Journal of Computer Science
https://dergipark.org.tr/en/pub/bbd

ISSN,e-ISSN: 2548-1304
Volume:7 , Issue:2 , pp:54-67, 2022
https://doi.org/10.53070/bbd.1113379
Research Paper

Konuşma Duygu Tanıma için Akustik Özelliklere Dayalı LSTM Tabanlı Bir Yaklaşım

An LSTM-Based Approach with Acoustic Features for Speech Emotion Recognition

Kenan DONUK*[1], Davut HANBAY[2]

[1]Bilgisayar Programcılığı Bölümü, Şırnak Üniversitesi, Cizre Meslek Yüksekokulu, Şırnak, Türkiye

[2]Bilgisayar Mühendisliği Bölümü, İnönü Üniversitesi, Malatya, Türkiye

(kenandonuk@sirnak.edu.tr, davut.hanbay@inonu.edu.tr)

Received:May 6, 2022 Accepted:Jun. 21, 2022 Published:Dec.7,2022

Özetçe— Konuşma duygu tanıma, konuşma sinyallerinden insan duygularını gerçek zamanlı olarak tanıyabilen aktif bir insan-bilgisayar etkileşimi alanıdır. Bu alanda yapılan tanıma görevi, duyguların karmaşıklığı nedeniyle zorlu bir sınıflandırma örneğidir. Etkili bir sınıflandırma işleminin yapılabilmesi yüksek seviyeli derin özelliklere ve uygun bir derin öğrenme modeline bağlıdır. Konuşma duygu tanıma alanında yapılmış birçok sınıflandırma çalışması mevcuttur. Bu çalışmalarda konuşma verilerinden duyguların doğru bir şekilde çıkarılması için birçok farklı model ve özellik birleşimi önerilmiştir. Bu makalede konuşma duygu tanıma görevi için bir sistem önerilmektedir. Bu sistemde konuşma duygu tanıma için uzun-kısa süreli bellek tabanlı bir derin öğrenme modeli önerilmiştir. Önerilen sistem ön-işlem, özellik çıkarma, özellik birleşimi, uzun-kısa süreli bellek ve sınıflandırma olmak üzere dört aşamadan oluşmaktadır. Önerilen sistemde konuşma verilerine ilk olarak kırpma ve ön-vurgu ön-işlemleri uygulanır. Bu işlemlerden sonra elde edilen konuşma verilerinden Mel Frekans Kepstrum Katsayıları, Sıfır Geçiş Oranı ve Kök Ortalama Kare Enerji akustik özellikleri çıkarılarak birleştirilir. Birleştirilen bu özelliklerin uzamsal bilgilerinin yanında zaman içindeki akustik değişimleri sistemde önerilen uzun-kısa süreli bellek ve buna bağlı bir derin sinir ağı modeliyle öğrenilir. Son olarak softmax aktivasyon fonksiyonu ile öğrenilen bilgiler 8 farklı duyguya sınıflandırılır. Önerilen sistem RAVDESS ve TESS veri setlerinin birlikte kullanıldığı bir veri kümesinde test edilmiştir. Eğitim, doğrulama ve test sonuçlarında sırasıyla %99.87 , %85.14 , %88.92 oranlarında doğruluklar ölçülmüştür. Sonuçlar, son teknoloji çalışmalardaki doğruluklarla kıyaslanmış önerilen sistemin başarısı ortaya konmuştur.

Anahtar Kelimeler : *Derin öğrenme, konuşma duygu tanıma, LSTM, ravdess, tess.*

Abstract— Speech emotion recognition is an area of active human-computer interaction that can recognize human emotions from speech signals in real time. The recognition task in this area is an example of a difficult classification due to the complexity of emotions. An effective classification process depends on high-level deep features and an appropriate deep learning model. There are many classification studies in the field of speech emotion recognition. In these studies, many different models and combinations of features have been proposed to accurately extract emotions from speech data. In this article, a system for speech emotion recognition task is proposed. In this system, a long-short-term memory-based deep learning model is proposed for speech emotion recognition. The proposed system consists of four stages: preprocessing, feature extraction, feature combination, long-short-term memory and classification. In the proposed system, the clipping and pre-emphasis pre-processes are applied to the speech data first. After these processes, Mel Frequency Kepstrum Coefficients, Zero Crossing Ratio and Root Mean Square Energy acoustic properties are extracted from the obtained speech data and combined. In addition to the spatial information of these combined features, their acoustic changes over time are learned with the proposed long-short-term memory and a deep neural network model associated with it. Finally, the information learned is classified into 8 different emotions by the softmax activation function. The proposed system has been tested on a dataset using RAVDESS and TESS datasets together. Accuracies of 99.87%, 85.14% and 88.92% were measured in training, validation and test results, respectively. The results were compared in terms of the accuracies in the recent studies and the success of the proposed system was revealed.

Keywords : *Deep learning, speech emotion recognition, long short-term memory, ravdess, tess.*

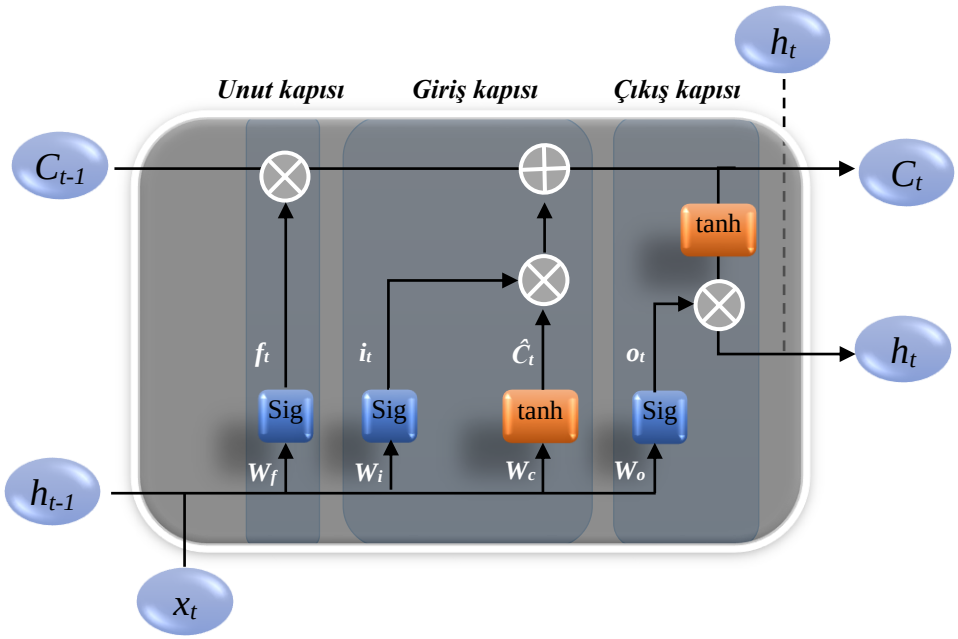
1. Giriş

Günümüz hayat koşullarında insanların bilgisayarlar ile olan etkileşimi hiç olmadığı kadar artmıştır. İnsan-bilgisayar etkileşimindeki bağların güçlendirilmesi ve birçok alanda insanların duygu durumları hakkında uzman sistemler tarafından otomatik tespitler yapılabilmesi için Konuşma Duygu Tanıma (KDT) sistemi ortaya çıkmıştır. İnsan-bilgisayar etkileşimi alanındaki önemli araştırma konularından biri olan KDT, konuşma sinyallerinin özniteliklerinin kullanılarak duygu sınıflandırmalarının yapıldığı bir sistemdir. KDT'de amaç insanların konuşma sinyallerinden duygusal durumları hakkında tespitler yapmaktır. KDT, insandaki duygusal durumların karmaşıklığının yanında insandan insana değişen duyguyu aktarmadaki ses farklılıkları nedeniyle zorlu bir sınıflandırma görevidir. Konuşma duygu tanıma probleminde başarılı bir tanımanın gerçekleştirilebilmesi için ses sinyallerinden uygun özniteliklerin çıkarılmasına ve etkili bir sınıflandırma modeline ihtiyaç vardır. Birçok farklı alanda uygulanan derin öğrenme yöntemlerinin konuşma tanıma alanında da başarılı bir şekilde uygulandığı mevcut çalışmalarla ortaya konmuştur. Derin öğrenme algoritmalarının önerildiği bazı KDT çalışmaları şunlardır.

