Derin Sinir Ağları ile Konuşma Tespiti ve Cinsiyet Tahmini

Farzad Kiani[1], Mehmet Ali Kutlugün[1], Mert Yılmaz Çakır[1]

[1] İstanbul Sabahattin Zaim Üniversitesi, Bilgisayar Bilimleri ve Mühendisliği Bölümü, İstanbul

farzad.kiani@izu.edu.tr, mehmet.kutlugun@std.izu.edu.tr, mert.cakir@std.izu.edu.tr

Özet: Konuşma tanıma, günümüzde insan hayatını kolaylaştıran teknolojiler göz önüne alındığında bilim dünyasında ağırlık verilen konulardandır. Bu alanda konuşma içerisinden konuşmacıya ait bilgilerin tespit edilmesi artan güvenlik önlemleriyle ilgi çeken bir konu olmuştur. Bu çalışmada da konuşma tanıma için gerekli aşamalar ve teknikler incelenmiştir. Ayrıca deneysel çalışma ile konuşmanın akustik özelliklerinden cinsiyet tahmini için derin sinir ağlarıyla verimli bir çözüm yolu önerilmektedir.

Anahtar Sözcükler: Yapay zeka, derin sinir ağları, konuşma tanıma

Abstract: Speech recognition is a subject that is given importance in the scientific world when considering the technologies that make human life easier today. Identification of the speaker's information in this area has become a topic of interest with increased security measures. In this study, necessary steps and techniques for speech recognition were examined. We also propose an efficient solution to the gender estimation of the acoustic features of speaking with the experimental work with deep neural networks.

Key words: Artificial intelligence, deep neural networks, speech recognition

1. Giriş

İnsan bilgisayar etkileşiminde önemli araştırma konularından birisi olan konuşma tanıma, konuşulan dilin bilgisayar tarafından anlamlı metin haline getirilmesini sağlayan metodolojileri ve teknolojileri içerir. Konuşma tanıma sistem tasarımı ve uygulaması için sinyal işleme, analiz yöntemleri ve karşılaştırma teknikleri kullanılmaktadır.

Konuşma tanıma alanında yaygın kullanıma sahip olan yapay sinir ağları yapay zekanın bir alt başlığıdır. Yapay zekan iki temel fikri içerir. Birincisi, insanoğlunun düşünce süreçlerini incelemektir. İkincisi ise makineler (bilgisayarlar, robotlar, vb.) vasıtasıyla bu süreçleri temsil etmekle ilgilidir. Bu adımları konuşma tanımaya uyarlarsak; konuşma tanıma sisteminin en önemli yararlarından birisi kullanıcının aynı anda başka işleri yapabilmesidir. Kullanıcı gözlem ve manuel operasyonlara konsantre olabilir ve sesli giriş komutlarıyla makineyi kontrol edebilir.

Yapay zeka, bilgisayar bilimlerinde çevresini algılayan ve bir hedefte başarı şansını maksimize eden eylemlere denilmiştir. Yapay zekanın araştırma konularından olan yapay sinir ağları insan sinir sisteminin doğal sinir ağından esinlenmiştir. Sinir ağları bilgiyi, harici girdilere karşı dinamik hal tepkileri ile işleyen birbirine bağlı işlem öğelerinden oluşan bir bilgi işlem sistemidir. Derin sinir ağları, derin inanç ağları ve tekrarlayan sinir ağları gibi derin öğrenme mimarileri, bilgisayar görme, konuşma tanıma, doğal dil işleme, sosyal ağ filtreleme, makine çevirisi ve biyoinformatik gibi alanlara uygulanmıştır. Derin sinir ağları çok üstün başarı oranı sağlamaktadır.

