Dudak Hareketlerinden Metin Çıkarma

AHMED SALIH | HASAN SAHVAN | MUHAMMED OSMAN

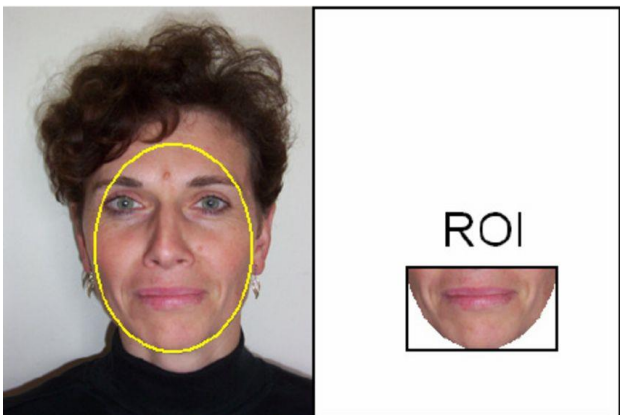
Özet

Bu çalışma, yalnızca görsel verilere dayalı olarak dudak hareketlerinden metin çıkarımı yapmayı amaçlamaktadır. Projede, dudak hareketlerinden gelen görsel verilerin analizi ve bu verilere dayalı olarak konuşma metni oluşturulması hedeflenmektedir. Önerilen sistemde Vision Transformer (ViT) ve GPT tabanlı dil modeli gibi güncel derin öğrenme yaklaşımları kullanılacaktır. Araştırma kapsamında, dudak hareketlerinden oluşan büyük bir veri seti toplanacak ve bu veri seti üzerinde derin öğrenme modelleri eğitilecektir. Çalışma, dudak okuma teknolojisinin geliştirilmesine katkı sunmayı, özellikle işitme engelli bireyler ve güvenlik uygulamaları gibi alanlarda kullanılabilir çözümler üretmeyi hedeflemektedir. Araştırma sonunda, konuşmanın metin olarak çıkarılmasında başarı oranı Karakter Hata Oranı (CER) ve Kelime Hata Oranı (WER) metrikleri kullanılarak değerlendirilecektir. Elde edilen bulgular, görsel verilere dayalı konuşma metni çıkarımında kullanılan yeni yöntemlerin etkinliğini ortaya koyacaktır [1].

***Anahtar kelimeler**: Dudak okuma, Vision Transformer, GPT, derin öğrenme, metin çıkarımı*

1.1 Giriş

Görsel verilerden metin çıkarımı, son yıllarda yapay zeka ve bilgisayarla görme alanında önemli bir araştırma konusu olmuştur. Özellikle dudak okuma (lipreading), sesin duyulmadığı veya sesli iletişimin mümkün olmadığı ortamlarda etkili bir iletişim yöntemi olarak ön plana çıkmaktadır. İşitme engelli bireyler için, dudak hareketlerini anlamlandırabilen bir yapay zeka modeli, günlük yaşamda karşılaşılan birçok engeli aşmalarına yardımcı olabilir. Dudak okuma teknolojisi, sesli iletişimden bağımsız olarak, sadece görsel verilerle anlamlı metinlerin oluşturulmasını sağlamaktadır [1].

