International Journal of Multidisciplinary Studies and Innovative Technologies

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

2024, 8 (2), 69-75
DOI: 10.36287/ijmsit.8.2.4
Received: July 31, 2024; Accepted: September 13, 2024

Türkçe Günlük Kelime ve İfadeler Kullanarak CNN ve LSTM ile Görsel Konuşma Tanıma

Nergis Pervan Akman[1*], Talya Tümer Sivri[2], Ali Berkol[3], Hamit Erdem[4]

[1*] *Defence and Information Systems, BITES, Ankara, TURKEY (nergis.pervan@bites.com.tr) (ORCID: 0000-0003-3241-6812)*
[2] *Informatics Institute, Middle East Technical University, Ankara, TURKEY (talya.tumer@gmail.com) (ORCID: 0000-0003-1813-5539)*
[3] *Defence and Information Systems, BITES, Ankara, TURKEY (ali.berkol@bites.com.tr) (ORCID: 0000-0002-3056-1226)*
[4] *Electrics and Electronics Department, Başkent University, Ankara, TURKEY (herdem@baskent.edu.tr) (ORCID: 0000-0003-1704-1581)*

Türkçe Özet – Dudak okuma; el hareketleri, jestler ve yüz ifadeleri gibi konuşma örüntülerini, hareketlerini ve mimiklerini değerlendirmek amacıyla bir konuşmacının yüzünü incelemek olarak tanımlanmaktadır. Bilgisayarlara dudak okuma yeteneği kazandırma çalışmaları, derin öğrenmede sınıflandırma ve örüntü tanıma alanında büyüyen bir araştırma alanıdır ve günümüzde hâlâ çözülmesi gereken açık problemler barındırmaktadır. Son yıllarda, farklı dillerde konuşmayı metne dönüştürmek ve sınıflandırmak için çeşitli yöntemler geliştirilmiş ve uygulanmıştır. Ayrıca, çoğu yöntemde çok modlu veriler, yani konuşma ve görüntü verileri birleştirilmiştir. Bu çalışma, görüntülerle yeni Türkçe dudak okuma verileri sağlamayı ve Türkçe günlük kelimeler için yüksek doğrulukta bir sınıflandırma yöntemi sunmayı amaçlamaktadır. Kullanılan veriler, YouTube platformundan toplanmıştır. Bu zorlu verilerle, günlük kelimeleri ve ifadeleri sınıflandırmak için Evrişimli Sinir Ağı (Convolutional Neural Network – CNN) ve Uzun Kısa-Süreli Bellek (Long Short-Term Memory – LSTM) eğitilmiştir. Birçok deney sonucuna göre, CNN modeli daha iyi performans göstermiştir. Çoklu model verileri kullanmadan yalnızca görüntüler kullanmak, belleğin yorgunluğunu önler ve hesaplama süresini azaltır. Ayrıca, literatürde sınırlı bir çeşitlilik olduğundan, bu çalışma çok sınıflı Türkçe bir veri kümesi sunmaktadır.

Anahtar Kelimeler *– dudak okuma, çoklu sınıf sınıflandırma, Türkçe veri kümesi, derin öğrenme, konuşma tanıma*

Atıf: Pervan Akman, N., Tümer Sivri, T., Berkol A., Erdem H., (2024). Türkçe Günlük Kelime ve İfadeler Kullanarak CNN ve LSTM ile Görsel Konuşma Tanıma. International Journal of Multidisciplinary Studies and Innovative Technologies, 6(2): 69-75.

Visual Speech Recognition Using CNN and LSTM with Turkish Daily Words and Phrases

Abstract – Contemplating a speaker's face to evaluate speech patterns, movements, gestures, and expressions can be described as lip reading. Gaining the ability to lip reading to computers is a growing research area and has open problems for classification and pattern recognition in deep learning. In the last years, various methods have been developed and applied in different languages to classify and convert speech to text. Moreover, most methods have combined multi-model data, i.e., speech and image. This study aims to provide new Turkish lip-reading data with only images and provide a high-accuracy classification method for Turkish daily words. Data was collected from the YouTube platform. Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models were trained to classify daily words and phrases with this challenging data. According to numerous experiment results, the CNN model worked better. Using only images, not multi-modal data, prevents the memory from fatigue and decreases the computation time. Furthermore, we provide a multiclass dataset in Turkish since there is a narrow variety in the literature.

Keywords *– lip reading, multiclass classification, Turkish dataset, deep learning, speech recognition*

Citation: Pervan Akman, N., Tümer Sivri, T., Berkol A., Erdem H., (2024). Visual Speech Recognition Using CNN and LSTM with Turkish Daily Words and Phrases. International Journal of Multidisciplinary Studies and Innovative Technologies, 8(2): 69-75

I. GİRİŞ

Beyin-bilgisayar arayüzleri (BBA), kullanıcı ve bilgisayar arasındaki yazılım bileşenlerine odaklanarak en işlevsel tasarım ve teknoloji uygulamalarını geliştirmeyi amaçlayan bir araştırma alanıdır. İnsan beyni ve bilgisayarlar, sinyaller aracılığıyla görsel örüntüleri yakalayabilir, öğrenebilir ve bu örüntüleri önceki deneyimlere dayalı olarak anlamlı sonuçlar çıkarmak için işleyebilir. Görsel konuşma tanıma, diğer adıyla dudak okuma, ses ve görsellerin BBA sistemlerinde veri olarak

kullanıldığı popüler bir araştırma alanıdır. Bir kişinin ne söylediğini sadece ağız hareketlerine bakarak anlamak, insanlar için oldukça karmaşıktır [1]. Dahası, insanların dudak okuma performansı oldukça düşüktür. Örneğin, işitme engelli ve işitme güçlüğü çeken yetişkinler, 30 tek heceli kelimeden oluşan küçük bir alt küme için sadece %17±12 ve 30 karmaşık kelime için %21±11 doğruluk oranına ulaşmaktadır [2]. Ayrıca, dudak okumanın verimli bir şekilde gerçekleştirilebilmesi için önemli olan bir başka konu da konuşmacılar arasındaki mesafedir. Tejedor'e göre [3], önerilen mesafe 50 santimetre ile 3 metre arasındadır.