Cai L. ve ark. (Cai ve diğerleri, 2020) IEMOCAP veri setini kullandıkları bu çalışmalarında çok modlu bir duygu tanıma sistemi önermişlerdir. Duygu tanıma için sisteme hem konuşma ifadesi hem de yüz ifadesi bilgisi verilmiştir. Konuşma bilgisinden 1B Evrişimli Sinir Ağı (ESA) ve Çift Yönlü Uzun Kısa Süreli Bellek (ÇY-UKSB) kullanılarak uzamsal ve zamansal akustik özellikler çıkarılmıştır. Yüz ifadesinin özellikleri ise küçük 2B evrişimli sinir ağları kullanılarak çıkarılmıştır. Konuşma ve yüz ifadesi öznitelikleri derin sinir ağında birleştirilerek softmax ile sınıflandırma gerçekleştirilmiştir. Elde edilen sonuçlar yüz ve konuşma özniteliklerinin birleştirilmesinin tek modlu sınıflandırmadan büyük ölçüde iyi performans verdiği görülmüştür. Issa D. ve ark. (Issa ve diğerleri, 2020) yaptıkları çalışmada konuşma duygu tanıma için girdi verisi olarak Mel-Frekans Kepstrum Katsayıları (MFKK), kromagram, mel-ölçek spektrogramı, tonnetz gösterimi ve spektral karşıtlık ses özelliklerinin kombinasyonu ile bir boyutlu (1B) ESA kullanan yeni bir yöntem önermişlerdir. Önerilen yöntem RAVDESS ve IEMOCAP veri kümeleri için son teknolojilere göre daha iyi performans gösterirken EMO-DB veri kümesi için ise önerilen modele dayalı olarak oluşturulan E modeli (Model E) önceki tüm çalışmalardan daha iyi bir performans göstermiştir. Atila O. ve Şengür A. (Atila ve Şengür, 2021) konuşma tabanlı duygu tanıma için dikkat katmanına sahip 3D ESA-UKSB modelini önermişlerdir. Bu model 28 katmana sahiptir. Modelde ses sinyallerinin spektrogram, mel-frekans kepstrum katsayısı, kokleagram ve fraktal boyut görüntü matrislerini içeren dört boyutlu bir hacim giriş verisi olarak kullanılmaktadır. 10 kat çapraz doğrulama ile önerilen modeldeki SAVEE, RML, RAVDESS veri setleri ile yapılan sınıflandırma deneylerindeki doğruluklar sırasıyla %87.5, %93.2, %96.18 olarak tahmin edilmiştir. Mujaddidurrahman A. ve ark. (Mujaddidurrahman ve diğerleri, 2021) yaptıkları çalışmada log-mel spektrogramının girdi olarak kullanıldığı 2D-ESA modeli önermişlerdir. EMODB veri setini kullandıkları çalışmalarında veri artırmanın önemine dikkat çekmişlerdir. Önerilen yöntemin EMODB veri seti ile veri arttırımı yapılmadan %74.0 doğruluğa ulaşılırken veri artırımı ile %88.0'lik sınıflandırma doğruluğuna ulaşılmıştır. Padi S. ve ark. (Padi ve diğerleri, 2020) konuşma duygu tanıma için çok pencereli bir veri arttırma yaklaşımı önermişlerdir. Önerdikleri yöntemde tek pencereli özellik çıkarımı ile birden fazla pencerenin kullanıldığı özellik çıkarmanın doğruluğa etkisi incelenmiştir. Duygu sınıflandırma için ESA tabanlı bir modelin kullanıldığı sistem SAVEE ve RAVDESS veri setleri üzerinde test edilmiştir. Her iki veri setinde de tek pencereli özellik çıkarımına kıyasla sırasıyla %70 ve %88 oranlarında daha iyi performans elde edilmiştir. Nasim A. S. ve ark. (Nasim ve diğerleri, 2021) RAVDESS ve TESS veri setlerini birleştirerek bu verilerden MFKK, Kroma ve Mel spektrogram özelliklerini sırasıyla 40, 12, 128 olmak üzere toplamda 180 özellik çıkarmışlardır. Bu özellikler birleştirilerek Gradient Boost, Decision Tree, Gaussian Naive Bayes, Multi-Layer Perception (MLP), Random Forest, AdaBoost, KNN, Logistic Regression, SVM, Linear SVM, Stokastik Gradient Descent gibi farklı algoritmalar üzerinde sınıflandırma doğruluğuna dayalı performans karşılaştırması yapmıştır. Karşılaştırmalar sonunda Gradient Boost algoritması ile 84.69% test doğruluğuna ulaşmışlardır. Asiya U. A. ve Kiran V. K. (Asiya ve Kiran, 2021) konuşma duygu tanıma görevi için RAVDESS ve TESS veri setlerini birleştirmişlerdir. Bu veri setlerinden sıfır geçiş oranı, mel-pektrogram, kroma, Mel Frekans Kepstrum Katsayıları ve Kök Ortalama Kare özelliklerini elde etmişlerdir. Bu özellikler kullanılarak 1D-ESA tabanlı bir mimari kullanarak sınıflandırma gerçekleştirmişlerdir. Modelin test doğruluğu 89% olarak ölçülmüştür. Öztürk Ö. F. ve Pashaei E. (Öztürk ve Pashaei, 2021) yaptıkları çalışmada RAVDESS ve TESS veri setlerini birleştirerek önerdikleri ESA-UKSB (Co-LSTM) ağı üzerinde konuşma duygu tanıma sınıflandırması gerçekleştirmişlerdir. Evrişimli Sinir Ağı ile Uzun Kısa Süreli Belleği birleştirdikleri ağın eğitimi için her konuşma kaydının MFKK özelliklerinin ortalaması alınarak 40 adet MFKK özelliği kullanmışlardır. Eğitilen ağın test doğruluğu 86.77% olarak ölçülmüştür.

2. Uzun Kısa-Süreli Bellek-LSTM

UKSB (Uzun Kısa Süreli Bellek- LSTM) 1997 yılında Sepp Hochreiter ve Jürgen Schmidhuber (Hochreiter ve Schmidhuber, 1997) tarafından ortaya atılmıştır. Sınıflandırma problemlerinde verilerin, zamansal süreçteki bağımlılığını öğrenebilen bir Tekrarlayan Sinir Ağı (TSA-RNN) türüdür. UKSB'ler TSA'dan farklı olarak kaybolan gradyan sorununu büyük oranda ortadan kaldırmıştır. Şekil 1'de UKSB biriminin yapısı gösterilmektedir.



Şekil 1. UKSB yapısı

UKSB birimleri temel olarak sırasıyla unutma, giriş ve çıkış kapılarından oluşmaktadır. Unutma kapısı (f_t), UKSB biriminin durumu olan ve önceki hücre çıkışını temsil eden (h_{t-1}) verileriyle, belirli bir zaman adımıyla ayrılmış giriş verilerinin (x_t) ağırlık matrisi (W_f) ile çarpılıp sapma vektörünün (b_f) eklenip Denklem 1'de verilen lojistik sigmoid işleminden geçmesiyle [0,1] arasında bir değer üretmesi olarak ifade edilir. Elde edilen değerler vektöründen 0'a yakın olan değerler girdi bileşeninin unutulması gerektiğini 1'e yakın değerler ise girdi bileşenin bellekte tutulması gerektiği bilgisini verir. Unutma kapısının çıkış vektörü, UKSB biriminin uzun süreli belleğini temsil eden önceki hücre durumu (C_{t-1}) ile noktasal bir çarpım işlemine tabi tutulmaktadır. Böylelikle uzun süreli bellekte hangi bilgilerin tutulmasının yararlı olacağı tespit edilir. Denklem 2'de verilen giriş kapısı (i_t), önceki hücre durumunu güncellemek için hangi bilgilerin ekleneceğini belirler. Giriş kapısı, h_{t-1} ve x_t girdi bileşenlerinin ağırlık matrisi (W_i) ile çarpılıp sapma vektörünün (b_i) eklenip sigmoid fonksiyonuna girmesi ile elde edilir. [0,1] arası değerlere sahip giriş kapısı vektörü Denklem 3'te verilen güncelleme vektörü ($\hat{C}_t$) ile noktasal çarpım işlemine girerek uzun süreli belleğe eklenecek yeni bilgileri belirler. Güncelleme vektörü, h_{t-1} ve x_t girdi bileşenlerinin ağırlık matrisi (W_c) ile çarpılıp sapma vektörünün (b_c) eklenip tanh fonksiyonuna girmesi ile elde edilir. Böylelikle unut kapısı filtrelemesiyle atılan bilgiler ile giriş kapısı filtrelemesiyle yeni eklenecek bilgiler göz önüne alınarak hücrenin yeni uzun süreli belleği yani yeni hücre durumu (C_t) elde edilir. Yeni hücre durumu Denklem 4'te verilmiştir. Son olarak Denklem 5'te verilen çıkış kapısı (o_t), h_{t-1} ve x_t girdi bileşenlerinin ağırlık matrisi (W_o) ile çarpılıp sapma vektörünün (b_o) eklenip sigmoid fonksiyonuna girmesi ile hücrenin yeni gizli durumunu (h_t) belirlemede kullanılır. Denklem 6'da verilen hücre yeni gizli durumu, h_{t-1} ve x_t girdi bileşenlerinin sigmoid fonksiyonuna girmesiyle elde edilen çıkış vektörü (o_t) hücrenin tanh fonksiyonu uygulanmış yeni uzun süreli belleğine yani yeni hücre durumuna (C_t) noktasal çarpım ile uygulanarak elde edilir.