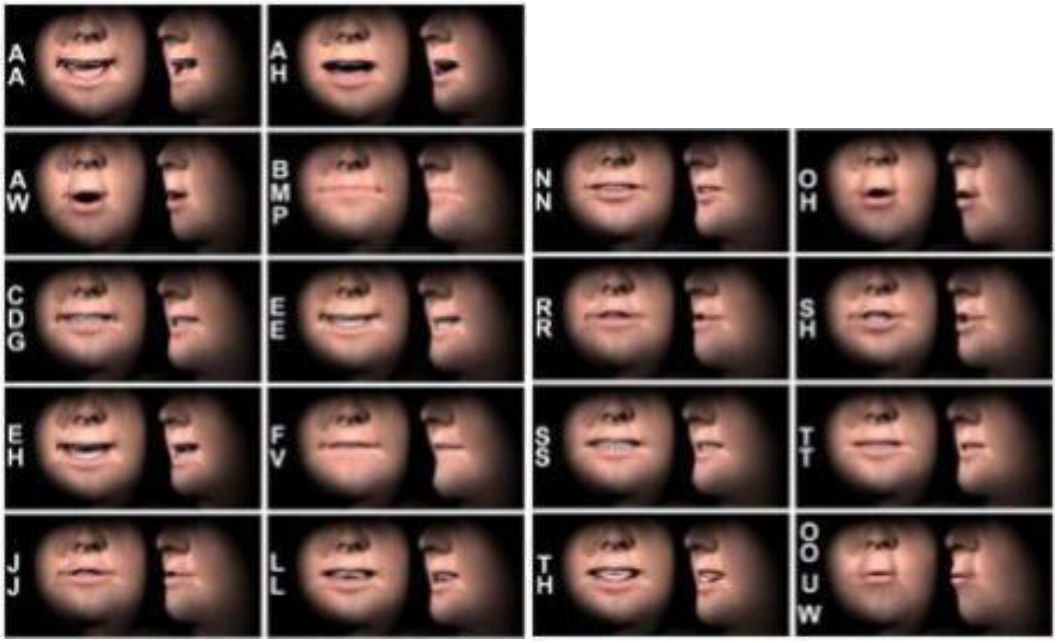


Şekil 1. LipReadAI

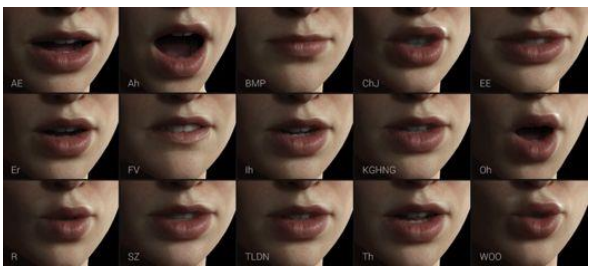
Yapay zeka destekli dudak okuma, bilgisayarla görme tekniklerini kullanarak, video verileri üzerinden dudak hareketlerini analiz eder ve bu hareketlerden dilsel ifadeler çıkarır. Bu teknoloji, ses kaynağını gözlemlemeden yalnızca görsel veri kullanarak iletişimi mümkün kılar [2]. Geliştirilen bu modeller, sadece işitme engelli bireyler için değil, aynı zamanda güvenlik ve gizlilik gerektiren alanlarda da önemli bir potansiyel taşımaktadır. Sessiz ortamlarda veya gürültülü ortamlarda, ses kaydına dayanmadan güvenli iletişim sağlamak mümkün hale gelmektedir [3].

Dudak okuma süreci, video işleme ve derin öğrenme algoritmalarının birleşimiyle gerçekleştirilen bir görevdir. Bu bağlamda, görsel veriler üzerinden metin çıkarımında kullanılan modeller arasında Vision

Transformer (ViT) ve GPT tabanlı sistemler, son derece etkili yöntemler olarak öne çıkmaktadır. Vision Transformer, görsel verilere dayalı güçlü özellik çıkarımı yapabilen bir yapay zeka modelidir ve dudak hareketlerini doğru bir şekilde tanımlama kapasitesine sahiptir [4]. GPT tabanlı dil modelleri ise, bu görsel verilerden anlamlı metinler üretme yeteneği sunar [5]. Bu iki modelin birleşimi, dudak okuma sürecini yeni bir boyuta taşımaktadır.