Bazı çalışmalar [6], [9], çok modlu verilerle dudak okuma üzerine yoğunlaşmaktadır. Çok modlu verilerle çalışmanın avantajları olsa da, önemli dezavantajları da bulunmaktadır. Özellikle birçok kişinin bulunduğu kalabalık günlük yaşam ortamlarından gelen ses kaynakları söz konusu olduğunda, veriden gürültüyü ayırmak zor bir problemdir. Ses verisini devre dışı bırakmak, günlük yaşam uygulamalarında dudak okuma için daha doğru modellerin geliştirilmesine yardımcı olacaktır. Ayrıca, hem görsel hem de ses verilerini kullanmak, aşırı veri kullanımı ve daha uzun eğitim süresi gerektirir. Derin öğrenme modellerini eğitirken bellek kullanımını dikkate almak önemlidir.

Ses-görüntü tabanlı dudak okuma, dikkate değer derecede iyi sonuçlar göstermiş olsa da, yalnızca görüntü tabanlı dudak okuma da etkinliğini kanıtlamıştır [10], [12], [17]. Tüm derin öğrenme uygulamaları gibi, bunun da bazı zorlukları bulunmaktadır. Yalnızca görüntü verisi içerdiğinden, benzer dudak hareketlerine sahip sesleri ayırt etmedeki zorluklar önemli bir problemdir. Ayrıca, birden fazla kişi varsa, algoritma çoğu uygulamada yalnızca bir kişinin verisini işleyebildiği için, kimin konuştuğunu ve algoritmanın kimi dikkate alacağını ayırt etmek gerçek dünya uygulamalarında zor olacaktır. Ancak, yukarıda belirttiğimiz gibi, görüntülerdeki kişilerin bilgilerini ayırmak ses verilerine göre nispeten daha kolaydır. Ayrıca, gerçek dünya problemlerinde, beyaz gürültüyü iptal etmek bir diğer önemli sorundur. Benzer şekilde, bu durum ses için nispeten zordur.

Bu çalışmada, sınıflandırma başarı oranını artırmak amacıyla yalnızca görüntü tabanlı dudak okuma modeli sunulmakta ve Ural-Altay dil ailesinin bir parçası olan Türkçe için yeni bir görüntü dudak okuma verisi literatüre kazandırılmaktadır. İlerleyen bölümlerde, veri ön işleme aşamalarına, Evrişimli Sinir Ağı ve Uzun Kısa-Süreli Bellek kullanılarak yapılan modelleme deneyine ilişkin problemleri ele alıyoruz.

Çalışmanın literatüre katkıları 1) Yeni bir dudak okuma verisi ile CNN ve LSTM gibi sık kullanılan yaklaşımların ele alınması, 2) Ural-Altay dil ailesinin bir parçası olan Türkçe görüntü veri kümesinin literatüre kazandırılması, şeklindedir.

II. LİTERATÜR ARAŞTIRMASI

Son yıllarda, dudak okuma problemi, sadece engelli bireyler ve onların yakınları değil, aynı zamanda yapay zeka araştırmacılarının da ilgisini çekmektedir. Bu bağlamda, ilk olarak ana diller üzerine gerçekleştirilmiş birçok çalışmadan bahsedilebilir [11], [7]. Daha genel uygulamalar geliştirmek için literatürdeki dil çeşitliliğini genişletmek çok önemlidir. İkinci olarak, veri türleri ve dillere göre son teknolojiye sahip birçok ileri düzey makale bulunmaktadır. Daha iyi doğruluk için Haar Feature-Based Cascade sınıflandırıcı ve CNN ağı kullanılmıştır [4]. Doğruluğu artırmak için geniş bir çalışma yelpazesi bulunmaktadır. Chitu ve Rothkrantz [7] ağız ve açıklığın yükseklikleri, genişlikleri ve alanları gibi görsel özelliklerin geometrik bilgilerini vurgulamışlardır. Tanıma problemi için Hidden Markov Model (HMM) kullanmışlardır. Articulated Feature Extraction yöntemlerinin kullanıldığı başka bir uygulamada da kısa cümlelerin tanınması için Dynamic Bayesian ağı ve sınıflandırma için Destek Vektör Makinesi (Support Vector Machine – SVM) kullanılmıştır [8]. Yan yüzeyden geometrik bilgi kullanan başka bir uygulama da HMM kullanmıştır [9]. Üst ve alt dudak konumlarından çıkarılan iki çizgi arasındaki açı, Lip Contour Features (LCGFs) olarak adlandırılmıştır. Yazarlar, dudak alanını tespit eder, dudakların merkez noktasını çıkarır ve LCGF adımları olarak dudak çizgilerini ve dudak açısını belirler.

Fenghour vd. [10], sinir ağları ve özellik çıkarma üzerine odaklanan çeşitli yöntemleri karşılaştırmak için iyi bir derlemedir. Yazarların en önemli çıkarımı Dikkat Dönüştürücülerinin (Attention-Transformers) ve Zamansal Evrişim Ağlarının (Temporal Convolutional Networks) Tekrarlayan Sinir Ağlarına (Recurrent Neural Networks – RNN) karşı avantajlarıdır. Çalışmada hem görsel-işitsel verilere hem de yalnızca görsel verilere odaklanmışlardır. Ayrıca, harf tabanlı, kelime tabanlı ve cümle tabanlı yöntemlerin İngilizce, Arapça, Çince ve Almanca gibi çeşitli dilleri kapsadığını belirtmişlerdir. Ozcan ve Basturk [5] AvLetters veri kümesinde AlexNet ve GoogleNet'in önceden eğitilmiş CNN modelini kullanmışlardır. Çalışmada veri boyutunu artırmak için veri artırma teknikleri kullanılmıştır. Makalede kullanılan teknikler, "gaussian", "salt and pepper" ve "speckle" ile gürültü ekleme, "unsharp" ile keskinleştirme ve "median" filtreleme ile yumuşatma şeklindedir. Lu ve Li [13] rakamların sınıflandırılması için yeni bir ağ önermişlerdir. Veriler, 3 kadın ve 3 erkek konuşmacının 100 defaya kadar telaffuz ettiği 0'dan 9'a kadar sayıları içermektedir. Uzamsal özellikleri çıkarmak için VGG19 ağı kullanılırken, zaman özelliklerini çıkarmak için Dikkat Tabanlı (Attention Based) LSTM kullanılmıştır.

