International Journal of Multidisciplinary Studies and Innovative Technologies

e-ISSN: 2602-4888
dergipark.org.tr/en/pub/ijmsit
Research Article

2025, 9 (1), 83-93
DOI: 10.36287/ijmsit.9.1.12
Received: March 3, 2025; Accepted: April 27, 2025

Yeni Türkçe Veri Seti ile Evrişimli Sinir Ağları Kullanarak Kelime Seviyesinde Otomatik Dudak Okuma

Nergis Pervan Akman[1*], Ali Berkol[2], Hamit Erdem[3]

1 BİTES Savunma Havacılık ve Uzay Teknolojileri A.Ş., Ankara, Türkiye (nergis.pervan@bites.com.tr) (ORCID: 0000-0003-3241-6812)*
2 BİTES Savunma Havacılık ve Uzay Teknolojileri A.Ş., Ankara, Türkiye (ali.berkol@bites.com.tr) (ORCID: 0000-0002-3056-1226)
3 Başkent Üniversitesi, Mühendislik Fakültesi, Elektrik Elektronik Mühendisliği Bölümü, Ankara, Türkiye (herdem@baskent.edu.tr) (ORCID: 0000-0003-1704-1581)

Türkçe Özet – Otomatik dudak okuma, son yıllarda önemli ölçüde gelişen bir araştırma problemidir. Dudak okuma, bazı durumlarda hem görsel hem de işitsel olarak değerlendirilmektedir. Bir güvenlik kamerasından istenmeyen bir kelimenin tespit edilmesi, görsel dudak okuma problemine bir örnektir. Bu tür salt görüntü içeren verilerin bulunduğu durumlarda, görsel-işitsel veri setleri uygulanamaz. Telaffuz edilen kelimenin ses girdisini her durumda elde etmek mümkün değildir. Bu çalışmada, yalnızca görüntü içeren yeni bir Türkçe veri seti toplandı. Yeni veri seti, kontrolsüz bir ortam olan YouTube videoları kullanılarak üretilmektedir. Bu nedenle, görüntüler ışık, açı, renk ve yüzün kişisel özellikleri gibi çevresel faktörler açısından zorlu parametrelere sahiptir. İnsan yüzündeki bıyık, sakal ve makyaj gibi dudakların tespit edilmesini zorlaştıran farklı özelliklere rağmen, veri üzerinde herhangi bir müdahale olmadan Evrişimli Sinir Ağları (CNN) kullanılarak tekil kelimeler ve iki kelimelik ifadeler dahil 10 sınıfta görsel konuşma tanıma problemine çözüm üretilmektedir. Ayrıca, bu çalışmada yalnızca görsel veri kullanıldığı için hesaplama maliyeti ve kaynak kullanımı çok modlu çalışmalara göre daha azdır. Aynı zamanda Ural-Altay dillerine ait yeni bir veri seti kullanılarak dudak okuma sorununu derin öğrenme algoritmasıyla ele alan bilinen ilk çalışmadır.

Anahtar Kelimeler *– Dudak Okuma, Çoklu Sınıf Sınıflandırma, Türkçe Dudak Okuma Veri Seti, Derin Öğrenme, Evrişimli Sinir Ağları, Dudak Tespiti*

Atıf: Pervan Akman N., Berkol A., Erdem H., (2025). Yeni Türkçe Veri Seti ile Evrişimli Sinir Ağları Kullanarak Kelime Seviyesinde Otomatik Dudak Okuma. International Journal of Multidisciplinary Studies and Innovative Technologies, 9(1): 83-93.

Automated Word-level Lip Reading using Convolutional Neural Networks with New Turkish Dataset

Extended Abstract

Research Problem/Questions – This study aims to develop a word-level automated lip-reading system using only visual data in Turkish, a language with limited representation in existing datasets. The primary goal is to recognize isolated words and phrases from lip movements using convolutional neural networks (CNNs), particularly in uncontrolled conditions like YouTube videos. It also emphasizes the challenge of reducing computational cost compared to multimodal systems.

Short Literature Review – In response, larger and more diverse datasets, such as LRW-1000 [32], were developed to simulate real-world variability more accurately with thousands of speakers and natural conditions. Beyond English, other languages, such as Czech, have also seen the development of audio-visual databases for continuous speech recognition, as demonstrated in the work by Trojanová et al. [22]. However, datasets in the Turkish language remain limited. Notably, Atilla and Sabaz [30] developed two Turkish datasets and utilized Bi-LSTM with CNNs, achieving strong results; however, their data was still collected in relatively controlled environments. Another Turkish study by Yargıç and Doğan [33] employed Kinect-based 3D lip feature extraction and utilized the KNN algorithm to classify spoken color names, achieving an accuracy of 78.22%. Many of these earlier datasets assume frontal face positioning and consistent lighting, which limits their robustness in natural settings. More recent research focuses on deep learning models that can process real-world complexity, such as visual-only lip reading under whisper or silent speech conditions [23], as well as robust face frontalization techniques [27]. The current study differentiates itself by introducing a new Turkish dataset collected from uncontrolled YouTube videos and applying CNNs directly to grayscale lip images without heavy preprocessing or multimodal input.

Methodology
– Dataset Collection: A new Turkish lip-reading dataset was created using video segments from YouTube, including TV series, films, and vlogs. The dataset consists of 10 classes — 5 single words and 5 two-word phrases (e.g., "günaydın", "teşekkür ederim") — which are among the most common expressions in daily Turkish communication. Each sample includes 40–60 frames extracted from 1–2 second clips, resulting in a total of 2335 samples. Manual cleaning was performed to remove irrelevant frames before and after the target word was spoken.
– Preprocessing and Feature Extraction: The Dlib library was used to detect frontal facial landmarks, followed by cropping the lip region using OpenCV. All frames were converted to grayscale to reduce computational cost. Two input formats were prepared: (1) discrete individual lip images, and (2) composite images created by concatenating 15 frames into a single 60x200 grayscale image. To ensure input consistency, shorter sequences were zero-padded and more extended sequences were truncated.
– Data Characteristics and Challenges: The dataset includes diverse visual conditions such as variations in lighting, angle, makeup, facial hair, and partial occlusions. Data analysis showed that 41% of the speakers had facial hair, 33% wore lipstick, and 11.7% had partially visible faces — all of which added to the complexity and realism of the dataset.
– CNN Model Design: Two distinct CNN architectures were developed for the two input types. Both models included convolutional layers (with ReLU activation), max-pooling layers, a flatten layer, dropout for regularization, and fully connected layers. The final output layer used Softmax activation to classify samples into 10 categories. The input sizes were adjusted for each architecture: the discrete model received 15 grayscale images of 50×50 pixels each, while the concatenated model used a single image sized 60×200 pixels.
– Training and Optimization: Training was carried out on an NVIDIA GTX 1650 Ti GPU using the Keras deep learning framework. To fine-tune performance, various hyperparameters, including learning rate, filter size, dropout rate, batch size, and number of epochs, were optimized using MLFlow and Hyperopt. Early stopping was employed to prevent overfitting, and training progress was monitored using accuracy and loss metrics.

Results and Conclusions – The CNN model using discrete lip images achieved an accuracy of 61.44%, while the concatenated version achieved 55.94%. Although the discrete model was more accurate, it required a higher computational cost. The study demonstrates that visual-only Turkish lip reading is feasible, highlighting the trade-off between accuracy and efficiency. This is the first known study using a Turkish dataset from natural video sources for CNN-based lip reading.

Keywords *– Lip Reading, Multiclass Classification, Turkish Lip Reading Dataset, Deep Learning, Convolutional Neural Networks, Lip Detection*

Citation: Pervan Akman N., Berkol A., Erdem H., (2025). Automated Word-level Lip Reading using Convolutional Neural Networks with New Turkish Dataset. International Journal of Multidisciplinary Studies and Innovative Technologies, 9(1): 83-93.

I. GİRİŞ

Konuşma, insanlar arasında kullanılan en yaygın iletişim yöntemidir. Konuşma işitsel olarak gerçekleştirilse de, görme de konuşulan ifadelerin anlaşılmasında büyük bir etkiye sahiptir. McGurk ve MacDonald, insanlara farklı metinlere işaret eden sesli ve görsel verileri göstererek insanların tepkilerini gözlemlemişlerdir [1]. Sesli anlatım ve görüntü birbirini destekleyen girdi verileridir. Otomatik görsel konuşma tanıma, genelleştirilebilir kelime çeşitliliği ve doğruluğunun sağlanması açısından sesli konuşma tanıma ve sesli-görüntülü konuşma tanımaya göre daha zorlu bir problem olduğundan doğruluk performansları daha düşüktür. Görsel konuşma tanımada zorluk yaratan durumlardan biri de sesteş sözcükler, yani benzer dudak hareketlerine sahip ifadelerdir. Ayrıca görüntünün kalitesi, kişinin yüzünün ve dudaklarının görüntüde yer almaması da zorlayıcı faktörlerdendir. Gürültülü ortamlarda akıllı telefonlara mesaj dikte etmek [2], [3], görsel sessiz şifreler kullanmak [4], [5], [6], sessiz filmleri yazıya dökmek [7], [8], konuşma engelli insanlar için dudak hareketlerine dayalı ses sentezlemek [9], [10], [11], [12] ve işitme engelli insanlara yardımcı olmak için dudak görünümlerini analiz etmek [13] otomatik dudak okuma sistemlerinin uygulama alanları arasında yer almaktadır.