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \quad (1)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (2)$$

$$\hat{c}_t = tanh(W_c[h_{t-1}, x_t] + b_c) \quad (3)$$

$$c_t = f_t * c_{t-1} + i_t * \hat{c}_t \quad (4)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * tanh(c_t) \quad (6)$$

3. Veri Seti ve Metodoloji

Çalışmada literatürde duygu tanımada yaygın olarak kullanılan RAVDESS (Livingstone ve Russo, 2018; Zenodo, 2022) ve TESS (University of Toronto Dataverse, 2022) olmak üzere iki farklı ses veri seti kullanıldı. Veri setleri detaylı bir şekilde açıklandıktan sonra, veri setlerinden KDT için özellik çıkarma aşamalarına yer verildi.

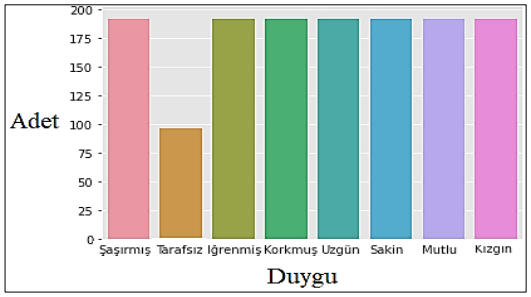
3.1. Veri Seti

RAVDESS: Ravdess veri seti Kuzey Amerika aksanıyla 2 farklı cümleyi seslendiren 12 erkek 12 kadın olmak üzere 24 profesyonel oyuncunun ses ve görsel kayıtlarını içerir. Bu kayıtlarda üzgün, mutlu, kızgın, sakin, korku, şaşırmış, tarafsız ve tiksinti olmak üzere 8 duygu sınıfına sahip konuşma örnekleri mevcuttur. Bu kayıtlar konuşma ve şarkı şeklinde hem ses hem de video şeklinde hazırlanmıştır. Ses formatındaki verilerde konuşma veri setinde 24 oyuncudan her birinin 60 deneme kaydı ile 1440 adet dosya mevcuttur (Livingstone ve Russo, 2018; Zenodo, 2022).

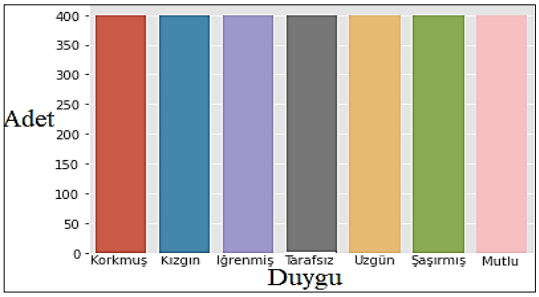
TESS: Anadilleri İngilizce olan üniversite ve müzik eğitimi almış 2 kadın aktris (24 ve 64 yaşlarında) tarafından "Say the word ___' ifadesindeki boşluğa gelecek 200 hedef kelimeden oluşan bir set söylenmiştir. İçinde kızgın, iğrenme, korku, mutlu, şaşırmış, üzgün ve tarafsız şeklinde 7 adet duygu ifadesi içeren toplamda 2800 kayıt oluşturulmuştur (University of Toronto Dataverse, 2022). Veri setlerine ait detaylı bilgiler Tablo 1'de verilmiştir. Ayrıca RAVDESS ve TESS veri setlerinin duygu dağılımları sırasıyla Şekil 2 ve Şekil 3'te verilmiştir.

Tablo 1. Veri setlerinin dağılımı

Veri seti	Konuşmacı sayısı	Veri seti kayıt sayısı	Duygu çeşidi adeti	Dil	Kayıt tipi
RAVDESS	12 kadın 12 erkek	1440	8	İng.	Ses
TESS	2 kadın	2800	7	İng.	Ses
TÜM SETLER	14 kadın 12 erkek	4240	7-8	İng.	Ses



Şekil 2. RAVDESS duygu dağılımı



Şekil 3. TESS duygu dağılımı

3.2. Özellik Çıkarma

Bu kısımda veri setlerinden konuşma duygu tanımada kullanılacak hangi özeliklerin nasıl çıkarıldığı anlatılmıştır. Çalışmada konuşma duygu tanıma için üç ses özelliği kullanılacaktır. Bunlar Kök Ortalama Kare Enerji (KOKE), Sıfır Geçiş Oranı (SGO), Mel Frekans Kepstrum Katsayısı (MFKK)'dır. Bu özellikler aşağıdaki kısımda detaylı bir şekilde açıklanmıştır.

3.2.1. Akustik Özellikler

3.2.1.1. MFKK (Mel Frekans Kepstrum Katsayısı) Özelliği

MFKK, 1980 yılında Davis ve Mermelstein tarafından tanıtılmıştır (Davis ve Mermelstein, 1980). Duygu ve konuşmacı tanımada yaygın olarak kullanılan MFKK ses sinyali bileşenlerini bir dizi katsayı olarak temsil etmektedir. MFKK'ların hesaplanmasında genellikle aşağıdaki adımlar uygulanmaktadır.

1.Ses sinyaline ön-vurgu uygulama

2.Ses sinyalini belirli adımla çerçeveleme

3.Her çerçeveye Hamming pencereleme uygulama

4.Elde edilen her çerçeve için Hızlı Fourier Dönüşümü (HFD) ile güç spektrumunu hesaplama

5.Güç spektrumlarını mel ölçeğine eşleyip filtre bankasını uygulama, her filtredeki gücü hesaplama

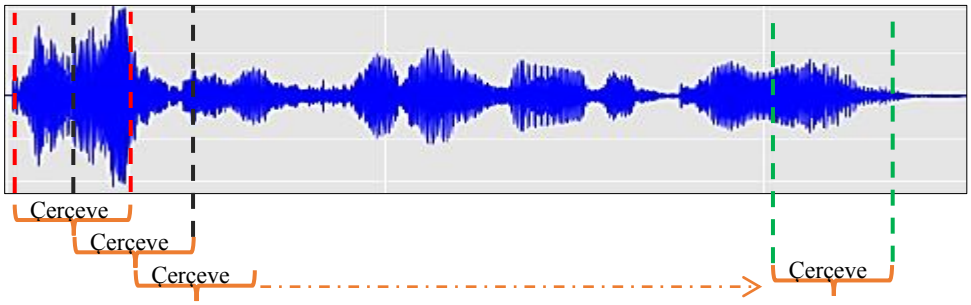
6.Filter bankası Mel-log güçlerinin AKD (Ayrık Kosinüs Dönüşümü)'sini alma

1. Ön-Vurgu: Ses sinyali, frekans spektrumunda incelendiğinde yüksek frekans bantları düşük güce sahiptirler. Ses sinyaline ön-vurgu uygulanarak yüksek frekans bantlarının genliği arttırılarak frekans spektrumu dengelenir. Ses sinyali için ön-vurgu işleminin bazı yararları şunlardır: Frekans spektrumunu dengelemek, hızlı fourier dönüşümündeki sayısal sorunları önlemek ve sinyal gürültü oranını (SNR) iyileştirmek (Chen ve Huang, 2021). Ön-vurgu Denklem 7'de verilmiştir.

$$x'(n) = x(n) - \alpha x(n-1) \quad (7)$$

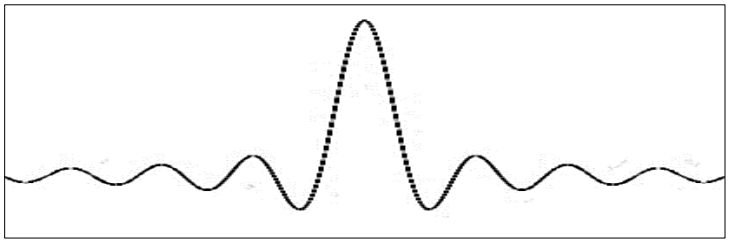
Denklem 7'de verilen $x(n)$ orijinal ses sinyalini, $x'(n)$ değeri ise ön-vurgu uygulanmış yeni ses sinyalini vermektedir. α değeri ise 0.9 ila 1 arasında değişen sabit ön-vurgu parametresidir.

