Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 36, S. 121-127, Mayıs 2022

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 36, pp. 121-127, May 2022

Research Article

Konuşma Tanımaya Uygulanan BiRNN, BiLSTM ve BiGRU Modellerinin Performans Değerlendirmesi

Halil İbrahim Yalman [1*], Zekeriya Tüfekci [2]

[1*] Çukurova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Adana, Türkiye, (ORCID: 0000-0003-0841-1309), halilyalman@hotmail.com.tr
[2] Çukurova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Adana, Türkiye, (ORCID: 0000-0001-7835-2741), ztufekci@cu.edu.tr

(1st International Conference on Engineering and Applied Natural Sciences ICEANS 2022, May 10-13, 2022)

(**DOI:** 10.31590/ejosat.1111314)

ATIF/REFERENCE: Yalman, H. İ. & Tüfekci, Z. (2022). Konuşma Tanımaya Uygulanan BiRNN, BiLSTM ve BiGRU Modellerinin Performans Değerlendirmesi. *European Journal of Science and Technology*, (36), 121-127.

Öz

Konuşma tanıma ses dalgalarının yazıya dönüştürülmesi işlemidir. Bu çalışmada sesli kitap veri seti üzerinde Çift Yönlü Basit Tekrarlayan Ağlar (BiRNN), Çift Yönlü Uzun Kısa Süreli Bellek (BiLSTM), Çift Yönlü Kapılı Tekrarlayan Hücreler (BiGRU) modellerinin konuşma tanıma performansı incelenmiş ve karşılaştırması yapılmıştır. Kullanılan modellerde Bağlantıcı Zamansal Sınıflandırma (CTC) ve Evrişimsel Sinir Ağları (CNN) kullanılmıştır. Ayrıca bu modellerin tek yönlü versiyonları ile karşılaştırması da yapılmıştır. Çalışmanın sonucunda en yüksek konuşma tanıma başarı oranına sahip model BiLSTM olduğu saptanmıştır. Bununla birlikte %33 daha az para metre ile %3 daha düşük konuşma tanıma oranına sahip BiGRU modeli de dikkate değer bulunmuştur. Çift yönlü modellerin tek yönlü modellere göre daha başarılı sonuçlar verdiği saptanmıştır.

Anahtar Kelimeler: Konuşma Tanıma, Derin Öğrenme, Evrişimsel Sinir Ağları, Çift Yönlü Uzun Kısa Süreli Bellek, Çift Yönlü Basit Tekrarlayan Ağlar, Çift Yönlü Kapılı Tekrarlayan Hücreler, Bağlantıcı Zamansal Sınıflandırma, Türkçe Sesli Kitap Veri seti.

Performance Evaluation of BiRNN, BiLSTM and BiGRU Models Applied to Speech Recognition

Abstract

Speech recognition is the process of converting sound waves into text. In this study, speech recognition performance of Bidirectional Recurrent Neural Network (BiRNN), Bidirectional Long Short Term Memory (BiLSTM), Bidirectional Gated Recurrent Units (BiGRU) models on the audiobook dataset was examined and compared. Connectionist Temporal Classification (CTC) and Convolutional Neural Networks (CNN) are used in the models. In addition, these models were compared with their unidirectional versions. As a result of the study, it was determined that the model with the highest speech recognition success rate was BiLSTM. However, the BiGRU model, which has 33% less parameters and 3% lower speech recognition rate, was also found to be remarkable. It has been determined that bidirectional models give more successful results than unidirectional models.

Keywords: Speech Recognition, Deep Learning, Convolutional Neural Networks, Bidirectional Long Short Term Memory, Bidirectional Recurrent Neural Networks, Bidirectional Gated Recurrent Units, Connectionist temporal classification, Turkish Audiobook Dataset.

[*] Sorumlu Yazar: halilyalman@hotmail.com.tr

1. Giriş

Konuşma tanıma seslerin metinlere dönüştürülmesidir. Gündelik hayatta kullanılan sistemler seslerin metinlere veya komutlara dönüştürülmesine ihtiyaç duyar. Bunun amacı teknolojinin hayatımızın her alanını kolaylaştırmayı hedeflemesinden dolayıdır.

Teknolojinin her alanında konuşma tanıma sistemlerine ihtiyaç duyulmaktadır. Otonom sistemlerin, giyilebilir teknolojik araçların ve mobil telefonların yaygınlaşması bu sistemlere ihtiyacı artırmıştır. Bu ihtiyaç, konuşma tanıma sistemlerinin önünü açmıştır. 1960'lı yıllarda Bell laboratuvarında yapılan çalışmalarla birlikte IBM'de 1961 yılında "Shoebox" çalışması konuşma tanıma adına yapılan ilk çalışmalardandır.

Konuşma tanımada 1970'lerde Markov modeli ve 1980'lerde Gizli Markov Modeli (HMM) kullanıldı. 2010'lu yıllarda Yapay Sinir Ağlarının konuşma tanımada kullanılması ve bilgisayarların işlem kapasitesinin artmasıyla birlikte çalışmalar hız kazanmıştır. Konuşma tanıma için Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM) ve Gated Recurrent Units (GRU) Sinir Ağı modelleri kullanılmaktadır.