Sesli veri setleri, görsel veri setleri ve görsel–işitsel veri setleri, hangi görevlerde kullanılabileceklerine, konuşmacı sayısına, kelime boyutuna, bilgiye ve kayıtların toplam süresine göre çeşitli özellikler içermektedir. Bu alandaki ilk veri tabanları alfabe ve rakam olarak oluşturulduğundan sınırlı bilgiye sahiptir [14], [15], [16]. Ancak, bu şekilde oluşturulan veri setleri, sürekli konuşma tanıma problemini çözmeye yönelik karmaşık bir yaklaşımdır. Harf ve rakam tanıma problemi sınırlı sayıda sınıfa sahip olduğundan, farklı algoritmaların sonuçlarını hızlıca değerlendirmek için çok sık kullanılmaktadır ve faydalı olmaktadır. Ancak, oluşturulan modeller sınırlı kapsamlara sahip olduklarından genelleştirilebilir değildir. Bu durum; kelimeler, ifadeler ve cümleler içeren gerçek dünya problemlerine yakın girdilere olan ihtiyacı ortaya çıkarmaktadır. İngilizce [17], [18], [19], [20], Fransızca [21] ve Çekçe [22] gibi birçok dilde bu tür kelime veya ifadeleri içeren veri tabanı örnekleri bulunmaktadır. Gerçek görüntülerde, kameranın insan yüzüne direkt olarak karşıdan baktığını garanti edemeyiz. Çoklu görüntü veri setlerinde, konuşmacı sayısı, sınıf sayısı, dil, kameranın çekim açısı gibi özelliklere sahip içerikler bulunur; bu içerikler, rakam [23], kelime [24] ve cümle [25] girdileri olarak yer almaktadır [26].

Bu çalışmada, temel katkımız, sadece görsel veri kullanarak hesaplama maliyetini azaltmak ve ilgili ses verisinin

bulunmadığı durumlarda bir dudak okuma sistemi geliştirmektir. Ayrıca, Türkçe dudak okuma çalışmalarında kullanılan doğal insan konuşma görüntülerinden elde edilen bilinen bir derin öğrenme modeli ve böyle bir veri seti bulunmamaktadır. Bu çalışmada farklı dudak temsillerini içeren girdilerle eğitilmiş, kelime düzeyinde bir Evrişimli Sinir Ağı (Convolutional Neural Network – CNN) modeli önerilmektedir. Makalenin geri kalanı şu şekilde düzenlenmiştir: 2. Bölüm'de literatür araştırması sunulmuştur. Daha sonra veri setinin detaylı analizi, dudakların tespiti, ağız temsili, CNN modeli ve eğitim parametreleri Bölüm 3'te açıklanmıştır. Kullanılan test verisi sayısı, performans puanları ve karmaşıklık matrisi gibi metrikler Bölüm 4'te gösterilmiştir. Son olarak bu çalışmanın değerlendirilmesi ve literatüre katkıları Bölüm 5'te özetlenmiştir.

II. LİTERATÜR ARAŞTIRMASI

Görsel konuşma tanıma problemi çalışmaları genel olarak öznitelik çıkarma ve sınıflandırma olmak üzere iki adımdan oluşmaktadır. Özellik çıkarma adımının ilk aşaması, doğru bir şekilde yüzü ve ardından dudakları tanımlamaktır. Doğal görüntülerden türetilen dudak okuma veri setlerinde, her görüntünün önden çekilmiş bir yüz içermesi her zaman mümkün değildir. Bazen görüntülerdeki yüzler sağa, sola, aşağıya veya yukarıya baktığı için, bu tür durumlarda doğru dudak görüntüsünü elde etmek zordur. Dudak tespiti adımından önce veri üzerinde yüzü öne bakacak şekilde hizalama (frontalizasyon) uygulamak etkilidir. Robust Face Frontalization (RFF) [27] yaklaşımı, girdi yüzünün konumunu ve 3 boyutlu şekil değişimini tahmin ederek, bunu önden görülen sentetik bir yüze dönüştürmektedir.

İlk dudak okuma yöntemleri, dudakların manuel olarak elde edilen geometrisini sınıflandırmak için Destek Vektör Makinelerine (Support Vector Machine – SVM) [28] veya Hidden Markov Model'lerine (HMM) [29] dayanmaktaydı. Mevcut yaklaşımlar ise çoğunlukla yenilikçi makine öğrenimi ve derin öğrenme yaklaşımlarından oluşmaktadır. Atila ve Sabaz [30] derin öğrenme modellerini kullanarak Türkçe dudak okuma problemlerini incelemişlerdir. Biri 111 kelime, diğeri 113 cümle içeren iki yeni Türkçe veri seti toplamışlardır. Çalışmada CNN kullanılarak özellik çıkarımı ve Çift Yönlü Uzun Kısa Süreli Bellek (Bidirectional Long Short-Term Memory – Bi-LSTM) kullanılarak sınıflandırma gerçekleştirilmiştir. Bi-LSTM %88,55 doğruluk elde edilerek cümle ve kelime veri setlerinde en iyi sonuçlara ulaşmaktadır. Cümle tespitinde dudak okumanın, kelime tespitine göre daha iyi performans gösterdiğini belirtilmektedir. Garg vd [31] araştırmalarında, görüntü çerçevelerinin tek bir büyük çerçeveye birleştirilmesini içeren Concatenated Frame Images adlı yeni bir yaklaşım tanıtmaktadırlar. Modellerinin ön yüzü olarak VGG mimarisine sahip bir 2 boyutlu evrişimli sinir ağı kullanılmaktadır. Görüntü çerçevelerini tek bir çerçeve içinde iç içe geçirerek, her veri noktası zamansal bilgisini uzamsal bilgiye dönüştürülmektedir ve bu sınıflandırma için bir LSTM ağına girdi olarak verilmektedir. Araştırmacılar, deneylerini MIRACL-VC1 veri setinden alınan videolar ile gerçekleştirmişlerdir. İlginç bir şekilde, en iyi performansın VGG ağının parametrelerinin dondurulması ve yalnızca LSTM'in eğitilmesiyle elde edildiğini bulmuşlardır.

Dudak okuma veri setleri, her bir veri tipinin yani temsil parçasının devamlı olarak ifade edilmesi ve oluşturulmasıyla elde edildiğinde sürekli olarak adlandırılmaktadır. Eğer her parça veri setindeki her bir örneğe karşılık geliyorsa izole edilmiş olarak adlandırılmaktadır. AVLetters veri seti [14], İngilizce'deki izole harflerden oluşmaktadır. 26 harf, 10 konuşmacı tarafından telaffuz edilmektedir ve görüntülerdeki yüzler önden pozlardır. Büyük bir veri seti olan LRW-1000 [32], sürekli kelimeler içermektedir. Bu veri seti, 2000'den fazla konuşmacı ve 7000'den fazla örnek içermektedir ve veri setindeki örnekler gerçek bir TV programından alındığı için doğal akışında olan kişilerin videoları yer almaktadır. Mevcut en büyük veri setlerinden biri olan AVDigits [23], izole edilmiş rakam sınıflandırması için toplanmıştır. Korpusdaki görüntüler, önden, 45 ve 90 derecelik dudak hareketlerini içermektedir. AVDigits'in sırasıyla rakam ve kelime öbeklerinden oluşan 53 ve 39 konuşmacılı 10 sınıfı vardır. Konuşmacılar cümleleri ve rakamları çeşitli ses seviyelerinde telaffuz etmektedirler. AVICAR veri seti [16], hareket halindeki bir arabadan elde edilmiştir. Bu veri seti, alfabe ve rakam segmentasyonu için sürekli, cümleler için izole olarak toplanmıştır. Her veri türü için 68 konuşmacı ve yüz açıları 4 farklı şekilde dahil edilmiştir. Alfabeler, rakamlar ve cümleler için sırasıyla 26, 10 ve 20 sınıf bulunmaktadır. AusTalk [20] da rakamları, kelimeleri ve cümle parçalarını içeren bir veri setidir. Bu parçalar, 1000 farklı konuşmacı tarafından telaffuz edilmiştir. Veri setindeki konuşmacıların yüzleri önden çekimlerdir. AusTalk, diğer birçok dudak okuma veri seti gibi İngilizce olarak oluşturulmuştur.

Literatürde Türkçe olarak yapılan ve iyi sonuçlar veren çok az dudak okuma çalışması bulunmaktadır. Bunlardan biri, Yargıç ve Doğan'ın çalışmasıdır [33]. MS Kinect cihazından alınan görüntülerle, Türkçe renk adlarının telaffuz edildiği bir veri seti oluşturulmuştur. Bu veri setinin 3 boyutlu olarak elde edilen dudak noktaları arasındaki mesafe verileri kullanılarak özellikler oluşturulmuştur. Toplanan bu veri seti ile, K-En Yakın Komşu (K-Nearest Neighbor – KNN) algoritması kullanılarak 15 sınıflı %78.22 doğruluk oranına sahip bir model üretilmiş ve bu model işitme engelli çocukları desteklemek amacıyla geliştirilmiştir. Bu Türkçe veri setine ek olarak, bilinen en büyük İngilizce görsel-işitsel veri seti olan Lip Reading Sentences (LRS) [7], sürekli konuşma tanıma için oluşturulmuştur. BBC yayınlarından derlenen bu veri tabanı, 100.000'den fazla ifade ve 1000'den fazla farklı kişi içermektedir. Özcan ve Baştürk [35] alfabe düzeyinde dudak okuma veri seti olan AvLetters'ı [14] sınıflandırmak için CNN algoritmasını kullanılmıştır. Bilindiği kadarıyla, transfer öğrenme kullanarak AlexNet ile dudak okuma çalışması bulunmamaktadır. Margam ve arkadaşları [36], GRID korpusundan konuşulan cümleleri tahmin etmek amacıyla ASCII karakterlerini çözmek için 3D-2D-CNN-BLSTM mimari konfigürasyonunu önermişlerdir. İlk yaklaşım olarak, önerilen ağ, Bağlantı Temelli Zaman Sınıflandırması (Connectionist Temporal Classification – CTC) kaybı (ch-CTC) kullanılarak karakterler üzerinde eğitilmiştir. Geleneksel otomatik konuşma tanıma eğitim hattında, BLSTM-HMM modeli, dar boğaz dudak özellikleri üzerinde eğitilmiştir. Aynı 3D-2D-CNN-BLSTM ağı, ikinci teknikte (w-CTC) kelime etiketleri üzerinde CTC kaybı ile eğitilmiştir. Görülmemiş konuşmacı test seti kullanılarak %24.5 Kelime Hata Oranı (Word Error Rate – WER) farkıyla LipNet'ten daha iyi bir sonuç elde edilmiştir.

