Bu da özellikle "meyve" kelimesinin 35 defa "perde" olarak sınıflandırılmasından kaynaklanmaktadır.

Recall ve precision değerleri ayrı ayrı olarak sistemin performansını ölçse de bakış açıları farklıdır. Bir kelimenin recall değeri yüksek iken precision değeri düşük olabilir. Bu da o kelimenin her ne kadar doğru tahmin edilse de başka kelimelerin de hatalı bir şekilde o kelimenin sınıfına dahil edildiğini gösterir. Bu sebeple tek başına bu 2 parametre, sistemin performansını ölçmeye yetmez. Tüm sistemin ve kelime bazlı sınıflandırma performans metrikleri için en geçerli olanı f1 skorudur. Çünkü bu değer hesaplanırken precision ve recall değerlerinin harmonik ortalaması alınır. Böylece f1 skoru, hem recall hem de precision değerinden etkilenir. Bu sayede recall değeri yüksek olan fakat precision değeri düşük olan veya recall düşük fakat precision yüksek olan kelimeler için tek bir metrik hesaplanabilir. Örneğin "biçimlendirmek" kelimesi için f1 skoru 1 olarak hesaplanmıştır. Bu da "biçimlendirmek" kelimesinin hem %100 (recall=1) oranda doğru sınıflandırıldığını hem de başka hiçbir kelimenin "biçimlendirmek" olarak sınıflandırılmadığını gösterir. Aynı şekilde "ama" kelimesinin recall değeri 1 iken f1 skoru ise 0.82'dir. Çünkü sistem başka kelimeleri de "ama" olarak sınıflandırdığı için precision değerini düşürür. Düşen precision da f1 skorunu düşürür. Aynı durum Çizelge 6.1'de verilen "sekreter", "sandalye" ve "banka" kelimeleri için de geçerlidir. Bu 3 kelimenin de recall değeri 0.9'un üzerinde olmasına rağmen precision değerleri en düşük kelimeler arasında yer aldığı için f1 skorları 0.7 civarlarındadır.

Çizelge 5.7'de en fazla başka kelimeyle karışan kelimelerin listesi verilmiştir. Bu listede dikkat çeken kelimelerden bir tanesi "perde" kelimesidir. "Perde" kelimesi metriklerin değerlendirilmesi açısından incelenmesi gerekir. "Meyve" kelimesi 35 defa "perde" kelimesiyle karışmıştır. Bu durum 90 adet "meyve" kelimesinin 35 defa "perde" kelimesi olarak sınıflandırıldığı anlamına gelir. Bu durum "perde" kelimesi için precision değerinin düşmesine sebep olur. Çünkü "perde" sınıfında olmayan

kelimeler de “perde” sınıfına dahil edilmiştir. Bu sebeple precision değeri düşer. Ayrıca yine aynı tabloda “perde” kelimesi “banka” kelimesi olarak 29 defa sınıflandırılmış. Bu durumda da “perde” sınıfının recall değeri düşer. Böylece “perde” için recall, precision ve hatta ikisi üzerinden hesaplanan f1 skor değeri oldukça düşük çıkmalıdır. Çizelge 5.9 incelendiğinde bu durumun gerçekleştiğini ve f1 skor değerinin açık ara farkla en düşük değere sahip olduğu görülebilir.

Sistemin genel performansı değerlendirilecek olursa recall değeri 0.8409, precision 0.8674 ve f1 skor 0.8432’dir. Sistemin genel performansını gösteren f1 skorunun recall değerinden az da olsa yüksek olması sistemin hem doğru sınıflandırma başarısının yüksek olduğunu hem de başka kelimelerin yanlış sınıflandırılması noktasında dengeli bir performans sergilediğini göstermektedir.

6.2. KELİMEDEKİ HARFLERİN PERFORMANSA ETKİSİ

Dudak okuma sistemlerinin düzgün çalışabilmesi için dudakların hareket etmesi gerekir. Dudak okuma sistemleri dudak hareketlerinden bir örüntü oluşturarak bunları kelimelerle eşleştirir. Bu yüzden dudak okuma sistemlerinin harfler ile de çok ciddi bir ilişkisi bulunur.

Türkçede harfler, ünlü harfler ve ünsüz harfler olarak ayrılmaktadır. Ünsüz harfler ise boğumlanma noktalarına göre 4’e ayrılır.

- Dudak Ünsüzleri
- Diş Ünsüzleri
- Damak Ünsüzleri
- Gırtlak Ünsüzleri

Gırtlak ünsüzleri, dudakların neredeyse hiç hareket etmeden ses tellerinin birbirine yaklaşarak veya dokunarak söylendiği ünsüzlerdir. Türkçede tek bir gırtlak ünsüzü bulunur o da h harfidir.

Diş ünsüzleri gırtlak ünsüzlerine göre biraz da olsa dudakları hareket ettirir. Diş ünsüzleri dil ucunun diş etine, damak sınırına veya üst dişlere değmesiyle veya yaklaşmasıyla boğumlanarak çıkan ünsüzlerdir. Bunlar da ikiye ayrılır. "Asıl diş ünsüzleri", 'd', 't', 'z', 's', 'n', 'r' diş etine değerek çıkar. Dudakta hafif bir açılmaya sebep olur. "Diş eti ünsüzleri" ise diş eti – damak noktasında boğumlanır. Bunlar 'c', 'ç', 'j', 'l', 'ş' gibi harflerdir. Asıl diş ünsüzlerine göre dudaklarda daha fazla hareket oluşturur.

Damak ünsüzleri, dilin sırt tarafının tümseklenerek ön veya arka damağa yakınlaşması ve dokunması sonucu çıkarılan ünsüzlerdir. Bunlar 'k', 'g', 'ğ', 'y' harfleridir. Tamamen dil üzerinden söylendiği için bulunduğu hecenin durumuna göre dudakta harekete hiç sebep olmayabilir.

Dudak ünsüzleri, iki dudağın birbirine veya alt dudağın üst dişlere dokunması ya da yaklaşmasıyla boğumlanarak çıkar. Eğer 2 dudak birbirine dokunarak çıkıyorsa bunlara "çift dudak ünsüzleri" denir. Çift dudak ünsüzleri 'b', 'p', 'm' harfleridir. Bu harfler dudağın açılıp kapanmasına sebep olduğu için bulunduğu kelimenin söylenmesi sırasında ciddi dudak hareketliliği sağlar. Bunun yanında "diş-dudak ünsüzlerinde" alt dudak üst dişlerle birleşir. Bu da sadece alt dudakta kısmi bir hareketlilik sağlar. Bunlar f ve v harfleridir.

Çizelge 6.2. Dudak ünsüzlerinin listesi.

Dudak Ünsüzleri	
Çift Dudak Ünsüzleri	**Diş Dudak Ünsüzleri**
/b/	/f/
/p/	/v/
/m/	

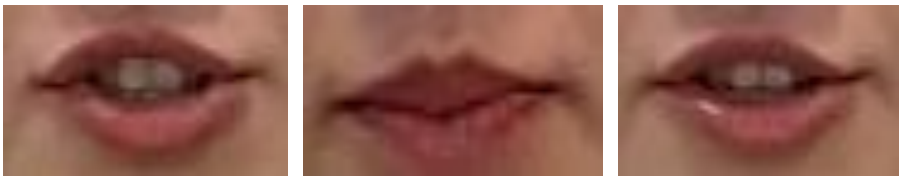
Ünlü harflerinde ise dudakların durumuna göre "düzlük" ve "yuvarlaklık" bulunur. Düz ünlüler dudakta hareket olmadan ağız açıklığından çıkar. Fakat yuvarlak ünlülerdeyse dudağın büzülüp yuvarlaklaşmış duruma gelmesi gerekir. Büzüşmesi

dudakta kısmi bir hareketlilik sağlar. Bu sebeple yuvarlak ünlüler, düz ünlülere göre dudakta daha fazla hareketlilik oluşturur [236].

Çizelge 6.3. Ünlü harflerin durumu.

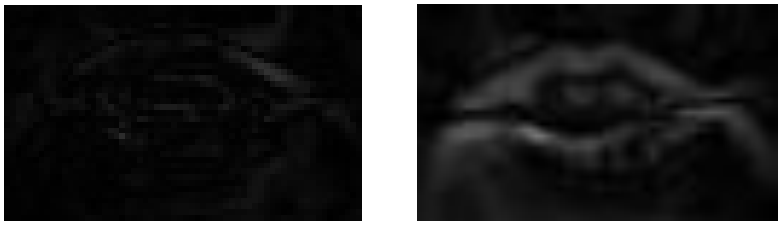
Ünlü Harfler	
Düz	**Yuvarlak**
/a/	/o/
/e/	/ö/
/ı/	/u/
/i/	/ü/

Harflerin ayrımına bakıldığında en çok çift dudak ünsüzlerinin (b, p ve m harfleri) dudakta harekete sebep olduğu görülmektedir. Yuvarlak ünlüler ve diş-dudak ünsüzleri ise çift dudak ünsüzleri kadar olmasa da dudakta hareketlilik sağlar.
Örneğin aynı konuşmacının c, m, ş harfleriyle -ce, -me, -şe hecelerini söylerkenki başlangıç dudak durumları Şekil 6.1'de sırasıyla verilmiştir.



Şekil 6.1. Hecelerin (ce, me, şe) söylenişi sırasında elde edilen kareler.

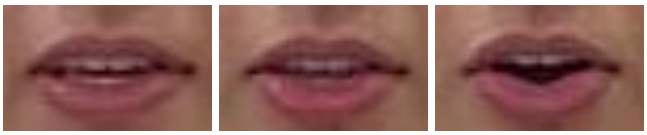
Dudakta hareket olup olmadığını ve harflerin söylenirken dudakta ne kadar fark oluştuğunu görmek için görüntülerin farkı alınmıştır. Görüntülerin farkının alınmasıyla elde edilen sonuçlar solda c ve ş harflerinin farkı ve sağda ş ve m harflerinin farkı olacak şekilde Şekil 6.2'de verilmiştir.



Şekil 6.2. C ve ş-m ve ş farkı.

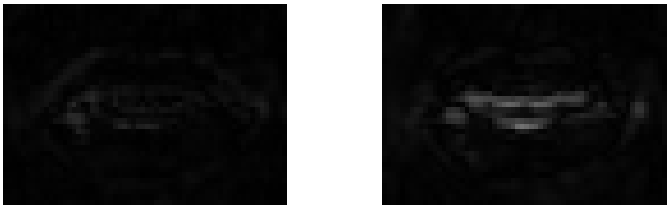
Şekil 6.2 de görüldüğü üzere (b) görüntüsünde görüntüdeki fark çok daha belirgindir. ‘c’ ve ‘ş’ ünsüzlerinin çıkışı noktasında ciddi fark oluşmazken dudak ünsüzlerinden biri olan ‘m’ harfi dudağın kapanmasını sağladığı için ciddi bir fark oluşturmuştur.

Yuvarlak ve düz ünlülerin kıyaslanmasındaysa bir konuşmacının video kelimesini söylerkenki ‘i’, ‘e’, ‘o’ harflerinin çıkışına ait kareler incelenmiştir. Şekil 6.3’te sırasıyla ‘i’, ‘e’ ve ‘o’ harflerine ait kareler verilmiştir.



Şekil 6.3. Video kelimesine ait i, e ve o harflerinin kareleri.

Özellikle o harfinin söylenişi sırasında dudaktaki açıklık dikkat çekmektedir. Ünlü harflerin söylenişinde önce ‘i’ ve ‘e’ harfine ait karelerin farkı alınmıştır. Daha sonra ‘e’ ve ‘o’ harflerinin farkı alınmıştır.



Şekil 6.4. ‘i’ ve ‘e’ harflerinin farkı (solda), ‘e’ ve ‘o’ harflerinin farkı (sağda).