CNN, LeCun ve diğerleri (1989) tarafından literatürde sunulan görüntü işleme için özel bir modeldir. CNN, konuşma tanıma sistemlerinde hem özellik çıkarımı hem de akustik modeller olarak kullanılmaktadır.

Rumelhart ve diğerleri (1986), sıralı verileri işlemek için RNN yapısını önerdi. Graves ve diğerleri (2013) ayrıca derin RNN modelini kullanarak akustik bir konuşma tanıma modeli önerdi. RNN modeli, LSTM ve GRU modellerinin temelini oluşturur. LSTM ve GRU modelleri RNN modelleridir. RNN modelinin çift yönlü bir versiyonu olan Bidirectional Recurrent Neural Networks (BiRNN), Schuster ve Paliwal (1997) tarafından tanıtıldı.

Hochreiter ve Schmidhuber (1997), RNN modelinin temel sorunlarını gidermek için LSTM modeli olan bir RNN modeli sundular. Sak ve diğerleri (2014), konuşma tanıma için LSTM modelini kullanmayı önerdi. LSTM modelinin RNN modelinden yaklaşık %50 daha az hata oranı ürettiğini gösterdiler.

Cho ve diğerleri (2014) LSTM modelinde sadeleştirmeler yaparak GRU modelini önermiştir. GRU modeli, LSTM modeline göre daha az kapıya sahip olan ve LSTM modeline benzer sonuçlar veren bir RNN modelidir. Shewalkar (2018) ayrıca LSTM modelinin GRU modelinden daha iyi sonuçlar elde ettiğini öne sürmüştür. Bu sonucu, GRU modeline kıyasla LSTM modelinin aynı sayıda düğüm ve daha uzun çalışma süresi ile elde etti.

RNN, LSTM ve GRU modellerinin çift yönlü modelleri sırasıyla Bidirectional RNN (BiRNN), Bidirectional LSTM (BiLSTM) ve Bidirectional GRU (BiGRU) olarak adlandırılır.

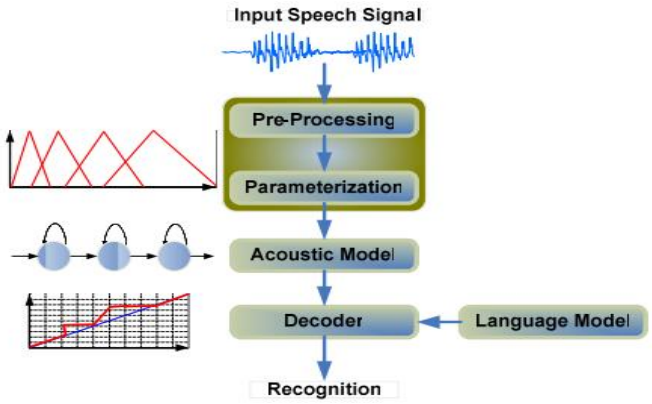
Zeyer ve diğerleri (2017), BiLSTM modelinin ileri beslemeli bir sinir ağından yaklaşık %8 daha iyi sonuçlar verdiğini öne sürmüştür. Arisoy ve diğerleri (2015), BiRNN modelinin tek yönlü RNN'den %0,2 daha iyi sonuçlar verdiğini açıkladı.

Ses dosyaları ve metin dosyaları sıralı verilerdir ve bu iki dosya arasında hangi harfin hangi sese karşılık gelme olasılığını bu iki dosya arasında bulmak modeller için bir problemdir. Graves ve diğerleri (2006), bu sorunun üstesinden gelmek için Connectionist Temporal Classification (CTC) modelini önerdi. CTC modeli, akustik modelin çıktılarını etiket dizisi üzerinde koşullu bir olasılık dağılımına dönüştürmek için bir sınıflandırıcı olarak kullanır.

Bu çalışmada konuşma tanıma için BiRNN, BiLSTM ve BiGRU akustik modellerinin Türkçe veri seti üzerindeki performansları karşılaştırıldı.

2. Materyal ve Yöntem

Konuşma tanıma işlemini gerçekleştirmek için bir veri kümesi gereklidir. Konuşma tanıma sisteminin blok şeması Şekil-1 de gösterilmektedir. Ön işleme adı verilen kısım öznitelik çıkarımıdır. Akustik model olarak kullanılan sinir ağları aracılığıyla eğitim verisiyle eğitilen model, test verisi ile test edilir ve performansı ölçülür. Bir dil modeliniz varsa, sürece dahil edilir.



Şekil 1. Konuşma Tanıma Sisteminin Blok Şeması (Mporas ve diğerleri, 2007)

Bu çalışmada kullanılan veri seti, öznitelik–çıkarımı, Sinir Ağlarının alt dalı Convolution Neural Network, Recurrent Neural Network, Long Short Term Memory, Gated Rate Unit sırasıyla açıklanacaktır.