Birçok çalışma, yüz tespiti ve görüntülerdeki konuşmacıların yüz özelliklerini çıkarmak için dlib kütüphanesini kullanmaktadır. Kastaniotis ve arkadaşları [39], yeni bir veri seti sunarak Yunanca kelimelerde görsel dudak okuma görevi için derin CNN'lerin yeteneklerini değerlendiren

iki deney üzerine odaklanmışlardır. Önerilen yöntem, bir CNN'in ardından çift yönlü bir LSTM ve sonrasında iki tam bağlantı katmanı ve basit bir dikkat mekanizmasından oluşan bir kombinasyona dayanmaktadır. Ayrıca, Leave-N samples out ve Leave-N persons out değerlendirme protokollerini kullanmışlardır. Her iki durumda da 50 kelime için rastgele sınıflandırma oranı %2'dir. Sarhan ve arkadaşları [40], videodan dudak okuma için hibrit dudak okuma HLR-Net adı verilen derin bir CNN modeli geliştirmiştir. Önerilen model üç aşamadan oluşmaktadır: ön işleme aşaması, kodlayıcı aşaması ve kod çözücü aşaması. HLR-Net, girdi olarak dudak hareketi videosunu kullanmakta, ardından bu videoyu OpenCV kütüphanesini kullanarak çerçevelere dönüştürmekte ve son olarak dlib kullanarak ağız özelliklerini çıkarmaktadır. Deney sonuçları, önerilen HLR-Net modelinin, LipNet, LCANet ve A-CTC gibi diğer üç son teknoloji modeli geride bıraktığını göstermektedir; görülmemiş konuşmacılar için karakter hata oranı %4.9, kelime hata oranı %9.7 ve BLEU skoru %92, örtüşen konuşmacılar için karakter hata oranı %1.4, kelime hata oranı %3.3 ve BLEU skoru %99 elde edilmiştir. Ting ve arkadaşları [41], bir dudak okuma veri setinin oluşturulmasını incelemişlerdir. Veri setindeki çerçeveler, scikit-video kullanılarak orijinal videolardan çıkarılmış ve yüz tespiti dlib uygulanarak gerçekleştirilmektedir. Dudak kırpma işlemini gerçekleştirmek için özellik noktaları işlenerek dudak görüntüleri yakalanmıştır. Sonuç olarak, önerilen veri setinde her biri 7.000 adet dudak resmine sahip 33 konuşmacı bulunmaktadır. Deney, dlib ve diğer kütüphanelerin kullanımının yüzü etkili ve doğru bir şekilde tanımlayıp videodaki dudak bölgesini çıkardığını kanıtlamaktadır. Fu ve arkadaşları [42] tarafından daha önce oluşturulan Databox adı verilen doğal olarak dağıtılmış kelime düzeyinde Çince veri setini genişletmişlerdir. Önerilen model, bir ResNet ağı ve Geçici Konvolüsyonel Ağdan oluşan mevcut son teknoloji yönteme dayalı olarak geliştirilmiştir. ResNet'i, Convolutional Block Attention adlı eklenti dikkat modülüne sahip hafif bir konvolüsyon ağı olan ShuffleNet ile değiştirmişlerdir. Deneylerde, ResNet-18 modeli ile %74.4 doğruluk oranı elde edilmektedir ki bu, deneylerdeki en yüksek doğruluk skorudur. Standart kanal genişliğine sahip ShuffleNet ile %68.4 doğruluk oranına ulaşılmaktadır.

III. MATERYAL VE METOT

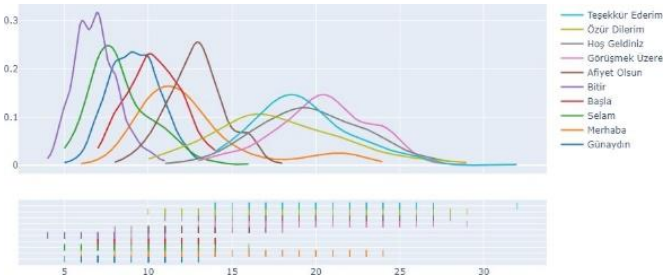
A. *Veri Seti*

Çalışmada kullanılan veri seti [44] daha önceki çalışmalarımızda [45], [46] kullandığımız ve YouTube üzerinden topladığımız Türkçe dudak okuma veri setidir. Veri seti, Türkçede günlük hayatta sık kullanılan 5 kelime ve 5 ifadeden oluşan 10 sınıf içermektedir. Bu sınıflar, "afiyet olsun", "başla", "bitir", "görüşmek üzere", "günaydın", "hoş geldiniz", "merhaba", "özür dilerim", "selam" ve "teşekkür ederim" ifadeleridir. Veri setinde bu ifadelerin seçilmesinin nedeni Türkçede selamlaşma anlamında en çok kullanılan ifadeler olmalarıdır. Ayrıca, sadece tek kelimelik sınıflar yerine karmaşıklığı artırmak için iki kelimelik ifadeler de seçilmiştir. İlgili kelimenin yer aldığı kısımların ekran görüntüleri, YouTube'da yayınlanan dizi, film ve blog videoları gibi çeşitli görüntülerden kaydedilmiştir. Ardından, kaydedilen görüntüler Python kullanılarak birer resim karesi olarak dosyalara kaydedilmiştir. İlgili kelimenin bulunduğu videolar 1 veya 2 saniye uzunluğunda olup, her bir örnek için elde edilen görüntü sayısı 40-60 aralığındadır. Toplanan görüntüler, kelimenin bir cümle akışı içinde telaffuz edildiği sırada kaydedildiğinden, bazı durumlarda bir kelimenin başı veya sonu başka bir kelimeyle birleşebilmektedir. Bu nedenle, kelimenin başlangıcından önceki çerçeveler ve kelime telaffuzunun sonundan sonraki çerçeveler, eleme yapılarak manuel olarak kaldırılmaktadır. Bu eleme işlemi sonrasında, her bir örnek için görüntü dizisinde önemli bir azalma gözlemlenmiştir. Veri toplanırken, her sınıf için mümkün olduğunca eşit sayıda örnekle denge sağlanmaya çalışılmıştır (bkz. Tablo 1).

Tablo 1 Sınıfların Veri Dağılımı

Sınıflar	Örnek Sayısı
afiyet olsun	235
başla	225
bitir	244
görüşmek üzere	224
günaydın	232
hoş geldiniz	226
merhaba	268
özür dilerim	209
selam	235
teşekkür ederim	237

Veri seti hem tek kelimelik hem de iki kelimelik sınıflar içerdiğinden, ifadelerin telaffuz süreleri değişkenlik göstermektedir. Örneğin "teşekkür ederim" ve "özür dilerim" gibi ifadeler daha uzun olduğundan telaffuz süreleri ve veri setinde kapladıkları çerçeve sayısı "selam" kelimesine göre daha fazladır. "Özür dilerim" kelimesinin çerçeve sayısı 30'u aşarken, "selam" kelimesinin çerçeve sayısı 15'i geçmemektedir (bkz. Şekil 1). Sınıflandırma yaparken dengeli bir temsil sağlamak için bu dağılımı dikkate almak kritik öneme sahiptir.



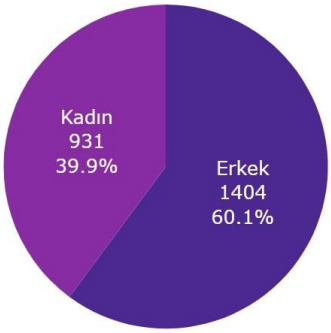
Şekil 1 Sınıfların Çerçeve Sayısı Frekansı

Dudak okuma problemi için toplanan veriler, doğal akışlarında konuşmalarına devam eden konuşmacıların videolarından elde edildiğinden, görüntüler çeşitlilik açısından zorluklar içermektedir (bkz. Şekil 2). Bazı durumlarda konuşmacılar yüzlerini doğrudan kameraya çevirmemesi, görüntüdeki ışık farklılıkları, görüntü kalitesi ve konuşmacının uzakta olması gibi durumlar da bulunmaktadır. Bunlara ek olarak, elde edilen görüntülerde mikrofon gibi nesnelerin konuşmacının önüne gelmesi, konuşmacının bıyığı veya ruj kullanımı gibi kişisel çeşitlilik yaratan problemler de mevcuttur. Bu veri setinin kullanımı, doğal görüntülerden elde edildiği için zordur, bu nedenle gerçek modelleme çalışmaları için araştırmacılar tarafından kullanılmaya uygundur.