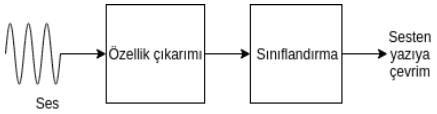
Teknoloji açısından konuşma tanımanın uzun bir geçmişi vardır. Son zamanlarda konuşma tanıma derin öğrenme ile yapılan ilerlemelerden yararlandı. Gelişmeler sadece bu alanda yayınlanan akademik yayınların artması ile değil daha önemlisi, konuşma tanıma sistemlerinin tasarımında ve performasının artmasında derin öğrenme yöntemlerinin dünya çapında endüstrinin benimsenmesiyle kanıtlandı. Bu konuşma endüstrisinde teknolojilerini derin öğrenmeye dayalı olarak tanıtan şirketlere örnek vermek gerekirse; Google, Microsoft, IBM, Baidu, Apple, Amazon, Nuance, SoundHound, IflyTek ve CDAC.

Bu çalışmada kadın ve erkek konuşmacılardan alınan konuşmalar ile sistem derin sinir ağları ile eğitilmektedir. Sistemin test aşaması ise farklı bir konuşmanın cinsiyet tahmini yapılmasıdır.

2. Çalışmanın Aşamaları

Konuşma ile cinsiyet tahmini için konuşmaların, öncelikle bu sürece hazırlanarak bilgisayar destekli olarak tanıma sürecine dahil edilmeleri gerekmektedir. Bu amaçla konuşmaların bir mikrofon aracılığıyla örneksel sinyallere dönüştürülmesi ve etiketlenmesi (örneğin sesler, fonemler, kelime ya da kelime grubu olarak) ve tanıma işlemlerine taban oluşturacak sınıflandırma teknikleriyle parametrik yapılar ya da yalın modellerle ifade edilen biçimlere dönüştürülmesi gerekmektedir [1] [2].

Bu sistem iki aşamadan oluşmaktadır. Eğitim ve test aşaması. Eğitim aşamasında kaydedilen konuşmaların özellik vektörleri çıkartılır ve sisteme cinsiyet etiketi ile kaydedilir. Sonrasında sınıflandırma aşaması ile iki cinsiyet arasında olasılıksal dağılım yapılır. Test aşamasında sisteme verilen konuşmanın özellik vektörleri çıkartılır ve sınıflandırma aşamasında en yüksek olasılıklı eşleşen cinsiyet ile tahmin yapılır.



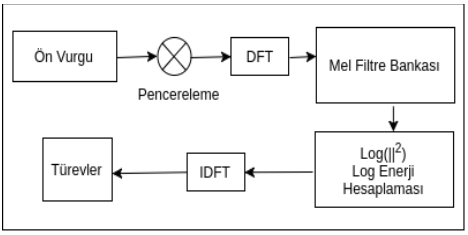
Şekil 1. Konuşmanın anlamlı metin haline dönüşümü

3. Özellik Çıkarımı

Makine öğrenmede, örüntü tanımada ve görüntü işleme yönteminde kullanılan ve boyutsallığı azaltmakla ilgili olan özellik çıkarımı, ilk ölçülen verilerle başlayarak bilgilendirici ve gereksiz olmaması amaçlanan türetilen değerleri (özellikleri) oluşturur. Sonrasında öğrenme ve genelleme aşamalarını kolaylaştırır ve ileriki süreçlere öncü olur. Bir algoritma için girilen veriler işlenemeyecek kadar büyük olduğunda ve gereksiz olduğu düşünülürse (örneğin, her iki birimde aynı ölçüm veya piksel olarak sunulan görüntülerin tekrarlanması), daha sonra azaltılmış bir sete dönüştürülebilir. İlk özelliklerin bir alt kümesine rastlanması, özellik seçimi olarak adlandırılır. Seçilen özelliklerin girilen verilerden ilgili bilgileri içermesi beklenir, böylece arzu edilen görev, başlangıç verisi yerine bu azaltılmış gösterimi kullanarak gerçekleştirilebilir [3] [4].