2.1.Konuşma Tanıma için Türkçe Veri Seti

Çalışmada kullanılan veri seti 42 kişiden alınan seslerden oluşmaktadır. Bu kişilerin %50'si kadın %50'si erkektir. Toplam veri seti 15.000 cümleden oluşmakta ve yaklaşık 20 saatlik ses verisi içermektedir. Ses dosyaları maksimum 10 saniye olacak şekilde sınırlandırılmış olup her ses dosyasının metin karşılığı (Label) oluşturulmuştur.

Veri setindeki tüm seslerin kullanılması ile oluşturulan yaklaşık 20 saatlik veri seti Veriseti-3 olarak adlandırılmıştır.

Veri setindeki seslerin yarısı kullanılarak oluşturulan 7.500 ses dosyası ve metin verisi içerin yaklaşık 10 saatlik veri seti Veriseti-2 olarak adlandırılmıştır.

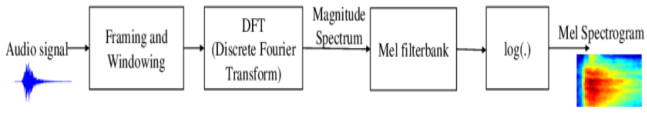
Veri setindeki seslerin dörtte biri kullanılarak oluşturulan 3.750 ses dosyası ve metin verisi içerek yaklaşık 5 saatlik veri seti Veriseti-1 olarak adlandırılmıştır.

Farklı boyutlarda veri seti oluşturulmasındaki maksat eğitim verisinin konuşma tanımadaki performansa etkisi ve modellerin başarım oranını incelemektir. Bu üç veri seti kullanılarak deneylerimiz yapılmıştır.

2.2.Öznitelik Çıkarımı

Öznitelik çıkarımı ham verideki belli bölüme odaklanmamız için yapılan bir bölümle ve odaklandığımız yerdeki nitelikleri ön plana çıkarma işlemine verilen genel terimdir. Konuşma

tanımanın öznitelik çıkarma adına birçok teknik bulunmakta olup bu çalışmada Mel Spectrogram kullanılacaktır. Mel Spectrogram özellik çıkarımının blok diyagramı Şekil 2'de sunulmuştur.



Şekil 2. Mel Spectrogram diyagramı *(Tak vd., 2017)*

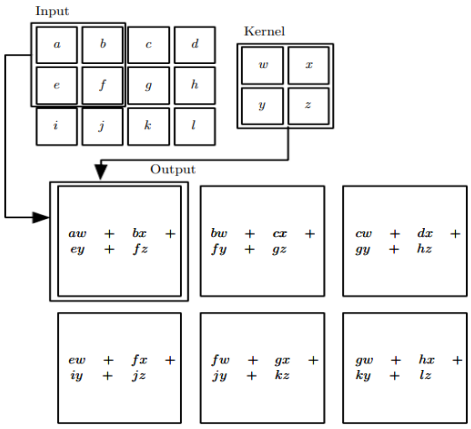
2.3.Sinir Ağları

Sinir ağları insan beynine benzer bir şekilde öğrendikleri bilgilerle yeni bilgileri algılama veya yeni bilgiler türetme amacı ile geliştirilen bilgisayar modelleridir. Konuşma tanımada sinir ağları var olan ses ve karşılığındaki metin verisini öğrenerek yeni gelen sesin metin verisini tahmin etmeye çalışır.

Sinir Ağları alt dalı olan 3 model üzerinde çalışmalar yapılacaktır.

2.3.1. Convolutional Neural Network

Evrişimsel sinir ağları veya CNN olarak da bilinen Evrişimsel ağlar, ızgara benzeri bilinen bir yapıya-sahip veriyi işlemek için kullanılan özel bir sinir ağıdır (Goodfellow vd., 2018).



Şekil 3. 2 boyutlu bir CNN örneği (Goodfellow vd., 2018)

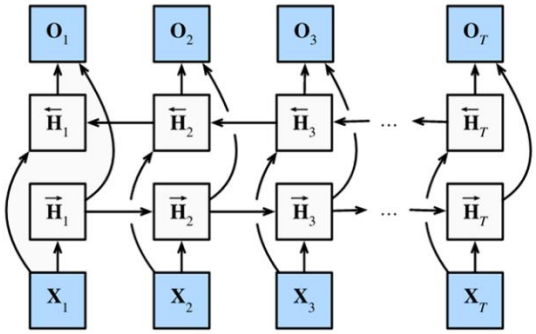
Şekil 3'te görüldüğü gibi, çekirdek her adımda kaydırılır ve girdi matrisinin çekirdek boyutu ile çarpılır. Bunu yaparken atlama (stride) parametresi ile kaydırılacak adım sayısını belirtiyoruz. Burada görüldüğü gibi 8x3 matris 6x2 matrise indirgenmiştir. Matrisin aynı boyutta kalmasını istiyorsak, kenarlarına belirli bir sayı ekleme işlemine dolgu (padding) denir. CNN, konuşma tanıma süreçlerinde hem akustik bir model hem de bir öznitelik çıkarma işlemi olarak kullanılabilir.

2.3.2. Recurrent Neural Network

Yinelemeli Sinir Ağları (Recurrent Neural Networks) sıralı verileri işlemek için özelleştirilmiş sinir ağları ailesidir (Rumelhart vd., 1986). Bu çalışmada kullanılan BiRNN, BiLSTM ve BiGRU modeli RNN alt dalıdır.

