GU J Sci, Part C, 10(4): 765-791 (2022)

Gazi University

Journal of Science

PART C: DESIGN AND TECHNOLOGY

http://dergipark.gov.tr/gujsc

A Literature Review on Speech Emotion Recognition Using Deep Learning Techniques

Emrah DİKBIYIK[1,*] Önder DEMİR[2] Buket DOĞAN[2]

[1]*İstanbul University-Cerrahpaşa, Vocational School of Technical Sciences, Department of Computer Technologies, 34500, Büyükçekmece/İSTANBUL*

[2]*Marmara University, Faculty of Technology, Department of Computer Engineering, 34840, Maltepe/İSTANBUL*

Article Info:

Review article
Received: 1.05.2022
Revision: 2.06.2022
Accepted: 12.06.2022

Highlights

- *Current deep learning techniques in Speech Emotion Recognition (SER) applications are reviewed.*
- *The review includes emotional speech datasets, features and classifiers.*
- *The review covers emotional speech datasets in Turkish language and SER applications developed with these datasets.*

Keywords

Speech Emotion Recognition
Deep Learning
Emotional Speech Datasets
Turkish Emotional Speech

Graphical/Tabular Abstract

This study includes the results of a literature research prepared by considering speech emotion recognition (SER) applications (we reviewed studies published between 2019 and 2021) in which deep learning methods are used. With the inclusion of deep learning applications in many research areas, the popularity it has gained has also been reflected in SER systems and has been the source of motivation for this literature research study.

Purpose: Within the scope of this study, current studies on SER systems in the literature were examined in detail, and the methods, datasets and requirements used were investigated. The effects of the methods used on the performance of the systems created were examined. Thus, it is aimed to assist researchers who will work to develop new models in this field.

Theory and Methods: SER is actually a pattern recognition application and the steps in a pattern recognition system can also be applied in emotion recognition systems over speech. A traditional SER system includes five main modules: the input speech signal, feature extraction, feature selection, classification, and emotion detection [21]. As can be seen in Figure A, deep learning can automatically learn the functions it needs to fulfill from the signs of picture, video and audio data, unlike classical machine learning methods. Thanks to its flexible structure, estimation accuracy increases depending on the data size.

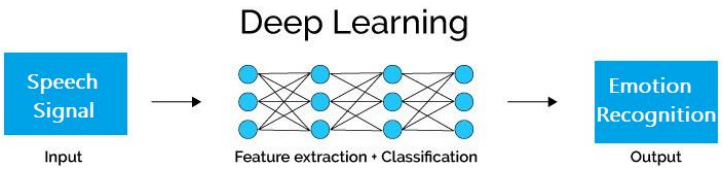


Figure A. Emotion recognition with deep learning

Results: In studies where deep learning models are applied, different input types such as raw speech waveform, spectrogram or log-mel spectrograms are used. It has been observed that especially Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) architectures are frequently applied methods and when the developed models are tested on different datasets, different accuracy in the classification of emotion emerges.

Conclusion: SER is a field of study related to speech processing, and besides speech processing, it has additional challenges that arise due to the extraction of emotion from sound. Because there are cultural and linguistic effects in the expression of emotions. Another challenging aspect of SER applications is adapting them to real-life applications. The increase in the number of datasets obtained from natural and noisy environments and consisting of records of different dialects of languages will contribute to the process of SER systems taking place in our lives in real time. On the other hand, although studies using Turkish datasets have increased recently, it can be said that it is still not at a sufficient level when compared with commonly used datasets. SER applications continue to be a research area with high potential that can take place in our lives in many fields such as education, security, health, marketing, IoT, virtual reality.

*Corresponding author, e-mail: emrahdikbiyik@iuc.edu.tr
DOI: 10.29109/gujsc.1111884

GU J Sci, Part C, 10(4): 765-791 (2022)

Gazi Üniversitesi

Fen Bilimleri Dergisi

PART C: TASARIM VE TEKNOLOJİ

http://dergipark.gov.tr/gujsc

A Literature Review on Speech Emotion Recognition Using Deep Learning Techniques

Emrah DİKBIYIK[1,*] Önder DEMİR[2] Buket DOĞAN[2]

[1]*İstanbul University-Cerrahpaşa, Vocational School of Technical Sciences, Department of Computer Technologies, 34500, Büyükçekmece/İSTANBUL*

[2]*Marmara University, Faculty of Technology, Department of Computer Engineering, 34840, Maltepe/İSTANBUL*

Makale Bilgisi

Derleme makalesi
Başvuru: 1.05.2022
Düzeltme: 2.06.2022
Kabul: 12.06.2022

Keywords

Speech Emotion Recognition
Deep Learning
Emotional Speech Datasets
Turkish Emotional Speech