4. Mel Frekanslı Kepstral Katsayılar

Konuşma tanımada, mel frekans cepstrum (MFC), bir frekansın doğrusal olmayan mel skalasındaki bir log güç spektrumunun doğrusal bir kosinüs transformasyonuna dayanan, bir sesin kısa dönemli güç spektrumunun gösterimi olan bir özellik çıkarımıdır. Mel frekanslı cepstral katsayıları (MFCC), bir MFC'yi topluca oluşturan katsayılardır. Ses klibinin bir tür cepstral gösteriminden türemiştir (doğrusal olmayan spektrumun bir spektrumu). Cepstrum ve mel-frekans cepstrum arasındaki fark, MFC'de frekans bantlarının, normal kepstrumda kullanılan doğrusal aralıklı frekans bantlarından daha yakından insan ses sistemi tepkisine yaklaşan mel ölçeğinde eşit aralıklarla yerleştirilmesidir. Bu frekans çarpıklığı, örneğin ses sıkıştırmasında daha iyi bir ses temsilini sağlayabilir [5].



Şekil 2. MFCC aşamaları

Frekanstan Mel ölçeğine dönüştürme formülü şöyledir:

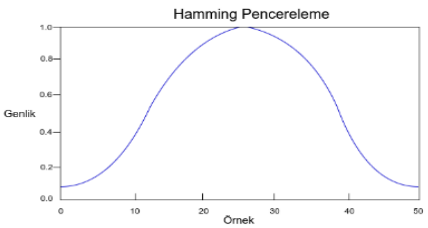
$$M(f) = 1125 \ln(1 + f/700) \quad (1)$$

Mel'den frekansa geri dönmek için:

$$M^{-1}(m) = 700(\exp(m/1125) - 1) \quad (2)$$

Pencereleme aşamasında, alınan sinyal içerisindeki devamsız kısımların dikkate alınmaması ses tanıma için kritik bir eşiktir. Bu işlem pencereleme ile gerçekleşmektedir. Pencereleme fonksiyonları arasında en yaygın olarak kullanılanlar dikdörtgen ve Hamming pencerelemedir. Önerilen sistemde yaygın olarak kullanılan, sesin frekans dönüşümü esnasındaki bozulmalarının azaltılabilmesinde ve konuşmanın spektral analizinde, Hamming pencereleme w(n) tercih edilmiştir. Her bir pencere içindeki örnek sayısı N olacak şekilde formulü şu şekildedir;

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{N-1}\right), & 0 \le n \le N-1 \\ 0, & n < 0, n > N-1 \end{cases}$$



Şekil 3. Hamming pencereleme

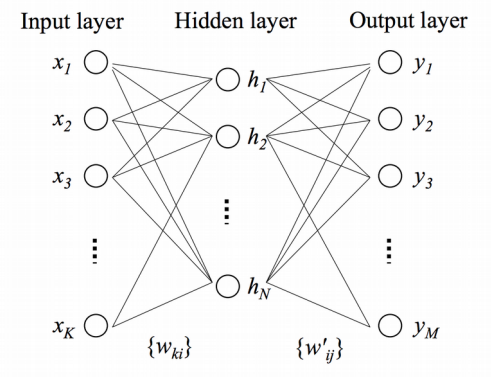
5. Sınıflandırma

Örüntü tanımada sınıflandırma, kategori üyeliği olarak bilinen gözlemleri (veya örnekleri) içeren bir eğitim seti temelinde yeni bir gözlemin ait olduğu bir grup kategorinin (alt popülasyonlar) tanımlanmasıdır. Bunun bir örneği, belirli bir e-postayı "spam" veya "spam olmayan" sınıflara atamak veya belirli bir hastaya gözlenen özelliklerle (cinsiyet, kan basıncı, belirli belirtilerin varlığı veya yokluğu vb.) açıklandığı gibi bir tanı tayin etmek olacaktır. Sınıflandırma, örüntü tanımanın bir örneğidir. Örüntü tanıma terminolojisinde sınıflandırma, denetlenen öğrenmenin bir örneği, yani doğru tanımlanmış gözlemlerin bir eğitim kümesinin bulunduğu yerde öğrenilir. İlgili denetimsiz işlem, kümeleme olarak bilinir ve veriyi, doğal bir benzerlik veya uzaklık ölçüsüne dayalı olarak kategorilere ayırmayı içerir.