Zamansal Konvolüsyonel Ağlar, LSTM'ye bir alternatiftir [14]. Martinez vd. [15] kelime düzeyinde sınıflandırma için Çok Ölçekli Zamansal Evrişim (Multi-Scale Temporal Convolution) yöntemini sunmuşlardır. Yalnızca işitsel, görsel-işitsel ve yalnızca görsel veriler üzerinde deneyler yapmışlardır. Amit vd. [16] sınıflandırma için CNN ve LSTM kullanılmış ve IMDB ve Google Görseller'den aldıkları ünlü insan yüzleri üzerinde önceden eğitilmiş VGGNet'i uygulamışlardır. Görüntüleri birleştirmek ve LSTM'den zamansal bilgi çıkarma işlemi, onların katkısı olmuştur.

Chung vd. [6] ağız hareketlerinin videolarını karakterlere dönüştürmeyi öğrenmek için Watch, Listen, Attend and Spell (WLAS) ağı geliştirilmiştir. Yalnızca görseller için çalışan WAS, WLAS modelinin bir parçasıdır. Ayrıca, eğitim süresini azaltmak ve aşırı uyumu önlemek için bir "curriculum learning strategy" önermişlerdir. Ek olarak İngiliz televizyonundan alınan 100.000'den fazla doğal cümle içeren Lip Reading Sentences (LRS) veri kümesi, görsel konuşma tanıma uygulamaları için yayınlanmıştır. LipNet [17], uçtan uca cümle ve ifade düzeyinde tahminler yapmak üzere geliştirilmiş ve eğitilmiştir. Model, karakter düzeyinde çalışmakta olup, uzay - zamansal CNN'ler, RNN'ler ve bağlantısal zamansal sınıflandırma (Connectionist Temporal Classification – CTC) kaybını kullanmaktadır [18]. Yazarlar, halka açık cümle düzeyinde bir veri kümesi olan GRID corpus üzerinde deneyler yapmışlardır [19].

LipType [12] ileri hız ve doğruluk için geliştirilmiş bir diğer modeldir. Yazarlar ayrıca zayıf ışık koşulları altında model sonuçlarının iyileştirilmesine de katkıda bulunmuşlardır. İlk aşama olarak yüz hatlarının Kalman Filtrelemesi, 3D-CNN ve 2D SE-ResNet ile düzeltilmesini içeren uzay-zamansal özellik çıkarma yöntemi (spatiotemporal feature extraction method) kullanılmış olup daha sonra CTC'li Çift Yönlü Geçitli Tekrarlayan Sinir Ağları (Bidirectional Gated Recurrent Neural Networks with CTC) uygulanmıştır.

Jittakoti ve Phumeechanya [23] CNN ve LSTM kullanarak Temporal Keyframe tekniği yoluyla dudak okuma performansını iyileştirmeye yönelik bir yöntem sunmaktır. Çalışmada kullanılan veri seti, tam ve yarım dudak görüntü verilerini içermektedir ve çalışma sırasında 3 kare, 5 kare ve 10 kare olmak üzere 3 gruba ayrılmıştır. Görülmemiş test seti değerlendirildiğinde, 10 kare, tam dudak görüntü veri seti için %87.9 doğruluk ve yarım dudak görüntü veri seti için %86.8 doğruluk ile en iyi tanıma oranını sağlamış ve karşılaştırılabilir bir performans sergilemiştir.

Shashidhar vd. [24] MIRACL VC1 veri seti için görsel konuşma tanıma amacıyla LSTM ve 3D CNN hibrit modeli önermiştir. Önerilen çalışma, hibrit modeli içermesiyle önceki çalışmalardan ayrılmaktadır. Sadece 3D CNN modelinin test doğruluğu %79 ve LSTM modelinin doğruluğu %85 iken, hibrit modelin eğitim, test ve doğrulama setlerinin doğruluğu sırasıyla %98, %85 ve %86 olmuştur.

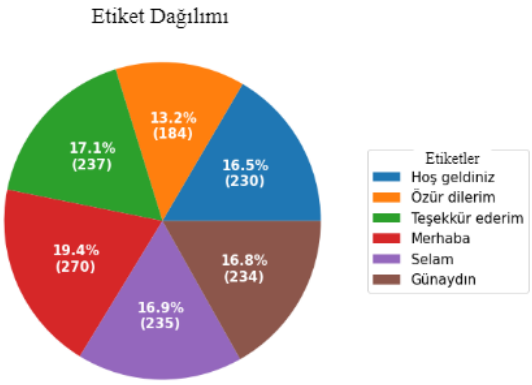
Pourmousa ve Özen [25] evrişimli sinir ağları kullanılarak Türkçe dilinde dudak okuma işlemi gerçekleştirilmiştir. Türkçe 20 sayının söylendiği video verileri kullanılmıştır. Bu videolar, sayıların dudak hareketlerini içerir. Veri seti, dudak hareketlerinin ve yüz ifadelerinin detaylı bir şekilde gösterildiği videoları kapsar. Bu kapsamda kişilerden sayıların videosunu (61 video) çekip göndermeleri istenmiş ve onun yanı sıra YouTube'tan 9 video toplanmıştır. CNN tabanlı yöntem, dudak okuma sistemlerinde yüksek doğruluk sağlar ve Türkçe sayıların tanınmasını geliştirir. Türkçe sayıların dudak hareketlerinden tanınması için CNN tabanlı yaklaşımlar üzerine yenilikçi bir bakış açısı sunar.