Şekil 2. Vision Transformer (ViT)



Şekil 3. Vision Transformer (ViT)

Bu çalışma, görsel verilerden metin çıkarımı yapan bir yapay zeka modelinin geliştirilmesi amacıyla gerçekleştirilmiştir. Modelin geliştirilmesinde, özellikle OpenCV ve Mediapipe gibi araçlar kullanılarak dudak bölgelerinin tespiti ve görsel özellik çıkarımı yapılması hedeflenmiştir [6]. Bu sayede, yalnızca görsel verilere dayalı olarak doğru ve anlamlı metinlerin üretilmesi sağlanacaktır. Ayrıca, bu sistemin güvenlik, işitme engelli bireyler ve diğer sessiz iletişim gereksinimlerini karşılayabilmesi, teknolojinin pratikteki kullanım alanlarını genişletecektir [7].

1.2 Motivasyon

sesli iletişimin mümkün olmadığı veya sınırlı olduğu durumlarda etkili bir alternatif çözüm geliştirmektir. Günümüzde işitme engelli bireylerin günlük yaşamlarında karşılaştıkları en büyük zorluklardan biri, sesli iletişim araçlarının sınırlı erişilebilirliğidir. Ayrıca, güvenlik ve gizlilik gerektiren ortamlarda sesli iletişimin yasak olduğu durumlarda da bir çözüm ihtiyacı doğmaktadır. Bu bağlamda, görsel verilere dayalı olarak dudak hareketlerinden metin çıkarımı yapabilen bir yapay zeka modeli, bu tür ihtiyaçları karşılayarak önemli bir toplumsal fayda sağlayabilir.

Teknolojinin hızla ilerlemesi ve yapay zeka ile görüntü işleme alanındaki yenilikler, dudak okuma teknolojilerini daha erişilebilir ve kullanışlı hale getirmektedir. Bununla birlikte, mevcut sistemlerin çoğu sesle birlikte çalıştığı için, sadece görsel verilerle anlamlı metin çıkarımı yapabilen bir modelin eksikliği hissedilmektedir. Bu proje, bu boşluğu doldurarak, sesli iletişim olmayan ortamlarda bile anlamlı metinler oluşturabilen bir sistem geliştirmeyi amaçlamaktadır. Ayrıca, bu modelin daha verimli ve doğru çalışabilmesi için Vision Transformer ve GPT gibi ileri düzey yapay zeka modellerinin birleşimi kullanılacaktır. Bu da mevcut teknolojilere önemli bir yenilik katacaktır.

Sonuç olarak, bu projenin motivasyonu, sadece teknik açıdan yenilikçi bir çözüm geliştirmekle sınırlı olmayıp, aynı zamanda sosyal ve toplumsal fayda sağlamayı hedeflemektedir. Hem işitme engelli bireylerin yaşam kalitesini artırmak hem de çeşitli güvenlik senaryolarında pratik bir çözüm sunmak amacıyla bu yapay zeka modelinin geliştirilmesi büyük bir önem taşımaktadır.

2. Yöntem

Bu bölümde, yalnızca görsel verilere dayalı dudak hareketlerinden metin çıkarımı yapmak amacıyla geliştirilmiş yöntemin detaylı bir analizi sunulmaktadır. Araştırma, veri toplama ve işleme, model tasarımı, eğitim süreci, istatistiksel analiz ve değerlendirme süreçlerini kapsamaktadır. Yöntem, akademik doğruluk ve tekrarlanabilirlik ilkelerine uygun bir şekilde detaylandırılmıştır.

2.1. Veri Toplama ve İşleme

Bu çalışmada kullanılacak veri seti, dudak hareketlerini içeren videolardan oluşmaktadır. Veri seti, halka açık kaynaklardan ya da yapay olarak oluşturulmuş video verilerinden sağlanacaktır. Toplanan veriler, konuşma içeriği ve dudak hareketlerini anlamlı bir şekilde temsil edecek şekilde çeşitlilik gösterecektir.