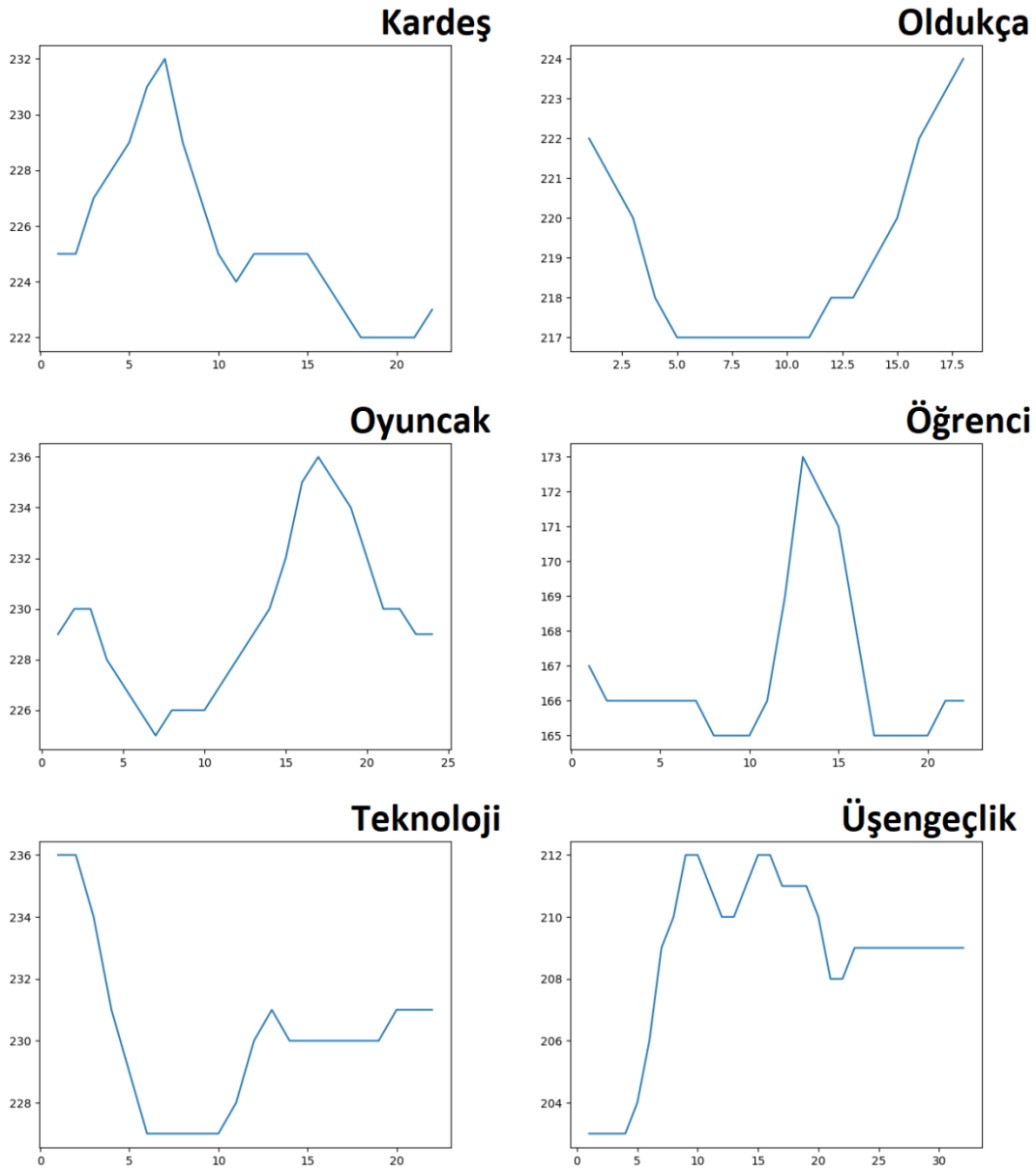
Şekil 6.4’te görüldüğü üzere düz ünlü olan ‘i’ ve ‘e’ harflerinin söylenişi arasında pek fark bulunmazken ‘o’ harfini söylerken dudak şekli değişmiştir. Ünlülerin durumundaysa yuvarlak ünlüler dudakların yuvarlanarak açılmasını sağladığı için

dişlerin ve dudakların görünümü değişir. Şekil 6.4'te verilen sağdaki görüntü farkı bunu net bir şekilde ortaya koymaktadır.

Dudak ünsüzlerin dudakta fark edilebilir harekete sebep olduğuna dair bir diğer ipucu landmarklar üzerinden gösterilebilir. Dudağın etrafını saran landmarkları veren mediapipe kullanılarak alt dudağın veya üst dudağın uç tarafına ait bir nokta seçilebilir. Bu landmarkların, dudağın yapısı gereği konuşurken yataydaki konumları neredeyse hiç yer değiştirmezken, dikeydeki konumları dudağın hareketine göre değişebilmektedir. Alt dudak için 17, üst dudak için 12. nokta kullanılabilir.

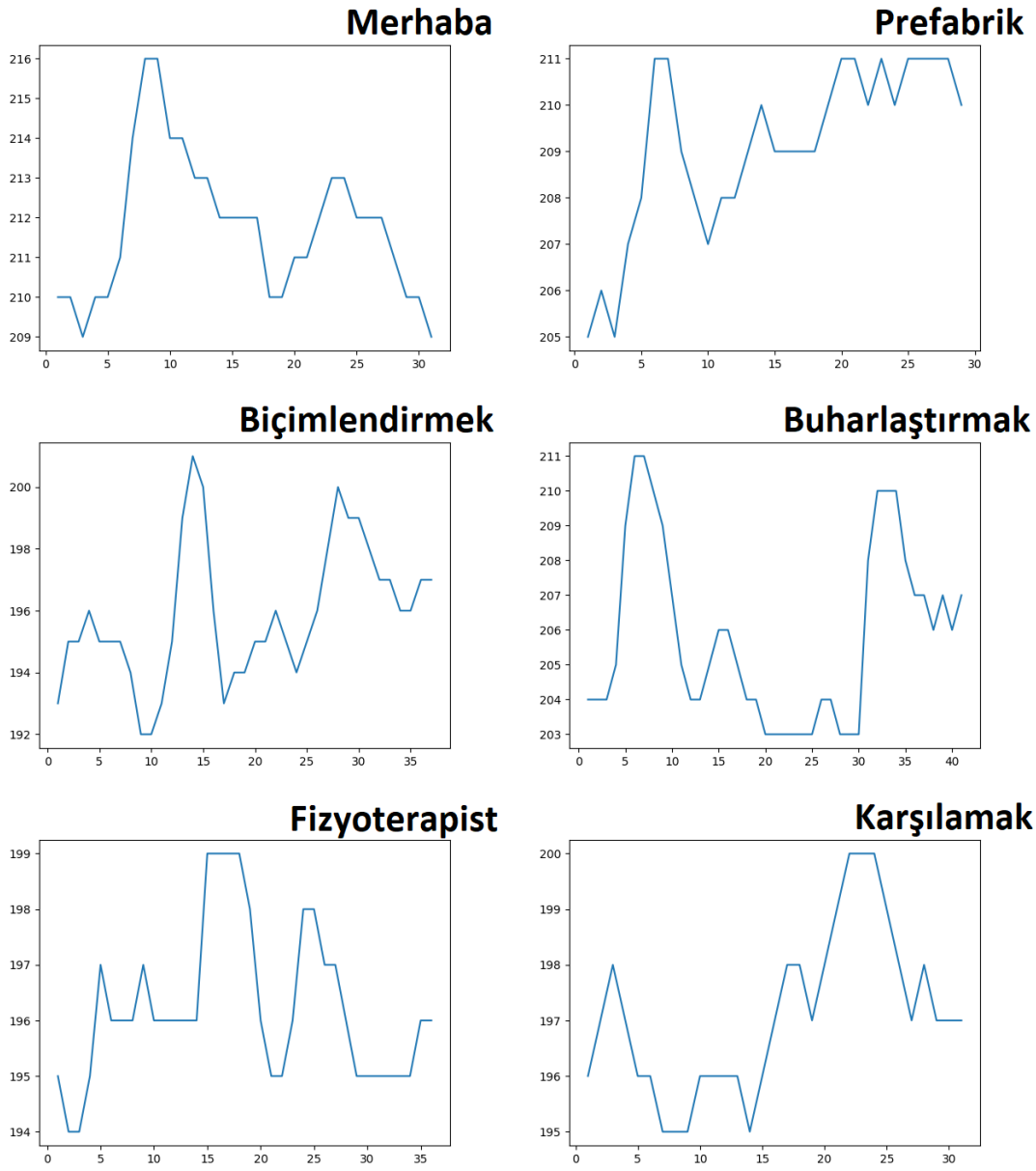
Şekil 6.5. 12 ve 17. noktaların temsil ettiği konumlar.

Benzer grafikler elde edildiği için 17. noktanın grafik değerleri verilecektir. İlk olarak çift dudak ünsüzü barındırmayan kelimelerdeki 17. Noktanın y eksenindeki konum değişimlerini gösteren grafikler Şekil 6.6'da verilmiştir.



Şekil 6.6. Çift dudak ünsüzüne sahip olmayan kelimelerin nokta konum grafiği.

Şekil 6.6'daki grafiklerde görüldüğü üzere y eksenindeki konumlarda çok fazla dalgalanma yoktur. Genelde doğrusala yakın grafikler elde edilmiştir. Şekil 6.7'de ise çift dudak ünsüzlerine sahip olan bazı kelimelerin söylenişi esnasında oluşan yer değişimin grafikleri verilmiştir.



Şekil 6.7. Çift dudak ünsüzüne sahip kelimelerin nokta konum grafiği

Şekil 6.6.'da ve 6.7'de verilen grafikler karşılaştırıldığında dudak ünsüzlerine sahip kelimelerin söylenişi esnasında özellikle Şekil 6.7'de verilen grafikler incelendiğinde dudakta ciddi hareketlenmelerle y ekseninde yer değişikliği meydana geldiği görülmektedir. Bu da çift dudak ünsüzlerine sahip harfleri barındıran kelimelerin söylenişi esnasında dudağın daha fazla hareket ettiğini gösteren bir başka veridir.

Kelime veri setindeki 111 adet kelimelerin 74 tanesi 'b', 'p' veya 'm' harflerinden en az bir tanesine sahipken 37 tanesi ise bu 3 harflerden hiçbirini barındırmaz. Dudak

ünsüzüne sahip olan ve olmayan kelimelerin ayrı ayrı recall, precision ve f1 skor değerleri Çizelge 6.4'te verilmiştir.

Çizelge 6.4. Dudak ünsüzlerin durumuna göre metrik değerleri.

Çift Dudak Ünsüzü	**Kelime Sayısı**	**Recall**	**Precision**	**F1 Skor**
Var	74	0.8823	0.895	0.88
Yok	37	0.7562	0.810	0.76

Çizelge 6.4'teki veriler incelendiğinde dudak okuma sisteminin çift dudak ünsüzüne sahip olan kelimeleri sınıflandırırken her 3 metrikte de çok daha iyi ve genel ortalamanın üstünde performans gösterdiği görülmüştür.

Çizelge 5.8'de en iyi sınıflandırılan 15 kelimenin listesi verilmiştir. Bu listedeki kelimelerin tamamı çift dudak ünsüzü barındırmaktadır. Çizelge 5.9'da ise en kötü sınıflandırılan 15 kelimenin listesi verilmektedir. 15 kelimeden sadece 6 tanesi çift dudak ünsüzü barındırmaktadır. En iyi sınıflandırılan kelimelerinin tamamında çift dudak ünsüzü bulunurken, en kötü sınıflandırılanların kelimelerin sadece %40'ında bulunması çift dudak ünsüzlerin sistemin performansını arttırdığını kanıtlayan bir diğer kanıt olmaktadır.

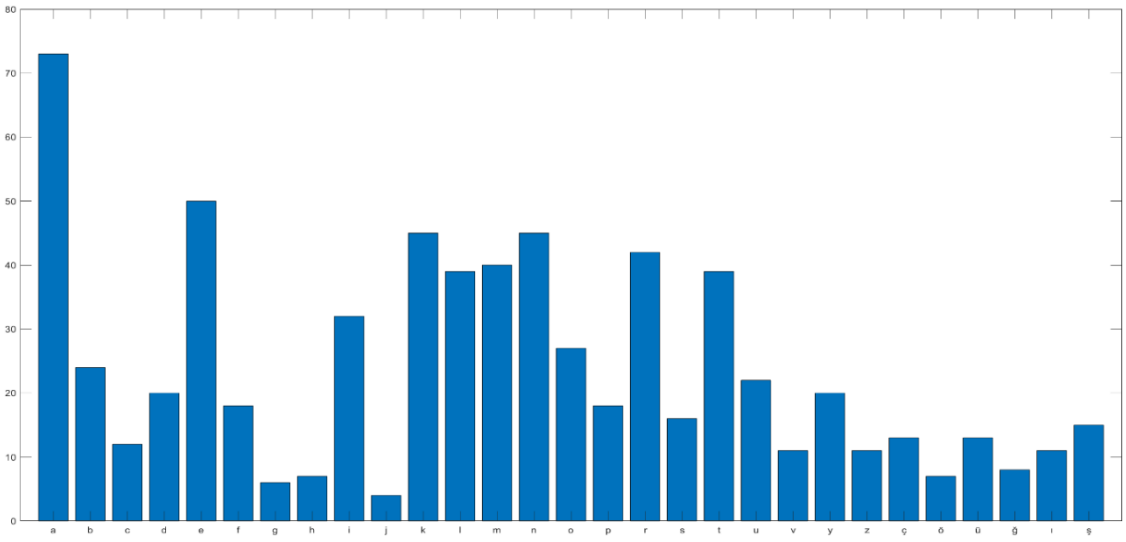
Ünlü harfler için de benzer yöntem izlenmiştir. 111 kelime arasından 59 tanesi yuvarlak ünlülerden en az birine sahiptir. 52 tanesiyse sadece düz ünlülere sahiptir.

Çizelge 6.5. Ünlü harflere göre performans bilgileri.