2. Çerçeveleme: Konuşma ses sinyali genellikle durağan olmayan bir forma sahiptir. Ses sinyalinin yapısı zaman içinde değiştiğinden ses sinyalinin tamamından ses bileşenlerini çıkarmak sinyalin özellik temsili açısından uygun bir yöntem değildir. Sürekli konuşma sinyallerinin sabit uzunlukta bölümlerde durağan ve kararlı olduğu bilinir. Bu nedenle sinyal kısa zaman çerçevelerine bölünür. Çerçeve süresinin dikkatli seçilmesi önemlidir çünkü çerçeve süresi çok uzunsa çerçeve içinde sinyal özellikleri değişir. Çerçeve süresi boyutu genellikle 20 ms'dir (Ancilin ve Milton, 2021). Şekil 4'te bir sinyale uygulanan çerçeveleme gösterilmiştir.



Şekil 4. Ses sinyaline uygulanan çerçeveleme

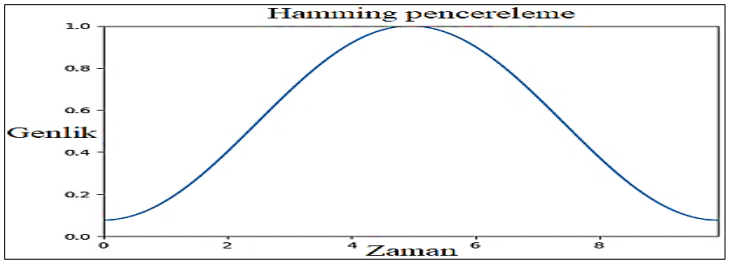
3. Hamming Pencereleme: Çerçevelenmiş ses sinyali parçalarının frekans bileşenleri elde edilirken keskin frekans değerleri yerine Şekil 5'te gösterilen dalgalı bir yapıya sahip bir frekans formuyla karşılaşılır. Burda gerçekleşen olay spektral sızıntıdır. Bu durumun sebebi ses sinyalinin bir parçası üzerinden frekans değerlerinin çıkarılıyor olmasıdır. Sesin orijinal frekans bileşenlerinin elde edilebilmesi için sinyal parçası üzerine Hamming pencereleme uygulanarak daha doğru bir frekans spektrumu elde edilir. Bu pencereleme ile sinyalin baş ve son kısımlarına sönümleme uygulanır. Dolayısıyla sinyalin orta kısımları daha çok vurgulanarak spektral sızıntının önüne geçilir. Hamming pencereleme fonksiyonu ve $x(n)$ çerçeve sinyali ile çarpımı sırasıyla Denklem 8 ve Denklem 9'da verilmiştir. Hamming fonksiyonunda yer alan *M*, Hızlı Fourier Dönüşümü (HFD) veri kümesindeki veri miktarını *n*, *0* ila *M* arasındaki bir değeri ifade etmektedir. Hamming fonksiyonu grafiği Şekil 6'da verilmiştir.



Şekil 5. Dalgalı frekans gösterimi

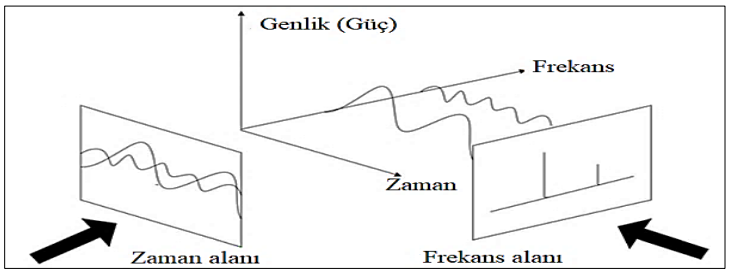
$$w(n) = 0.54 - 0.46cos(\frac{2\pi n}{M-1}) \quad (8)$$

$$x'(n) = x(n) * (0.54 - 0.46\,cos\left(\frac{2\pi n}{M-1}\right)) \quad (9)$$



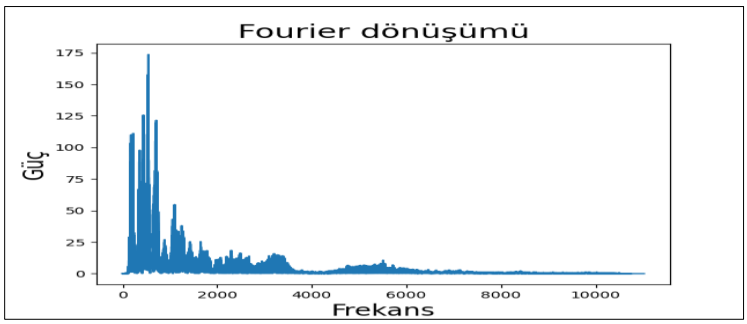
Şekil 6. Hamming pencereleme

4. Hızlı Fourier Dönüşümü (HFD) ve Frekans Spektrogramı: Ses sinyalleri farklı genlik ve frekanslara sahip sinyallerin evrişimi olarak düşünülebilir. Şekil 7'de ses sinyalinin zaman alanından frekans alanına geçişi gösterilmektedir.



Şekil 7. Zaman alanından frekans alanına geçiş (Sun, 2019)

Sesi oluşturan bu farklı sinyallerin frekans ve genliklerinin tespit edilmesi ses sinyali analizinde önemli bir yere sahiptir. Ses sinyalini frekans bileşenlerine ayırmak için Hızlı Fourier Dönüşümü kullanılır. Şekil 8'de bir ses sinyaline HFD ugyulanması ile elde edilen frekans spektrumu gösterilmiştir.

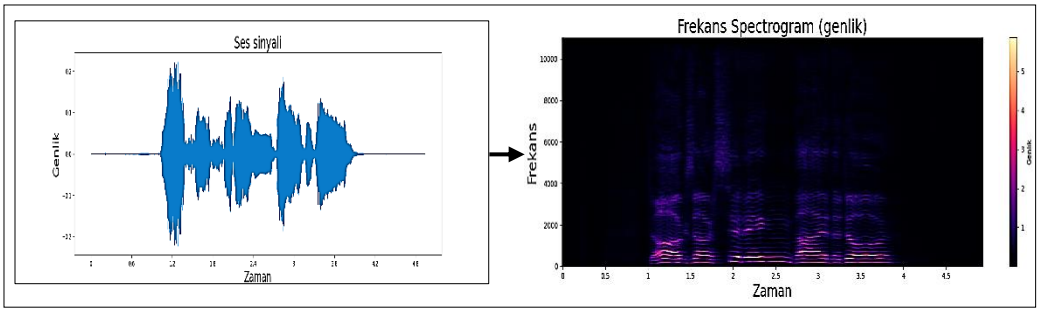


Şekil 8. Frekans spektrumu

HFD'nin hesaplanması Denklem 10'da verilmiştir. Denklemde yer alan *N* çerçevedeki örnek sayısını, *n* mevcut örneği, *x(n) n* örneğindeki sinyalin değerini, *k* mevcut frekansı [*0, N-1 hz*], *X(k)* ses sinyalindeki *k* frekansına ait genlik ve faz değerlerini veren karmaşık bir sayıyı ifade etmektedir. *X(k)* ile elde edilen bir dizi katsayının karesi alınarak mevcut ses sinyalinde yer alan frekanslara ait güçler hesaplanır. Güç hesabı Denklem 11'de verilmiştir. P_n ifadesi ses sinyalinin zaman içindeki frekanslara ait güç değerlerini temsil eden 2B matristir. Ses sinyalinin zamana bağlı frekans spektrogramına (P_n) dönüşümü grafiği Şekil 9'da verilmiştir.