Anahtar Kelimeler

Konuşmadan Duygu Tanıma
Derin Öğrenme
Duygusal Konuşma Veri Setleri
Türkçe Duygusal Konuşma

Abstract

People's speech varies according to the emotions they are in and contains information about these emotions. Carrying out studies on speech emotion recognition (SER) systems to discover this information has been a remarkable research area. Different datasets were created with the studies, many features of speech were discussed, different classification algorithms were applied for emotion recognition. This study includes the results of a literature research prepared by considering SER applications (we reviewed studies published between 2019 and 2021) in which deep learning methods are used. In addition, the emotional datasets used in these applications were examined and the features used in emotion recognition were included. Unlike other studies, emotional datasets created in Turkish and studies on these data sets are also discussed as a separate section.

Derin Öğrenme Yöntemleri İle Konuşmadan Duygu Tanıma Üzerine Bir Literatür Araştırması

Öz

İnsanların konuşmaları, içinde bulundukları duygulara göre değişiklik gösterir ve bu duygularla ilgili bilgiler içerir. Bu bilgileri keşfetmek için konuşmadan duygu tanıma sistemleri üzerine çalışmalar gerçekleştirmek dikkat çeken bir araştırma alanı olmuştur. Yapılan çalışmalarla farklı veri setleri ortaya çıkmış, konuşmaya ait birçok özellik dikkate alınmış ve duygu tanıma için farklı sınıflandırma algoritmaları uygulanmıştır. Bu çalışma, derin öğrenme yöntemlerinin kullanıldığı konuşmadan duygu tanıma uygulamaları (2019-2021 yılları arasında yapılan çalışmalar) dikkate alınarak hazırlanmış bir literatür araştırmasının sonuçlarını içerir. Bununla birlikte bu uygulamalarda kullanılan duygusal veri setleri incelenmiş, duygu tanımada kullanılan özelliklere yer verilmiştir. Diğer çalışmalardan farklı olarak Türkçe dilinde hazırlanmış duygusal veri setleri ve bu veri setleri üzerinde yapılan çalışmalar da ayrı bir bölüm olarak ele alınmıştır.

1. GİRİŞ (INTRODUCTION)

Duygu kavramı “belirli nesne, olay veya bireylerin insanın iç dünyasında uyandırdığı izlenim” olarak tanımlanmaktadır [1]. Her duygu farklı bir durumun işaretçisi olarak kendine has bilgiler içerir. İç dünyamızda şekillendirdiğimiz duygular, içinde bulunduğumuz durumla ilgili herhangi bir mesajı iletmek için çeşitli şiddet düzeylerinde ortaya çıkmaktadırlar [2]. İnsanlarda oluşan bu izlenimi bilgisayar ile keşfetmek, ölçmek ve sınıflandırmak günümüzde yaygın araştırma konularından biridir [3]. İnsanların sesinden, yazdıkları metinlerden veya video görüntülerinden elde edilen veriler bu araştırmalar için önemli bir kaynak oluşturmaktadır. İnsan duygusunun nötr, kızgın, mutlu, üzgün, şaşkınlık veya korku gibi kategorilerden hangisine ait olduğunun tespit edilmesini mümkün kılan bu çalışmaların pazarlama [4], sağlık [5] , güvenlik [6] ,eğitim [7] , müzik [8] gibi birçok uygulama alanı bulunmaktadır. İnsanlar genellikle içinde bulundukları duygu durumları konuşurken ses özelliklerine, bir metin yazarken veya yine konuşurken kullandıkları kelimelere , kurdukları cümlelere, bir resim veya videoda ise mimiklerine, vücut

hareketlerine yansımaktadır. Duygu analizi üzerine yapılan çalışmalar 1990'ların ortalarından bu yana ilgi çeken bir araştırma konusu olmuştur. Literatürde Duygu analizi ve Fikir/Düşünce madenciliği kavramlarının birbirinin yerine kullanıldığı da görülmektedir. Fikir madenciliği bir öğenin veya varlığın farklı yönleriyle ilgili, öznel kararları çıkarmak ve analiz etmek için otomatik bir yöntem önerir . Duygu analizi ise temel olarak görüşlerin kutuplaşmasına, yani olumlu veya olumsuz otomatik olarak tanınmasına odaklanır [9].

Duygu analizi için hesaplamalı yaklaşımlar dört grupta incelenebilir [9].