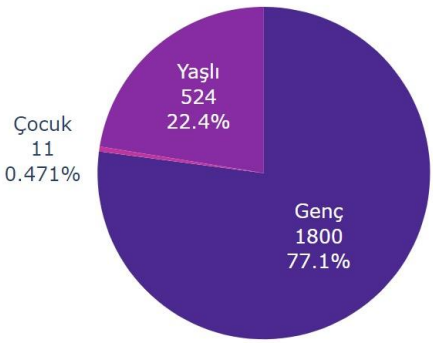
Şekil 2 Veri Kümesinde Karşılaşılan Zorlu Durumlar

Tablo 2, Şekil 3, 4, 5, 6 ve 7, veri setindeki cinsiyet, yaş, bıyık, sakal ve ruj gibi fiziksel özellikler ve ayrıca yüzün tamamen görünür olup olmadığı hakkında bilgi vermektedir. Bu veri setinde toplam 2335 örnek/konuşmacı bulunmaktadır. Bunların çoğunluğu genç ve erkektir. Konuşmacıların %41'inde bıyık ve/veya sakal bulunmaktadır. Bıyık ve sakal, dudak okuma problemi açısından dudakları net bir şekilde tanımlama noktasında zorluk yaratmaktadır. Ayrıca, konuşmacıların %33.3'ünde belirgin şekilde ruj bulunmaktadır. Bu, net bir dudak hattı oluşturacağından kadın konuşmacıların dudaklarının tespit edilmesini kolaylaştıran bir durumdur. Veri setindeki bir diğer zor durum ise bazı konuşmacıların yüzlerinin önden net olarak görünmemesidir.

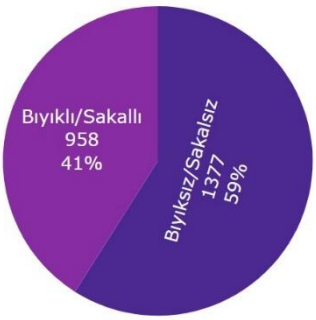
Şekil 7'ye göre, tüm konuşmacıların %11.7'si kısmi yüzlerin bulunduğu çerçeveleri içeren örnekler barındırmaktadır.



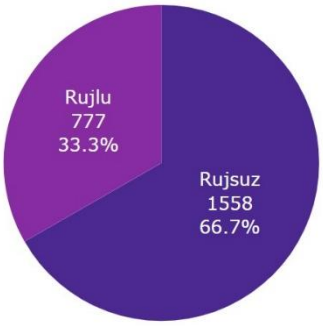
Şekil 3 Cinsiyete Göre Veri Dağılımı



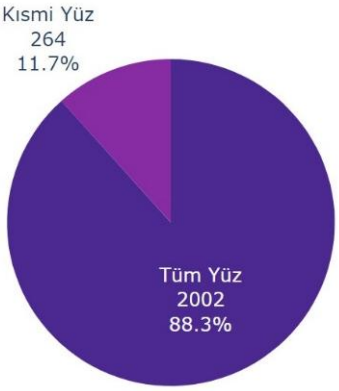
Şekil 4 Yaşa Göre Veri Dağılımı



Şekil 5 Bıyıklı veya Sakallı İnsan Sayısına Göre Veri Dağılımı



Şekil 6 Rujlu İnsan Sayısına Göre Veri Dağılımı

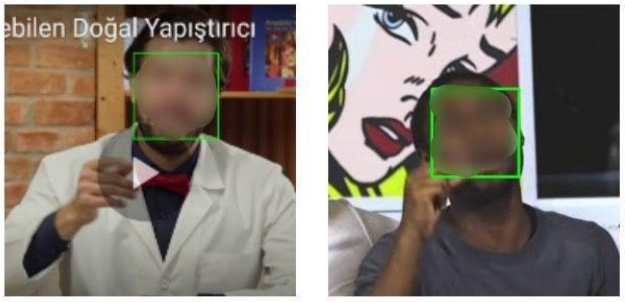


Şekil 7 Kısmi Yüze Sahip İnsan Sayısına Göre Veri Dağılımı

B. *Dudak Tespiti*

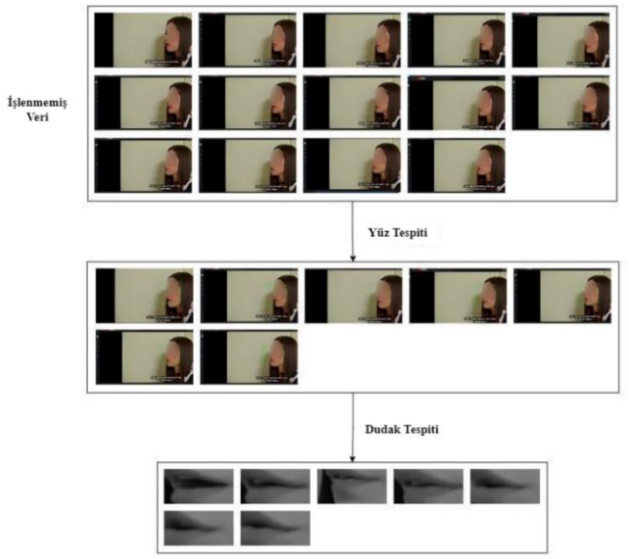
Dudak okuma probleminde, görüntülerin RGB olmasına her zaman ihtiyaç yoktur, bu sebeple yüz ve dudak tespitinde ve sonrasındaki çalışmalardan derin öğrenme model eğitimi sırasında hesaplama ve zaman maliyetlerini azaltmak amacıyla görüntüler gri tonlamalı hale dönüştürülür. Bu çalışmada ilk olarak hazır bir kütüphane olan dlib kütüphanesi kullanılarak veri setindeki insan görüntülerinden yüzler kesilmektedir. Kullanılan get_frontal_face_detector() fonksiyonu çağrıldığında, dlib'in önceden eğitilmiş HOG+Linear SVM yüz detektörünü döndürmektedir.

HOG+LINEAR SVM hızlı ve etkili bir şekilde çalışmaktadır. HOG'un doğası gereği, döndürme ve görüş açısı durumlarına uyum sağlamaktadır. Hızlı tespiti nedeniyle gerçek zamanlı yüz tespiti için uygun bir yöntemdir. Şekil 8'de görülebileceği gibi, yüzler açılı olsa veya yüzün önünde bir engel olsa dahi dudakları da içeren hassas bir yüz tespiti yapılabilmektedir.



Şekil 8 HOG+SVM ile Yüz Tespiti

Dudak kesme çalışmalarında, OpenCV kütüphanesi kullanılarak, dudak kısmını almak için belirli bir dizi nokta ile ilgili bölgenin konturu çizilir. Yer işaretlerinde 49-68 aralığı dudak bölgesine karşılık geldiğinden elde edilen yüz görüntüsü üzerinde bu aralık kesilir. Ardından, boundingRect() fonksiyonu yardımıyla belirlenen bölge dikdörtgen bir çerçeve içine alınır. Şekil 9'da öncelikle ham görüntüler, ikinci aşamada tespit edilen yüzler ve sonda ise kesilmiş dudak görüntüleri gösterilmektedir. Her ham görüntüye karşılık bir yüz tespiti yapılmadığı için tespit edilen dudak görüntüleri ham görüntülerin sayısından daha azdır. Her görüntüde açı ve ışık açısından benzer görüntüler bulunsa da her biri için yüz tanıma işleminin gerçekleştirilemediği gözlemlenmiştir. Son olarak, elde edilen dudak görüntüleri, daha sonraki aşamalarda kullanılmak üzere 100x200 boyutunda kaydedilmektedir.



Şekil 9 Dudak Tespiti

C. *Dudakların* Temsili

Geliştirilen ilk yaklaşımda, bir örneğin her bir ardışık görüntüsü derin öğrenme modelinde kullanılarak akışın korunması sağlanmaktadır. Dudakların ayrık olarak modele verilmesi durumunda, bu görüntüler 15 görüntüden oluşan bir seri halinde kullanılarak tek bir görüntü yerine bir akış sağlanmaktadır. İkinci yaklaşımda ise, 15 görüntü birleştirilerek tek bir daha küçük görüntü olarak kullanılmaktadır. Birleştirme işleminden sonra 100x200 boyutundaki her bir görüntü, çok büyük bir görüntü elde edilmemesi için 20x40 olarak yeniden boyutlandırılmaktadır. Bir dudak örneğinin çerçeve sayısı 15'ten azsa, gri tonlamalı olarak 0 değerleriyle doldurularak 15'e tamamlanacak şekilde görüntü eklenmektedir. Eğer çerçeve sayısı 15'ten fazla ise, fazla çerçeveler çıkarılır. 15 çerçeve, 3 satır ve 5 sütun şeklinde ardışık olarak birleştirildiğinde, 60x200 boyutunda bir görüntü elde edilmektedir. Şekil 10, bu birleştirilmiş 15 ardışık görüntüyü göstermektedir.