2.3.2.1.Bidirectional Recurrent Neural Network

Tekrarlayan sinir ağları (Recurrent Neural Network) sıralı verileri işlemek için bir sinir ağları ailesidir (Rumelhart ve diğerleri, 1986). Bidirectional Recurrent Neural Network (BiRNN), t zamanında hem önceki hem de sonraki zaman düğümünden alınan bilgiler tarafından mevcut bilginin kararlaştırıldığı bir RNN türüdür.



Şekil 4. BiRNN yapısı

RNN modeli sadece önceki zaman düğümünden gelen bilgilerle işlenirken, Şekil 4'te görüldüğü gibi çift yönlü RNN versiyonunda hem önceki hem de sonraki zaman düğümünü işler. BiRNN modelinin denklemleri Denklem (1)'den Denklem (3)'e gösterilmektedir: (Zhang ve diğerleri, 2021)

$$\overrightarrow{H}_t = \tanh(X_t W_{xh}^{(f)} + \overrightarrow{H}_{t-1} W_{hh}^{(f)} + b_h^{(f)}) \quad (1)$$

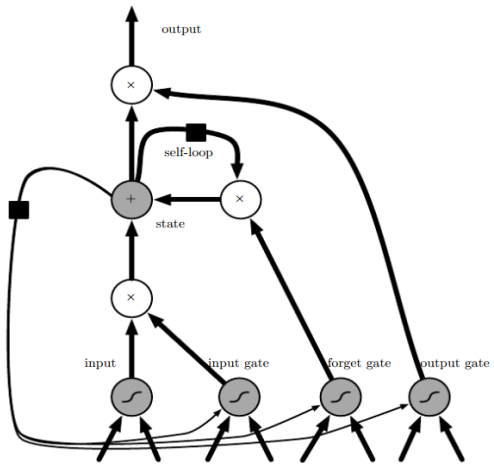
$$\overleftarrow{H}_t = \tanh(X_t W_{xh}^{(b)} + \overleftarrow{H}_{t+1} W_{hh}^{(b)} + b_h^{(b)}) \quad (2)$$

$$O_t = \tanh(H_t W_{hq} + b_q) \quad (3)$$

W ağırlık matrisini ifade eder. b yanlılık vektörünü ifade eder. Tanh tanjant hiperbolik fonksiyonudur. Çift yönlü modeller, ileri ve geri yapılar eklendiğinden, tek yönlü modellere göre matris işlemlerinde farklı parametrelere sahiptir. H_t simgesinin üzerindeki oklar ileri ve geri yapılarını temsil eder. İleri ve geri yapılarından elde edilen matrisler birleştirilerek H_t matrisi elde edilir.

2.3.2.2.Bidirectional Long Short Term Memory

Uzun Kısa Süreli Bellek dış yinelemeleri ve kapılarla sağladığı iç yinelemeli olan bir modeldir RNN mimarisinin bir alt türüdür



Şekil 5. LSTM yapısı

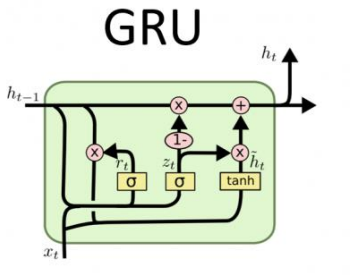
LSTM'in blok diyagramı Şekil 5'te verildiği gibi 4 kapının kontrolü ile çalışmaktadır. LSTM kapıları arasında unutma kapısı(f_t), hangi bilgilerin unutulacağına, giriş kapısı(i_t) hangi bilgilerin ekleneceğine, çıkış kapısı(o_t) ise hangi bilgilerin diğer duruma aktarılacağına karar verir. Ağırlık matrisleri, her bir kapının ne kadar etkileneceğine karar verir. Aşağıda, bu sistemin işlem adımlarını açıklayan LSTM denklemleri bulunmaktadır:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (4)$$
$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \quad (5)$$
$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (6)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (7)$$
$$h_t = o_t tanh(c_t) \quad (8)$$

Yukarıdaki denklemlerde W ağırlık matrisini, b bias vektörünü ve σ sigmoid aktivasyon fonksiyonunu ifade eder. BiLSTM modelinin çift yönlü olması nedeniyle bu denklemlerin aynısı hem ileri hem geri yönlü olmak üzere uygulanır.

2.3.2.3.Bidirectional Gated Recurrent Unit

Kapılı tekrarlayan üniteler, LSTM sisteminin basitleştirilmiş bir versiyonu olan RNN'nin bir alt tipidir.