1. Metin üzerinde gerçekleştirilen duygu analizi
2. Konuşma üzerinde gerçekleştirilen duygu analizi
3. Görüntü tabanlı duygu analizi
4. Çok modlu duygu analizi

Metin üzerinde duygu analizi gerçekleştirmeye yönelik çalışmalar 2000'lerin başından itibaren ilgi görmeye başlamış özellikle sosyal medya uygulamaları üzerinde yapılan çalışmalarla birlikte sayılarında bir artış yaşanmıştır [10]. Metin üzerinde duygu analizi; dijital metinlere uygulanan, doğal dilin otomatik olarak işlenmesine ait dilsel işlemler kümesidir. Amacı bir metinde ifade edilen duyguları tanımlamak ve verilen bir konuya karşı polariteyi (olumlu ya da olumsuz) tahmin etmektir. Metin üzerinde duygu analizi kelime düzeyinde, cümle düzeyinde ve doküman düzeyinde olmak üzere farklı seviyelerde uygulanabilir ve özellikle sosyal ağ platformlarında insanların görüşlerini kolayca ifade edebilmeleri ile oluşan çok sayıda görüşün otomatik olarak işlenmesi sürecinde oldukça kullanışlıdır. Makine öğrenmesi ve derin öğrenme modelleri, sosyal ağlardan elde edilen veri setlerinin çoğalması ve fikirlerin tanımlanmasını kolaylaştıran kelimelerin sayısının azalması sayesinde kısa metinlere uygulandığında yüksek performanslara ulaşmaktadır. Bununla birlikte doküman düzeyinde duygu analizi, dokümanların daha çok kelime içermesi ve cümleler arasındaki anlamsal bağlantılar sebebiyle daha karmaşıktır [11].

Konuşma ve duygular üzerine yapılan çalışmalar, geçmişi 1930'lara kadar uzanan önemli bir literatüre sahiptir [12]. Bu uzun geçmişe dayanarak gelişen hesaplamalı yaklaşımlar aracılığıyla konuşma üzerinde duygu analizi gerçekleştirmek ses işleme ile ilgili bir çalışma alanı olmuştur. Bu amaçla geliştirilen sistemlerin giriş verisi bir konuşma içeren ses sinyalidir. İnsanların sahip olduğu duyguyu bu ses sinyali üzerinden tespit etmek için sesin prozodik özellikler (prosodic features) [13] spektral özellikler (spectral features) [14] gibi farklı özellikleri kullanılabilir. Bu özellikler, konuşmalara ait duyguların sınıflandırılmasında önemli bir rol oynar.

Görüntü temelli duygu analizi**,** bilgisayarlı görme ile ilgili yeni ve popülerliği artan bir araştırma alanıdır. Görüntü tabanlı duygu analizindeki başlıca araştırma görevleri, yüz veya bedensel hareketler [15,16] ile ilişkili olarak duygunun tespit edilmesi etrafında dönmektedir. Amaç, bir bireyin sergilediği hissi görsel olarak gözlemlenebilir ifadelerinden yola çıkarak modellemek ve tespit etmektir [9].

Çok modlu duygu analizi yaklaşımları metin, ses/konuşma ve görüntü üzerinden yapılan analizlerin duygu yoğunluğunu daha doğru tespit edebilmek için bir arada kullanıldığı yöntemlerdir [17]. Çok modlu duygu analizi ile ilgili literatür daha çok video görüntülerinin de dahil edildiği uygulamaları içermektedir. Zadeh ve ark. 2017 yılında yayınladıkları bir çalışmada çok modlu duygu analizi için tek modlu (unimodal), iki modlu (bimodal) ve üç modlu (trimodal) olmak üzere gruplandırmalar gerçekleştirmiştir. Tek modlu analiz sadece metni, sesi veya görüntüyü kullanmaktadır. Metin, ses ve videonun ikişerli kullanıldığı analizler iki modlu, üçünün bir arada kullanıldığı analizler ise üç modlu olarak tanımlanmıştır [18].

Bu çalışma kapsamında literatürde konuşmadan duygu tanımaya yönelik güncel çalışmalar detaylı olarak incelenmiş, kullanılan yöntemler ve gereklilikler araştırılmıştır. Kullanılan yöntemlerin oluşturulan sistemlerde başarımları üzerindeki etkisi incelemiştir. Böylece bu alanda yeni modeller geliştirmek üzere çalışacak araştırmacılara yardımcı olunması amaçlanmıştır. Makalenin geri kalan bölümü şu şekilde düzenlenmiştir. İkinci bölümde konuşmadan duygu tanıma kavramı, bu amaçla geliştirilen sistemlerde kullanılan konuşmaya ait özellikler ve kullanılan duygusal veri setlerine yer verilmiştir. Üçüncü bölümde derin öğrenme yöntemleri kullanılarak geliştirilen güncel konuşmadan duygu tanıma çalışmaları incelenmiştir. Dördüncü bölümde ise Türkçe dilinde oluşturulmuş duygusal konuşma veri setleri ve bu veri setleri üzerinde yapılan çalışmalara (derin öğrenme yöntemlerinden bağımsız olarak) yer verilmiştir.