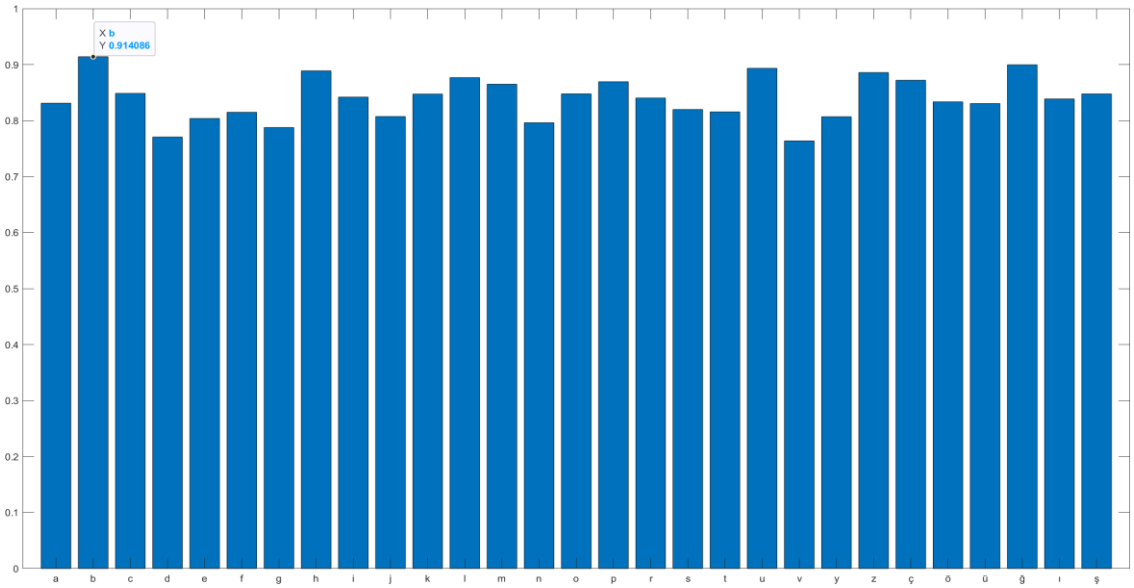
	Kelime Sayısı	**Recall**	**Precision**	**F1 Skor**
En az bir yuvarlak ünlü	59	0.851	0.91	0.87
Sadece düz ünlüler	52	0.829	0.82	0.80

Çizelge 6.5 incelendiğinde kelimelere yuvarlak ünlüler eklendiğinde precision değerinin ortalamada yükseldiği görülmektedir. Bu durum yuvarlak ünlülerin düz

ünlülere göre kelimelerin ayırt edilmesinde ve yanlış sınıflara dahil edilmemesini sağlamaktaki başarısını gösterir. Çizelge 6.5'te sadece yuvarlak ünlülere sahip olup düz ünlülere sahip olmayan kelimelerin değerleri eklenmemiştir. Çünkü buna uyan 3 kelime bulunmaktadır. Bunların sınıflandırma oranındaki yüksek başarısı her ne kadar yorumumuzu desteklese de örnek sayısının azlığından dolayı yanlış yönlendirmemesi açısından eklenmemiştir. Şekil 6.8'de her harfin veri setindeki bulunduğu kelime sayılarının histogramını verilmektedir. Şekil 6.9'da harflerin bulunduğu kelimeye göre ortalama f1 skorlarına ait grafik verilmiştir.



Şekil 6.8. Harflerin kelimelerde kullanım histogramı.



Şekil 6.9. Harfleri barındıran kelimelerin ortalama f1 skorları.

Şekil 6.9 incelendiğinde çift dudak ünsüzü olan ‘b’ harfi f1 skoru en yüksek değere sahiptir. Şekil 6.9’daki grafik incelenirken Şekil 6.8’deki histogramda da harflerin kullanım sıklığına bakılmalıdır. Aksi takdirde az sayıda kullanılan ‘h’ gibi harflerin başarı oranlarındaki yüksekliğin sebebi tam olarak anlaşılamaz. Örneğin h harfi sadece 7 kelimede kullanılmakta olup bunların 6’sında çift dudak ünsüzü olan kelimelerdir. Yani h harfindeki yüksekliğin sebebi sık ve gerçekten gösterdiği yüksek başarı değil de az kullanılması ve çift dudak ünsüzlerle birlikte kullanılmasından kaynaklanmaktadır. Benzer durum ‘j’ harfi ve benzeri az kullanılan harfler için de geçerlidir.

6.3. KELİMELERDEKİ BENZERLİKLERİN PERFORMANSA ETKİSİ

Bu bölümde kelimeler arasındaki “Levenshtein Algoritmasından” elde edilen mesafenin veya kelimeler arasındaki benzerliğin dudak okuma sistemlerindeki performansına değinilmektedir.

İlk olarak veri setindeki her bir kelimenin geri kalan 110 kelimeyle uzaklık değeri hesaplanıp ortalaması alınmıştır. Bu şekilde her kelime için uzaklık değeri elde edilmiştir. Veri setindeki her kelime için hesaplanan uzaklık değerlerinin de ortalaması alındığında veri setini temsil eden ortalama bir levenshtein değeri elde edilmiş olur. Tüm veri setinin ortalama uzaklık değeri 7.4219 olarak hesaplanmıştır.

Çizelge 5.6’da sistemin tüm veri seti üzerinde karıştırdığı kelimelerin listesi verilmektedir. Bu çizelgede ilk sütundaki kelimeler, ikinci sütundaki kelimelerin sınıfına dahil edilmiştir. Karıştırılan bu kelimeler arasındaki uzaklık ortalaması 6.01’dir. Bu değer veri setine ait ortalama değerin altındadır. Yani kötü sınıflandırılan kelimeler Levenshtein benzerliğine göre daha benzer kelimelerdir.

Çizelge 5.7’de ise Çizelge 5.6’daki verilerden yola çıkarak en çok karıştırılan 5 kelimenin listesi verilmiştir. Bu listedeki 5 kelimenin benzerlik değeri daha da düşük (veya benzer) çıkarak 4.6 elde edilmiştir. Bu durum kelime benzerliğiyle, dudak

okuma sistemlerinin performansı arasında ciddi bir ilişki olduğunu göstermektedir. Bu değerlerin özeti Çizelge 6.6'da verilmiştir.

Çizelge 6.6. Dudak ünsüzlerin durumuna göre metrik değerleri.

	Levenshtein Benzerlik Ortalaması
Tüm Veri Setinin	7.4219
Karıştırılan Kelimeler (Çizelge 5.6)	6.01
En çok karıştırılan 5 Kelime (Çizelge 5.7)	4.6

Veri setindeki her bir kelime için ayrı ayrı hesaplanmış olan levenshtein değeri en yüksek olan kelimelerin yani diğer kelimelerden daha farklı olan kelimelerin benzerlik oran ortalamaları ve performans metrikleri Çizelge 6.7'de verilmektedir.

Çizelge 6.7. Levenshtein uzaklık ortalaması en yüksek kelimelerin listesi.

Kelime	**Levenshtein Değeri**	**Recall**	**Precision**	**F1 Skor**
Türkçeleştirmek	12.80	1	0.96	0.98
Biçimlendirmek	11.81	1	1	1
Buharlaştırmak	11.58	1	0.97	0.98
Fizyoterapist	11.40	1	0.95	0.97
Şereflendirme	11.06	0.9778	0.90	0.94

Çizelge 6.7'de verilen 5 kelimenin 4'ü, Çizelge 5.8'de verilen en iyi sınıflandırılan kelimeler listesinde yer almaktadır. Böylece Çizelge 6.7 gösteriyor ki kelimeler farklılaştıkça sistemin kelimeleri tanıma performansı oldukça yükselmektedir. Benzerlik uzaklığı en düşük ortalamaya sahip 5 kelimenin listesi Çizelge 6.8'de verilmiştir.

Çizelge 6.8. Levenshtein uzaklık ortalaması en düşük kelimelerin listesi.

Kelime	Levenshtein Değeri	Recall	Precision	F1 Skor
Banka	6.14	0.911	0.52	0.66
Çanta	6.19	0.778	0.79	0.78
Fakat	6.19	0.567	0.68	0.61
Baba	6.21	0.888	0.97	0.93
Kaplan	6.29	0.788	0.83	0.81

Çizelge 6.8'daki değerler incelendiğinde benzerlik uzaklıklarının en düşük olduğu kelimelerin performans metrikleri ortalamanın altında kalmaktadır. Özellikle precision değerinin düşük olması, bir kelimenin diğer kelimelerle benzerliği arttıkça karıştırılma olasılığının arttığını gösteren bir diğer somut veridir. Buna aykırı olan bir kelime var. O da "baba" kelimesidir. Bu kelimede bulunan 2 adet çift dudak ünsüzü olan 'b' harfi kelimenin doğru sınıflandırmasına ciddi oranda katkı sağladığı için bu durum meydana gelmiştir. Diğer yandan özellikle "banka" kelimesinin 1 tane bile olsa çift dudak ünsüzüne sahip olduğu halde neden bu şekilde en düşük değere sahip olduğu farklı bir açıdan Bölüm 6.6'da tartışılmaktadır.

Bölüm 5'te yapılan 6. testin sonuçlarında "baban" kelimesinin "babaanne" kelimesiyle karıştığı görülmektedir. Fakat 4. testin sonuçlarında "baba", "babaanne" ile karışmamaktadır. "baba"–"babaanne" kelimeleri arası uzaklık 4 birimken "baban"-"babaanne" arası 3 birimdir. Kelime benzerlik olarak yakınlaştıkça sistemin tanıması zorlaşmıştır. Aynı şekilde "türkiye" kelimesi 4. testte "teknoloji" kelimesiyle karıştırılmamışken "türkiyeyi" ve "teknoloji" karışmıştır. Bu örnekteki durumda da uzaklık 8'den 7'ye düşmüştür.

Ayrıca kelimelerin benzerliği her zaman bu şekilde performansa etki etmemiştir. Örneğin Çizelge 5.6'da "merdiven", "meyve" ile uzaklığı 4 birimken 10 defa "meyve" olarak sınıflandırılmıştır. Yine aynı "merdiven" kelimesi, aralarında 7 birim uzaklık bulunan "akraba" kelimesi olarak 11 defa sınıflandırmıştır. Yani "merdiven" kendisine

daha uzak olan bir kelimeyle daha fazla karıştırılmıştır. Bir başka örnekteyse "perde" 4 farklı kelimeyle karışmıştır. Fakat en çok karıştığı 4 kelime arasında karışma oranı olarak en yüksek olduğu kelime kendisine benzerlik olarak en uzak olan kelimedir. Bu duruma ait bazı örneklerin değerleri Çizelge 6.9'da verilmektedir.

Çizelge 6.9. Benzerlik oranıyla ters sonuç üreten bazı kelimeler.

Kelime	**Sınıflandırma Sonucu**	**Levenshtein Değeri**	**Yanlış Sınıflandırma Adeti**
Merdiven	Meyve	4	10
	Akraba	***7***	***11***
Perde	Merdiven	4	7
	Veya	4	10
	Banka	***5***	***29***
	Anne	4	6
Stajyer	Sandalye	5	10
	Anne	***6***	***12***

Çizelge 6.9'da verilen kelimeler ve değerleri, dudak okuma sistemlerinde kelimelerin benzerliklerinin performansa etkisine olan doğrusal ilişkisini zedelemektedir. Benzerlik ve performans arasında bir ilişki olsa da bu ilişki dudak ünsüzü, kelimelerin uzunluğu gibi diğer performansı değiştirebilen durumlardan oldukça fazla etkilenmektedir.

6.4. KELİME UZUNLUKLARININ PERFORMANSA ETKİSİ

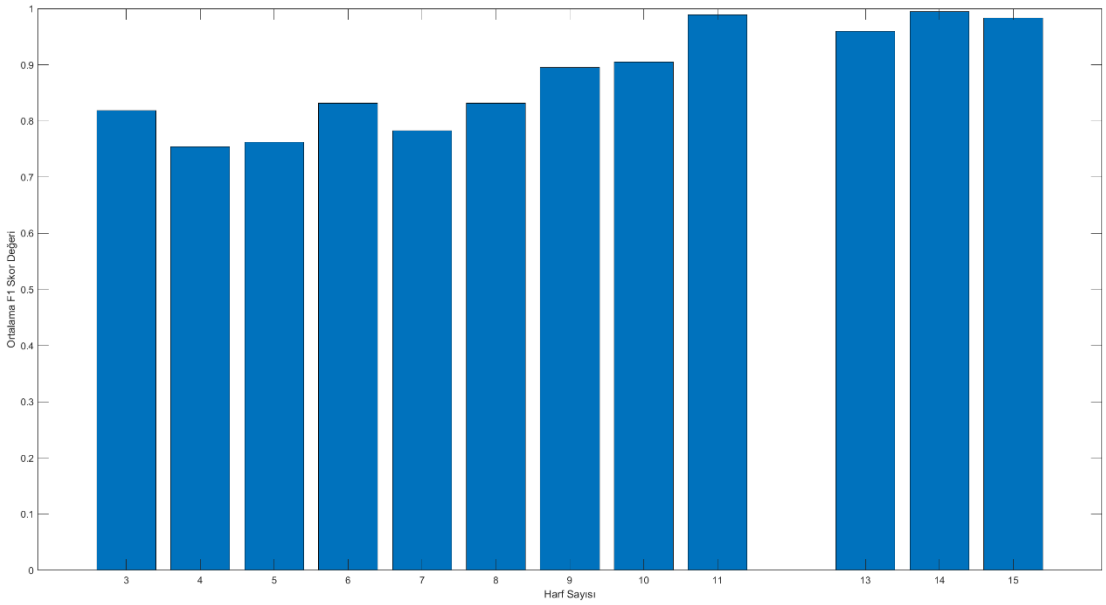
Bu bölümde kelimelerdeki harf sayılarının dudak okuma performansı değerlendirilmektedir. Çizelge 6.10'da veri setine ait kelime uzunluk istatistikleri verilmektedir. Veri setinde 111 kelimenin ortalama harf sayısı 7.3'tür. En uzun kelime 15 harfli "türkçeleştirmek" kelimesidir. En kısa kelime 3 harfli "ama" kelimesidir. Çizelge 5.7'deki en çok karıştırılan 5 kelimenin harf ortalaması 5.6 ve aynı zamanda karıştırıldıkları kelimelerin de harf ortalaması 5.6'dır ve bu değerler ortalamanın

altındadır. Çizelge 5.9'da verilen en kötü sınıflandırılan 15 kelimenin uzunluk ortalaması 6.46'dır ve ortalamanın altındadır. Çizelge 5.8'de verilen en iyi sınıflandırılan 15 kelimenin ortalama uzunluğu 9.66'dır ve ortalamanın üstündedir. Tüm veri setinin ortalamasının 7.3 olduğu göz önünde bulundurulduğunda kısa kelimelerin daha zor sınıflandırıldığı ve daha uzun kelimelerin daha iyi sınıflandırıldığı sonucuna varılabilir.

Çizelge 6.10. Veri setine ait kelime uzunluk istatistikleri.