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-i\frac{2\pi kn}{N}} \tag{10}$$

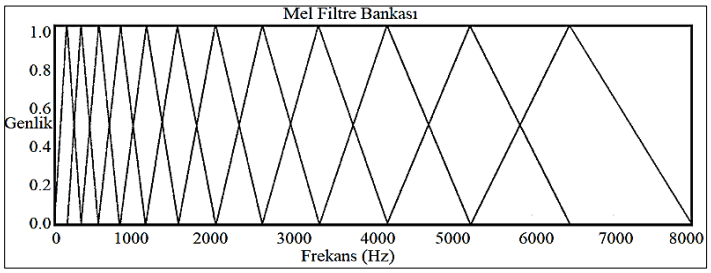
$$P_n = \frac{(HFD(x_n))^2}{N_{HFD}} \tag{11}$$



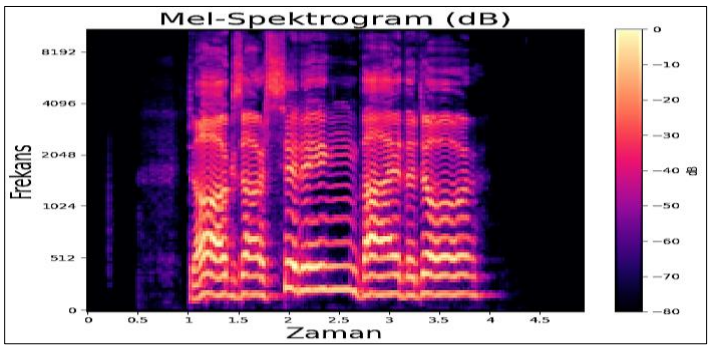
Şekil 9. Zamana bağlı frekans spektrogramı

5.Frekans-Mel Çevrimi ve Filtre Bankası Mel ölçeği: 1937'de Stevens, Volkman ve Newman tarafından önerilmiştir (Stevens ve diğerleri, 1937). İnsan kulağı seslerin frekans değerlerini 1000 Hz'e kadar doğrusal, 1000 Hz den sonraki değerler için logaritmik olarak algılamaktadır. Dolayısıyla insan kulağının algıladığı ses frekansı "Mel" ölçeği adı verilen ampirik bir frekans ölçeği ile ölçülmektedir. Bu nedenle ses sinyallerinden elde edilen frekansların insan kulağının algısına benzer şekilde bir dönüşümden geçmesi gerekmektedir. Bu dönüşüm doğrusal olarak değişen frekans spektrogramının doğrusal olmayan Mel-spektrogramına dönüştürülmesiyle gerçekleşir. Ses frekansının (f) Mel ölçeği türünden hesaplanması Denklem 12'de verilmiştir (O'Shaughnessy, 1987). Mel spektrogramını elde etmek için frekans spektrumuna Mel üçgen filtreleri uygulanır. Her biri farklı frekans aralığındaki Mel filtreleri, HFD uygulanmış çerçevelerdeki frekans değerleri ile çarpılıp toplanarak her bir Mel bandının gücü elde edilir. Şekil 10'da verilen ve üst üste binmiş 12 adet Mel-filtresi başlangıçta birbirlerine yakın durumdalarken yüksek frekanslara çıkıldığında yayılmaya başlarlar. Bu durum Mel ölçeğinin ses algımıza benzerliğini ifade etmektedir. Şekil 11'de ses sinyalinin her çerçevesindeki frekans spektrumuna Mel ölçeğinin uygulanması ile elde edilen Mel spektrogramını göstermektedir.

$$Mel(f) = 2595 * log_{10}(1 + \frac{f}{700}) \quad (12)$$



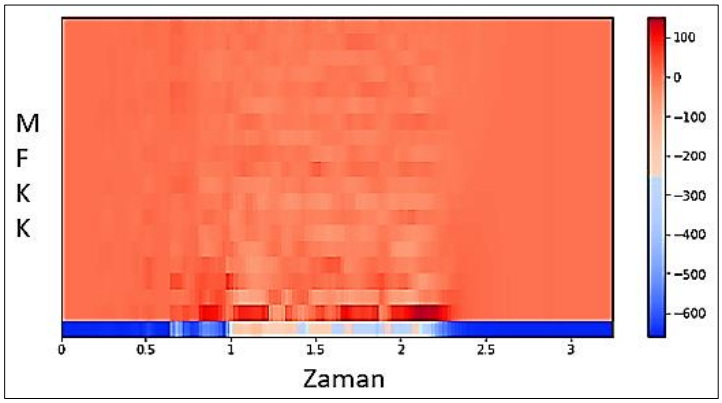
Şekil 10. Mel filtre bankası



Şekil 11. Mel spektrogramı

6. Ayrık Kosinüs Dönüşümü Uygulama (AKD): Son olarak MFKK özelliklerini elde etmek için Log-Mel spektrogramına ayrık kosinüs dönüşümü uygulanır. Ayrık Kosinüs Dönüşümü (AKD), 1972 yılında Nasir Ahmed tarafından önerilen sinyal işleme ve veri sıkıştırmada kullanılan bir dönüştürme tekniğidir (Wikipedia Discrete Cosine Transform, 2022; Ahmed ve diğerleri, 1974). MFKK öznitelik çıkarımında son adım olarak AKD uygulanır. Bu adımda her sinyal çerçevesi için genellikle 13 MFKK katsayısı elde edilir. Bu sayı ses sınıflandırma görevlerinde belirleyici olarak kabul edilir (Silva ve diğerleri, 2020). Her çerçevede elde edilen katsayılar sütun boyunca eklenerek MFKK özellik haritası çıkarılır. Denklem 13'te MFKK katsayısı çıkarım formülü verilmiştir. Formülde yer alan $C_t(n)$, t çerçevesinin n. MFKK katsayısını temsil etmektedir. M, MFKK'ların sayısını gösterir. $X'_n(m)$ değeri ise m. mel filtresinin logaritmik enerjisini göstermektedir (Chen ve Huang, 2021). MFKK katsayılarının spektrogram olarak gösterimi Şekil 12'de verilmiştir.

$$C_t(n) = \sum_{m=0}^{M-1} X'_n(m)\, cos(\frac{\pi n(m-0.5)}{M}) \quad (13)$$

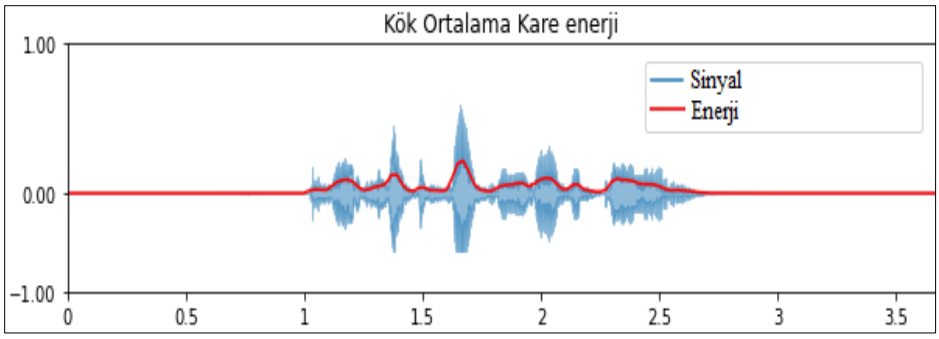


Şekil 12. MFKK spektrogramı

3.2.1.2. Kök Ortalama Kare Enerji-KOKE

Kök Ortalama Kare Enerji ses sinyalinin zaman içinde değişen genlik durumunu veya kısa süreli enerji değişimi hakkında bilgi vermektedir. Enerji ne kadar yüksek olursa sesin o kadar yüksek olduğu anlaşılır. Bu özellik, ses segmentasyonu ve müzik türü sınıflandırma görevlerinde yaygın olarak kullanılmaktadır. KOKE şu şekilde hesaplanır. Çerçevelere ayrılan ses sinyallerinin çerçeve içindeki genliklerinin kareleri alınır ve toplanır. Daha sonra elde edilen değer çerçeve uzunluğuna bölünüp karekökü alınır böylelikle mevcut çerçevenin KOKE'si elde edilir. KOKE formülü ve grafiği sırasıyla Denklem 14 ve Şekil 13'te verilmiştir.

$$KOKE = \sqrt{\frac{1}{n}(x_1{}^2 + x_2{}^2 + x_3{}^2 \dots\dots\dots + x_n{}^2)} \quad (14)$$