Exarchos vd. **Error! Reference source not found.** 3D Evrişimli Sinir Ağları ve Uzun Kısa Vadeli Hafıza ağlarını birleştiren yenilikçi bir yaklaşım önerilmektedir. Araştırmada; farklı konuşma kalıplarını, konuşmacıları ve çevresel koşulları kapsayan titizlikle oluşturulmuş "MobLip" adlı bir veri setinden yararlanmaktadır. Çalışmada kullanılan veri seti, dudak hareketlerini içeren video verileri içermektedir. 3D CNN'ler tarafından çıkarılan mekansal bilgiler ile LSTM'ler tarafından yakalanan zamansal dinamikler arasındaki sinerji, aydınlatma değişikliklerine ve konuşmacı çeşitliliğine karşı dayanıklılık sergileyerek %87,5'e varan bir doğruluk oranına ulaşarak etkileyici sonuçlar vermektedir.

III. Materyal ve Metot

A. *Veri Toplama*

Dudak okuma uygulamalarına yönelik literatürde, İngilizce, Almanca gibi çeşitli dillerde yalnızca görüntü, yalnızca ses ve ses-görüntü verileri için çok sayıda çalışma bulunmaktadır. Bu çalışmada Türkçe için yeni bir kelime düzeyinde ve ifade düzeyinde çok sınıflı bir veri kümesi [20] öneriyoruz. Veri kümesi, "selam", "merhaba", "günaydın" olmak üzere üç kelime sınıfı ve "hoş geldiniz", "özür dilerim", "teşekkür ederim" olmak üzere üç kelime öbeği ile toplam altı sınıf içermektedir. Her sınıf yaklaşık olarak aynı sayıda veri içermektedir; bkz. Şekil 1. Veriler Youtube platformu üzerinden oluşturulmuştur. İlgili kelimeler söylenirken kısa videolar kaydedilerek daha sonra, kelimenin başladığı ve bittiği dudak hareketlerine göre çerçeveleri belirlenmiştir. Veri kümesini toplarken konuşmacı sayısı, konuşmacı ile kamera arasındaki mesafe ve dudaklar ile kamera arasındaki açı gibi açılardan veri çeşitliliğine özellikle dikkat edilmiştir. Sentetik verilerle eğitilen modeller iyi tahmin ve sınıflandırma sonuçlarına sahip olsa da, eğitilen modelde kullanılan verilerin toplandığı ortam fazlasıyla kontrollüdür. Bu nedenle mümkün olduğunca gerçek dünyaya benzetmeye çalışılmıştır. Her sınıfın çerçeve sayılarına ilişkin dağılımlarına bakmak önemlidir, çünkü bu hem konuşmacının konuşma hızı hem de kelimenin telaffuz edilme uzunluğu göz önüne alındığında, çerçeve sayıları açısından zaman zaman dengeli olabilirken zaman zaman çerçeve sayıları farklılaşmaktadır. Her sözcükteki çerçeve sayıları önemlidir ve model çerçeve sayılarını etkileyebileceğinden sözcüğün ne kadar hızlı telaffuz edildiğine bağlıdır. Şekil 2'de bazı sınıfların dengeli veri sağlayan normal dağılımlara sahip olduğu görülmektedir. "selam", "özür dilerim" ve "tüm sınıflar" sağa eğik dağılımlardır, bkz. Şekil 2e,2d,2g.



Şekil. 1. Sınıfların Etiket Dağılımı

Yani verilerin çoğu ortalamanın altında çerçeve sayısına sahiptir. Diğerleri tahmin edilebilir normal dağılımlardır.

B. *Veri Ön İşleme*

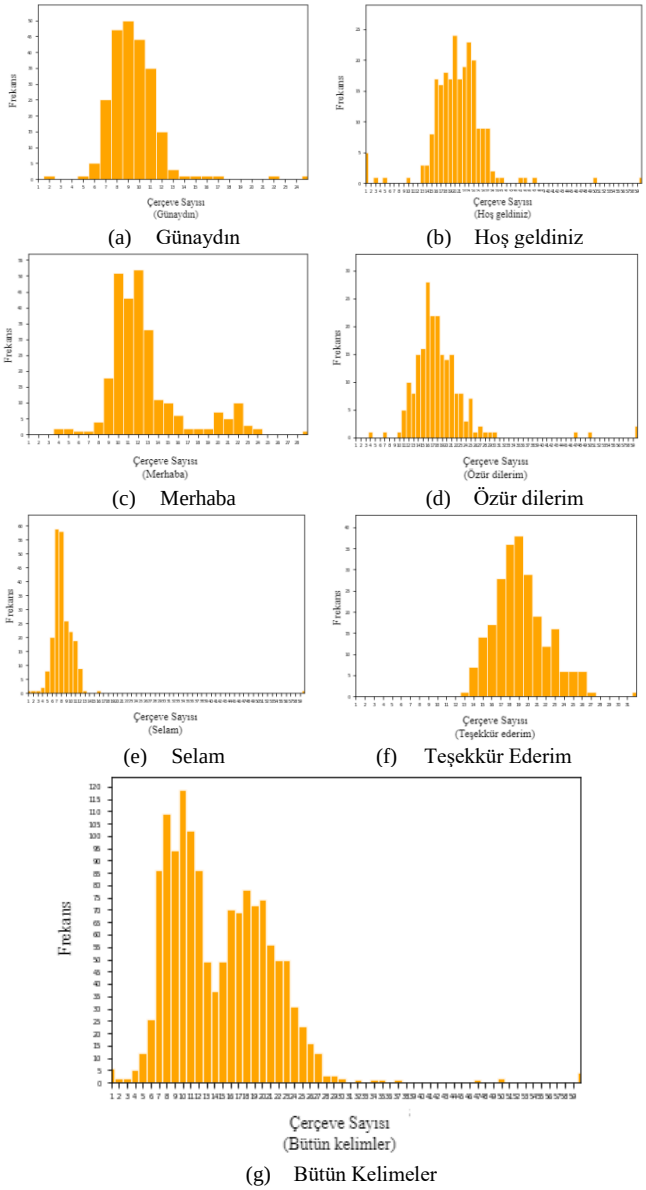
Toplanan veriler oldukça ham görüntüler olduğundan, bazı görüntü düzenlemelerine ihtiyaç duyulmuştur. İlk ve en temel işlem, görüntülerin gri tonlamaya dönüştürülmesidir. Görseldeki kişinin gözleri, burnu veya görüntüdeki diğer kısımları dudak okuma problemi için gerekli olmadığından, öncelikle yüzü, ardından dudak bölümleri kesilerek modele dahil edilir. Daha sonra, OpenCV [21] ve dlib [22] kütüphaneleri kullanılarak görseldeki yüz ve ağız kesimi için yüz işaret noktaları tespit edilir (Bkz. Şekil 3). Son olarak her bir dudak görüntüsü sabit bir boyut olarak yeniden boyutlandırılır. Bu boyut, hesaplama maliyetini azaltmak amacıyla mümkün olduğunca küçük tutulur.