	Değer
En Kısa Kelime	"Ama" – 3 Harfli
En Uzun Kelime	"Türkçeleştirmek" – 15 Harfli
Tüm Veri Setinin Uzunluk Ortalaması	7.30
En İyi Sınıflandırılan 15 Kelimenin Ortalaması	9.66
En Kötü Sınıflandırılan 15 Kelimenin Ortalaması	6.46
En Çok Karıştırılan 5 Kelimenin Ortalaması	5.60

En çok karıştırılan 5 kelime incelendiğinde orijinal kelime ve yanlış sınıflandırıldığı kelimelerin harf sayılarının ya aynı ya da çok yakın olduğu bunun yanında uzunluklarının hep ortalamanın altında kaldığı görülmektedir. Bu da kısa kelimelerin yine kısa kelimelerle karışma olasılığını arttırdığını göstermektedir. Şekil 6.10'da verilen harf sayılarına göre f1 skor performans grafiği de uzun kelimelerin daha iyi sınıflandırıldığını göstermektedir.



Şekil 6.10. Kelime uzunluklarına göre sınıflandırma F1 skor ortalama grafiği.

Şekil 6.10'da 3 harfli kelimeleri temsil eden f1 skor değerinin yüksek olduğu görülmektedir. Bu durum yanıltıcıdır. Çünkü 3 harfli tek bir kelime bulunur. O da "ama" kelimesidir. Tek kelimenin getirdiği yanıltıcı f1 skor değeri göz ardı edilebilir.

6.5. SAKAL BIYIK GİBİ FİZİKSEL ETKENLER

Dudak okumanın temel parametresi dudakların takibidir. Tez kapsamında Bölüm 2 ve 3'te literatürde yapılan birçok çalışma göstermiştir ki eğer dudakları kapatan herhangi bir nesne, açısal değişiklik veya erkeklerdeki sakal-bıyık durumları varsa bunlar dudak okumayı zorlaştırmaktadır.

Oluşturulan veri setinde dudağı kapatacak kadar sakal ve bıyığa sahip 2 konuşmacı bulunmaktadır. Bunlara ait görseller Şekil 6.11'de verilmiştir.

Şekil 6.11. Sakal/bıyık sahibi konuşmacıların görüntüsü.

Bu durumun etkisini göstermek açısından ilk olarak modelin eğitim aşaması, bu 2 konuşmacının verileri gösterilmeden gerçekleştirilmiştir. Eğitimin amacı sisteme daha önce böyle "problemli" verilerin gösterilmediği durumlarda performansın ölçülmesini sağlamaktır. Sistem eğitildikten sonra 2 konuşmacıya ait toplamda 3330 adet örnek veri test aşamasında gösterilmiştir. Elde edilen sonuçlar Çizelge 6.11'de verilmiştir.

Çizelge 6.11. Sakalsız konuşmacıların test sonuçları.

	Recall	Precision	F1 Skor
Konuşmacı-1	0.8078	0.821	0.814
Konuşmacı-2	0.7946	0.782	0.788

Çizelge 6.11'de de görülebileceği üzere eğer sistem eğitim aşamasında hiç bu tarz bir veri görmediyse test sırasında performansı düşmüştür. Daha sonra konuşmacı-1'in verileri eğitim aşamasında gösterildikten sonra sadece konuşmacı-2'nin verileri test edilmiştir. Buna dair test sonuçları da Çizelge 6.12'de verilmiştir.

Çizelge 6.12. İkinci testin sonuçları.

	Recall	Precision	F1 Skor
Konuşmacı-2	0.8156	0.819	0.821

Çizelge 6.13'te görüldüğü üzere sistem eğitim aşamasında sakallı ve bıyıklı 1 konuşmacıyla eğitildiğinde f1 skorunda %3.3'lük bir artış olmuştur. Bu veriler, veri setine yeterince bu şekildeki verilerin eklenmesi durumunda sakal ve bıyık tarzı problemlerin oluşturduğu olumsuzlukların azaltılabileceğini yönelik ipuçları vermektedir.

6.6. KARIŞTIRILAN KELİMELER ve SORUNLAR

Harflerin analiz edildiği Bölüm 6.2.'de çift dudak ünsüzlerinin başarıyı arttırdığı belirtilmiştir. Elde edilen sonuçların birçoğu bunu destekler niteliktedir. Hatta en iyi sınıflandırılan 15 kelimenin tamamında en az 1 tane çift dudak ünsüzü bulunmaktadır. Fakat detaylı incelendiğinde Çizelge 5.7'de verilen en çok karıştırılan 5 kelimenin verildiği listede 2 kelime çifti dikkat çekmektedir. Hatta en çok karıştırılan kelime, çift dudak ünsüzü barındıran "meyve" kelimesidir. "Meyve" kelimesi 35 defa "perde" olarak sınıflandırılmıştır. Diğer kelime çiftinde de "perde" kelimesi 29 defa "banka" olarak sınıflandırılmıştır. Her ne kadar çift dudak ünsüzlerinin performansı arttırdığı söylense de bir dezavantajı var ki o da 'p', 'b' ve 'm' harflerinin dudaktan çıkışının neredeyse birebir aynı olmasıdır. "Meyve" kelimesi 'm' harfiyle, karıştırıldığı "perde" kelimesi ise 'p' harfiyle başlamaktadır. Yani her ikisi de çift dudak ünsüzüyle başladığı için dudağın kapatılıp tekrardan açılmasıyla söylenir.

Çift dudak ünsüzlerinin diğer ünsüzlerle arasında dudaktaki söyleniş açısından ciddi farklar olduğu Bölüm 6.2'de zaten belirtilmiştir. Fakat bu sefer ikisi de dudak ünsüzü olan 2 harf kıyaslanacaktır. İlk olarak ikisi de dudak ünsüzü olan fakat biri çift dudak ünsüzü olan 'p' harfiyle, diş dudak ünsüzü olan 'v' harfi arasındaki fark incelenmelidir.

Şekil 6.12. Konuşmacının p (solda) ve v(sağda).

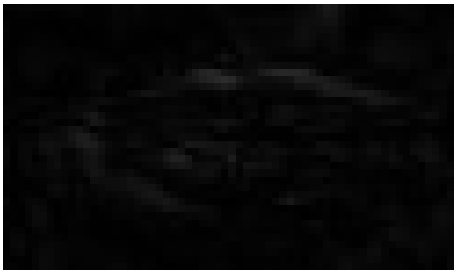
Görüntülerin farkı alındığında Şekil 6.13'teki gibi bir görüntü oluşur.

Şekil 6.13. Görüntülerin (Şekil 6.12.) farkı.

Şekil 6.13'te görüldüğü üzere her ikisi de dudak ünsüzü olan 'p' ve 'v' harflerinin söylenişi esnasında dudaktaki hafif açıklık ve dişlerin konumu dahil pek çok fark meydana gelmektedir. Çift dudak ünsüzü olan ve karıştırıldığı düşünülen 'p', 'b' ve 'm' harfleri için bir konuşmacının "meyve" ve "perde" kelimesini söylerkenki videonun ilk kareleri üzerinden değerlendirme yapılabilir. Şekil 6.14'te bir konuşmacının p harfini ve m harfini söylerken alınan kareler verilmiştir.

Şekil 6.14. Konuşmacının 'p' (solda) ve 'm' (sağda) söyleyişi.

Bu 2 görüntünün farkını aldığımızda bu harflerin dudaktan ne kadar benzer çıktığı görülebilir. Görüntülere ait fark Şekil 6.15'te verilmiştir.

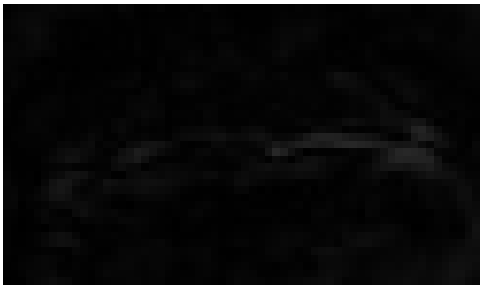


Şekil 6.15. Görüntülerin (Şekil 6.14) farkı.

Diğer kelime çifti olan "perde" ve "banka" kelimesi için farklı bir konuşmacıdan rastgele seçilen videoların ilk karesi Şekil 6.14'te verilmiştir.

Şekil 6.16. Konuşmacının p (solda) ve b (sağda) söyleyişi.

Görüntülerin farkı alındığında Şekil 6.13'tekine benzer bir durumla karşılaşılmaktadır ve harfler çok benzer şekilde dudaktan çıkmaktadır.



Şekil 6.17. Görüntülerin (Şekil 6.16) farkı.

Ayrıca "meyve"-"perde" çiftinin "perde"-"banka" çiftine göre daha fazla karışmış olmasının sebebi de sahip oldukları ünlü harfler. "Meyve" ve "perde" kelimelerinin

ikisi de aynı ünlü harfler ve benzer dudak hecelerine sahip olduğu için görece daha fazla karışmıştır.

Dudaklardan benzer şekilde çıkan tek grup elbette çift dudak ünsüzleri değil. Aslında her gruptaki harfler benzer şekilde çıkar. Zaten bu şekilde gruplanmasının sebebi dudakların ve dilin aldığı pozisyon olduğu için, benzer şekilde çıkan harfler aynı gruba alınmıştır. Bu yüzden o, ö, u, ü harfleri de benzer şekilde çıkar. Çünkü bunlar "yuvarlak ünlüler" olarak gruplandırılmıştır. Yuvarlak denmesinin sebebi de dudağın bu harflerin söylendiği esnada yuvarlak hale gelmesidir [236]. Bu yüzden en çok karıştırılan kelimeler listesinde "uçurtma" ve "dondurma" kelimeleri yer almaktadır. Kelimeler incelendiğinde her ikisinin de ilk 2 ünlü harfi yuvarlak dudak ünlüsü ve her ikisi de "-ma" hecesiyle bitmiştir.

Literatürdeki pek çok çalışma uzun kelimelerin dudak okuma sistemleri tarafından daha iyi sınıflandırıldığını ve tanındığını belirtmektedir [40,188,237]. Tez kapsamında elde edilen buna dair sonuçlar Bölüm 6.4'te detaylı bir şekilde değerlendirilmiş olup elde edilen bulgular bu ifadeyi doğrular niteliktedir. Kısacası uzun kelimelerin sınıflandırılma oranı, kısa kelimelere oranla daha fazladır. "Meyve"-"perde" ve "perde"-"banka" kelime çiftlerinin karışma sebeplerinden biri de bu kelimelerin ortalamanın altında harf sayılarına sahip olmalarıdır.

Kelimelerin benzerlikleri de kelimelerin karışmasına sebep olmaktadır. Çizelge 5.6'daki kelime çiftlerinin %74'ü ortalama benzerlik uzaklığının altındadır. Kelimelerin çoğunun ortalamadan daha fazla birbirine benziyor olduğu anlamına gelmektedir. Bu da kelimelerin sınıflandırılmasını zorlaştıran en önemli unsurlardandır.

Son olarak Bölüm 6.5'te belirtildiği üzere erkekler için dudağı kapatacak kadar sakal ve bıyık dudak okuma performansını oldukça düşürür. Fakat yine çalışmalardan elde edilen bilgilerden yola çıkarak veri setinin yeterince genişletildiği ve çeşitlendirildiği bir durumda performans düşünün minimuma indirilmesi söz konusu olabilir.

6.7. CNN MODELLERİNİN PERFORMANSI

Hem cümle hem de kelime veri setinde ResNet-18 modelinin diğer modellere göre daha yüksek sınıflandırma başarısı elde ettiği görülmüştür. Çizelge 5.7'e göre kelime veri setinde ResNet-18'in eğitimi GoogleNet'ten 1 saat 15 dakika daha uzun sürmesine rağmen ResNet-18 yaklaşık %15 daha iyi sonuç vermiştir. ResNet-18 modelindeki eğitim süresi GoogleNet modeline göre %26 artmıştır. ResNet-18 ve GoogleNet modellerinin eğitim süreleri arasında oluşan bu kadar büyük farklılığın sebebinin ResNet-18'in sahip olduğu 11.5M parametrenin ve model derinliğinin etkisi olduğu düşünülmektedir. Bu sayı, GoogleNet'in parametre sayısından %40 daha fazladır.

Cümle veri seti için de benzer sonuçlar görülmektedir. ResNet-18'in eğitim süresi GoogleNet'ten %7 daha uzundur. ResNet-18 ve GoogleNet arasındaki sınıflandırmadaki başarı oranı farkı %16'dır. İki modelin kıyaslanması sırasında eğitim süresindeki bu farkın veri setinden ziyade modelin yapısından, karmaşıklığından ve hesaplama maliyetlerinden kaynaklandığını göstermektedir. Çünkü her ikisinde de benzer sayıda videolar kullanılmıştır. Her ne kadar videodan gelecek olan kare sayısı daha fazla olsa da bu kadar dramatik bir farkın oluşmasına sebep olmayacaktır.

Veri setinin boyutu nedeniyle her yönüyle işlenmesi zordur ve tüm bu süreçler oldukça çok zaman almaktadır. Bu nedenle özellikle yöntemlerin eğitim aşamasında veri boyutuna bağlı olarak çalışma sürelerinde ciddi artışlar gözlemlenmiştir. ResNet-18 modelinin 40 kelimelik ve 9 konuşmacıyla oluşturulan veri setinin eğitim süresi, 111 kelime ve 24 konuşmacıyla yapılan eğitimde 64 dakikadan, 20 katına yükselerek 1315 dakikaya çıkmıştır. Bunlara ek olarak, çalışma zamanı grafikleri, her iki veri setinde yer alan verilerin büyüklüğüyle orantılı olarak benzer sonuçlar vermiştir. Bu sonuçlar bize yöntemlerin çalışma süresinin çoğunlukla veri setinin büyüklüğüne bağlı olduğunu göstermiştir.

Test işlemlerinde önceden eğitilmiş bir CNN modeli tercih edilmiştir. Bunun nedeni, veri kümesinin durumunu kontrol etmek ve diğer model ve veri setlerinin başarı performansını karşılaştırmalı analiz etmektir.