Şekil 6. GRU yapısı

Şekil 6 incelendiğinde GRU yapısının LSTM'den farkı, tek bir kapı hem unutma hem de güncelleme kararını aynı anda kontrol etmesidir. (Goodfellow ve diğerleri, 2016)

BiGRU modeli Denklem (9)'ten Denklem (17)'e formüle edilmiştir (Bhuvaneswari ve diğerleri, 2019):

Right Direction:

$$\overrightarrow{z_t}^{(i)} = \sigma\left(\vec{W}_{(i)}^{(z)}x_t^i + \vec{U}_{(i)}^{(r)}h_{t-1}^{(i)}\right) \quad (9)$$
$$\overrightarrow{r_t}^{(i)} = \sigma\left(\vec{W}_{(i)}^{(r)}x_t^i + \vec{U}_{(i)}^{(r)}h_{t-1}^{(i)}\right) \quad (10)$$
$$\tilde{\vec{h}}_t^{(i)} = \tanh(\vec{W}_{(i)}x_t + r_t^{\circ}\vec{U}_{(i)}h_{t-1}) \quad (11)$$
$$\overrightarrow{h_t}^{(i)} = z_t^{(i)\circ}h_{t-1}^{(i)} + \left(1 - z_t^{(i)}\right)^{\circ}\tilde{h}_t^{(i)} \quad (12)$$

Left Direction:

$$\overleftarrow{z_t}^{(i)} = \sigma\left(\overleftarrow{W}_{(i)}^{(z)}x_t^i + \overleftarrow{U}_{(i)}^{(r)}h_{t-1}^{(i)}\right) \quad (13)$$
$$\overleftarrow{r_t}^{(i)} = \sigma\left(\overleftarrow{W}_{(i)}^{(r)}x_t^i + \overleftarrow{U}_{(i)}^{(r)}h_{t-1}^{(i)}\right) \quad (14)$$
$$\overleftarrow{\tilde{h}_t}^{(i)} = \tanh\left(\overleftarrow{W}_{(i)}x_t + r_t^{\circ}\overleftarrow{U}_{(i)}h_{t-1}\right) \quad (15)$$
$$\overleftarrow{h_t}^{(i)} = z_t^{(i)\circ}h_{t-1}^{(i)} + \left(1 - z_t^{(i)}\right)^{\circ}h_t^{(i)} \quad (16)$$

Output:

$$y_t = \text{softmax}(U\left[\overrightarrow{h_t}^{(top)}, \overleftarrow{h_t}^{(top)}\right] + a) \quad (17)$$

Denklemlerde, güncelleme birimi (z_t), sıfırlama birimi (r_t) ve aday gizli durum ($\tilde{h}_t$) formüle edilir. Mevcut girişi x_t, önceki zaman durumundaki gizli katmanını h_{t-1} temsil eder. Sembollerin üzerindeki ok işareti sağ ve sol yönleri gösterir. W sembolleri ağırlık matrisini ve b sembolleri bias vektörünü ifade eder. Çıkışta sağ ve sol yön matrisleri birleştirilmiştir. Softmax, softmax aktivasyon fonksiyonunu temsil eder.

2.4.Connectionist Temporal Classification

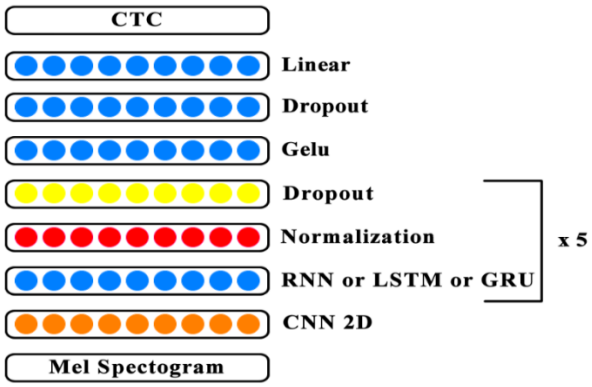
RNN sınıflandırma için kullanılan bir katman olan Bağlantıcı Zamansal Sınıflandırma (Connectionist Temporal Classification) modeli, akustik modelin çıktılarını etiket dizisi üzerinde koşullu bir olasılık dağılımına dönüştürmek için bir sınıflandırıcı olarak kullanılır. CTC modeli Türk alfabesinde kullanılan 29 harf, "x", "w"," **q**" , " **.** ","," ve boşluk dahil olmak üzere 35 karakterden birini çıktı olarak verir. Hedeflenen metin bilgisi ile tahmin edilen metin bilgisi arasındaki farka hata oranı denir. Etiket hata oranı (Label Error Rate) formülü aşağıda verilmiştir.

$$LER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \quad (18)$$

Denklem 18'de iki metin arasında yapmak dönüşüm yapmak için eklenecek harf adedi (I), silinecek harf adedi (D), değiştirecek harf adedi (S) ve doğru olarak bilinen harf adedi (C) ile iki dize arasındaki benzerlik ölçümüdür. N ise iki dizi arasındaki referans ifadenin boyutunu gösterir.

3. Deneysel Kurulum ve Sonuçlar

Bu çalışmada Şekil-7'de gösterildiği üzere konuşma tanıma sistemi kullanılmıştır. Şekil 7'den de gösterildiği gibi bu çalışmada akustik model olarak kullanılan BiRNN, BiLSTM ve BiGRU modellerinin performansları karşılaştırılmıştır. Kullanılan modellerde model haricindeki katmanlar ve parametreler sabit bırakılmıştır.