6. Derin Sinir Ağları

2000'lerin başında konuşma tanıma işleminde, ileri beslemeli yapay sinir ağları gibi geleneksel yaklaşımlar hakimdi. [6] Ancak konuşma tanımada derin öğrenme yöntemi, 1997 yılında Sepp Hochreiter & Jürgen Schmidhuber tarafından yayınlanan bir tekrarlayan sinir ağı [7] tarafından bugün yaygın kullanıma sahiptir.

Derin bir ileri beslemeli sinir ağı (Deep Neural Network), girdi ve çıktı katları arasında çok sayıda gizli katman içeren yapay bir sinir ağıdır. [8] Sığ sinir ağlarına benzer şekilde, DNN'ler karmaşık doğrusal olmayan ilişkileri modelleyebilir. DNN mimarileri, alt katmanların özelliklerinden kompozisyona olanak tanıyan kompozisyonel modeller üretir, böylece büyük bir öğrenme kapasitesi ve böylece konuşma verisinin karmaşık modellerini modelleme potansiyeli sağlanır. [9]



Şekil 4. Temel derin öğrenme

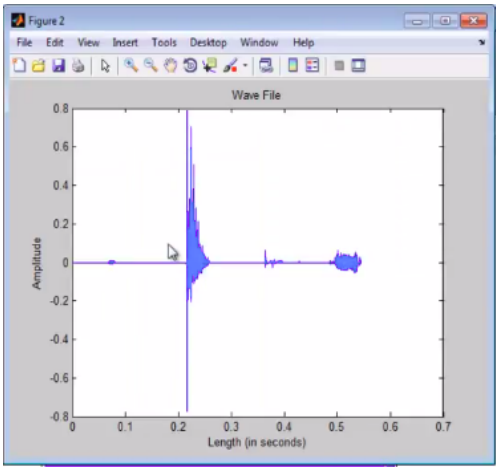
Bir sinir ağı, bir giriş katmanı, çıktı katmanı ve gizli katmandan oluşur. Giriş katmanı, x = $\{x_1, ..., x_K\}$ vektöründen oluşur. Gizli katman, bir sinir ağı nöron vektörü, h = $\{h_1, ..., h_N\}$ 'lerden oluşur. Son olarak, çıkış vektörü, y = $\{y_1, ..., y_M\}$ 'nin her elemanı için bir nöron içeren bir çıktı katmanı vardır.

w ağırlık vektörü olmak üzere gizli katmandaki rastgele nöronun çıkışı h_i aşağıdaki şekilde olmaktadır.

$$h_i = f(u_i) = f(\sum_{k=1}^{K} w_{ki} x_k)$$

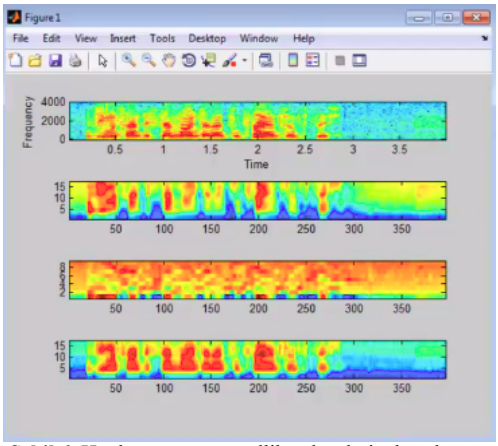
7. Uygulama

Uygulama, Matlab'da 10 erkek ve 10 kadından alınan konuşmalar wav formatında cinsiyet etiketiyle kaydedilmiştir.



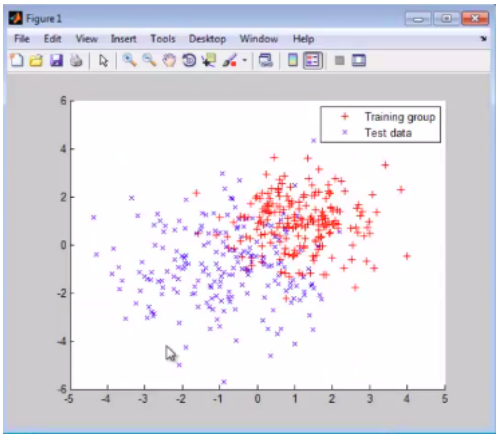
Şekil 5. Her konuşma .wav olarak kaydedilir

Her bir konuşmanın özellik vektörleri MFCC tekniği ile çıkartılmıştır.