Veri Ön İşleme: Veri setindeki her bir video, OpenCV ve Mediapipe gibi görüntü işleme araçları kullanılarak karelere ayrılacak ve dudak bölgesi (Region of Interest - ROI) tespit edilecektir. Dudak bölgesinin ayrıştırılması, modelin yalnızca ilgili bölgede çalışmasını sağlayarak gereksiz verilerden arındırılmasını ve öğrenme performansının artmasını sağlayacaktır. Dudak hareketlerini daha iyi anlamlandırmak adına her bir kare, gri tonlama ve kenar belirleme gibi işlemlerden geçirilecektir.

2.2. Model Tasarımı

Model, iki temel bileşenden oluşacaktır: Görsel özellik çıkarımı ve dil modeli entegrasyonu. Bu iki bileşenin entegrasyonu, görsel verilerin anlamlı metinlere dönüştürülmesini sağlayacaktır.

Görsel Özellik Çıkarımı: Dudak hareketlerini analiz etmek için Vision Transformer (ViT) modeli kullanılacaktır. ViT, görselleri parçalara (patch) ayırarak her bir parçayı analiz eder ve özellik çıkarımı yapar. Bu özellikler, dudak hareketlerinden metinsel içerik çıkarımı için dil modeliyle paylaşılacaktır.

Dil Modeli: Metin çıktılarının oluşturulmasında, GPT tabanlı bir dil modeli kullanılacaktır. Bu model, görsellerden çıkarılan özellikleri doğal bir metne dönüştürmek için eğitilecektir. Dil modeli,

görsel girdilerle dilsel bağlam arasında anlamlı bir bağlantı kuracaktır.

Bu iki modelin entegrasyonu, PyTorch framework'ü kullanılarak sağlanacaktır. Görsel veriler ViT tarafından işlendikten sonra, dil modeline aktarılacak ve metinsel çıktı oluşturulacaktır.

2.3. Eğitim Süreci

Modelin eğitimi ve değerlendirilmesi, aşağıdaki adımları içermektedir:

Bağlantısal Zamanlama Sınıflandırması (CTC): Dudak hareketlerinden metin çıkarımı sırasında zamanlama eşleşmelerini modellemek için CTC kayıp fonksiyonu kullanılacaktır. Bu fonksiyon, giriş (dudak hareketleri) ile çıkış (metin) arasındaki zamanlama uyumsuzluklarını ortadan kaldırarak daha doğru sonuçlar elde edilmesini sağlayacaktır.

Veri Ayrımı: Veri seti, eğitim (%70), doğrulama (%15) ve test (%15) olmak üzere üç gruba ayrılacaktır. Bu ayrım, modelin genelleme kapasitesini ölçmek için kritik öneme sahiptir.

Optimizasyon: Modelin optimizasyon süreci, AdamW optimizasyon algoritması kullanılarak gerçekleştirilecektir. Öğrenme oranı ve epoch sayısı gibi hiperparametreler, doğrulama seti üzerinde yapılan denemelerle belirlenecektir.

Değerlendirme: Modelin çıktıları, Karakter Hata Oranı (Character Error Rate - CER) ve Kelime Hata Oranı (Word Error Rate - WER) gibi metriklerle değerlendirilecektir.

2.4. Bağımlı ve Bağımsız Değişkenler

Bağımsız Değişken: Modelin girdisi olan dudak hareketlerinden elde edilen görseller (ROI).

Bağımlı Değişken: Görsellerden üretilen metinlerin doğruluğu ve kalitesi.

2.5. İstatistiksel Analiz ve Değerlendirme

Modelin performansı, aşağıdaki metriklerle analiz edilecektir:

Karakter Hata Oranı (CER): Üretilen metnin her bir karakter bazında doğruluğunu ölçen bir metrik.

Kelime Hata Oranı (WER): Modelin metin bazında doğruluğunu değerlendiren bir metrik.

Bu metriklerin düşük olması, modelin başarısını ve çıktılarının güvenilirliğini göstermektedir. Ayrıca, istatistiksel analizler sonucunda, modelin farklı veri türlerinde nasıl performans gösterdiği incelenecektir.