Şekil 7. Konuşma Tanıma Modeli

3.1. Model Hiperparametreleri

Kullanılan modeller Python programlama dilindeki PyTorch ve Torchaudio kütüphaneleri kullanılarak oluşturulmuştur. Kullanılan modellerde, Şekil 2'de gösterildiği gibi Torchaudio kütüphanesi kullanılarak Mel Spectrogram öznitelikleri elde edilmiştir. Şekil 7'de gösterildiği gibi, Mel Spectrogram çıktısı CNN modelinin girişi olarak kullanılır ve CNN'nin çıkışı beş katmanlı BiRNN, BiLSTM veya BiGRU modelin girişidir. Modelin son katmanında sınıflandırıcı olarak CTC kullanılmıştır.

Bu çalışmadaki modeller birçok parametre kullanmaktadır. Bu hiperparametrelerin değerleri modelin performansını değiştirebilir. Modelde kullanılan hiperparametreler şu şekilde açıklanmıştır:

- **Öznitelik Çıkarımı:** Torchaudio kitaplığını kullanarak, konuşma sinyalini 20 ms'lik bir Hanning penceresi (%50 örtüşme ile) kullanarak çerçevelere bölünmüştür. Her çerçeveden Mel Spectrogram özniteliği çıkarılmıştır.
- **Katman Sayısı:** Tek 2D-CNN katmanı ve 32 filtre kullanılmıştır. Her bir filtrenin kernel boyutu (3,3) olup atlama(stride) boyutu (2,2)'dir. BiRNN, BiLSTM ve BiGRU modellerinde ise 5 katman kullanılmıştır.
- **Eğitim Döngüsü (Epoch):** 300 döngü çalıştırılmıştır.
- **Düğüm Sayısı (Node)**: BiRNN, BiLSTM ve BiGRU modellerinde her katmanda 256 düğüm kullanılmıştır.
- **Öğrenme oranı**: 0.0005 öğrenme adımı kullanmıştır.
- **Eğitim ve Test Oranı:** %97'si eğitim ve %3'ü test için ayrılmıştır.
- **Grup Boyutu (Batch Size):** Grup boyutu 4 olarak alınmıştır.
- **Aktivasyon Fonksiyonu:** Modellerin yapısı gereği sigmoid, tanh ve softmax kullanılmıştır. Ayrıca gelu aktivasyon fonksiyonu bir katman olarak kullanılmıştır.

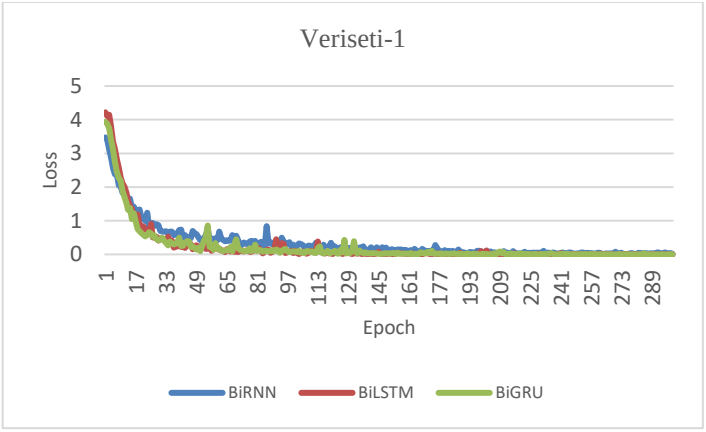
Karşılaştırılan BiRNN, BiLSTM ve BiGRU modellerinin farklı iç yapılarından dolayı katman ve düğüm sayıları aynı olmasına rağmen parametre sayıları birbirinden farklıdır. Akustik modelin toplam parametre sayısı Tablo 1'de gösterilmektedir. Tablo 1'de görüldüğü gibi BiRNN modeli en az toplam parametre sayısına sahip model, en fazla toplam parametre sayısına sahip model ise BiLSTM'dir.

Tablo 1. Toplam parametre sayıları

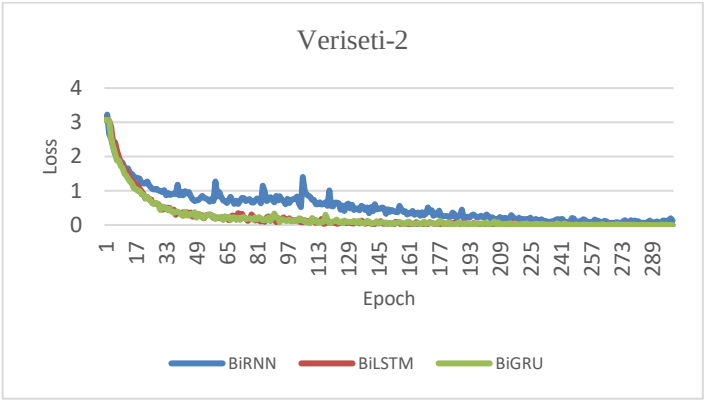
BiRNN	2906467
BiLSTM	11179363
BiGRU	8421731