Her görüntü için yapılan ön işlemeye ek olarak, her örnek için kare sayısı sabit bir değere oturtulmuştur. Bazı kelimelerin video sekanslarındaki kare sayıları, her konuşmacının konuşma hızına ve kelimenin uzunluğuna göre değişiklik gösterdiğinden, tutarlılığı sağlamak amacıyla sabit bir boyuta getirilmiştir. Yapılan deneylerde, kelime uzunluklarına göre en iyi sonuçlar 15 değeri için elde edilmiştir. Eğer örneklerin

çerçeve sayısı 15'ten büyükse devam eden çerçeveler göz ardı edilir; 15'ten küçükse, boş çerçeveler ile doldurulmaktadır.

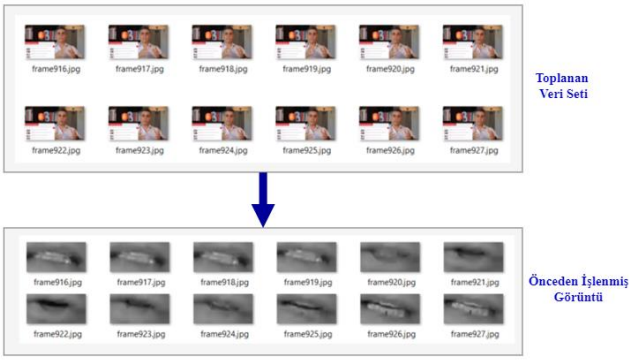
C. *Modeller*

Dudak okuma sınıflandırma problemi üzerinde CNN ve LSTM kullanarak çalışmalar gerçekleştirilmiştir. CNN, görüntü işleme problemlerinde en sık kullanılan yaklaşımlardan biridir. CNN mimarisi, ReLU aktivasyon fonksiyonunu kullanan iki evrişim katmanı ve ardından gelen evrişimlerde iki maksimum havuzlama (max pooling) katmanı içermektedir. Evrişim katmanında pencere boyutu 3, filtrelerin adım sayısı ise 2 olarak belirlenmiştir. Sınıflandırma kısmı, tam bağlantı (fully connected) katmanları ve aşırı uyumu (over-fitting) önlemek için kullanılan bırakma (dropout) katmanını içermektedir (Bkz. Şekil 4). Son olarak softmax katmanı, Türkçe'deki üç kelime ve üç kelime öbeği için bir skor döndürür.

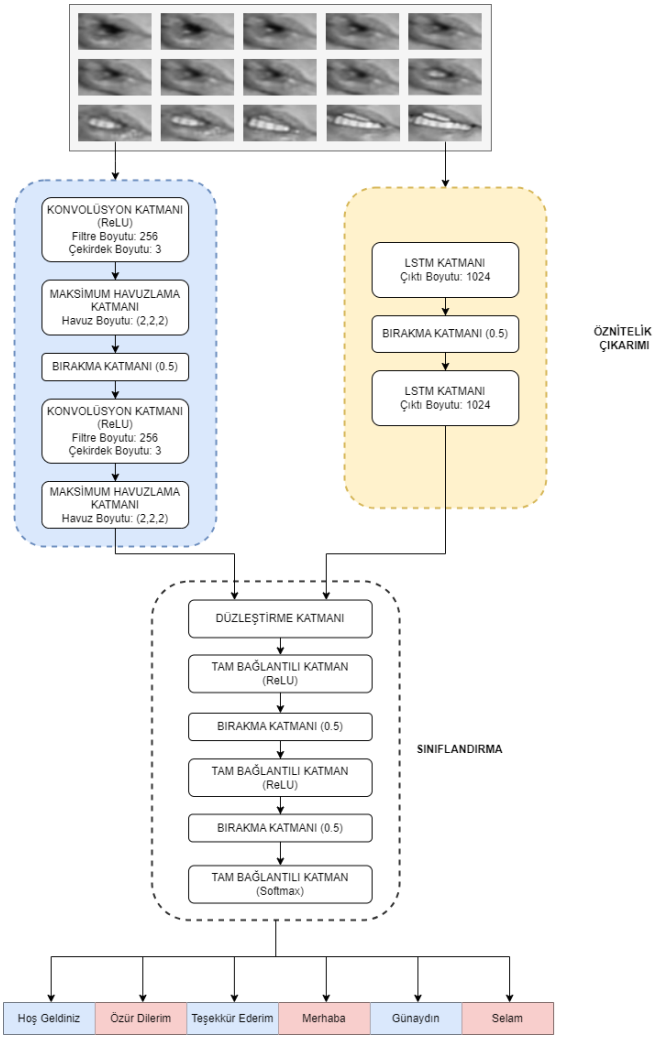


Şekil 2. Tüm Sınıflara Ait Çerçeve Sayıları

Önerilen derin öğrenme modelleri, sınıflandırma katmanında aynı mimariyi kullanırken, özellik çıkarım katmanında CNN ve LSTM kullanılarak oluşturulmuştur (Bkz. Şekil 4). Kullandığımız bir diğer derin öğrenme mimarisi de LSTM'dir. LSTM, zaman serisi yaklaşımlarında genel olarak çok iyi sonuçlar üretmektedir. Yapılan çalışmada dudak görselleri bir sekans oluşturduğundan LSTM'deki girdi, unutma ve çıktı kapıları sayesinde probleme uygun bir algoritmadır. Dudak okuma problemlerine zaman serisi yaklaşımı uygulandığından, teorik olarak LSTM kullanmak mantıklıdır. Temel LSTM modelimiz, 1024 çıktı boyutuna sahip iki LSTM katmanından oluşmaktadır.