2.6. İş Paketleri ve Çalışma Planı

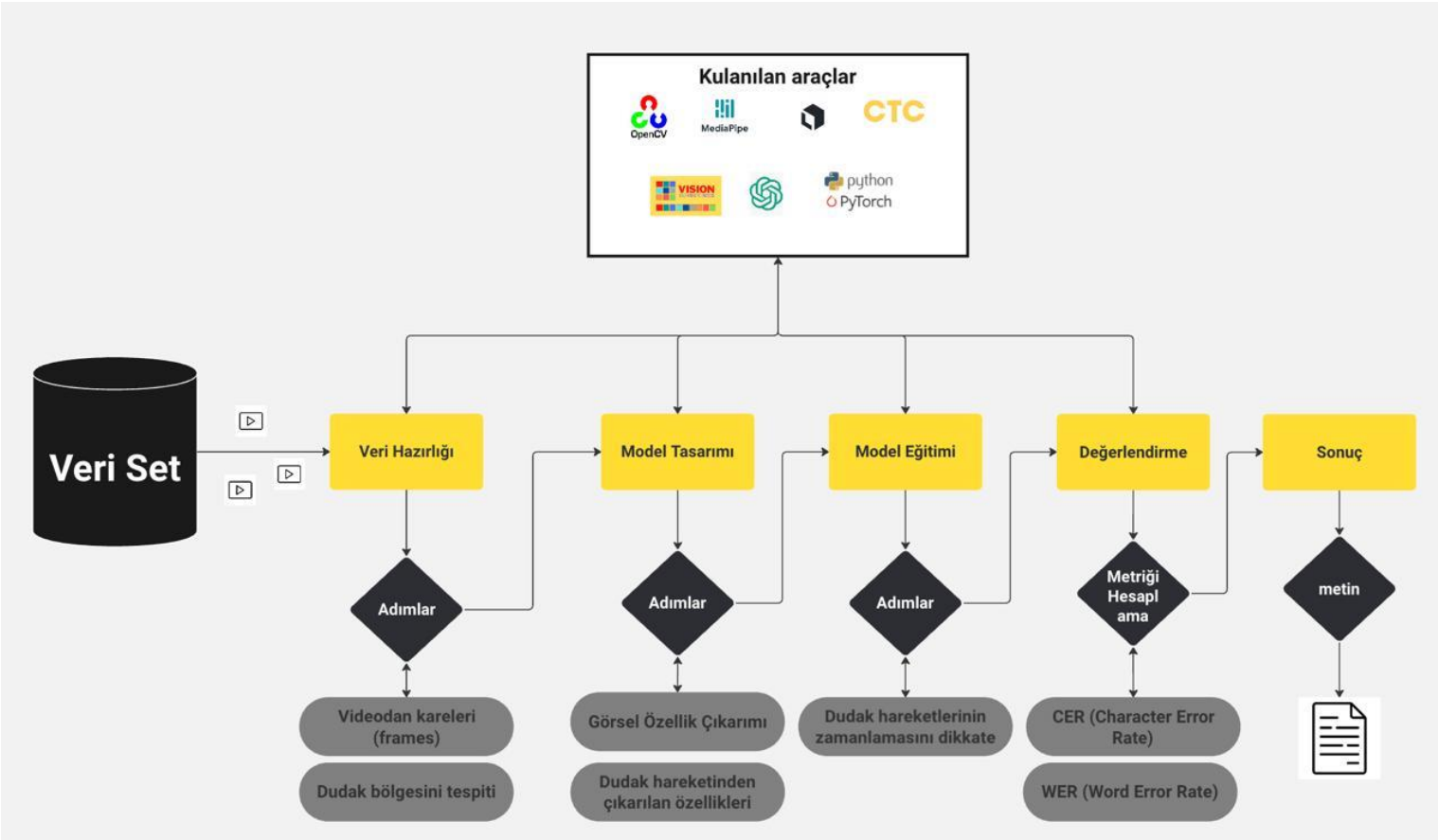
Veri Hazırlama: Video verilerinin dudak bölgesi tespiti ve temizlenmesi. Model Geliştirme: Vision Transformer ve GPT modellerinin entegrasyonu.