3.2. Deneysel Sonuçlar ve Tartışma

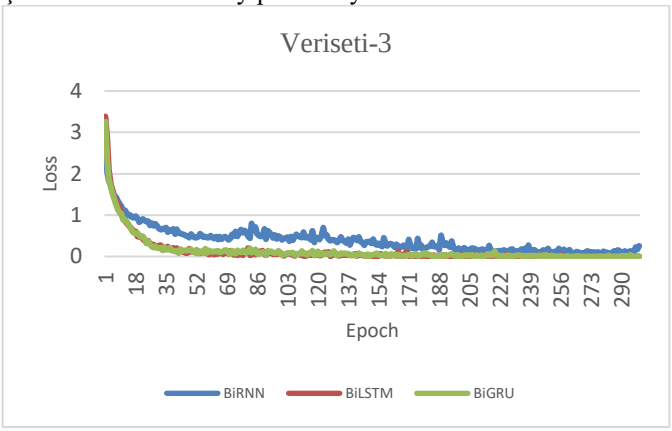
Şekil 8, 9 ve 10 sırasıyla Veriseti-1, Veriseti-2 ve Veriseti-3 için BiRNN, BiLSTM ve BiGRU modellerinin kayıp değerlerinin epoch değerine göre değişimini verilmiştir. Üç şekilden de anlaşılacağı üzere BiLSTM ve BiGRU akustik modelleri için epoch değeri arttıkça kaybın düzenli olarak azaldığı görülmektedir. BiRNN için aynı kararlılıkla kayıp değerinin azalmadığı görülmektedir. Özellikle veri setinin boyutu büyüdükçe BiRNN akustik modeli BiLSTM ve BiGRU modellerine göre daha kötü bir loss değeri vermektedir.



Şekil 8. Veriseti-1 Kayıp Fonksiyonu



Şekil 9. Veriseti-2 Kayıp Fonksiyonu



Şekil 10. Veriseti-3 Kayıp Fonksiyonu

Tablo 2, 3 ve 4'te görüldüğü üzere Veriseti-1, Veriseti-2 ve Veriseti-3 için tüm modellerin LER, toplam loss ve eğitim süresi

verilmiştir. İlgili tablolarda da anlaşılacağı üzere her bir veri setinde BiRNN modelinin en kötü LER, BiLSTM modelinin ise en iyi LER verdiği görülmektedir. Bu sonuçtan BiLSTM akustik modelinin tüm veri setleri için en başarılı model olduğu ve en düşük LER değerini verdiği sonucuna ulaşılabilir. Ancak tüm veri setleri için BiGRU'nun LER'i BiLSTM'in LER oranına yakın sonuçlar olduğu tablolardan görülebilir.

Eğitim sürelerini karşılaştırılan modellerden BiGRU Tablo 2'den de görüleceği gibi, yaklaşık 5 saatlik veri seti için en düşük eğitim süresine sahiptir. BiRNN, yaklaşık 10 saat ve 20 saatlik veri setleri için en kısa eğitim süresine sahip görünüyor. Her bir veri seti için en uzun eğitim süresine BiLSTM'nin sahip olduğu görülmektedir.

Kullanılan modellerden BiRNN, LER oranı diğer modellere göre çok yüksek olduğundan BiRNN akustik modelinin tercih edilmesinin anlamsız olduğu sonucuna varılabilir. BiLSTM ve BiGRU modellerinin LER'i karşılaştırıldığında Tablo 2, 3 ve 4'te BiLSTM'in her bir veri seti için BiGRU'ya göre yaklaşık %3 daha düşük LER oranı olduğu görülmektedir. Bu nedenle LER değeri çok önemliyse akustik model olarak BiLSTM kullanmak daha uygun görünmektedir.

Tablo 2, 3 ve 4'te Çalışma süresileri karşılaştırıldığında veri setleri için BiLSTM modelinin çalışma süresinin BiGRU modeline göre yaklaşık %9 daha uzun olduğu görülmektedir. Ayrıyeten Tablo 1'den de görüldüğü üzere BiLSTM'nin parametre sayısı BiGRU'nundan takribi %33 fazladır. BiGRU'nun akustik model olarak kullanılması, çalışma süresi ve parametre sayısı göz önüne alındığında BiLSTM'den daha avantajlıdır. BiLSTM daha iyi LER sonuçları vermesine rağmen, BiGRU daha hızlı eğitilebilir ve BiLSTM'den daha az parametreye sahiptir. Bu nedenle BiGRU, BiLSTM'den %3 daha kötü sonuçlar vermesine rağmen BiLSTM yerine BiGRU tercih edilebilir.

Tablo 2, 3 ve 4'ten çıkarılabilecek sonuçlardan bir diğeri, veri seti boyutu arttıkça tüm akustik modellerde LER'in düştüğüdür.