Şekil 3. Uçtan Uca Veri Ön İşleme Örneği



Şekil 4. Model Mimarisi: Diyagramın mavi kısmı CNN modelinin özellik çıkarma katmanını; sarı kısım LSTM modelinin özellik çıkarma katmanını göstermektedir.

D. *Eğitim*

Eğitim süreci boyunca birçok değer için hiperparametre ayarlaması (hyperparameter tuning) deneysel olarak gerçekleştirilmiştir (bkz. Tablo 1). Deneysel çalışmalar, her bir örnekteki dudak görüntüleri üzerinde CNN katmanlarının filtre boyutu, LSTM çıktı katmanları ve boyutları, öğrenme oranı, giriş boyutu ve eğitime dahil edilecek çerçeve sayısı için farklı değerleri tekrar etmiştir. Ayrıca doğrulama kaybı değerinin iyileştirilmemesi durumunda erken durdurma stratejisinin eklendiği eğitimi de deneye dahil ettik. 1390 örnek, eğitim ve model testi için %70 eğitim, %15 test ve %15 doğrulama olarak bölünmüştür.

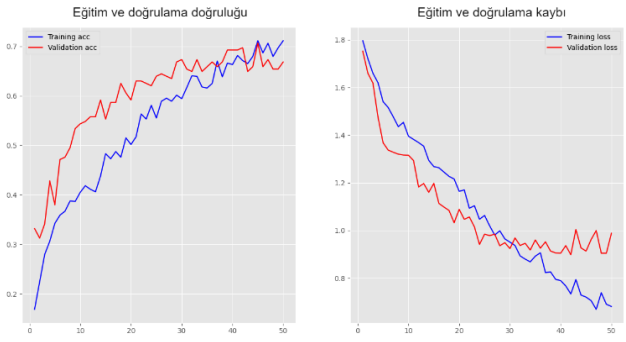
Tablo 1 CNN ve LSTM modelleri hiperparametreleri

	CNN	**LSTM**
Epok	50	65
Yığın Boyutu (Batch Size)	16	4
Öğrenme Oranı (Learning Rate)	2e-4	2e-4
Katman Sayısı	2	2
Filtre Boyutu	3x3	-
Bırakma katmanı	0.5	0.5
Aktivasyon Fonksiyonu	ReLU	ReLU
Optimizasyon Fonksiyonu	Adam	Adam
Kayıp Fonksiyonu (Loss Function)	Kategorik Çapraz Entropi (Categorical Cross Entropy)	Kategorik Çapraz Entropi (Categorical Cross Entropy)
Kelime Boyutu	15	15
Girdi Boyutu	50	50

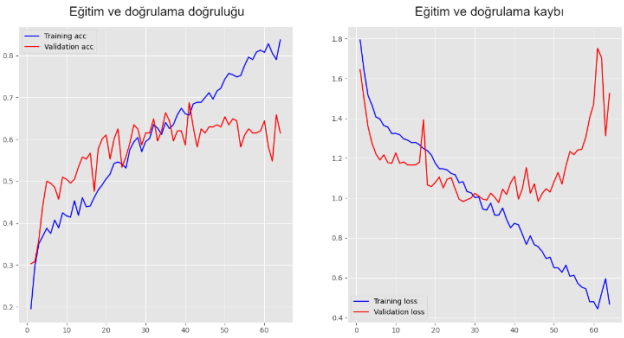
Çalışmalar Python dilinde CUDA 11.2 kullanılarak gerçekleştirilmiştir. Ekran kartı olarak NVIDIA GeForce GTX 1650 Ti 4GB kullanılmıştır.

IV. SONUÇLAR

Erken durdurma (early stopping) stratejisine sahip eğitilmiş CNN ve LSTM modelleri, CNN için 50 epok ve LSTM için 65 epok sonunda elde edilmiştir. Doğruluk ve kayıp grafiği eğitim sürecinden elde edilir. Şekil 5 ve 6'da görülebileceği gibi eğitimin daha fazla devam etmesi durumunda model verilerden daha fazla şey öğrenecektir.



Şekil 5. CNN Modelinin Eğitim ve Doğrulama Kaybı ve Doğruluğu



Şekil 6. LSTM Modelinin Eğitim ve Doğrulama Kaybı ve Doğruluğu

CNN ve LSTM modellerini, veri çeşitliliğini göz ardı etmemek adına, duyarlılık (recall), kesinlik (precision), f1-skor (f1-score) ve doğruluk (accuracy) performans metriklerini kullanarak değerlendirdik ve karşılaştırdık. Mikro doğruluk, modelin toplamda ne kadar doğru tahminde bulunduğunu ölçen bir metriktir. Mikro kesinlik, modelin pozitif olarak tahminlediği örneklerin ne kadarının gerçekten pozitif olduğunu ölçerken, duyarlılık modelin gerçek pozitifleri ne kadar iyi yakaladığını ölçer. Yani, gerçek pozitiflerin doğru tahminlere oranıdır. F1 skor ise, kesinlik ve duyarlılığın harmonik ortalamasıdır, yani iki metrik arasındaki dengeyi sağlar.