Çalışma içerisinde birçok yerde değinildiği üzere sistemlerin performansını karşılaştırmak pek çok parametreye tabi olduğu için direkt bir karşılaştırma yapılamamaktadır. Bunun en temel sebebi de kullanılan veri seti tarafımızca oluşturulmuş olup ilktir. Türkçede daha önce yapılmış bir veri seti olmadığı için modelin diğer Türkçe veri setleriyle performansı analiz edilememektedir. Fakat diğer dillerde yapılmış çalışmalar Çizelge 3.1 ve 3.2'de başarı oranları açısından incelendiğinde %70 ve %92 arasında sonuçlar elde edilmiştir. Tez kapsamında elde edilen başarı oranları da kelime ve cümle veri seti için sırasıyla 0.8409 ve 0.8855 olduğu düşünülecek olursa performans değerleri gayet yerinde ve olumludur.

BÖLÜM 7

GELECEK ÇALIŞMALAR ve TAVSİYELER

Her ne kadar tez kapsamında Türkçe dudak okumaya yönelik detaylı çalışmalar yapılmış olsa da karşılaşılan zorluklara yönelik yeni değerli çalışmalar yapılabilir. İlk olarak harf gruplarını sınıflandırmada yaşanılan zorlukları çözebilecek bir yöntem geliştirilebilir. Örneğin 'p', 'b' ve 'm' harflerinin kendi aralarındaki sınıflandırmada yaşanılan zorluk bu sayede giderilir. Bunun için sadece dudak okuma yapmak yerine olası kelime adayları belirlendikten sonra doğal dil işleme yöntemlerini kullanarak bu olası kelime adaylarından uygun kelime seçilebilir.

Diğer bir çalışma da harf sayısı az olan kelimelerin tahmini konusunda yaşanan kısmi performans düşüklüğü konusunda yapılabilir. Bu sayede özellikle Türkçede "ve", "veya", "ki" gibi sık kullanılan fakat harf sayısı az olan edat veya bağlaçların tahminine ciddi bir katkı olacaktır.

Literatürdeki dudak okuma çalışmaları incelendiğinde CNN'e ciddi bir bağımlık söz konusudur. Fakat dudak okuma gibi problemlerde dudakta çok ufak değişikliklerin bile anlamı olduğu için CNN modelleri bu tarz konularda eksik kalmaktadır. CNN modellerinden ziyade mediapipe'ın sunduğu landmarklar veya keypoint algoritmaları üzerinden dudaktaki noktaların takibi ve analizi yapılabilir. Bu sayede dudaktaki ufak hareketlenmeler bile takip edilip değerlendirilir.

Son olarak Türkçe dudak okumaya yönelik daha fazla kelimenin ve konuşmacının yer aldığı yeni veri setlerinin oluşturulması, bu tarz uygulamaların gerçek hayatta da kullanılabilmesini sağlar. Veri seti konusunda 3 şeye dikkat edilmesi önerilir. (1) Konuşmacı sayısı yüksek olmalı ve mümkünse konuşmacılar farklı yüz, dudak profillerine sahip olmalı. (2) Çoklu görünümlü yani değişik açılardan alınmış veri seti literatüre de ciddi katkı sağlayacaktır. (3) Gerekirse telaffuz sayısı daha az olabilir fakat kelime havuzunun çok daha geniş olması modellerin performansını ölçmeyi daha gerçekçi kılacaktır.

BÖLÜM 8

SONUÇLAR

Bu tez çalışması kapsamında, otomatik Türkçe dudak okuma için derin öğrenme tabanlı bir model geliştirilmiştir. Ayrıca literatürde olmayan Türkçe dudak okuma veri seti de bilimsel araştırmalara açık halde paylaşılmıştır. Türkçe dudak okuma için literatüre katkı yapmanın yanında veri setinin diğer dillerde oluşturulmuş veri setleri baz alınarak kıyaslamaları yapılmıştır. Geçmişte oluşturulan veri setlerinin eksiklikleri tez kapsamında oluşturulan veri setiyle giderilmeye çalışılmıştır. Sistem performansı, tezin oluşturulduğu yıl itibariyle Türkçe dilinde yapılmış ilk çalışma olduğu için performans açısından değerlendirilebilecek başka bir çalışma yoktur. Fakat diğer dillerdeki çalışmaların performans metrikleri detaylı verilmediği için sadece WRR değerine göre bakıldığında oldukça yeterli sonuçlar üretmiştir.

Türkçe dudak okumaya yönelik detaylı analizler sayesinde yaşanılan zorluklar özellikle Bölüm 6'da ele alınmıştır. Ayrıca sistemin performansına etki eden parametreler somut veriler sunularak sıralanmıştır. Bu sayede Türkçe dudak okuma için yapılmış bu tez çalışması ilerde yapılacak çalışmalarda araştırmacılara yol gösterecektir. Karşılaşacakları olası zorlukları daha kolay aşabileceklerdir.

Dudak Okuma çalışmalarında ulaşılmak istenen nokta, konuşmaları gerçek zamanlı olarak en iyi şekilde tanımlamaktır. Bir cümlede kelime seviyesinde dudak okuma modelini kullanmak için tüm kelimelerin ayrı ayrı bölünmesi gerekir. Çoğu kelime bir bakıma birlikte söylenmekte ve dudaklar bir kelimeden diğerine çok hızlı geçmektedir. Hatta bazen dudak hiç hareket etmeden dil hareketleriyle kelimeler söylendiği için bu görevin tek başına görüntü işleme teknikleriyle yapılabilmesi şimdiki teknolojiyle oldukça zordur [37].

Gelecek çalışmalarda, veri setinin genişletilerek önce bu tez kapsamında oluşturulan modelin performansı tekrar test edilebilir. Daha sonra derin öğrenme modellerinden örneğin kapsül ağlar gibi daha farklı ağlar veya landmarklar üzerinden bir model geliştirilebilir. Aynı zamanda bunun gerçek zamanlı bir şekilde yapılmasını sağlayacak masaüstü veya mobil uygulama insanların hizmetine sunulabilir. Gerçek zamanlı dudak okuma sistemleri için hala ciddi performans gösteren bir çalışma bulunmamaktadır. Gerçek zamanlı çalışan bir yaklaşım veya model önerisi getirilebilir.

KAYNAKLAR

1. McGurk H., and MacDonald J., "Hearing lips and seeing voices", ***Nature***, 264: 746-748 (1976).

2. Potamianos, G., Neti, C., Luettin, J., and Matthews, I., "Audio-Visual Automatic Speech Recognition: An Overview", ***Issues in Audio-Visual Speech Processing,*** (2004).

3. Potamianos, G., Neti, C., and Gravier, G., "Recent advances in the automatic recognition of audiovisual speech", ***Proc IEEE,*** 91: 1306-1325 (2003).

4. Skipper, J. I., Wassenhove, V., and Nusbaum, H. C., "Hearing Lips and Seeing Voices: How Cortical Areas Supporting Speech Production Mediate Audiovisual Speech Perception", ***Cerebral Cortext,*** New York, 17: 2387–2399 (2007).

5. Erber, NP., "Auditory-Visual Perception of Speech", ***Hearing Aid Jornal,*** 32: 32-33 (1979).

6. Submy, W. H., and Pollack, I., "Visual Contribution to Speech Intelligibility in Noise", ***EURASIP Journal on Audio, Speech, and Music Processing,*** 2007: (2006).

7. Hilder, S., Harvey, R., and Theobald, B. J., " Comparison of human and machine-based lip-reading", ***Proceedings of the International Conference on Auditory-Visual Speech Processing,*** 86-89 (2009).

8. Ronquest, R. E., Levi, S. V., and Pisoni, D. B., Language identification from visual-only speech signals", ***Attention, Perception & Psychophysics,*** 72: 1601-1613 (2010).

9. Seymour, R., Stewart, D., and Ming, J., "Comparison of Image Transform-Based Features for Visual Speech Recognition in Clean and Corrupted Videos", ***EURASIP J Image Video Process,*** 1-9 (2008).

10. Antonakos, E., Roussos, A., and Zafeiriou, S., "A survey on mouth modeling and analysis for Sign Language recognition", ***11th IEEE Int Conf Work Autom Face Gesture Recognition,*** Ljubljana, (2015).

11. Bowden, R., Cox, S., and Harvey, R., " Recent developments in automated lip-reading. Opt Photonics Counterterrorism", ***Crime Fight Def IX*** (2013).

12. McQuillan, L., "Is lip-reading the secret to security", ***Biometric Technol Today***, 5-7, (2019).

13. Lesani, F. S., Ghazvini, F. F., and Dianat, R., "Mobile phone security using automatic lip reading", ***9th Int Conf e-Commerce Dev Ctries With Focus e-Business***, (2015).

14. Hassanat, A., "Automatic lip reading for security", ***1st Mosharaka International Conference on Biomedical Engineering, Electronics and Nanotechnology,*** 11-16, (2011).

15. Ephrat, A., and Peleg, S., "Vid2speech: Speech reconstruction from silent video", ***ICASSP, IEEE Int Conf Acoust Speech Signal Process,*** 5095–5099 (2017).

16. Gabbay, A., Shamir, A., and Peleg, S., "Visual Speech Enhancement", ***Proc Annu Conf Int Speech Commun Assoc INTERSPEECH,*** 1170-1174 (2018).

17. Chuang, S. Y., Wang, H. M., and Tsao, Y., "Improved Lite Audio-Visual Speech Enhancement", ***IEEE/ACM Transactions on Audio, Speech, and Language Processing,*** 30: 1345-1359 (2020).

18. Werda, S., Mahdi, W., and Hamadou, A. B., "Lip Localization and Viseme Classification for Visual Speech Recognition", ***International Journal of Computing & Information Sciences,*** 5 (1): 62-75 (2013).

19. Teferi, D., and Bigun, J., "Damascening video databases for evaluation of face tracking and recognition - The DXM2VTS database", ***Pattern Recognition Letters ,*** 28 (15): 2143-2156 (2007).

20. Shin. J., Lee J., and Kim. D., "Real-time lip reading system for isolated Korean word recognition", ***Pattern Recognition,*** 44 (3): 559-571 (2011).

21. Dupont, S., and Luettin, J., "Audio-visual speech modeling for continuous speech recognition", ***IEEE Transactions on Multimedia,*** 2 (3): 141-151 (2000).

22. Nefian, A. V., Liang, L., and Pi, X., "Dynamic Bayesian Networks for Audio-Visual Speech Recognition", ***International Conference on Biometrics,*** 3832: 539-545 (2002).

23. Lin, B. S., and Yaho, Y. H., “Development of novel lip-reading recognition algorithm”, ***IEEE Access,*** 5: 794-801 (2017).

24. Morade, S. S., and Patnaik, S., “A novel lip reading algorithm by using localized ACM and HMM: Tested for digit recognition”, ***Optics,*** 125 (18): 5181-5186 (2014).

25. Mustafa, R., and Zhu, D., “A novel lip-reading method using RGB-D camera”, ***Image Analysis and Recognition,*** 8815: 21-28 (2014).

26. Jain, A., and Rathna, G. N., “Lip reading using simple dynamic features and a novel ROI for feature extraction”, ***ACM Int Conf Proceeding,*** 73-77 (2018).

27. Bear, H. L., Harvey, R. W., Theobald, B. J., and Lan, Y., “Which Phoneme-to-Viseme Maps Best Improve Visual-Only Computer Lip-Reading?”, ***Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics),*** 8888: 230–239 (2014).

28. Glotin, H., Vergyr,, D., Neti, C., Potamianos, G., and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," ***2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221),*** 1: 173-176 (2001).

29. Ortiz, I. R., “Lipreading in the Prelingually Deaf: What makes a Skilled Speechreader?”, ***The Spanish journal of psychology,*** 11 (2): 488-502, (2008).

30. Yargic, A., Dogan, M, “A lip reading application on MS Kinect camera”, ***2013 IEEE Int Symp Innov Intell Syst Appl IEEE INISTA,*** 1-5 (2013).

31. Sarhan, A. M., Elshennawy, N. M., and Ibrahim, D. M., “HLR-Net: A hybrid lip-reading model based on deep convolutional neural networks”, ***Computers, Materials and Continua,*** 68(2): 1531-1549 (2021).

32. Stafylakis, T., and Tzimiropoulos, G., “Combining Residual Networks with LSTMs for Lipreading”, ***Proc Annu Conf Int Speech Commun Assoc INTERSPEECH,*** 3652-3656 (2017).

33. Sterpu, G., and Naomi, H., “Towards Lipreading Sentences with Active Appearance Models”, ***The 14th International Conference on Auditory-Visual Speech Processing,*** Stockholm, 70-75 (2017).

34. Thangthai, K., and Harvey, R., “Improving computer lipreading via DNN sequence discriminative training techniques”, ***INTERSPEECH 2017,*** Stockholm, 3657-3661 (2017).

35. Petridis, S., Wang, Y., Li, Z., and Pantic, M., “End-to-End Audiovisual Fusion with LSTMs”, ***The 14th International Conference on Auditory-Visual Speech Processing,*** Stockholm, 36-40 (2017).

36. Thangthai, K., Harvey, R. W., Cox, S. J., and Theobald, B. J., “Improving lip-reading performance for robust audiovisual speech recognition using DNNs”, ***AVSP,*** Vienna, 127–131 (2015).

37. Huyen, C.T., “German Word Level Lip Reading with Deep Learning. Doctoral dissertation”, ***Hochschule für angewandte Wissenschaften Hamburg,*** (2019).

38. Chen, X., Du, J., and Zhang, H., “Lipreading with DenseNet and resBi-LSTM. Signal”, ***Image Video Process,*** 14: 981-989 (2020).