Bu literatür araştırması çalışmasında kullanılan terminoloji listesi ise Tablo 1'de açıklanmıştır.

Tablo 1. *Bu çalışmada kullanılan terminoloji listesi (Türkçe kısaltmalara göre sıralama yapılmıştır.)*

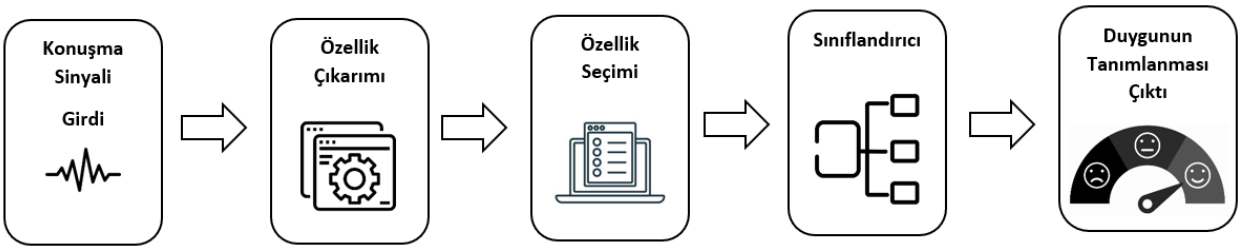
İngilizce Kısaltma	**Açılımı**	**Türkçe Karşılığı**	**Türkçe Kısaltması**
1D CNN	1 Dimensional Convolutional Neural Networks	1 Boyutlu Evrişimsel Sinir Ağı	1B ESA
2D CNN	2 Dimensional Convolutional Neural Networks	2 Boyutlu Evrişimsel Sinir Ağı	2B ESA
WAvgRc	Weighted Average Recall	Ağırlıklı Ortalama Duyarlılık	AOD
UAR	Unweighted Average Recall	Ağırlıksız Ortalama Duyarlılık	AszOD
DBM	Deep Boltzmann Machine	Derin Boltzmann Makinesi	DBM
DBN	Deep Belief Network	Derin İnanç Ağı	DİA
DTL	Deep Transfer Learning	Derin Öğrenme Aktarımı	DÖA
LPCC	Linear Predictive Cepstral Coefficient	Doğrusal Öngörü Kepstral Katsayıları	DÖKK
DNN	Deep Neural Networks	Derin Sinir Ağları	DSA
LPC	Linear Predictor Coefficient	Doğrusal Tahmin Katsayısı	DTK
SVM	Support Vector Machine	Destek Vektör Makinesi	DVM
DTW	Dynamic Time Warping	Dinamik Zaman Bükme	DZM
CNN	Convolutional Neural Networks	Evrişimsel Sinir Ağları	ESA
ConvLSTM	Convolutional LSTM	Evrişimsel Uzun Kısa Süreli Bellek	EUKSB
DCNN	Dilated Convolutional Neural Networks	Genişletilmiş Evrişimsel Sinir Ağır	GESA
GMM	Gaussian Mixture Model	Gaussian Karışım Modeli	GKM
GFLB	Global Feature Learning Block	Global Özellikleri Öğrenme Bloğu	GÖÖB
HNR	Harmonic to Noise Ratio	Harmonik Gürültü Oranı	HGO
BiLSTM	Bi-directional Long-Short Term Memory	İki Yönlü Uzun Kısa Süreli Bellek	İY-UKSB
SER	Speech Emotion Recognition	Konuşmadan Duygu Tanıma	KDT
K-NN	K-Nearest Neighbor	K-En yakın Komşuluk	K-EK
GRU	Gated Recurrent Unit	Kapı Yinelemeli Birimler	KYB
STFT	Short Time Fourier Transform	Kısa Zamanlı Fourier Dönüşümü	KZFD
LFPC	Log Frequency Power Coefficient	Log Frekans Güç Katsayısı	LFGK
MFCC	Mel Frequency Cepstral Coefficient	Mel-Frekans Kepstral Katsayısı	MFKK
AE	Auto Encoder	Otomatik Kodlayıcı	OK
RvNN	Recursive Neural Network	Özyinelemeli Sinir Ağı	ÖSA
RBF	Radial Based Function	Radyal Tabanlı Fonksiyon	RTF