$$Mikro\ Doğruluk = \frac{Toplam\ doğru\ tahimler}{Toplam\ örnekler} \quad (1)$$

$$Mikro\ Kesinlik = \frac{Toplam\ gerçek\ pozitifler}{Toplam\ gerçek\ pozitifler + Toplam\ yanlış\ pozitifler} \quad (2)$$

$$Mikro\ Duyarlılık = \frac{Toplam\ gerçek\ pozitifler}{Toplam\ gerçek\ pozitifler + Toplam\ yanlış\ negatifler} \quad (3)$$

$$\text{Mikro F1} - \text{skor} = 2 \text{ x } \frac{\text{Mikro kesinlik x Mikro duyarlılık}}{\text{Mikro kesinlik+Mikro duyarlılık}} \quad (4)$$

Dudak okuma modellerinin 6 sınıfından elde ettiğimiz test doğrulukları (mikro), CNN ve LSTM için sırasıyla %60 ve %56 (Bkz. Tablo 2 ve Tablo 3). Ancak bazı kelimeler için tespit performansı genel doğruluktan daha iyi olurken bazı kelimeler için bu skor daha düşüktür. Örneğin, "özür dilerim" sınıfı için kesinlik skoru, CNN modeli için %74 ve LSTM modeli için %82 olup, bu değerler genel doğruluktan daha yüksektir. Buna karşın, "özür dilerim" sınıfı için duyarlılık değeri, CNN modeli için %59 ve LSTM modeli için %28 olup, bu değerler genel doğruluktan daha düşüktür. Ayrıca, bazı sınıflar için CNN ve LSTM modeli sonuçlarında skorun daha düşük ve daha yüksek olduğu durumlar mevcuttur. Örneğin, "özür dilerim" sınıfı için doğruluk, LSTM modeli için genel doğruluğa düşük iken, CNN modeli için genel doğruluğa denktir. Bu durum, farklı YouTube videolarından elde edilen veri kümesinin çeşitliliğinden ve sınıf örnek setinin farklı olmasından kaynaklanmaktadır. CNN ve LSTM modelleri toplamda 209 örnek ile test edilmiştir.

Tablo 1. CNN Model Sonuçları

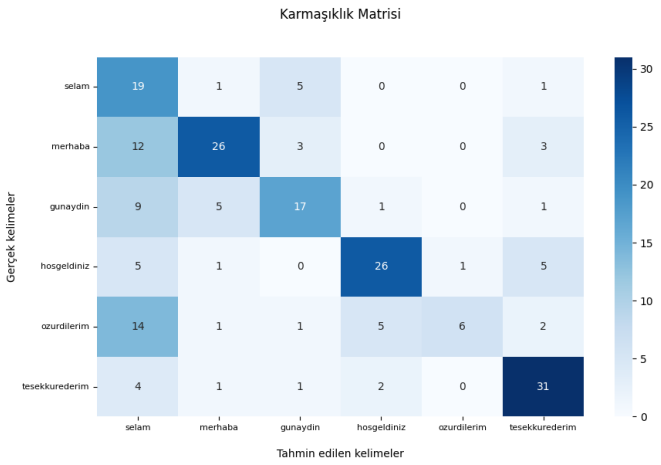
Kelime	Doğruluk	Kesinlik	Duyarlılık	F1-Skor	Boyut
Hoş geldiniz	0,73	0,30	0,73	0,43	26
Özür dilerim	0,59	0,74	0,59	0,66	44
Teşekkür ederim	0,51	0,63	0,52	0,57	33
Merhaba	0,68	0,76	0,68	0,72	38
Selam	0,20	0,86	0,21	0,33	29
Günaydın	0,79	0,72	0,79	0,76	39
Tüm Kelimeler	0,60	0,69	0,60	0,60	209

Tablo 2. LSTM Model Sonuçları

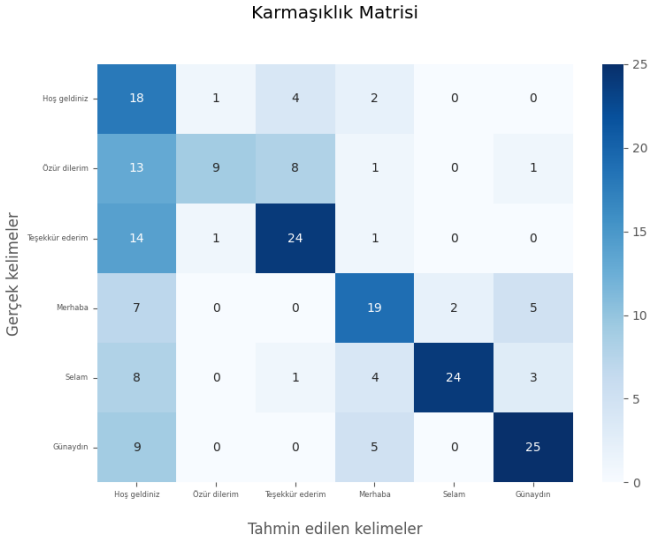
Kelime	Doğruluk	Kesinlik	Duyarlılık	F1-Skor	Boyut
Hoş geldiniz	0,72	0,26	0,72	0,38	25
Özür dilerim	0,28	0,82	0,28	0,42	32
Teşekkür ederim	0,60	0,65	0,60	0,62	40
Merhaba	0,57	0,59	0,58	0,58	33
Selam	0,60	0,92	0,60	0,73	40
Günaydın	0,64	0,74	0,64	0,68	39
Tüm Kelimeler	0,56	0,69	0,57	0,59	209

Şekil 7 ve 8'de her kelime için yapılan tahminler ve yanlış tahminle sonuçlanan kelimeler gösterilmektedir. Köşegendeki yoğunluğa bakıldığında çoğunlukla her kelime için iyi bir performansın olduğu görülmektedir. Yanlış tahmin edilen "günaydın" kelimesi için "merhaba" ve "selam" kelimelerine odaklanılmaktadır. "özür dilerim" ve "teşekkür ederim" yanlış tahmin edildiğinde, doğru tahminin "hoş geldiniz" olması gerektiği belirtilmiştir. Bu iki duruma göre yorum yapılırsa, kelimelerin ve ifadelerin kendi içlerinde tahmin karışıklığına neden olduğu söylenebilir.