Şekil 1 Birleştirilmiş Dudaklar

D. *CNN Modeli*

Son yıllarda, makinelerin hesaplama gücünün artması, donanım kaynaklarının genişlemesi ve GPU'ların kullanımı gibi teknolojik gelişmeler sayesinde, CNN modellerinin kullanımı önemli ölçüde artmıştır. CNN'lerin klasik yapay sinir ağlarından en önemli farklarından biri, görüntünün belirli bölgelerini alarak parametre sayısını azaltmasıdır. Son yıllarda yapılan çalışmalarda görüntü sınıflandırma, nesne tespiti, görüntü segmentasyonu gibi girdi verisinin görüntü olduğu birçok problemin yanı sıra girdinin metin olduğu çalışmalarda da kullanılmaktadır. Bu alandaki karmaşık problemlerin kolaylıkla çözülebilmesinin en büyük nedenlerinden biri, görüntü üzerinde herhangi bir öznitelik çıkarımı yapmadan, bir uzman bilgisine ihtiyaç duymadan ve görüntü üzerindeki herhangi bir nesne için konum ve şekil gibi öznitelikleri tespit etmeden CNN ile model geliştirilebilmesidir. Evrişimli sinir ağları, doğal sinyallerin özelliklerinden yararlanmak için dört temel prensipten yararlanmaktadır: yerel bağlantılar, paylaşılan ağırlıklar, havuzlama ve çok sayıda katmanın kullanımı CNN'i diğer sinir ağı yaklaşımlarından ayıran en önemli katman, evrişim katmanıdır. Temel olarak girdi, evrişim, havuzlama ve tam bağlantılı katmanlardan oluşmaktadır. Sınıflandırıcı CNN modelleri kabaca öznitelik çıkarma ve sınıflandırma katmanlarından oluşmaktadır. Öznitelik çıkarma katmanları, evrişim ve havuzlama katmanları sayesinde girdi görüntüsünün anlamlı çıktılarını üreterek diğer görüntülerle ilişkilendirilmektedir. Sınıflandırma katmanında, sınıflandırma problemi için sınıf sayıları ve bunların anlamlı görüntülerine dayalı olasılıklar üretilmektedir. Buna dayanarak çok sınıflı bir modelde son katmandaki softmax fonksiyonu ile tahmin yapılmaktadır.

E. *Önerilen Model*

Önerilen CNN mimarisi, birleştirilmiş dudak görüntüleri ve ayrık olarak eğitilmiş dudak görüntüleri için iki tanedir.

Tablo 2 Verinin Öznitelikleri

Sınıf	Kadın	Erkek	Yaşlı	Genç	Çocuk	Bıyık/Sakal	Ruj	Kısmi Yüz	Tüm Yüz
afiyet olsun	125	110	102	132	1	79	100	71	164
başla	55	170	66	157	2	126	43	57	119
bitir	90	154	73	171	0	122	71	35	209
görüşmek üzere	87	137	24	199	1	83	78	2	222
günaydın	102	130	60	172	0	92	86	26	206
hoş geldiniz	109	117	26	198	2	66	82	9	217
merhaba	113	155	39	227	2	106	106	6	262
özür dilerim	76	133	42	166	1	89	65	22	187
selam	52	183	67	167	1	116	37	25	210
teşekkür ederim	122	115	25	211	1	79	109	11	226
TOPLAM	931	1404	524	1800	11	958	777	264	2022

Ayrık Dudak Girdileri ile CNN Modeli

Bölüm C'de bahsedildiği gibi, ayrık dudak görüntüleri girdi katmanına bir dizi olarak verilmektedir. Mimari, iki evrişim ve iki maksimum havuzlama katmanı içermektedir. Evrişim katmanları, aktivasyon fonksiyonu olarak ReLU (Rectified Linear Unit) kullanmaktadır, filtre boyutları 128 ve 64 olup, filtrelerde kullanılan adım 1'dir ve dolgu (padding) bulunmamaktadır. Maksimum havuzlama katmanlarının havuzlama boyutları 3x3x3 ve adımı 2'dir. Bu dört katmanı düzleştirme (flatten) katmanı takip etmektedir ve mimari, bırakma (dropout) ile tam bağlantılı katmanlarla devam etmektedir. Girdi vektörü, 50x50 sabit boyutlu 15 görüntüden oluşmaktadır. Bu görüntülere dolgu olmadan ve 1 adımlık adımlarla evrişim katmanında rastgele 128 filtre uygulanmaktadır. Evrişim işleminden sonra 13x48x48x128 boyutunda bir çıktı üretilmektedir. Evrişim katmanını takip eden maksimum havuzlama katmanının çıktısında 3x3 havuzlama boyutu olduğundan, çıktı 6x24x24x128 boyutunda olur. Sırasıyla conv3d, maksimum havuzlama ve düzleştirme katmanları uygulandıktan sonra, 15488 boyutlu bir vektör elde edilmektedir. Özellikle CNN modellerinde aşırı öğrenmeyi (overfitting) önlemek için ReLU aktivasyon fonksiyonuna ve %0.5 oranında bırakma (dropout) katmanlarına sahip iki tam bağlantılı katman uygulanmaktadır. Son olarak çok sınıflı bir sınıflandırma problemi üzerinde çalışıldığı için, mimari, softmax aktivasyon fonksiyonu ile 10 boyutlu bir vektör çıktısı üreten tam bağlantılı bir katmanla sonlandırılır. Çıktıda, "afiyet olsun", "başla", "bitir", "görüşmek üzere", "günaydın", "hoş geldiniz", "merhaba", "özür dilerim", "selam" ve "teşekkür ederim" şeklindeki 10 sınıf için olasılıklar üretilmektedir.

Birleştirilmiş Dudak Girdileri ile CNN Modeli

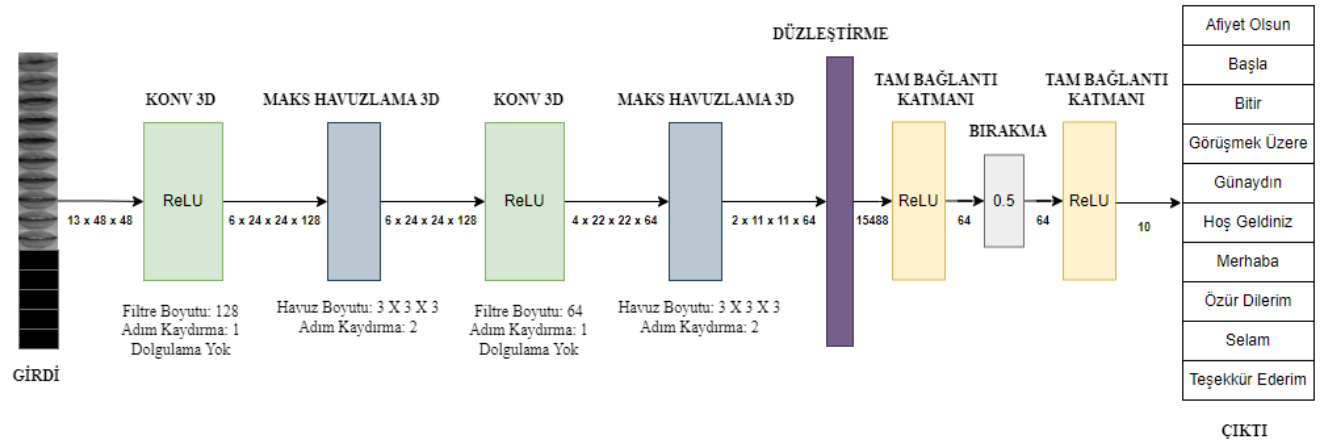
Bu yaklaşımda; ayrık dudak görüntülerinin girdi olarak alındığı durumdan farklı olarak dudaklar birleştirilmektedir ve 15 görüntü, tek bir görüntü girdisi oluşturacak şekilde kullanılmaktadır. Bu nedenle, hesaplama maliyeti açısından oldukça uygun bir yöntemdir. Tek bir görüntü bir görüntü dizisini temsil ettiğinden veri karmaşıklığını azaltmaktadır. Girdi olarak 50x50 boyutunda bir görüntü kullanılarak evrişim katmanına gönderilmektedir. Havuzlama boyutu 2x2 olduğu için, düzleştirme (Flatten) katmanının girdisi 11x11x16 boyutundadır. Dudakların ayrı ayrı verildiği yaklaşımdaki mimariden farklı olarak, Düzleştirme katmanının ardından 9216 boyutlu bir vektör çıktısı olan 1 tam bağlantı katman ve bırakma katmanı bulunmaktadır. Son olarak, 10 sınıfa sahip bir çıktı vektörü üretilmektedir.