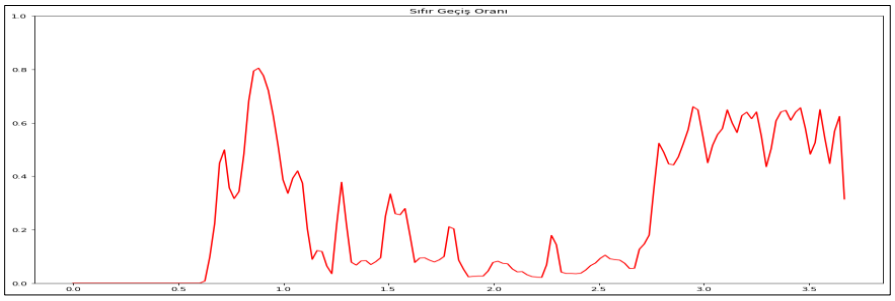


Şekil 13. KOKE grafiği

3.2.1.3. Sıfır Geçiş Oranı-SGO

Bu özellik vurmalı sesleri sınıflandırmada, konuşma tanımada ve müzik bilgilerinin alınmasında yaygın olarak kullanılmaktadır. SGO aynı zamanda bir sinyalin gürültüsünün bir ölçüsü olarak ta yorumlanabilir. Bir ses çerçevesinin Sıfır Geçiş Oranı, çerçevedeki sinyalin işaret değişikliklerinin hızı veya yatay zaman eksenini geçme sayısıdır. Bu sayı çerçevenin uzunluğuna bölünerek SGO hesaplanır (Giannakopoulos ve Pikrakis, 2014; Wikipedia Zero-crossing rate, 2022). SGO formülü ve grafiği sırasıyla Denklem 15 ve Şekil 14'te verilmiştir.

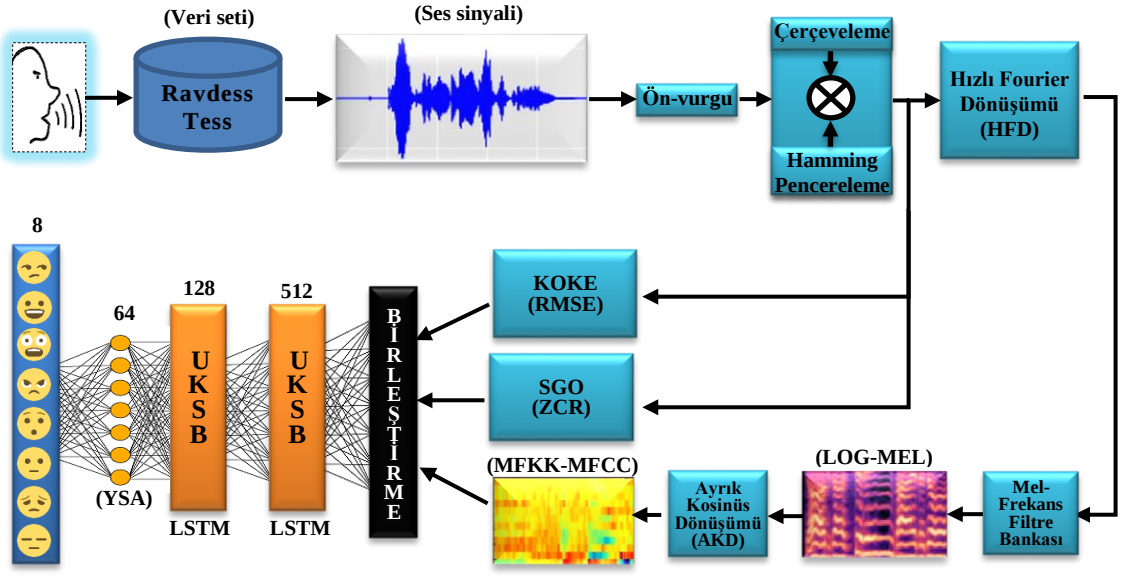
$$SGO = \frac{1}{2N}\sum_{n=1}^{N} |sinyal(x[n]) - sinyal(x[n-1])| \quad (15)$$



Şekil 14. SGO grafiği

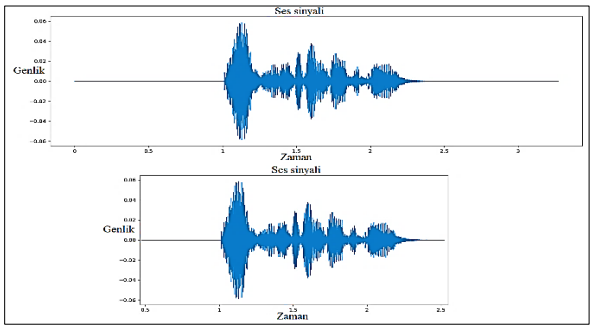
4. Önerilen Sistem

Konuşma duygu tanıma için önerilen sistem Şekil 15'te verilmiştir. Modelin eğitim ve test aşamaları Ravdess, Tess veri setlerinin birleşiminden oluşturulan bir veri kümesi ile gerçekleştirilmiştir. Duygu tanıma için veri setlerindeki konuşma kayıtlarından konuşma sinyallerinin karakteristik özelliklerinin çıkarılması gerekmektedir. Ses verilerinin analizinde kullanılan birçok farklı ses özelliği mevcuttur (Alías ve diğerleri, 2016). Bu çalışmada konuşma tabanlı duygu tanıma için Kök Ortalama Kare Enerji (KOKE), Sıfır Geçiş Oranı (SGO) ve Mel Frekans Kepstrum Katsayısı (MFKK) ses özelliklerinin birleşimi kullanılmıştır. Sistem yapısı 4 kısımdan oluşmaktadır. İlk kısım veri setlerini ön-işlemden geçirme, çerçeveleme ve hamming pencerelemedir. İkinci kısım MFKK, KOKE ve SGO özelliklerinin elde edilmesi ve birleştirilmesidir. Üçüncü kısım birleştirilen bu özelliklerden zamana dayalı bağlamsal bilgi öğreniminin gerçekleştirilebilmesi için sistemde önerilen UKSB tabanlı derin öğrenme ağı ve son kısım olarak softmax aktivasyonuna sahip sınıflandırma katmanından oluşmaktadır.



Şekil 15. Önerilen sistem

Önerilen sistemin ilk kısmında sinyallere ön-işlem uygulanmaktadır. Bu işlemlerden ilki, iki verisetindeki ses sinyallerinin 0.5-2.5 sn dışındaki sessiz kısımlar ses analizlerinde yaygın olarak kullanılan librosa kütüphanesi kullanılarak çıkarılmıştır (Librosa, 2022). Ses sinyallerindeki sessiz kısımların çıkarılmasının sebebi sessiz kısımların model eğitimine katkısının olmamasıdır. Kırpılmış ses sinyali örneği Şekil 16'da gösterilmiştir.



Şekil 16. Kırpılmış ses sinyali

İkinci işlem olarak kırpılmış ses sinyallerine bu defa ön-vurgu işlemi gerçekleştirilir. Ön vurgu işlemi ile yüksek frekansa sahip genlikler daha güçlü temsil edilerek özellik çıkarımına katkıda bulunulur. Bu işlemden sonra ses sinyaline çerçeveleme işlemi gerçekleştirilir. Bu işlem ile ses sinyali kendi içinde kararlı küçük ses sinyali parçalarına ayrılır. Çalışmada 2 veri setinde de ses sinyallerine uygulanan bir çerçeveleme için örnek sayısı 2048, atlama (örtüşme) örnek sayısı ise 512 olarak alınmıştır. Bir ses sinyaline ait çerçeve sayısının nasıl hesaplandığı Denklem 16'da verilmiştir.

$$Çerçeve\ sayısı = \left(\frac{Örneklem\ sayısı - Çerçeve\ uzunluğu}{Atlama\ uzunluğu}\right) + 1 \quad (16)$$

Çalışmada farklı ses kayıtlarının örneklem sayıları farklı olduğundan her ses kaydı aynı örneklem sayısına sahip olacak şekilde ayarlanmıştır. Her ses kaydı için örneklem sayısı olarak 44.100 olarak alınmıştır. Denklem 16 göz önüne alındığında kırpılmış bir ses sinyalinden 84 adet çerçeve üzerinde özellik çıkarımı yapılmıştır. Sinyaller belirli uzunluktaki ses sinyallerine ayrıldıktan sonra her bir ses sinyali parçasına Hamming pencereleme işlevi uygulanır. Bu işlev ile parçalara ayrılan ses sinyallerinde meydana gelebilecek spektral sızıntının önüne geçilerek ileriki aşamalarda elde edilecek frekans spektrumu hakkında daha doğru sonuçlar elde edilmektedir.