39. Kurniawan, A., and Suyanto, S., “Syllable-Based Indonesian Lip Reading Model”, ***8th Int Conf Inf Commun Technol ICoICT,*** 1-6 (2020).

40. Fernandez-Lopez, A., and Sukno, F. M., “Survey on automatic lip-reading in the era of deep learning”, ***Image and Vision Computing,*** 78: 53-72 (2018).

41. Obukhov, A., “Haar Classifiers for Object Detection with CUDA”, GPU ***Computing Gems Emerald Edition,*** 517-544 (2011).

42. Grishchenko, I., Ablavatski, A., Kartynnik, Y., Raveendran, K., and Grundmann, M., “Attention Mesh: High-fidelity Face Mesh Prediction in Real-time”, ***CVPR Workshop on Computer Vision for Augmented and Virtual Reality,*** (2020).

43. Lecun, Y., Bottou, L., and Haffner, P., “Gradient-Based Learning Applied to Document Recognition”, ***Proceedings of the IEEE,*** 86 (11): 2278–2324 (1998).

44. LeCun, Y., Bengio, Y., and Hinton, G., “Deep learning”, ***Nature,*** 521: 436–444 (2015).

45. Maas, A., and Ng, A., “Rectifier Nonlinearities Improve Neural Network Acoustic Models”, ***ICML Workshop on Deep Learning for Audio, Speech, and Language Processing,*** 28 (3): (2013).

46. Hochreiter, S., and Schmidhuber, J., “Long Short-Term Memory”, ***Neural Computation,*** 9 (8): 1735–1780 (1997).

47. Yu, Y., Hu, C., Si, X., Zheng, J., and Zhang, J., "Averaged Bi-LSTM networks for RUL prognostics with non-life-cycle labeled dataset", ***Neurocomputing,*** 402: 134-147 (2020).

48. Li, S., Yang, S., and Liang, J., "Recognition of ships based on vector sensor and bidirectional long short-term memory networks", ***Applied Acoustics,*** 164 (8): 107248 (2020).

49. Zhao, Y., Xu, R., Song, M., "A cascade sequence-to-sequence model for Chinese Mandarin lip reading", ***Association for Computing Machinery,*** New York, 32: 1-6 (2019).

50. Petridis, S., Wang, Y., Ma, P., Li, Z., and Pantic, M., "End-to-end visual speech recognition for small-scale datasets", ***Pattern Recognition Letters,*** 131: 412-427 (2020).

51. Rahmani, M. H., and Almasganj, F., "Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features", ***3rd International Conference on Pattern Recognition and Image Analysis,*** Shahrekord, 195-199 (2017).

52. Xu, K., Li, D., Cassimatis, and N., Wang, X., LCANet: End-to-end lipreading with cascaded attention-CTC, ***13th IEEE Int Conf Autom Face Gesture Recognition,*** Xi'an, 548-555 (2018).

53. Wand, M., Schmidhuber, J., and Vu, N. T., Investigations on End- to-End Audiovisual Fusion, ***IEEE International Conference on Acoustics, Speech and Signal Processing,*** 3041 - 3045 (2018).

54. Petridis, S., Stafylakis, T., and Ma, P., "End-to-End Audiovisual Speech Recognition", ***IEEE International Conference on Acoustics, Speech, and Signal Processing,*** Calgary 6548-6552 (2018).

55. Afouras, T., Chung, J. S., and Zisserman, A., "Deep Lip Reading: a comparison of models and an online application", ***Interspeech,*** 3514–3518 (2018).

56. Saitoh, T., Zhou, Z., Zhao, G., and Pietikäinen, M., "Concatenated Frame Image Based CNN for Visual Speech Recognition", ***Computer Vision – ACCV 2016 Workshops***, 277–289 (2016).

57. Chung, J. S., and Zisserman, A., "Lip Reading in the Wild", ***Asian Conference on Computer Vision,*** 87-103 (2017).

58. Zimmermann, M., Mehdipour Ghazi, M., Ekenel, H. K., and Thiran, J. P., "Visual Speech Recognition Using PCA Networks and LSTMs in a Tandem GMM-HMM System", ***Asian Conference on Computer Vision,*** Taiwan, 264-276 (2016).

59. Lee, D., Lee, J., and Kim, K. E., "Multi-view Automatic Lip-Reading Using Neural Network" ***Asian Conference on Computer Vision,*** Taipei, 290–302 (2016).

60. Koumparoulis, A., and Potamianos, G., "Deep View2View Mapping for View-Invariant Lipreading", ***2018 IEEE Spoken Language Technology Workshop (SLT),*** Athens, 588-594 (2019).

61. Han, H., Kang, S., and Yoo, C. D. "Multi-view visual speech recognition based on multi task learning" ***IEEE International Conference on Image Processing (ICIP),*** Beijing, 3983–3987 (2018).

62. Bakry, A., and Elgammal, A., "MKPLS: Manifold kernel partial least squares for lipreading and speaker identification", ***IEEE Conference on Computer Vision and Pattern Recognition,*** Portland, 684–691 (2013).

63. Puviarasan, N., and Palavinel S., "Lip reading of hearing impaired persons using HMM" ***Expert Systems with Applications,*** 38 (4): 4477-4481 (2011).

64. Hua, M., Liu, L., Chen, Z., He, Q., Li, B., and Yi Z., "FaceEraser: Removing Facial Parts for Augmented Reality" ***Workshop on Computer Vision for AR/VR,*** (2021).

65. Kelley, J. L., Chapuis, L., Davies, W. I. L., and Collin S. P., "Sensory System Responses to Human-Induced Environmental Change", ***Frontiers in Ecology and Evolution,*** 6 (95): (2018).

66. Adamczyk, K., Górecka Bruzda, A., Nowicki, J., Gumułka, M., Edyta, M., Schwarz, T., Earley, B., and Czesław, K., "Perception of environment in farm animals – A review", ***Annals of Animal Science,*** 15(3):565-589 (2015).

67. Lei Z., and Yi, D., "Lip movements enhance speech representations and effective connectivity in auditory dorsal stream", ***NeuroImage,*** 257: 119311 (2022).

68. Sebastian, P., Mareike, D., Maren, S., Bojana, M., Stephanie, R., Christiane, M. T., and Stefan, D., "Hearing-impaired listeners show increased audiovisual benefit when listening to speech in noise", NeuroImage, 196: 261-268 (2019).

69. Eun, J. P., Laura, L. M., and Sharlene, D. N., "Effects of concurrent action and object naming treatment on naming skills and functional brain activation patterns

in primary progressive aphasia: An fMRI study with a case-series design", ***Brain and Language,*** 218: 104950 (2021).

70. Rene, L. U., Hugo, B., Peter, R. M., Christopher, G. S., Joseph, R. D., Heather, M. C., Mary, M. M., Alissa, M. B., Val, J. L., Clifford, R. J., Matthew, L. S., Anthony, J. S., Jennifer, L. W., Keith, and A. J., Clinical and neuroimaging characteristics of clinically unclassifiable primary progressive aphasia, ***Brain and Language,*** 197: 104676 (2019).

71. Shindo, M., Kaga, K., and Tanaka, Y., "Speech discrimination and lip reading in patients with word deafness or auditory agnosia", ***Brain and Language,*** 40 (2): 153-161 (1991).

72. Eleanor, M. S., "Wernicke's Area", ***Encyclopedia of the Human Brain,*** 805-818 (2002).

73. David, B., and Donna, B., "Neuroscience: Music and the Brain", ***Encyclopedia of Creativity (Third Edition)***, 225-232 (2020).

74. Andrew, J. K., and David, R. M., "Plasticity of auditory maps in the brain", ***Trends in Neurosciences,*** 14 (1): 31-37 (1991).

75. Aleksi, J. S., Noelia, M. M., and Teppo, S., "Music Perception and Amusia", ***Encyclopedia of Behavioral Neuroscience, 2nd edition (Second Edition),*** 678-685 (2022).

76. Bourguignon, M., Baart, M., Kapnoula, E. C., and Molinaro, N., "Lip-Reading Enables the Brain to Synthesize Auditory Features of Unknown Silent Speech", ***The Journal of Neuroscience : The Official Journal of the Society for Neuroscience,*** 40 (5): 1053-1065 (2020).

77. Jérôme, T., "Temporal Resolution", ***Encyclopedia of GIS. Springer,*** Boston, 1404–1411 (2008).

78. Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., and Harvey, R., "Extraction of visual features for lipreading", IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (2): 198-213 (2002).

79. Cox, C., and Harvey, R., "The Challenge of multispeaker lip-reading", ***International Conference on Auditory-Visual Speech Processing,*** Queensland, 179-184 (2008).

80. Lee, B., Hasegawa Johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., and Huang, T., "AVICAR: Audio-visual speech corpus in a car environment", ***INTERSPEECH,*** 2489-2492 (2004).

81. Messer, K., Matas, J., Kittler, J., Luettin, J., and Maître, G., " XM2VTSDB The Extended M2VTS Database", ***AVBPA'99,*** 72-77 (1999).

82. Fox, N. A., OMullane, B. A., and Reilly, R. B., "VALID: A new practical audio-visual database, and comparative results", ***International Conference on Audio- and Video-Based Biometric Person Authentication,*** Berlin, 777-786 (2005).

83. Bailly-Bailliere, E., S. Bengio, Bimbot, F., Hamouz, M., Kittler, J., Mariethoz, J., Matas, J., Messer, K., Popovici, V., and Poree, F., "The BANCA database and evaluation protocol", ***Audio- and Video-Based Biometric Person Authentication,*** 625-638 (2003).

84. Patterson, E. K., Gurbuz, S., Tufekci, Z., and Gowdy, J. N., "CUAVE: A new audio-visual database for multimodal human-computer interface research", ***International Conference on Acoustics, Speech, and Signal Processing,*** 2: 2017-2020 (2002).

85. Goecke, R., and Millar, J. B., "The audio-video australian english speech data corpus AVOZES", ***International Conference on Spoken Language Processing,*** Jeju Island, 2525-2528 (2004).

86. Estival, D., Cassidy, S., Cox, F., and Burnham, D., "AusTalk: an audiovisual corpus of Australian English", ***International Conference on Language Resources and Evaluation,*** Reykjavik, 3105-3109 (2014).

87. Igras, M., Ziołko, B., Jadczyk, T., "Audiovisual database of Polish speech recordings", ***Studia Informatica,*** 33 (2B): 163-172 (2012).

88. Tamura, S., Miyajima, C., Kitaoka, N., Yamada, T., Tsuge, S., Takiguchi, T., Yamamoto, K., Nishiura, T., Nakayama, M., and Denda, Y., "CENSREC-1-AV An audio-visual corpus for noisy bimodal speech recognition", AVSP, Hakone, 181-187 (2010).

89. Ortega, A., Sukno, F., Lleida, E., Frangi, A. F., Miguel, A., Buera, L., and Zacur, E., "AV@CAR A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition", ***Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04),*** Lisbon, 763-766 (2004).

90. Huang, J., Potamianos, G., Connell, J., and Neti, C., "Audio-visual speech recognition using an infrared headset", ***Speech Communication,*** 44 (1): 83-96 (2004).

91. Lucey, P. J., Potamianos, G., Sridharan, S., "Patch-based analysis of visual speech from multiple views", ***International Conference on Auditory-Visual Speech Processing,*** 69-74 (2008).

92. Anina, I., Zhou, Z., Zhao, G., and Pietikainen, M., "OuluVS2 A multiview audiovisual database for non-rigid mouth motion analysis", ***International Conference on Automatic Face and Gesture Recognition***, Ljubljana, 1-5 (2015).

93. Petridis, S., Shen, J., Cetin, D., and Pantic, M., "Visual-only recognition of normal, whispered and silent speech", ***International Conference on Acoustics, Speech and Signal Processing,*** 6219-6223 (2018).

94. Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., and Mashari, A., "Audio visual speech recognition", ***Audio-Visual Speech Recognition Final Workshop 200 Report,*** 11 (1): 1274-1288 (2000).

95. Sanderson, C., "The VidTIMIT database", ***A motion based approach for audio-visual automatic speech recognition***, 89-95 (2002).

96. Hazen, T. J., Saenko, K., La, C. H., and Glass, J. R., "A segment-based audio-visual speech recognizer data collection, development, and initial experiments", ***International Conference on Multimodal Interfaces,*** Pennsylvania, 235-242 (2004).

97. Petrovska-Delacretaz, D., Lelandais, S., Colineau, J., Chen, L., Dorizzi, B., Ardabilian, M., Krichen, E., Mellakh, M. A., Chaari, A., and Guerfi, S., "The IV2 Multimodal Biometric Database (Including Iris, 2D, 3D, Stereoscopic, and Talking Face Data), and the IV2-2007 Evaluation Campaign", ***IEEE Second International Conference on Biometrics: Theory, Applications and Systems***, Washington, 1-7 (2008).

98. Trojanova, J., Hruz, M., Campr, P., and Zelezny, M., "Design and recording of Czech audio-visual database with impaired conditions for continuous speech recognition", ***International Conference on Language Resources and Evaluation,*** Marrakech, 1239-1243 (2008).

99. Zhao, G., Barnard, M., and Pietikainen, M., "Lipreading with local spatiotemporal descriptors", IEEE Transactions on Multimedia, 11 (7): 1254-1265 (2009).

100. Lan, Y., Theobald, B. J., Harvey, R., Ong, E. J., and Bowden, R., "Improving visual features for lip-reading", ***International Conference on Auditory-Visual Speech Processing,*** Kanagawa, 142-147 (2010).

101. Rekik, A., Ben-Hamadou, A., and Mahdi, W., "A new visual speech recognition approach for RGB-D cameras", **International Conference on Image Analysis and Recognition**, Portugal, 21–28 (2014).