Şekil 6. Her konuşmanın özellik vektörleri çıkartılmıştır

Çıkartılan özellik vektröleri sınıflandırma aşamasına giriş parametresi olarak verilir. Sınıflandırma aşamasında DNN Tool kullanılmıştır. Kadın ve erkek konuşmalarını gruplamıştır.



Şekil 5. DNN ile tanınması istenilen konuşmanın karşılaştırılması yapılır

8. Değerlendirme

Yapılan uygulama 5 kadın ve 5 erkek konuşmacı üzerinde test edilmiştir. Deney sonucunda 9 kez doğru cinsiyet tahmini sağlanmıştır. Yüksek başarımıyla derin sinir ağlarının iyi performans verdiği gözlenmiştir.

Bu çalışmayla konuşmanın akustik özelliklerinden cinsiyet, yaş aralığı gibi istenilen durum tespitlerinin derin sinir ağları ile yüksek verimli olarak yapılabileceği gösterilmiştir.

Konuşma tanıma, amacına yönelik konuşulanlar ile ilgilidir. Konuşma tanıma teknolojisi büyüyen bir alandır. Otomatik konuşma tanıma teknolojisi konusunda büyük ilerleme kaydedilmektedir. Yine de, bu alanda tanıma oranı, arka plan gürültüsü, konuşmacı değişkenliği, konuşma oranı, aksan vb. açısından birçok engel vardır. Konuşma tanıma oranı esas olarak özelliklerin seçimi ve sınıflandırma yöntemlerine bağlıdır [10].

Kaynaklar

[1] Mark A. Eckert, Lois J. Matthews, and Judy R. Dubno, Self-Assessed Hearing Handicap in Older Adults With Poorer-Than-Predicted Speech Recognition in Noise, Journal of Speech, Language, and Hearing Research, January 2017, Vol. 60, 251-262. doi:10.1044/2016_JSLHR-H-16-0011

[2] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D. Castillo, Rama Chellappa, An All-In-One Convolutional Neural Network for Face Analysis, 2017, DOI: 10.1109/FG.2017.137, IEEE

[3] K. M. Shiva Prasad, G. N. Kodanda Ramaiah and M. B. Manjunatha, Speech Features Extraction Techniques for Robust Emotional Speech Analysis/Recognition, Indian Journal of Science and Technology, Vol 10(3), January 2017

[4] Masanobu Nakamura, Takashi Masuko, Speech feature extraction apparatus and speech feature extraction method, 2017, US9754603.

[5] Trima Mustofa, Implementation Speech Recognition for Robot Control Using MFCC and ANFIS, Journal of Telematics and Informatics, 2017, Vol: 5, No: 2, DOI: 10.12928/jti.v5i2.

[6] Herve Bourlard and Nelson Morgan, Connectionist Speech Recognition: A Hybrid Approach, The Kluwer International Series in Engineering and Computer Science; v. 247, Boston: Kluwer Academic Publishers, 1994.

[7] Hochreiter, S; Schmidhuber, J (1997). "Long Short-Term Memory". Neural Computation. 9 (8): 1735–1780.

[8] Hinton, Geoffrey; Deng, Li; Yu, Dong; Dahl, George; Mohamed, Abdel-Rahman; Jaitly, Navdeep; Senior, Andrew; Vanhoucke, Vincent; Nguyen, Patrick; Sainath, Tara; Kingsbury, Brian (2012). "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups". IEEE Signal Processing Magazine. 29 (6): 82–97.

[9] Deng, Li; Yu, Dong (2014). "Deep Learning: Methods and Applications" (PDF). Foundations and Trends in Signal Processing. 7 (3–4): 197–387.

[10] Kaur, G., Srivastava, M., & Kumar, A. (2017). Analysis of Feature Extraction Methods for Speaker Dependent Speech Recognition. International Journal of Engineering and Technology Innovation, Vol 7, No 2. IJETI.