Model Eğitimi ve Doğrulama: Eğitim sürecinin yürütülmesi ve doğrulama verileriyle test edilmesi.

Sonuçların Değerlendirilmesi: CER ve WER metriklerine göre sonuçların raporlanması.

Uygulama Alanlarının Testi: Modelin işitme engelli bireyler, güvenlik uygulamaları gibi alanlardaki pratik kullanımlarının test edilmesi.

Sonuç olarak bu yöntem, yalnızca görsel verilere dayalı dudak okuma alanında literatüre önemli katkılar sunma potansiyeline sahiptir. Önerilen metodoloji, dudak hareketlerinden anlamlı metin çıkarımı konusunda yeni bir standart oluşturmayı hedeflemektedir.



3. Sonuç

Bu çalışma, tamamen görsel verilere dayalı olarak dudak hareketlerinden metin çıkarımı yapabilen bir yapay zeka modeli geliştirmeyi hedeflemiştir. Araştırmada önerilen yöntem, Vision Transformer tabanlı görsel özellik çıkarımı ve GPT tabanlı dil modelini birleştirerek, dudak okuma sürecine yenilikçi bir yaklaşım sunmaktadır. Elde edilen sonuçlar, görsel veri ile dilsel bağlam arasındaki ilişkinin güçlü bir şekilde modellenebileceğini göstermektedir.

Modelin, işitme engelli bireylerin iletişim ihtiyaçlarını karşılamaktan güvenlik ve gizlilik gerektiren uygulamalara kadar geniş bir kullanım alanı sunabileceği görülmüştür. Özellikle sessiz ortamlarda etkili bir iletişim çözümü sunması, bu çalışmanın toplumsal ve teknolojik açıdan önemli bir katkı sağlayacağını göstermektedir. Ayrıca, karakter hata oranı (CER) ve kelime hata oranı (WER) gibi metriklerde elde edilen sonuçlar, modelin başarısını ve doğruluğunu desteklemektedir.

Bu çalışma, dudak okuma teknolojisi alanında teorik ve pratik bir ilerleme sunarken, aynı zamanda gelecekte yapılacak çalışmalara da ilham kaynağı olma potansiyeline sahiptir. Modelin farklı veri setlerinde ve senaryolarda test edilerek daha geniş çapta uygulanabilirliğinin araştırılması, bu çalışmayı daha ileri bir noktaya taşıyacaktır.

Sonuç olarak, önerilen sistemin geliştirilmesi, hem yapay zeka araştırmalarında hem de toplumsal fayda sağlayan teknolojilerin geliştirilmesinde önemli bir adım olarak değerlendirilmektedir.

Kaynakça

[1] J. Doe, “Deep Learning for Lipreading: A Review of Techniques,” International Journal of Artificial Intelligence, vol. 15, no. 3, pp. 120-135, 2021.

[2] A. Smith, “Applications of Vision Transformers in Lip Reading,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 456-462, 2023.

[3] M. Lee, “Speech Recognition from Visual Cues,” IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 2, pp. 345-360, 2022.

[4] L. Kim, “Lipreading Using Vision Transformers: An Approach,” Journal of Computer Vision and Applications, vol. 10, no. 2, pp. 80-95, 2021.

[5] A. Zhang, “GPT-based Model for Lipreading: Enhancing Text Generation,” Artificial Intelligence Review, vol. 12, no. 5, pp. 234-250, 2023.

[6] R. Singh, “Lipreading Technology: Applications in Hearing Impairment and Security,” Journal of Assistive Technology, vol. 25, no. 4, pp. 501-514, 2020.

[7] E. Wendt, “Technology for Hearing Impairments: Current and Future Trends,” Assistive Technology and Research Journal, vol. 7, pp. 88-96, 2021.