Şekil 7. CNN Modelinin Karmaşıklık Matrisi



Şekil 8. LSTM Modelinin Karmaşıklık Matrisi



V. TARTIŞMA

Bu çalışmada, yeni bir Türkçe dudak okuma veri kümesinin kazandırılmasının yanı sıra bu veri ile yapılan CNN ve LSTM modellerinin değerlendirilmesi yapılmıştır. Gerçek dünyaya uyan bir model geliştirmeye çalışıldığından, bu veri kümesinin oluşturulmasındaki en önemli kısım, tamamen gerçek dünya verilerinden elde edilmesi oldu. Veri kümesi hem ön işleme hem de eğitim açısından zorlu olsa da, CNN modelinin LSTM'den daha iyi olduğu çok sınıflı sınıflandırma problemlerinde iyi bir sonuç elde edilmiştir. Ayrıca kelime ve cümleleri kendi aralarında sınıflandırmada yanlış pozitif ve yanlış negatif değerlerin daha yaygın olduğunu gördük ki bunun da en doğal sonuçlardan biri olduğunu düşünüyoruz.

Çalışma kapsamında yapılan değerlendirmeler için her bir kelime sınıfı test edilirken daha fazla veriye sahip olunması durumunda daha iyi sonuçlara ulaşması beklenmektedir.

VI. GELECEK ÇALIŞMALAR

Gelecek çalışmalarda, farklı ön işleme stratejileri uygulamayı ve yüz ve ağız bölgelerinin kesilmesi için yeni algoritmalar geliştirmeyi planlıyoruz. Böylece, yeni hibrit modellerle sınıflandırma skorunu artırmayı hedefliyoruz. Ayrıca, 10 kelime ve kelime öbeği sınıfı oluşturduk. Ancak, model yalnızca bunların 6'sı ile test edilmiştir. Gelecekteki diğer bir çalışma ise genişletilmiş veri kümesi ile yeni modeller denemek olacaktır.

KAYNAKLAR

[1] C. G. Fisher. “Confusions among visually perceived consonants.” Journal of Speech, Language, and Hearing Research, 11(4) pp. 796–804, Dec. 1968.

[2] R. D. Easton and M. Basala. “Perceptual dominance during lipreading”. Perception and Psychophysics, 32(6) pp.562–570, Nov. 1982.

[3] Cecilia Tejedor, A. Leer en los labios. Manual práctico para entrenamiento de la comprensión labiolectora. Madrid: CEPE, 2000.

[4] Shrestha, K. (n.d.). “Lip Reading using Neural Network and Deep learning.” 1802.

[5] T. Ozcan, and A. Basturk, “Lip Reading Using Convolutional Neural Networks with and without Pre-Trained Models.” Balkan Journal of Electrical and Computer Engineering, vol. 7(2) pp. 195-201, Apr. 2019.

[6] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild.” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6447-6456.

[7] Chitu, A., Rothkrantz, L. "Visual Speech Recognition Automatic System for Lip Reading of Dutch". Journal on Information Technologies and Control, vol. 7, no. 3, pp. 2-9, 2009.

[8] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, T. Darrell "Visual Speech Recognition with Loosely Synchronized Feature Streams," in Proceedings of the 10th International Conference on Computer Vision, 2005, pp.1424–1431.

[9] K. Iwano, T. Yoshinaga, S. Tamura, S. Furui. "Audio-Visual Speech Recognition Using Lip Information Extracted from Side-Face Images", Hindawi Publishing Corporation EURASIP Journal on Audio, Speech, and Music Processing vol. 2007, pp.1-9, 2007

[10] S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, "Deep learning-based automated lip-reading: A survey," IEEE Access, vol. 9, pp. 121184–121205, 2021.

[11] M. Faisal, and S. Manzoor, "Deep Learning for Lip Reading using Audio-Visual Information for Urdu Language". CoRR, 2018. DOI: https://doi.org/10.48550/arXiv.1802.05521

[12] L. Pandey and A. S. Arif, "LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model." in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, 2021 Article 1, pp. 1–19. DOI: https://doi.org/10.1145/3411764.3445565

[13] Y. Lu and H. Li, "Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory". Applied Sciences. 9(8) 1599. 2019. DOI: https://doi.org/10.3390/app9081599

[14] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv:1803.01271, 2018.

[15] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks." in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. pp. 6319-6323 https://doi.org/10.1109/icassp40776.2020.9053841

[16] G. Amit, J. Noyola, and S. Bagadia. "Lip reading using CNN and LSTM". Stanford University, CS231n project report, 2016.

[17] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas. "Lipnet: End-to-End Sentence-Level Lipreading." Dec. 2016. http://arxiv.org/abs/1611.01599

[18] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." in ICML, 2006 pp. 369–376.

[19] M. Cooke, J. Barker, S. Cunningham, and X. Shao. "An audio-visual corpus for speech perception and automatic speech recognition." The Journal of the Acoustical Society of America, vol. 120(5) pp. 2421–2424, 2006. DOI: https://doi.org/10.1121/1.22290.

[20] https://doi.org/10.17632/4t8vs4dr4v.1.

[21] OpenCV Team. (2024). *OpenCV: Open Source Computer Vision Library*. Version 4.6.0. https://opencv.org/

[22] King, D. E. (2009). *dlib: A C++ Library for Machine Learning*. Version 19.24. http://dlib.net/

[23] Jittakoti, A., & Phumeechanya, S. (2024, March). Temporal Keyframe Technique based on CNN and LSTM for Enhancing Lip Reading Performance. In 2024 12th International Electrical Engineering Congress (iEECON) (pp. 1-5). IEEE.

[24] Shashidhar, R., Shashank, M. P., & Sahana, B. (2023). Enhancing visual speech recognition for deaf individuals: a hybrid LSTM and CNN 3D model for improved accuracy. Arabian Journal for Science and Engineering, 1-17.

[25] Pourmousa, H., & Özen, Ü. (2022). LIP READING USING CNN FOR TURKISH NUMBERS. Journal of Business in The Digital Age, 5(2), 155-160.