F. *Eğitim*

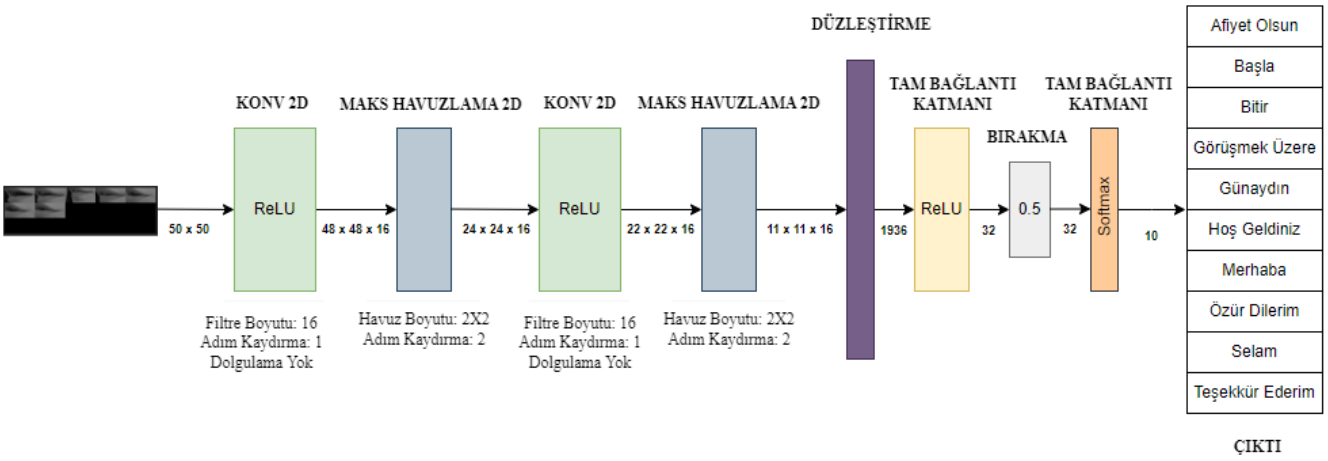
Eğitim sürecinde, dudakların ayrı ayrı eğitilmesi ve birleştirilmesi yaklaşımlarına yönelik iki farklı çalışma için farklı hiperparametreler üzerinde deneyler yapılmıştır. Deneylerin sonuçlarını değerlendirmek ve hiperparametre ayarlarını yapmak için makine öğrenimi yaşam döngüsünü yönetmek amacıyla geliştirilen bir Python kütüphanesi olan mlflow kullanılmaktadır. Dudakların birleştirildiği ve ayrıldığı durumlar için farklı hiperparametreler uygulanmaktadır (Tablo 3). İki yaklaşımın model kapasitesi ve karmaşıklığı farklı olduğundan, öğrenme oranı, yığın boyutu (batch size) ve epok (epoch) sayısı gibi parametreler değişiklik göstermektedir. Eğitim sırasında, her iki model için de 1, 2 ve 3 olmak üzere farklı evrişim katmanları denenmiştir. Model doğruluk grafikleri ve test sonuçlarına dayanarak, aşırı öğrenme veya yetersiz öğrenme (underfitting) olmayacak şekilde model kapasitesi göz önünde bulundurularak evrişim katmanlarının sayısında değişiklikler yapılmaktadır. Öğrenme oranı değerleri 0.0002, 0.002, 0.001 ve 0.0001, her iki katman için evrişim filtre boyutları: 16, 32, 64, 128 ve 256, bırakma oranları ise 0.3, 0.4, 0.5 ve 0.6 olarak belirlenmiştir. Optimum sonuçlar için Hyperopt kütüphanesi ile deneyler tekrarlanmış ve mevcut parametrelere ulaşılmıştır. Epok sayısı, kayıp değerine bağlı olarak belirlenmiştir.

IV. SONUÇLAR VE TARTIŞMA

Dudak okuma gibi dilin de işin içine girdiği çalışmalarda dilin farklı telaffuzları ve varyasyonları olduğundan doğru bir değerlendirme yapmak zordur. Bu kapsamda yapılan çalışmalarda kelime düzeyinde karşılaştırılabilir Türkçe bir veri seti mevcut değildir. Çalışmalarımızda temel olarak Türkçe dudak okuma problemine yönelik bir CNN mimarisi geliştirilmesi amaçlanmaktadır. CNN mimarisini temel alan tüm deneyler GPU üzerinde çalıştırılmaktadır. 4GB bellekli NVIDIA 1650 Ti grafik kartı kullanılmaktadır. İyileştirmeler Python Keras kütüphanesi kullanılarak yapılmaktadır. Bunlara ek olarak görselleştirme için Plotly ve Seaborn kütüphanelerinden, görüntü işleme çalışmaları için ise OpenCV kütüphanelerinden yararlanılmaktadır. Dudakların birleşik ve ayrık olduğu iki gösterim aşamasında da deneyler Tablo 4'teki örnek sayısı ile gerçekleştirilmektedir.



Şekil 11 Ayrık Dudakların Temsilinde Kullanılan CNN Modeli



Şekil 12 Birleştirilmiş Dudakların Temsilinde Kullanılan CNN Modeli

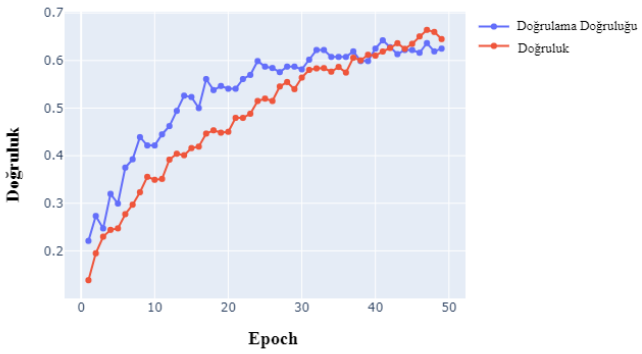
Tablo 3. Her Sınıfa Ait Test Örneklerinin Sayısı

Sınıf	Örnek Sayısı
afiyet olsun	29
başla	35
bitir	46
görüşmek üzere	23
günaydın	32
hoş geldiniz	34
merhaba	43
özür dilerim	31
selam	37
teşekkür ederim	35

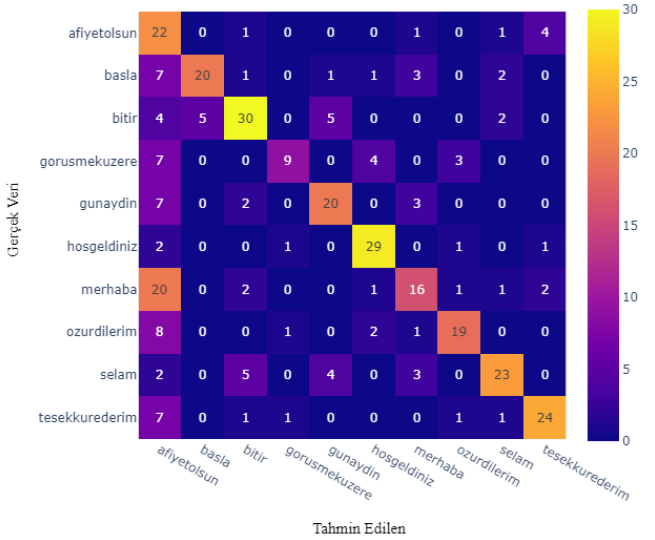
A. *Ayrık Dudak Görüntülerine Ait Eğitim Sonuçları*

Eğitim sonuçlarına bakıldığında, erken durdurma (early stopping) kullanılarak belirlenen epok sayısına göre doğruluk değişimleri Şekil 13'te gösterilmektedir. Doğruluk değerinde 50 epok boyunca iyileşme görülmediği için eğitim süreci durdurulmuştur. Eğitimin devam etmesi durumunda model aşırı öğrenmeye maruz kalacağından başka hesaplamalar yapılmasına gerek yoktur. Test verilerinde tahmin edilen sınıfların sonuçları incelendiğinde, yanlış olarak tespit edilen sınıflar genellikle "afiyet olsun" sınıfında olduğu görülmektedir (Şekil 14). Bu sınıfa ait verilerin içerisinde diğer sınıflara oranla daha fazla kısmi yüz görüntüsü yer almaktadır, dolayısıyla daha zorlu olabileceği yorumu yapılabilir.

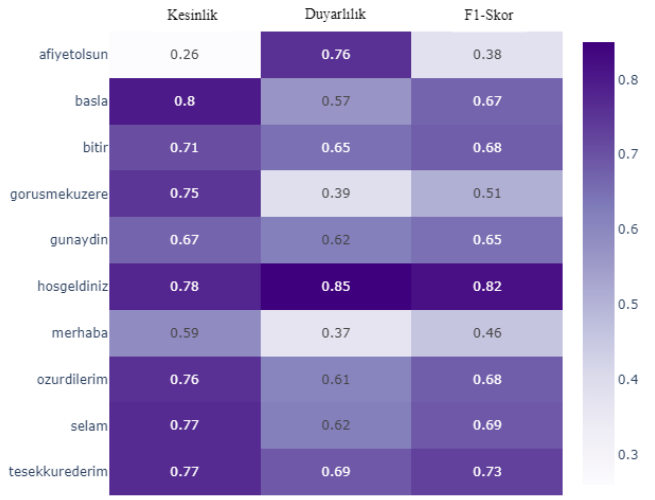
Özellikle sınıfı "başla", "günaydın" ve "özür dilerim" olan örneklerde yanlış tahminler "afiyet olsun" sınıfında yoğunlaşmaktadır. "Afiyet olsun" sınıfında yapılan hatalar, genellikle "merhaba" ifadesindeki 20 örnekte yapılmıştır. Buna karşılık, "hoş geldiniz" ve "selam" ifadelerinde yanlış tahmin edilen "afiyet olsun" örneği sayısı 2'dir. Bunları takip eden "başla", "hoş geldiniz", "teşekkür ederim", "selam", "özür dilerim", "görüşmek üzere" ve "bitir" sınıflarında ise kesinlik (precision) skorlarının yüksek olduğu görülmektedir, (Şekil 15). Buradan hareketle, bu sınıflara ilişkin pozitif tahminlerin çoğunluğunun doğru olduğunu söyleyebiliriz. Genel olarak, karmaşıklık matrisine dayanarak "afiyet olsun" sınıfındaki hata oranının yüksek olduğu görülmektedir. Bu ifadelerin veri setindeki seslendirilmesinde belirgin bir dudak hareketi olmayabilir veya Türkçe dilbilgisi kurallarına göre daha zorlayıcı ifadelerden biri olarak yorumlanabilir. Bu veri setinde örnek sayısı açısından sınıf dengesizliği olmamasına rağmen, tahmin performansları sınıflara göre değişiklik göstermektedir çünkü tıpkı gerçek yaşam sahnelerinde olduğu gibi konuşmacı farklılıkları, görüş açıları ve ışık farkları gibi her sınıf için çeşitlilik yaratan durumlar mevcuttur.