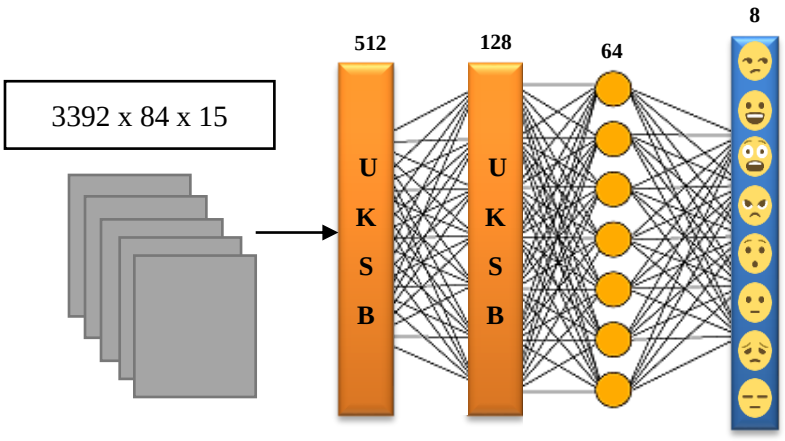
İkinci kısımda Hamming pencereleme işlemi uygulanan çerçeve sinyal parçalarının KOKE, SGO, MFKK özellik değerleri elde edilmiştir. Bu özelliklerin nasıl çıkarıldığı Bölüm 3.2'de detaylı bir şekilde anlatılmıştır. Çalışmada 84 adet çerçeveye bölünen ses sinyalinin her parçası için sırasıyla KOKE için 1 adet özellik, SGO için 1 adet özellik ve MFKK için 13 adet özellik elde edilir. Bu özellikler birleştirilerek 15 birim uzunluğunda bir özellik vektörü elde edilir. Elde edilen bu vektör ile beraber 84 x 15'lik bir 2B özellik matrisi elde edilir. Elde edilen 2B özellik matrisi UKSB-YSA tabanlı ağa hazır forma getirilerek eğitime başlanır. Tablo 2'de ses sinyallerinden çıkarılan özelliklere ait bilgiler verilmektedir.

Tablo 2. Ses sinyallerinden çıkarılan özellik bilgileri

Özellik adı	Özellik miktarı (çerçeve başına)	Toplam özellik miktarı (ses kaydı başına)	Toplam özellik miktarı (Tüm kayıtlar)
KOKE (RMSE)	1	84 x 1	4240 x 84 x 1
SGO (ZCR)	1	84 x 1	4240 x 84 x 1
MFKK (MFCC)	13	84 x 13	4240 x 84 x 13
BİRLEŞME	**15**	**84 x 15**	**4240 x 84 x 15**

Son olarak üçüncü ve dördüncü kısımda, konuşma duygu tanımanın en önemli işlemleri gerçekleşir. Çünkü bu kısımda ses sinyallerinden duygu tanımanın gerçekleştirilebilmesi için duygular arasındaki farkların bir derin öğrenme yöntemiyle öğrenilmesi gerekmektedir. Ses sinyalleri arasındaki farkların öğrenilmesinin yanı sıra her

duyguya ait ses sinyalinin bir zaman serisi içerisindeki değişimide göz önüne alınarak Şekil 17'de verilen UKSB tabanlı model önerilmiştir.

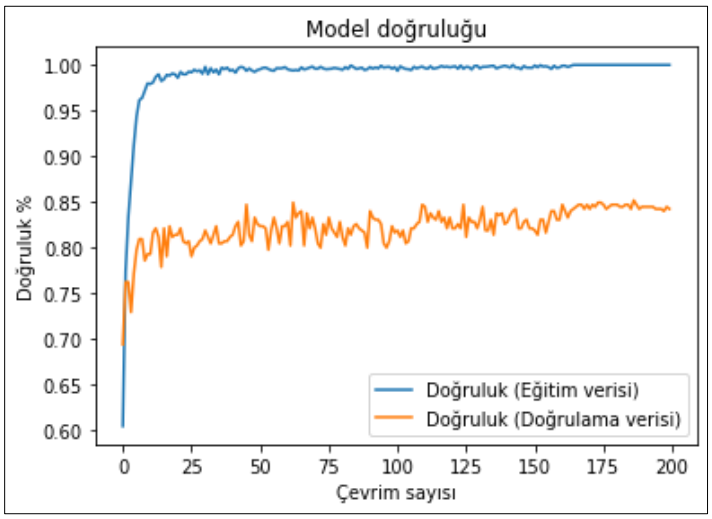


Şekil 17. UKSB tabanlı derin ağ ve sınıflandırma

Şekil 17'de 3392 x 84 x 15 adet eğitim verisi UKSB tabanlı derin öğrenme modeline gönderilir. Eğitim verisi ilk olarak 512 adet UKSB birimiyle bağlamsal bilgiyi öğrenir. Bu birimin çıkışları tam bağlantılı olarak 128 adet UKSB birimine bağlanır. Daha sonra son 128 birimli UKSB çıkışı yine 64 adet nöron ile bağlanır. Bu nöronlar ReLU fonksiyonu ile aktive edilerek sınıflandırma katmanına iletilir. Sınıflandırma katmanı 8 nöron içerir, nöronların aldıkları değere göre softmax aktivasyon fonksiyonu ile 8 farklı duyguya (üzgün, mutlu, kızgın, sakin, korku, şaşırmış, tarafsız ve tiksinti) sınıflandırma gerçekleştirilir.

5. Deneysel Sonuçlar

Eğitime hazır hale getirilen veriler eğitim, doğrulama ve test verilerine sırasıyla %80, %10, %10 oranlarında ayrılmıştır. Bu oranlar veri miktarı olarak 3392, 424, 424 olarak hesaplanmıştır. Eğitimde hatanın minimizasyonu için kullanılan kayıp fonksiyonu "categorical crossentropy" seçilmiştir. Optmizasyon algoritması olarakta "RMSProp" kullanılmıştır. Şekil 17'de önerilen model üzerinde 3392 adet eğitim verisi ve 424 adet doğrulama verisi ile gerçekleştirilen yaklaşık 200 çevrimlik eğitim sonunda eğitim verisi için %99.87, doğrulama verisi için %85.14 oranlarında doğruluklar ölçülmüştür. Eğitim ve doğrulama verisi eğitim grafiği Şekil 18'de verilmiştir. Eğitim sürecince hiç kullanılmayan 424 adet test verisinin eğitilen model üzerinde test edilmesiyle %88.92'lik bir test doğruluğu elde edilmiştir. Her duyguya ait kesinlik, duyarlılık, f1 puan ve doğruluk gibi ölçütlerin sonuçları Tablo 3'te verilmiştir.

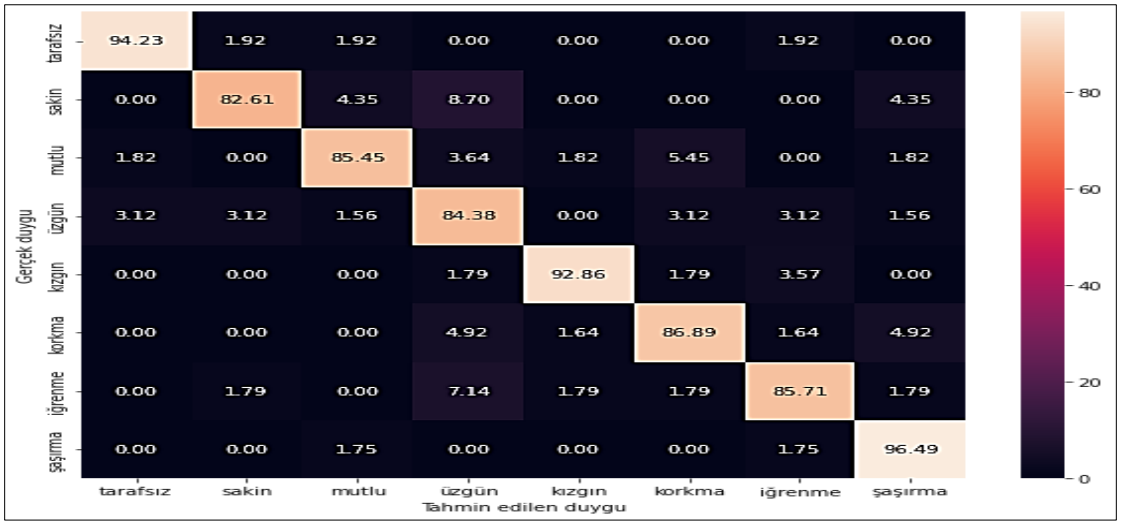


Şekil 18. Model doğruluğu

Tablo 3. Her duyguya ait doğruluk ölçüt değerleri

Duygu	Test Veri miktarı	Kesinlik(%)	Duyarlılık(%)	F1 puan(%)	Doğruluk(%)
Tarafsız	52	94	94	94	**~ 94**
Sakin	23	83	83	83	**~ 83**
Mutlu	55	92	85	89	**~ 85**
Üzgün	64	82	84	83	**~ 84**
Kızgın	56	95	93	94	**~ 93**
Korkma	61	88	87	88	**~ 87**
İğrenme	56	87	86	86	**~ 86**
Şaşırma	57	89	96	92	**~ 96**