Tablo 2. Veriseti-1 Model Karşılaştırması

Veriseti-1

MODELS	LER	EĞİTİM SÜRESİ (dakika)	TOPLAM LOSS
CNN-BiRNN	0.3336	356	101,4044
CNN-BiLSTM	0.2586	378	68,6523
CNN-BiGRU	0.2669	346	66,9942

Tablo 3. Veriseti-2 Model Karşılaştırması

Veriseti-2

MODELS	LER	EĞİTİM SÜRESİ (dakika)	LOSS
CNN-BiRNN	0.2935	644	164,193
CNN-BiLSTM	0.2109	761	68,7557
CNN-BiGRU	0.2282	710	67,8758

Tablo 4. Veriseti-3 Model Karşılaştırması

Veriseti-3

MODELS	LER	EĞİTİM SÜRESİ (dakika)	LOSS
CNN-BiRNN	0.2509	1274	119,4506
CNN-BiLSTM	0.1797	1594	39,8228
CNN-BiGRU	0.1968	1467	41,3766

Yalman ve Tüfekçi (2022) önerdiği tek yönlü RNN, LSTM ve GRU modeli ile bu çalışmada bulunan sonuçların karşılaştırılması tablo 5'te gösterilmiştir.

Tablo 5. Model Karşılaştırması

	Veriseti-1(Yaklaşık 5 Saat)	Veriseti-2(Yaklaşık 10 Saat)	Veriseti-3(Yaklaşık 20 Saat)
MODELS	LER	LER	LER
CNN-RNN	0.4176	0.3592	0.4561
CNN-LSTM	0.3284	0.2825	0.2329
CNN-GRU	0.3398	0.2920	0.2485
CNN-BiRNN	0.3336	0.2935	0.2509
CNN-BiLSTM	**0.2586**	**0.2109**	**0.1797**
CNN-BiGRU	0.2669	0.2282	0.1968

Çift yönlü modellerin tek yönlü modellere göre konuşma tanıma başarısında daha iyi olduğu görülmektedir.

4. Sonuç

Yaklaşık olarak 5, 10 ve 20 saatlik veri seti kullanılan bu çalışmada veri seti büyüklüğünün konuşma tanıma oranına etkisi incelenmiştir. Ayrıca RNN, LSTM ve GRU modellerinin çift yönlü versiyonlarının tek yönlü versiyonları ile karşılaştırması yapılmıştır.

Deneysel çalışmalardan çıkarılan sonuçlara göre veri seti boyutunun artışı konuşma tanıma başarı oranına olumlu katkıda bulunmuştur ve LER değerinin düştüğü gözlemlenmiştir. Bu durumdan daha büyük veri seti kullanmanın önemli olduğu sonucu çıkarılabilir.

Deneyler, BiLSTM modelinin en yüksek konuşma tanıma başarı oranına sahip olduğunu, BiRNN modelinin ise diğer iki modele kıyasla çok düşük bir tanıma oranına sahip olduğunu gösteriyor. Bu nedenle BiRNN'i akustik model olarak kullanmanın hiçbir avantajı yoktur. BiLSTM modeli BiGRU modeline göre yaklaşık %33 daha fazla parametreye sahip olmasına rağmen BiGRU ve BiLSTM modellerinin konuşma tanıma başarı oranları benzer sonuçlar vermektedir. BiLSTM modeli BiGRU modeline göre daha yüksek eğitim süresi olduğunu da gösterilmiştir. Bu nedenle akustik model olarak BiLSTM yerine BiGRU kullanılması daha az parametre ve daha kısa eğitim süresi nedeniyle daha avantajlı olabilir.

BiRNN, BiLSTM ve BiGRU modellerinin RNN, LSTM ve GRU modellerine göre daha başarılı LER sonucu vermesi çift yönlü modellerin daha başarılı olduğunu göstermektedir.

Daha büyük veri seti ile daha çok katman ve düğüme sahip modellerin denemesi gelecekte hedeflenmektedir.

Kaynakça

Arisoy, E., Sethy, A., Ramabhadran, B., & Chen, S. (2015, April). Bidirectional recurrent neural network language models for automatic speech recognition. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5421-5425). IEEE.

Bhuvaneswari, A., Thomas, J. T. J., & Kesavan, P. (2019). Embedded Bi-directional GRU and LSTMLearning Models to Predict Disasterson Twitter Data. Procedia Computer Science, 165, 511-516.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning (pp. 369-376).

Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). Ieee.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Mporas, I., Ganchev, T., Siafarikas, M., & Fakotakis, N. (2007). Comparison of speech features on the speech recognition task. Journal of Computer Science, 3(8), 608-616.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv preprint arXiv:1402.1128.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 45(11), 2673-2681.

Shewalkar, A. (2019). Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. Journal of Artificial Intelligence and Soft Computing Research, 9(4), 235-245.

Tak, R. N., Agrawal, D. M., & Patil, H. A. (2017, December). Novel phase encoded mel filterbank energies for environmental sound classification. In International Conference on Pattern Recognition and Machine Intelligence (pp. 317-325). Springer, Cham.

Yalman, H. İ., & Tüfekci, Z. (2022). Yeni Bir Türkçe Sesli Kitap Veri Seti Üzerinde Convolutional RNN+ CTC, LSTM+ CTC ve GRU+ CTC Modellerinin Karşılaştırılması. Avrupa Bilim ve Teknoloji Dergisi, (34), 321-327.

Zeyer, A., Doetsch, P., Voigtlaender, P., Schlüter, R., & Ney, H. (2017, March). A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 2462-2466). IEEE.

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into deep learning. arXiv preprint arXiv:2106.11342.