Şekil 13 Ayrık Dudakların Epok Başına Düşen Eğitim ve Doğrulama Doğruluğu



Şekil 14 Ayrık Dudak Görüntüleriyle Eğitilmiş Modelin Karmaşıklık Matrisi

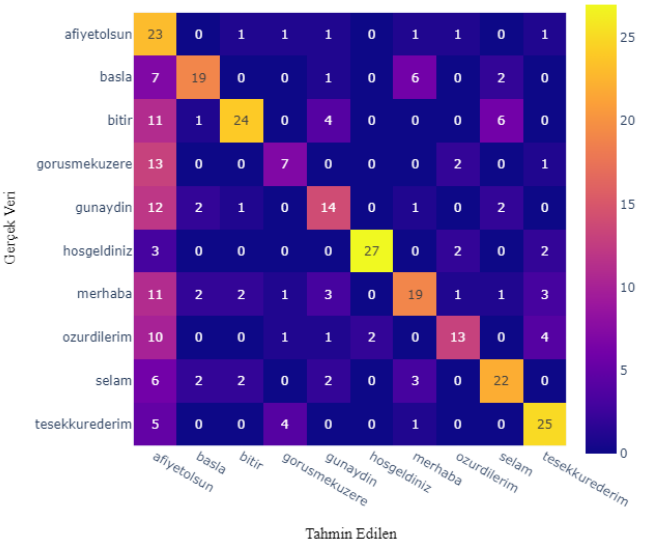


Şekil 15 Ayrık Dudak Görüntüleriyle Eğitilmiş Modelin Sınıflandırma Raporu

B. *Birleştirilmiş Dudak Görüntülerine Ait Eğitim Sonuçları*

Birleştirilmiş dudaklar kullanılarak gerçekleştirilen CNN model eğitimi, erken durdurma ile 42 epokta sona ermiştir (bkz. Şekil 16). Bu nedenle eğitimin daha da devam etmesi halinde düşük test doğruluğu elde etmek kaçınılmaz olmaktadır. Benzer şekilde, ayrık dudak görüntülerinin yer aldığı veri setinde olduğu gibi, birleştirilmiş dudak görüntülerinin sonuçlarında da hatalı tahminler "afiyet olsun" sınıfında toplanmaktadır (Şekil 17). Bunun dışında, tahminlerin genellikle "başla", "hoş geldiniz", "selam" ve "teşekkür ederim" sınıflarında yüksek olduğu ve diğer sınıflarla karışmadığı görülmektedir. Karmaşıklık matrisinde görüldüğü üzere, sınıflandırma raporu grafiğinde (Şekil 18) "afiyet olsun" sınıfının kesinlik (precision), duyarlılık (recall) ve f1-skor (f1-score) değerlerinin düşük olduğu görülmektedir. Diğer sınıfların doğruluk yüzdelerini yorumlamak için, dudakların ayrık olarak kullanıldığı eğitimle karşılaştırıldığında daha dengeli sonuçlar görülmektedir.

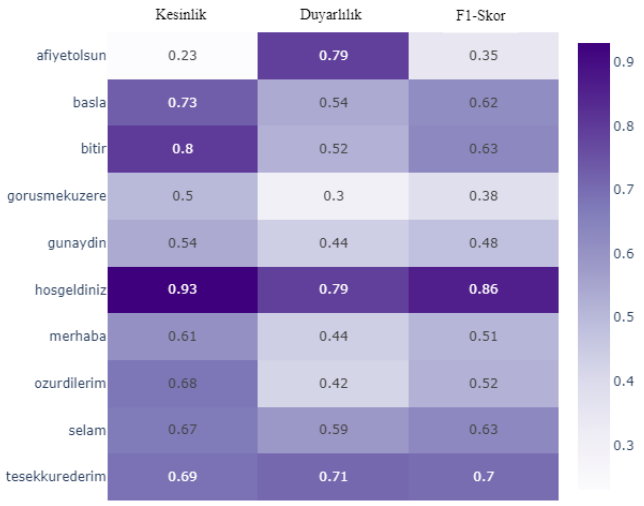
Şekil 16 Birleştirilmiş Dudakların Epok Başına Düşen Eğitim ve Doğrulama Doğruluğu



Şekil 17 Birleştirilmiş Dudak Görüntüleriyle Eğitilmiş Modelin Karmaşıklık Matrisi

Tablo 4. İki CNN Modeline Ait Doğruluk Değerleri ve Eğitim Süreleri

	Doğruluk	Eğitim Süresi
Birleştirilmiş Dudaklar	%55.94	15 saniye
Ayrık Dudaklar	%61.44	32 dakika



Şekil 18 Birleştirilmiş Dudak Görüntüleriyle Eğitilmiş Modelin Sınıflandırma Raporu

C. *Tartışma*

Toplam deney sonuçları karşılaştırıldığında, doğruluk oranı birleştirilmiş dudaklar için %55.94 ve ayrık dudaklar için %61.44'tür (bkz. Tablo 4). Süreler ise sırasıyla 15 saniye ve 32 dakikadır. Ayrık dudak görüntülerinde yüksek doğruluğun nedeni, daha net ve yüksek çözünürlüklü ayrık görüntülerin kullanılmasıdır. Ayrık dudakların eğitim süresi uzun olduğundan, hesaplama maliyeti açısından daha külfetli olarak değerlendirilebilir; ancak yüksek doğruluğu nedeniyle performans açısından tercih edilebilir. Görüntü temsili açısından girdilerden birinde 50x50 boyutunda 15 adet görüntü kullanılırken, diğerinde 60x200 boyutunda 1 adet görüntü kullanılmaktadır. Eş zamanlı tahmin gerektiren durumlarda, birleştirilmiş dudakların kullanıldığı temsili kullanmak daha uygun olacaktır, ancak doğru tespit önemli olduğunda, ayrık dudakları kullanan CNN modelini tercih etmek uygundur. Her iki çalışmanın doğruluk oranları değerlendirildiğinde, hala bir artış potansiyeli olduğu görülebilir. Bu potansiyele ulaşmak için yapılabilecek iyileştirmelerden biri, veriyi çeşitlendirerek artırmak, yani daha fazla veri toplamaktır. Bir diğer seçenek ise veri artırma (data augmentation) tekniklerini kullanarak bu çalışmayı yürütmektir. Benzer çalışmalarla karşılaştırıldığında, kullanılan veri seti içeriklerinin oldukça zorlu olduğu görülmektedir. Bu yeni veri setinde, yüzler önden görüntülenmemektedir, bazı görüntüler çok karanlıkken diğerleri oldukça parlaktır ve arka planın çok karmaşık olması nedeniyle bazen yüzü doğru bir şekilde tespit etmek mümkün olmamaktadır.

V. DEĞERLENDİRME

Bu çalışmada, yeni bir Türkçe veri seti için bir CNN modeli önerilmektedir. Ayrıca, iki farklı girdi temsili ile doğruluk ve hesaplama maliyeti karşılaştırılmaktadır. Bu temsil yöntemlerinden ilkinde, ardışık dudak görüntüleri modelin girdisini ayrı ayrı oluştururken, diğerinde dudaklar birleştirilerek tek bir görüntü oluşturulmaktadır. Performans açısından, ayrık dudaklar daha iyi sonuç verirken, birleştirilmiş dudaklar zaman maliyeti açısından daha iyi performans göstermektedir. Ayrıca, doğal YouTube görüntülerinden toplanan Türkçe veri seti, diğer çalışmalara kıyasla gerçek dünya görüntülerine daha yakın olması nedeniyle zorludur. Literatürdeki çalışmalarda toplanan görüntüler, kontrollü bir ortam oluşturularak sabit bir arka plan ve sabit bir insan duruşu ile elde edilmiştir. Doğal videolar üzerinden otomatik dudak okuma, işitme engelli bireyler için otomatik altyazı oluşturma açısından da büyük önem taşımaktadır. Bu çalışmada, doğal video görüntülerinden dudak okuma yaparak bir CNN modeli önerilmektedir. Ural-Altay dillerinde doğal dil ve dudak okuma çalışmaları sınırlı olduğundan, bu alana özgün bir çalışma ile katkıda bulunulmaktadır. Gelecek çalışmalarda daha geniş bir veri seti toplanarak dudak tespiti ile daha yüksek doğrulukta sınıflandırma yapılması planlanmaktadır.

TEŞEKKÜR

Yazarların Katkıları

Veri düzenleme, metodoloji, görselleştirme, N.P.A.; denetim ve danışman, H.E.; proje yönetimi, A.B. Tüm yazarlar, makalenin yayımlanan versiyonunu okumuş ve kabul etmiştir.

Çıkar Çatışması Beyanı

Yazarlar arasında çıkar çatışması bulunmamaktadır.

Araştırma ve Yayın Etiği Beyanı

Yazarlar bu çalışmanın Araştırma ve Yayın Etiğine uygun olduğunu beyan etmektedirler.

Verinin Erişilebilirlik Durumu

Veri şu adreste mevcuttur:
https://data.mendeley.com/datasets/4t8vs4dr4v/1

KAYNAKLAR

[1] H. McGurk, J. MacDonald, "Hearing lips and seeing voices." *Nature*, 264, pp. 746–748, 1976.