Tablo 3 incelendiğinde test veri seti "sakin" duygusu hariç 8 duygu sınıfına dengeli bir şekilde dağılmıştır. "sakin" duygusu diğer duygulara göre en düşük doğrulukla sonuçlanmıştır. Bu durum "sakin" duygusunun veri miktarının azlığından ve "üzgün" duygu ifadesi ile benzer ses özelliklerine sahip olduklarından kaynaklanabilir. En yüksek doğruluk "şaşırma", "tarafsız" ve "kızgın" duygularında ölçülmüştür. Bu durumun sebebi ise bu duyguların belirleyici özelliklerinin ses sinyallerinde daha iyi temsil edilmesi olarak düşünülebilir. Sistemin performansını daha iyi anlamak için Şekil 19'da test verilerine ait karışıklık matrisi verilmiştir.



Şekil 19. Karışıklık matrisi

Önerilen sistem RAVDESS, TESS veri setlerinin birleştirildiği diğer çalışmalarla karşılaştırılmıştır. Tablo 4 incelendiğinde yaklaşık %89 doğruluğa sahip (Asiya ve Kiran, 2021) referanslı çalışmada 5 adet farklı özellik kullanılmıştır. Oysa önerilen sistemde özellik çeşidi sayısı 3'tür. Bu nedenle daha az özellik çeşidi ile aynı doğruluk elde edilmiştir. Karşılaştırma sonuçları Tablo 4'te verilmiştir.

Tablo 4. Önerilen sistemin diğer çalışmalarla performans karşılaştırması

Veri seti	Ref.	Özellik	Method	Doğruluk
RAVDESS + TESS	(Nasim ve diğerleri, 2021)	MFKK Kroma Mel spektrogram	Gradyan artış (Gradient Boost)	**~ %85**
	(Öztürk ve Pashaei, 2021)	MFKK	ESA-UKSB (Co-LSTM)	**~ %87**
	(Asiya ve Kiran, 2021)	Sıfır geçiş oranı Mel-spektrogram Kroma Mel Frekans Kepstrum Katsayısı Kök Ortalama Kare	1-D ESA (1-D CNN)	**~ %89**
	Önerilen Sistem	Kök Ortalama Kare enerji Sıfır Geçiş Oranı Mel Frekans Kepstrum Katsayı	UKSB (LSTM)+YSA (ANN)	**~ %89**

6. Sonuç

Bu makalede, konuşma duygu tanıma için UKSB tabanlı bir sistem önerilmiştir. Önerilen sistemde literatürdeki diğer çalışmaların aksine ESA kullanılmamış bunun yerine UKSB birimden sonra bir YSA kullanılmıştır. Sistemin sonuçları ve özellik çeşidi sayısı da dikkate alındığında literatürdeki diğer çalışmalardan daha etkili olduğu görülmektedir. Önerilen sistemin test verisi üzerindeki doğruluğu %88.92 olarak ölçülmüştür.

Çalışmada önerilen konuşma duygu tanıma sistemi RAVDESS ve TESS veri setleri birleşiminin uygulandığı literatürdeki diğer çalışmalardan genel olarak daha iyi sonuçlar elde etmiştir. Sonraki çalışmalarda daha iyi bir konuşma duygu tanıma performansı için Çy-UKSB (Çift Yönlü LSTM), 2D ESA-UKSB (CNN-LSTM), 1D ESA-UKSB, 2D ESA mimari kombinasyonları kullanılabilir. Ayrıca verisetine farklı veri arttırım yöntemleri uygulanarak performans arttırılabilir. Eğitilen ağın genelleme yeteneğinin arttırılması için veri seti SAVEE, CREMA-D, IEMOCAP vd. veri setleriyle genişletilebilir.

Kaynaklar

Cai L, Dong J & Wei M. (2020) Multi-Modal Emotion Recognition from Speech and Facial Expression Based on Deep Learning. Proceedings - 2020 Chinese Automation Congress, CAC 2020, pp. 5726–5729.

Issa D, Fatih Demirci M, Yazici A (2020) Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control* 59:101894.

Atila O, Şengür A (2021) Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition. *Applied Acoustics* 182:108260.

Mujaddidurrahman A, Ernawan F, Wibowo A, Sarwoko E. A, Sugiharto A, Wahyudi M. D. R. (2021) Speech Emotion Recognition Using 2D-CNN with Data Augmentation. 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), pp. 685–689.

Padi S, Manocha D, Sriram R. D (2020) Multi-Window Data Augmentation Approach for Speech Emotion Recognition. http://arxiv.org/abs/2010.09895

Nasim A. S, Chowdory R. H, Dey A, Das A. (2021) Recognizing Speech Emotion Based on Acoustic Features Using Machine Learning. 2021 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2021. https://doi.org/10.1109/ICACSIS53237.2021.9631319

Asiya U. A, Kiran V. K. (2021) Speech Emotion Recognition-A Deep Learning Approach. Proceedings of the 5th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2021, pp. 867–871.

Öztürk Ö. F, Pashaei E (2021) Konuşmalardaki duygunun evrişimsel LSTM modeli ile tespiti. Convolutional LSTM model for speech emotion recognition. *DUJE (Dicle University Journal of Engineering)* 12:581–589.

Hochreiter S, Schmidhuber J. (1997) Long Short-Term Memory. *Neural Computation* 9(8):1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

Livingstone S. R, Russo F. A (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE* 13(5):e0196391. https://doi.org/10.1371/JOURNAL.PONE.0196391

Zenodo (2022) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) | Zenodo. https://zenodo.org/record/1188976#.YiypnHpBy71. Accessed 12 March 2022.

University of Toronto Dataverse (2022) Toronto emotional speech set (TESS). https://dataverse.scholarsportal.info/dataset.xhtml?persistentId=doi:10.5683/SP2/E8H2MF. Accessed 6 May 2022.

Davis S. B, Mermelstein P (1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28(4):357–366.

Chen Q, Huang G (2021) A novel dual attention-based BLSTM with hybrid features in speech emotion recognition. *Engineering Applications of Artificial Intelligence* 102:104277.

Ancilin J, Milton A (2021) Improved speech emotion recognition with Mel frequency magnitude coefficient. *Applied Acoustics 179*:108046.

Sun J (2019) Research on vocal sounding based on spectrum image analysis. *Eurasip Journal on Image and Video Processing* 2019(1). https://doi.org/10.1186/S13640-018-0397-0

Stevens S. S, Volkmann J, Newman E. B (1937) A Scale for the Measurement of the Psychological Magnitude Pitch. *Journal of the Acoustical Society of America*, 8(3):185–190.

O'Shaughnessy D. (1987) Speech communication : human and machine. In *Wikipedia*. Addison-Wesley.

Wikipedia (2022) Discrete Cosine Transform. https://en.wikipedia.org/wiki/Discrete_cosine_transform. Accessed 10 March 2022.

Ahmed N, Natarajan T, Rao K. R (1974) Discrete Cosine Transform. *IEEE Transactions on Computers* C–23(1):90–93. https://doi.org/10.1109/T-C.1974.223784

Silva A. C. M. da, Coelho M. A. N, Neto R. F (2020) A Music Classification model based on metric learning applied to MP3 audio files. *Expert Systems with Applications*, 144:113071.

Giannakopoulos T, Pikrakis A. (2014) Introduction to Audio Analysis: A MATLAB Approach, pp. 1–266.

Wikipedia (2022) Zero-crossing rate. https://en.wikipedia.org/wiki/Zero-crossing_rate. Accessed 26 April 2022.

Alías F, Socoró J. C, Sevillano X (2016) A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds. *Applied Sciences* 6(5):143.

Librosa (2022) librosa 0.9.1 documentation. https://librosa.org/doc/latest/index.html. Accessed 16 April 2022.