102. Wong, Y. W., Chng, S. I., Seng, K. P., Ang, L. M., Chin, S. W., Chew, W. J., and Lim, K. H., "A new multi-purpose audiovisual UNMC-VIER database with multiple variabilities", ***Pattern Recognition Letters,*** 32 (13): 1503–1510 (2011).

103. Benezeth, Y., Bachman, G., Le-Jan, G., Souviraa-Labastie, N., and Bimbot, F., "BL-database A French audiovisual database for speech driven lip animation systems Ph.D. thesis", ***INRIA,*** (2011).

104. Fernandez-Lopez, A., Martinez, O., and Sukno, F. M., "Towards estimating the upper bound of visual-speech recognition The visual lip-reading feasibility database", ***International Conference on Automatic Face and Gesture Recognition,*** Washington, 208–215 (2017).

105. Howell, D. L., Confusion modelling for lip-reading, Ph. D. Thesis, ***University of East Anglia,*** (2015).

106. Harte, N., and Gillen, E., "TCD-TIMIT An audio-visual corpus of continuous speech", ***Transactions on Multimedia,*** 17 (5): 603-615 (2015).

107. Verkhodanova, V., Ronzhin, A., Kipyatkova, I., Ivanko, D., Karpov, A., and Zelezny, M., "HAVRUS corpus high-speed recordings of audio-visual Russian speech", ***International Conference on Speech and Computer,*** Budapest, 338–345 (2016).

108. Mroueh, Y., Marcheret, E., and Goel, V., "Deep multimodal learning for audio-visual speech recognition", ***International Conference on Acoustics, Speech and Signal Processing,*** Queensland, 2130–2134 (2015).

109. Cooke, M., Barker, J., Cunningham, S., and Shao, X., "An audio-visual corpus for speech perception and automatic speech recognition", ***Journal of the Acoustical Society of America,*** 120 (5): 2421-2424 (2006).

110. Vorwerk, A., Wang X., Kolossa, D., Zeiler, S., and Orglmeister, R., "WAPUSK20 - A database for robust audiovisual speech recognition", ***International Conference on Language Resources and Evaluation,*** Valletta, 3016-3019 (2010).

111. Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J., and Szykulski, M., "An audio-visual corpus for multimodal automatic speech recognition", ***Journal of Intelligent Information Systems,*** 1-26 (2017).

112. Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A., "Lip reading sentences in the wild", ***Conference on Computer Vision and Pattern Recognition,*** Honolulu, 3444-3453 (2017).

113. Lan, Y., Theobald, B. J., and Harvey, R., "View independent computer lip-reading", ***International Conference on Multimedia and Expo,*** Melbourne, 432-437 (2012).

114. Kumar, K., Chen, T., and Stern, R. M., "Profile view lip reading", International Conference on Acoustics, ***Speech and Signal Processing,*** Honolulu, 429-432 (2007).

115. Almajai, I., Cox, S., Harvey, R., and Lan, Y., "Improved speaker independent lip reading using speaker adaptive training and deep neural networks", ***International Conference on Acoustics, Speech and Signal Processing,*** Shanghai, 2722–2726 (2016).

116. Pass, A., Zhang, J., and Stewart, D., "An investigation into features for multi-view lipreading", ***International Conference on Image Processing,*** Hong Kong, 2417-2420 (2010).

117. Estellers, V., and Thiran, J. P., "Multipose audio-visual speech recognition", ***European Conference on Signal Processing,*** Barcelona, 1065-1069 (2011).

118. Yao, X. L. H., and Wang, X. H. Q., "HIT-AVDB-II A new multi-view and extreme feature cases contained audio-visual database for biometrics", ***Proceedings of the 11th Joint Conference on Information Sciences,*** Vienna, 357-363 (2008).

119. Sahu, V., and Sharma, M., "Result based analysis of various lip tracking systems", ***International Conference on Green High Performance Computing,*** Nagercoil, 1–7 (2013).

120. Cappelletta, L., and Harte, N., "Viseme definitions comparison for visual-only speech recognition", ***European Conference on Signal Processing,*** Barcelona, 2109-2113 (2011).

121. Luettin, J., Thacker, N. A., and Beet, S. W., "Visual speech recognition using active shape models and hidden markov models", ***International Conference on Acoustics, Speech, and 1935 Signal Processing***, 817-820 (1996).

122. Gowdy, J. N., Subramanya, A., Bartels, C. and Bilmes, J., "DBN based multi-stream models for audio-visual speech recognition", ***IEEE International Conference on Acoustics, Speech, and Signal Processing,*** Montreal, 993-996 (2004).

123. Eveno, N., Caplier, A., and Coulon, P. Y., "Accurate and quasi- automatic lip tracking", ***Circuits and Systems for Video Technology,*** 14 (5): 706-715 (2004).

124. Mase, K., and Pentland, A., "Automatic lipreading by optical-flow analysis", ***Systems and Computers in Japan,*** 22 (6): 67- 76 (1991).

125. Hong, X., Yao, H., Wan, Y., and Chen, R., "A PCA based visual DCT feature extraction method for lip-reading", ***International Conference on Intelligent Information Hiding and Multimedia,*** Pasadena, 321-326 (2006).

126. Lucey, P., and Potamianos, G., "Lipreading using profile versus frontal views", ***International Workshop on Multimedia Signal Processing,*** Victoria, 24-28 (2006).

127. Zhou, Z., Zhao, G., Hong, X., and Pietikainen, M., "A review of recent advances in visual speech decoding", ***Image and Vision Computing,*** 32 (9): 590-605 (2014).

128. Lucey, P. J., Potamianos, G., and Sridharan, S., "A unified approach to multi-pose audio-visual ASR", ***Proceedings of Interspeech,*** Antwerp, 650-653 (2007).

129. Gurban, M., and Thiran, J. P., "Information theoretic feature extraction for audio-visual speech recognition", ***Signal Processing,*** 57 (12): 4765-4776 (2009).

130. Huang, J., and Kingsbury, B., "Audio-visual deep learning for noise robust speech recognition", ***International Conference on Acoustics, Speech and Signal Processing,*** Vancouver, 7596-7599 (2013).

131. Papandreou, G., Katsamanis, A., Pitsikalis, V., and Maragos, P., "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition", ***IEEE-ACM Transactions on Audio, Speech, and Language Processing,*** 17 (3): 423-435 (2009).

132. Papandreou, G., Katsamanis, A., Pitsikalis, V., and Maragos, P., "Adaptive multimodal fusion by uncertainty compensation with application to audio-visual

speech recognition", ***International Conference on Multimodal Processing and Interaction,*** Chania, 1-15 (2008).

133. Wang, S. L., Liew, A. W. C., Lau, W. H., and Leung, S. H., "An automatic lipreading system for spoken digits with limited training data", ***Circuits and Systems for Video Technology,*** 18 (12): 1760-1765 (2008).

134. Lucey, P. J., Sridharan, S., and Dean, D. B., "Continuous pose invariant lipreading", ***Proceedings of Interspeech***, Brisbane, 2679–2682 (2008).

135. Pachoud, S., Gong, S., and Cavallaro, A., "Macro-cuboïd based probabilistic matching for lip-reading digits", ***IEEE Conference on Computer Vision and Pattern Recognition,*** Anchorage, 1-8 (2008).

136. Rekik, A., Ben-Hamadou, A., and Mahdi, W., "An adaptive approach for lip-reading using image and depth data", ***Multimedia Tools and Applications,*** 75 (14): 8609–8636 (2016).

137. Estellers, V., Gurban, M., and Thiran, J. P., "On dynamic stream weighting for audio-visual speech recognition", ***IEEE ACM Transactions on Audio, Speech, and Language Processing***, 20 (4): 1145–1157 (2012).

138. Pei, Y., Kim, T. K., and Zha, H., "Unsupervised random forest manifold alignment for lipreading", ***IEEE International Conference on Computer Vision,*** Sydney, 129–136 (2013).

139. Cox, S. J., Harvey, R., Lan, Y., Newman, J. L., and Theobald, B. J., "The challenge of multispeaker lip-reading", ***International Conference on Auditory-Visual Speech Processing,*** Queensland, 179-184 (2008).

140. Petridis, S., Wang, Y., Ma, P., Li, Z., and Pantic, M., "End-to-end visual speech recognition for small-scale datasets", ***Pattern Recognition Letters,*** 131: 421-427 (2020).

141. Stewart, D., Seymour, R., Pass, A., and Ming, J., "Robust audio-visual speech recognition under noisy audio-video conditions", ***IEEE Transactions on Cybernetics,*** 44 (2): 175–184 (2014).

142. Wand, M., Koutnık, J., and Schmidhuber, J., "Lipreading with long short-term memory", ***International Conference on Acoustics, Speech and Signal Processing,*** Shanghai, 6115– 6119 (2016).

143. Lan, Y., Harvey, R., Theobald, B., Ong, E. J., and Bowden, R., "Comparing visual features for lipreading", ***The International Conference on Auditory-Visual Speech Processing (AVSP),*** Norwich, 102-106 (2009).

144. Kolossa, D., Zeiler, S., Vorwerk, A., and Orglmeister, R., "Audiovisual speech recognition with missing or unreliable data", ***International Conference on Auditory-Visual Speech Processing,*** Norwich, 117–122 (2009).

145. Zhou, Z., Zhao, G., and Pietikainen, M., "Lipreading a graph embedding approach", I***nternational Conference on Pattern 1945 Recognition,*** Istanbul, 523–526 (2010).

146. Zhou, Z., Zhao, G., and Pietikainen, M., "Towards a practical lipreading system", ***Conference on Computer Vision and Pattern Recognition,*** Colorado, 137–144 (2011).

147. Ong, E. J., and Bowden, R., "Learning temporal signatures for lip reading", ***International Conference on Computer Vision Workshops,*** Barcelona, 958–965 (2011).

148. Ong, E. J., and Bowden, R., "Learning sequential patterns for lipreading", ***British Machine Vision Conference,*** Dundee, 1-10 (2011).

149. Zhou, Z., Hong, X., Zhao, G., and Pietikainen M., "A compact representation of visual speech data using latent variables", ***IEEE Transactions on Pattern Analysis and Machine Intelligence,*** 36 (1): 181-187 (2014).

150. Sui, C., Togneri, R., and Bennamoun, M., "A cascade gray-stereo visual feature extraction method for visual and audio-visual speech recognition", ***Speech Communication,*** 90 (C): 26–38 (2017).

151. Bear, H. L., and Harvey, R., "Decoding visemes improving machine 2025 lip-reading", ***International Conference on Acoustics, Speech and Signal Processing,*** Shanghai, 2009–2013 (2016).

152. Chung, J. S., and Zisserman, A., "Out of time automated lip sync in the wild", ***Asian Conference on Computer Vision,*** Taipei, 251–263 (2016).

153. Wu, P., Liu, H., Li, X., Fan, T., and Zhang, X., "A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion", ***IEEE Transactions on Multimedia,*** 18 (3): 326– 338 (2016).

154. Lee, D., Lee, J., and Kim, K. E., Multi-view automatic lip-reading using neural network", ***Asian Conference on Computer Vision,*** Taipei, 290–302 (2016).

155. Howell, D., Cox, S., and Theobald, B., ***"Visual units and confusion modelling for automatic lip-reading",*** Image and Vision Computing, 51 (C): 1–12 (2016).

156. Lan, Y., Harvey, R., and Theobald, B. J., "Insights into machine lip reading", ***International Conference on Acoustics, Speech and Signal Processing,*** Kyoto, 4825- 4828 (2012).

157. Fisher, C. G., "Confusions among visually perceived consonants", ***Journal of Speech, Language, and Hearing Research,*** 11 (4): 796–804 (1968).

158. Fernandez-Lopez, A., and Sukno, F. M., Automatic viseme vocabulary construction to enhance continuous lip-reading, ***International Conference on Computer Vision Theory and Applications***, Porto, 52–63 (2017).

159. Price, P., Fisher, W. M., Bernstein, J., and Pallett, D.S., "The DARPA 1000-word resource management database for continuous speech recognition", ***The International Conference on Acoustics, Speech, and Signal Processing (ICASSP)***, New York, 651–654 (1998).

160. Baker, C., "Foundations of Bilingual Education and Bilingualism", ***Multilingual Matters Ltd,*** (2006).

161. Jacob, N., and Stephen, C.,. "Language Identification Using Visual Features", ***Audio, Speech, and Language Processing, IEEE Transactions,*** 20 (1): 1936-1947 (2012).

162. Fu, Y., Zhou, X., Liu, M., Hasegawa-Johnson, M., and Huang, T. S., "Lipreading by locality discriminant graph", ***International Conference on Image Processing,*** San Antonio, 325– 328 (2007).

163. Marcheret, E., Libal, V., and Potamianos, G., "Dynamic stream weight modeling for audio-visual speech recognition", ***International Conference on Acoustics, Speech and Signal Processing,*** Honolulu, 945–948 (2007).

164. Jeffers, J., and Barley, M., "Speechreading (lipreading)", ***Thomas,*** (1971).

165. Saitoh, T., and Konishi, R., "Profile lip reading for vowel and word recognition", ***International Conference on Pattern Recognition,*** Istanbul, 1356–1359 (2010).

166. Navarathna, R., Kleinschmidt, T., Dean, D. B., Sridharan, S., and Lucey, P. J., "Can audio-visual speech recognition outperform acoustically enhanced speech recognition in automotive environment", ***Interspeech,*** Florence, 2241– 2244 (2011).

167. Ngiam J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y., "Multimodal deep learning", ***International Conference on Machine Learning,*** Bellevue, 689-696 (2011).

168. Chitu, A. G., and Rothkrantz, L. J., "Automatic visual speech recognition", ***Speech Enhancement, Modeling and Recognition- Algorithms and Applications,*** 95-120 (2012).

169. Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., and Ogata, T., "Lipreading using convolutional neural network", ***Interspeech,*** Singapore, 1149–1153 (2014).

170. Bear, H. L., Cox, S. J., and Harvey, R. W., "Speaker-independent machine lip-reading with speaker-dependent viseme classifiers", ***Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing,*** Vienna, 190-195 (2015).

171. Bear, H. L., Harvey, R. W., Lan, Y., "Finding phonemes improving machine lip-reading", ***Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing,*** Vienna, 115-120 (2017).

172. Biswas, A., Sahu, P. K., and Chandra, M., "Multiple camera in car audio-visual speech recognition using phonetic and visemic information", ***Computers & Electrical Engineering,*** 47 (C): 35-50 (2015).

173. Moon, S., Kim, S., and Wang, H., "Multimodal transfer deep learning with applications in audio-visual recognition", ***MMML Workshop at Neural Information Processing Systems,*** 1-6 (2016).

174. Ninomiya, H., Kitaoka, N., Tamura, S., Iribe, Y., and Takeda, K., "Integration of deep bottleneck features for audio-visual speech recognition", ***Interspeech,*** Dresden, 563– 567 (2015).

175. Gers, F. A., Schmidhuber, J. A., and Cummins, F. A., "Learning to forget Continual prediction with LSTM", ***Neural Computation,*** 12 (10): 2451–2471 (2000).

176. Graves, A., Fernandez, S., Gomez, F., and Schmidhuber, J., "Connectionist temporal classification labelling unsegmented sequence data with recurrent neural

networks", ***International Conference on Machine Learning,*** Dalian, 369-376 (2006).

177. Graves A., and Jaitly, N., "Towards end-to-end speech recognition with recurrent neural networks", ***International Conference on Machine Learning,*** Beijing, 1764–1772 (2014).

178. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., and Coates, A., "Deep speech Scaling up end-to-end speech recognition", ***International Conference on Machine Learning,*** San Francisco, 1-34 (2014).

179. Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks", ***Conference on Advances in Neural Information Processing Systems,*** Nevada, 1097-1105 (2012).

180. Srivastava, R. K., Greff, K., and Schmidhuber, J., "Training very deep networks", ***in Advances in neural information processing systems,*** Montreal, 2377-2385 (2015).

181. Petridis, S., and Pantic, M., "Deep complementary bottleneck features for visual speech recognition", ***International Conference on Acoustics, Speech and Signal Processing,*** Shanghai, 2304-2308 (2016).

182. Sui, C., Bennamoun, M., and Togneri, R., "Listening with your eyes Towards a practical visual speech recognition system using deep boltzmann machines", ***International Conference on Computer Vision,*** Santiago, 154–162 (2015).

183. Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., and Ogata, T., "Audio-visual speech recognition using deep learning", ***Applied Intelligence,*** 42 (4): 722–737 (2015).

184. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., "Going deeper with convolutions", ***Conference on Computer Vision and Pattern Recognition,*** Boston, 1-9 (2015).

185. Ordonez, F. J., and Roggen, D., "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition", ***Sensors,*** 16 (1): 1-25 (2016).

186. Bengio, Y., Simard, P., and Frasconi, P., "Learning long-term dependencies with gradient descent is difficult", ***Neural Networks,*** 5 (2): 157-166 (1994).

187. Chung, J. S., Zisserman, A., "Out of time automated lip sync in the wild", ***Asian Conference on Computer Vision,*** Taipei, 251-263 (2016).

188. Assael, Y. M., Shillingford, B., Whiteson, S., and Freitas, N. , "Lipnet Sentence-level lipreading", ***GPU Technology Conference,*** California, 1-13 (2017).

189. Graves, A., and Schmidhuber, J., "Framewise phoneme classification with bidirectional LSTM and other neural network architectures", ***Neural Networks,*** 18 (5): 602-610 (2005).

190. Bengio, Y., Simard, P., and Frasconi, P., "Learning long-term dependencies with gradient descent is difficult", ***Neural Networks,*** 5 (2): 157–166 (1994).

191. He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition", ***Conference on Computer Vision and Pattern Recognition,*** Las Vegas, 770-778 (2016).

192. Fung, H. L., and Mak, B., "End-to-end low-resource lip-reading with maxout CNN and LSTM", ***International Conference on Acoustics, Speech and Signal Processing,*** Calgary, 2511-2515 (2018).

193. Estellers, V., and Thiran, J. P., "Multi-pose lipreading and audiovisual speech recognition", ***Journal on Advances in Signal Processing,*** Barcelona, 1065-1069 (2012).

194. Wand, M., and Schmidhuber, J., "Improving speaker-independent lipreading with domain-adversarial training", ***Interspeech,*** Stockholm, 3662-3666 (2017).

195. Ganin, Y. and Lempitsky, V., "Unsupervised Domain Adaptation by Backpropagation", ***ICML,*** Lille, 1180-1189 (2015).

196. Ioffe, S., Szegedy, C., "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", ***Proceedings of the 32nd International Conference on International Conference on Machine Learning,*** Lile, 448-456 (2015).

197. Sergey, I., "Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models", ***Proceedings of the 31st International Conference on Neural Information Processing Systems,*** California, 1942-1950 (2017).

198. Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A., "How Does Batch Normalization Help Optimization? (No, It Is Not About Internal Covariate Shift)", ***Proceedings of the 32nd International Conference on Neural Information Processing Systems,*** New York, 2488-2498 (2018).

199. Lin, M., Chen, Q., and Yan, S., "Network in Network", ***2nd International Conference on Learning Representations,*** Banff, 1-10 (2014).

200. Newman, J.,. "Language identification using visual features", Ph. D. Thesis, ***School of Computing Sciences University of East Anglia,*** Norwich (2011).

201. Zhou, Z., Zhao, G., Hong, X., and Pietikainen, M., "A review of recent advances in visual speech decoding", ***Image and Vision Computing,*** 32 (9): 590-605 (2014).

202. Petridis, S., Li, Z., and Pantic, M., "End-to-end visual speech recognition with LSTMs", ***International Conference on Acoustics, Speech and Signal Processing,*** Orleans, 2592– 2596 (2017).

203. Petridis, S., Wang, Y., Li, Z., and Pantic M., "End-to-end multi-view lipreading", ***British Machine Vision Conference,*** London, 1-27 (2017).

204. Janke, M., Wand, M., and Schultz T., "Impact of lack of acoustic feedback in EMG-based silent speech recognition", ***Interspeech,*** Chiba, 2686-2689 (2010).

205. Sharfuddin, A. A., Tihami, M. N., and Islam, M. S., "A Deep Recurrent Neural Network with BiLSTM model for Sentiment Classification", ***International Conference on Bangla Speech and Language Processing,*** Sylhet, 1-4, (2018).

206. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y., "Multimodal deep learning", ***International Conference on Machine Learning (ICML),*** Washington, 1-8 (2011).

207. Hu, D., and Li, X., "Temporal multimodal learning in audiovisual speech recognition", ***Conference on Computer Vision and Pattern Recognition,*** Las Vegas, 3574–3582 (2016).

208. Takashima, Y., Aihara, R., Takiguchi, T., Ariki, Y., Mitani, N., Omori, K., and Nakazono, K., "Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss", ***Interspeech,*** San Francisco, 277–281 (2016).

209. Zimmermann, M., Ghazi, M. M., Ekenel, H. K., and Thiran, J. P., "Visual speech recognition using PCA networks and LSTMs in a tandem GMM-HMM system", ***Asian Conference on Computer Vision,*** Taipei, 264–276 (2016).

210. Bear, H. L., and Harvey, R., "*Phoneme-to-viseme mappings the good, the bad, and the ugly*", ***Speech Communication,*** Stockholm, 40–67 (2017).

211. Thangthai, K., Bear, H. L., and Harvey, R., "Comparing phonemes and visemes with DNN-based lipreading", ***British Machine Vision Conference,*** London, 1-11 (2017).

212. Mesbah, A., Berrahou, A., Hammouchi, H., Berbia, H., Qjidaa, H., and Daoudi, M., "Lip reading with Hahn Convolutional Neural Networks", ***Image and Vision Computing***, Dunedin, 76-83 (2019).

213. Huang, H., Song, C., Ting, J., Tian, T., Hong, C., Di, Z., and Gao, D., "A Novel Machine Lip Reading Model", Procedia Computer Science, 199 (2022): 1432-1437 (2022).

214. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S., "End-to-End Object Detection with Transformers", ***Computer Vision - ECCV,*** Glasgow, 213-229 (2020).

215. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., "Attention is All You Need", Advances in neural information processing systems, California, 1-15 (2017).

216. Kuwabara, H., Takeda, K., Sagisaka, Y., Katagiri, S., Morikawa, S., and Watanabe, T., "Construction of a large-scale Japanese speech database and its management system", ***International Conference on Acoustics, Speech, and Signal Processing,*** Glasgow, 560-563 (1989).

217. Obukhov, A., "Haar Classifiers for Object Detection with CUDA", ***GPU Computing Gems Emerald Edition,*** 517-544 (2011).

218. Viola, P., and Jones, M., "Rapid object detection using a boosted cascade of simple features", ***Computer Vision and Pattern Recognition Conference,*** Kauai, 1-9 (2001).

219. Huang, C. L., and Huang, Y.M., "Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification", ***Journal of Visual Communication and Image Representation,*** 8 (3): 278- 290 (1997).

220. Pantic, M., and Rothkrantz, L., "Expert System for Automatic Analysis of Facial Expression", ***Image and Vision Computing,*** 18 (11): 881-905 (2000).

221. Subramanyam, P. S., and Fegade, S. A., "Face and Facial Expression Recognition - A Comparative Study", ***International Journal of Computer Science and Mobile Computing,*** 2 (1): 1-13 (2013).

222. Kobayashi, H., and Hara, F., "Facial Interaction between Animated 3D Face Robot and Human Beings", ***IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation,*** Orlando, 3732-3737 (1997).

223. Rowley, H., Baluja, S., Kanade, T., "Neural Network-Based Face Detection", ***IEEE Transactions on Pattern Analysis and Machine Intelligence,*** 20 (1): 23-38 (1998).

224. Essa, I. A., and Pentland, A. P., "Facial Expression Recognition using a Dynamic Model and Motion Energy", ***International Conference on Computer Vision,*** Cambridge, 360-367 (1995).

225. Hefenbrock, D., Oberg, J., Thanh, N. T. N., Kastner, R. and Baden, S. B., "Accelerating Viola-Jones Face Detection to FPGA-Level Using GPUs", ***18th IEEE Annual International Symposium on Field-Programmable Custom Computing Machines,*** Charlotte, 11-18 (2010).

226. Jia, H., Zhang, Y., Wang, W., and Xu, J., "Accelerating Viola-Jones Face Detection Algorithm On GPUs", ***IEEE 14th International Conference on High Performance Computing and Communications,*** Liverpool, 396-403 (2012).

227. Wai, A. W. Y., Tahir, S. M., and Chang, Y. C., "GPU Acceleration of Real Time Viola-Jones Face Detection", ***IEEE International Conference on Control System, Computing and Engineering,*** Penang, 13-18 (2015).

228. Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., and Grundmann, M., "BlazeFace Sub-millisecond Neural Face Detection on Mobile GPUs", ***CVPR Workshop on Computer Vision for Augmented and Virtual Reality,*** Long Beach, 1-4 (2019).

229. Kartynnik, Y., Ablavatski, A., Grishchenko, I., and Grundmann, M., "Real-time facial surface geometry from monocular video on mobile GPUs", ***CVPR Workshop on Computer Vision for Augmented and Virtual Reality,*** Long Beach, 1-4 (2019).

230. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C. L., Yong, M., Lee, J., Chang, W. T., Hua, W., Georg, M., and Grundmann, M., "MediaPipe A Framework for Building Perception Pipelines", ***Google Research,*** 1-9 (2019).

231. Lecun, Y., Haffner, P., and Bengio, Y., "Gradient-Based Learning for Object Detection, Segmentation and Recognition", ***The IEEE,*** 1-43 (2001).

232. Sabour, S., Frosst, N., Hinton, G., "Dynamic Routing Between Capsules", ***31st Conference on Neural Information Processing Systems,*** Long Beach, 1-11 (2017).

233. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", ***Journal of Machine Learning Research,*** 15 (56): 1929−1958 (2014).

234. Atila, Ü., and Sabaz, F., "Turkish lip-reading using Bi-LSTM and deep learning models", **Engineering Science and Technology, and International Journal**, 1-10 (2022).

235. Mingfeng, H., Mutallip, M., Yadikar, N., Aysa, A., and Ubul, K., "A Survey of Research on Lipreading Technology", ***IEEE Access, 8***: 204518-204544 (2020).

236. Ankara Üniversitesi, "TUR147 Türkiye Türkçesi Ses Bilgisi", https://acikders.ankara.edu.tr/pluginfile.php/111341/mod_resource/content/0/3.%20Hafta.pdf (2017).

237. Easton, R. D., and Basala, M., "Perceptual dominance during lipreading.", **Perception & Psychophysics**, 32(6): 562–570 (1982).

ÖZGEÇMİŞ

Furkan SABAZ ilk ve orta öğrenimini Siirt'te tamamladı. Siirt Fen Lisesi'nden 2009 yılında mezun oldu. 2013 yılında Çukurova Üniversitesi Bilgisayar Mühendisliği bölümünden mezun oldu. 2015 yılında Karabük Üniversitesi Bilgisayar Mühendisliği bölümünde başladığı yüksek lisans programından 2017 yılında mezun olarak Bilgisayar Mühendisliği Anabilim Dalı'nda doktora eğitimine başladı. Doktora eğitimini Karabük Üniversitesinde devam ettirmektedir. Halen aynı üniversitede araştırma görevlisi olarak çalışmaktadır.