[2] A. Gabbay, A. Ephrat, T. Halperin, S. Peleg, "Seeing through noise: Speaker separation and enhancement using visually-derived speech." *arXiv preprint arXiv:1708.06767,* 4, 2017.

[3] D. Stewart, R. Seymour, A. Pass, J. Ming, "Robust audio-visual speech recognition under noisy audio-video conditions." *IEEE transactions on cybernetics,* 44, pp. 175–184, 2013.

[4] F.S. Lesani, F.F. Ghazvini, R. Dianat, "Mobile phone security using automatic lip reading." in Proceedings of the 2015 9th International Conference on e-Commerce in Developing Countries: With focus on e-Business (ECDC). IEEE, 2015, pp. 1–5.

[5] S. Mathulaprangsan, C.Y. Wang, A.Z. Kusum, T.C. Tai, J.C. Wang, "A survey of visual lip reading and lip-password verification." in *Proceedings of the 2015 International Conference on Orange Technologies (ICOT)*. IEEE, 2015, pp. 22–25.

[6] S. Sengupta, A. Bhattacharya, P. Desai, A. Gupta, "Automated lip reading technique for password authentication." *International Journal of Applied Information Systems (IJAIS)* 2012, pp. 18–24.

[7] J. Son Chung, A. Senior, O. Vinyals, A. Zisserman, "Lip reading sentences in the wild." in *Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition,* 2017, pp. 6447–6456.

[8] Y.M. Assael, B. Shillingford, S. Whiteson, N. De Freitas, "Lipnet: Sentence-level lipreading." *arXiv preprint arXiv:1611.01599,* 2, 2016.

[9] A. Ephrat, T. Halperin, S. Peleg, "Improved speech reconstruction from silent video." in *Proceedings of the Proceedings of the IEEE*

International Conference on Computer Vision Workshops, 2017, pp. 455–462.

[10] A. Jaumard-Hakoun, K. Xu, C. Leboullenger, P. Roussel-Ragot, B. Denby, "An articulatory-based singing voice synthesis using tongue and lips imaging." in *Proceedings of the ISCA Interspeech 2016,* 2016, vol. 2016, pp. 1467–1471.

[11] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, B. Yvert, "Real-time control of an articulatory-based speech synthesizer for brain computer interfaces." *PLoS computational biology* 12, e1005119, 2016.

[12] A. Gabbay, A. Shamir, S. Peleg, "Visual speech enhancement." *arXiv preprint arXiv:1711.08789* 2017.

[13] A.B. Mattos, D.A.B. Oliveira, "Multi-view mouth renderization for assisting lip-reading." in *Proceedings of the Proceedings of the 15th International Web for All Conference,* 2018, pp. 1–10.

[14] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, R. Harvey, "Extraction of visual features for lipreading." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2002*,* 24, pp. 198–213.

[15] S.J. Cox, R.W. Harvey, Y. Lan, J.L. Newman, B.J. Theobald, "The challenge of multispeaker lip-reading." in *Proceedings of the AVSP*, 2008, pp. 179–184.

[16] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, T. Huang, "AVICAR: Audio-visual speech corpus in a car environment." in *Proceedings of the Eighth International Conference on Spoken Language Processing,* 2004. pp. 2489-2492.

[17] Y.W. Wong, S.I. Ch'ng, K.P. Seng, L.M. Ang, S.W. Chin, W.J. Chew, K.H. Lim, "A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities." *Pattern recognition letters* 2011, *32*, pp. 1503–1510.

[18] C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matejka, J. Cernocký, "Bi-Modal Person Recognition on a Mobile Phone: Using Mobile Phone Data" 2012 IEEE international conference on multimedia and expo workshops. IEEE, 2012.

[19] A. Rekik, A. Ben-Hamadou and W. Mahdi, "A new visual speech recognition approach for RGB-D cameras." in Proceedings of the International Conference Image Analysis and Recognition, Springer, 2014, pp. 21-28

[20] D. Estival, S. Cassidy, F. Cox, D. Burnham, "AusTalk: An Audio-Visual Corpus of Australian English." in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 3105-3109.

[21] D. Petroyska-Delacrétaz, S. Lelandais, J. Colineau, L. Chen, B. Dorizzi, M. Ardabilian, E. Krichen, M.A. Mellakh, A. Chaari, S.Guerfi, et al., "The iv 2 Multimodal Biometric Database (Including Iris, 2d, 3d Stereosopic, and Talking Face Data), and Applications and Systems", *IEEE Second International Conference on Biometrics: Theory, Applications and Systems, IEEE,* 2008, pp.1-7.

[22] J. Trojanová, M. Hrúz, P. Campr, M. Železny, "Design and Recording of Czech Audio-Visual Databsase with Impaired Conditions for Continuous Speech Recognition" in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08),* 2008, pp.1-5.

[23] S. Petridis, J. Shen, D. Cetin, M. Pantic, "Visual-only Recognition of Normal, Whispered and Silent Speech." in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASP).* IEEE 2018, pp. 6219-6223.

[24] K. Kumar, T. Chen, R.M. Stern, "Profile View Lip Reading", in Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07), 2007, pp. IV-429.

[25] T. Afouras, J.S. Chung, A. Senior, O. Vinyals, A. Zisserman, "Deep Audio-Viusal Speech Recognition" in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2018 paper 44.12, pp. 8717-8727.

[26] A. Fernandez-Lopez, F.M. Sukno, "Survey on Automatic Lip-Reading in the Era of Deep Learning" Image and Vision Computing 78, pp.53-72.

[27] Z. Kang, R. Horaud, M. Sadeghi, "Robust face frontalization for visual speech recognition." in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2485-2495.

[28] G. Zhao, M. Barnard, M. Pietikainem, "Lipreading with local spatiotemporal descriptors." *IEEE Transactions on Multimedia*, vol. 11, pp. 1254-1265, 2009.

[29] M. Gurban, J.P. Thiran, "Information theoretic feature extraction for audio-visual speech recognition." *IEEE Transactions on Signal Processing*, vol. 57, pp. 4765-4776, 2009.

[30] Ü. Atilla, F. Sabaz, "Turkish lip-reading using Bi-LSTM and deep learning models." *Engineering Science and Technology, an International Journal*, p. 101206, 2022.

[31] A. Garg, J. Noyola, S. Bagadia, "Lip reading using CNN and LSTM" Univ. of Stanford, CS231, Tech. Rep. 2016.

[32] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K, Long, S. Shan, X. Chen, "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild." in *Proceedings of the 2019 14th IEEE International Conference on Automatic face & gesture recognition (FG 2019). IEEE*, 2019, pp.1-8.

[33] A. Yargıç, M. Doğan, "A lip reading application on MS Kinect camera." in *Proceedings of the 2013 IEEE INISTA*. IEEE, 2013, pp. 1-5.

[34] I. Fung, B. Mak, "End-to-end low-resource lip-reading with maxout CNN and LSTM." in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2018, pp. 2511-2515.

[35] T. Ozcan, A. Basturk, "Lip reading using convolutional neural networks with and without pre-trained models." *Balkan Journal of Electrical and Computer Engineering*, vol. 7, pp.195-201, 2019.

[36] D.K. Margam, R. Aralikatti, T. Sharma, A. Thanda, S. Roy, S.M. Venkatesan, et al., "LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models." *arXiv preprint arXiv:1906.12170*, 2019.

[37] S. Petridis, Z. Li, M. Pantic, "End-to-end visual speech recognition with LSTMs." in *Proceedings of the 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP).* IEEE, 2017, pp.2592-2596.

[38] B. Martinez, P. Ma, S. Petridis, M. Pantic, "Lipreading using temporal convolutional networks." in *Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* IEEE, 2020, pp.6319-6323.

[39] D. Kastaniotis, D. Tsourounis, A. Koureleas, B. Peev, C. Theoharatos, S. Fotopoulos, "Lip Reading in Greek words at unconstrained driving scenario." in *Proceedings of the 2019 10th International Conference on Information Intelligence Systems and Applications (IISA).* IEEE, 2019, pp.1-6.

[40] A.M. Sarhan, N.M. Elshennawy, D.M. Ibrahim, "HLR-net: a hybrid lip-reading model based on deep convolutional neural networks." *Computers, Materials & Continua* vol. 68, pp.1531-1549, 2021.

[41] J. Ting, C. Song, H. Huang, T. Tian, "A Comprehensive Dataset for Machine-Learning-based Lip-Reading Algorithm." *Procedia Computer Science* vol. 199, pp. 1444-1449, 2022.

[42] Y. Fu, Y. Lu, R. Ni, "Chinese Lip-Reading Research Based on ShuffleNet and CBAM." *Applied Sciences* vol. 13, 2023.

[43] N. Deshmukh, A. Ahire, S.H. Bhandari, A. Mali, K. Warkari, "Vision based Lip Reading System using Deep Learning." in *Proceedings of the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI).* IEEE, 2021, pp. 1-6.

[44] A. Berkol, T. Tümer-Sivri, N. Pervan-Akman, M. Çolak, H. Erdem, "Visual Lip Reading Dataset in Turkish." *Data*, vol. 8, 2023.

[45] A. Berkol, N. Pervan-Akman, H. Erdem, "Türkçe Günlük Kelime ve İfadeler Kullanarak CNN ve LSTM ile Görsel Konuşma Tanıma." *International Journal of Multidisciplinary Studies and Innovative Technologies 8.2: 69-75.*

[46] A. Berkol, N. Pervan-Akman, H. Erdem, "Lip reading multiclass classification by using dilated CNN with Turkish dataset." *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET). IEEE, 2022.*