ZCR	Zero Crossing Rate	Sıfır Geçiş Oranı	SGO
HMM	Hidden Markov Model	Saklı Markov Modeli	SMM
SMO	Sequential minimal optimization	Sıralı Minimal Optimizasyon	SMO
FCN	Fully Connected Network	Tam Bağlantılı Ağ	TBA
TEO	Teager Energy Operator	Teager Enerji Operatörü	TEO
RNN	Recurrent Neural Network	Tekrarlayan Sinir Ağı	TSA
LSTM	Long Short-Term Memory	Uzun Kısa Süreli Bellek	UKSB
DiSA	Directional Self-Attention	Yönlü Öz Dikkat	YÖD
LFLB	Local Feature Learning Block	Yerel Özellikleri Öğrenme Bloğu	YÖÖB
ANN	Artificial Neural Networks	Yapay Sinir Ağları	YSA
Veri seti İsimleri			
AFEW5	Acted Facial Expression in the Wild		
BAUM	BAhçeşehir University Multimodal Database		
CASIA	Chinese Emotional Speech Corpus		
CREMA-D	Crowd-sourced Emotional Multimodal Actors Dataset		
Berlin Emo-DB	Berlin Database of Emotional Speech		
eNTERFACE	The eNTERFACE’05 Audio-Visual Emotion Database		
GEMEP	Geneva Multimodal Emotion Portrayal		
IEMOCAP	Interactive Emotional Dyadic Motion Capture		
RAVDESS	The Ryerson Audio-Visual Database of Emotional Speech and Song		
RML	Ryerson Multimedia Lab		
SAVEE	Surrey Audio-Visual Expressed Emotion		
TurES	Turkish Emotional Speech		
TurEV-DB	Turkish Emotion Voice Database		

2. KONUŞMADAN DUYGU TANIMA-KDT (SPEECH EMOTION RECOGNITION-SER)

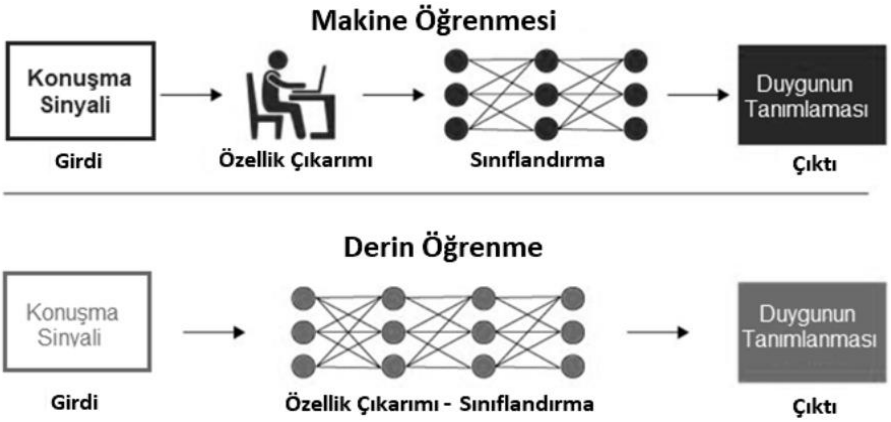
Duygu tespiti basit bir sınıflandırma sorunu değildir. Duyguların ifadesi kültürden kültüre değişiklik gösterebilir, hatta aynı kültür içinde kişiden kişiye göre de değişebilir. Bununla birlikte aynı konuşma, konuşmanın bağlamına göre farklı duyguları içinde barındırabilir. Konuşma üzerinden duygu analizi temelde insanın duygusal durumunu kendi sesinden tanımlayan bir sistemdir. Duygu sınıflandırılmasının etkin bir şekilde yapılabilmesi için birçok akustik özellik (vurgu, perde, tonlama, duraklama, vb. gibi özellikler) araştırılmış ve kullanılmıştır [19]. Konuşmadan duygu tanıma *(KDT)* işlemleri bu akustik özelliklerin değişiklik göstermesi, konuşmaya eklenen şive-ağız ve çevresel faktörler gibi etkenlerden dolayı birtakım zorluklara sahiptir [20].

KDT aslında bir örüntü tanıma uygulamasıdır ve bir örüntü tanıma sisteminde bulunan aşamalar konuşma üzerinden duygu analizi sistemlerinde de uygulanabilir. Geleneksel bir KDT sistemi; Şekil 1'de görüldüğü üzere girdi olan konuşma sinyali, özellik çıkarma, özellik seçimi, sınıflandırma ve duygunun belirlenmesi olmak üzere beş ana modül içerir [21].



Şekil 1. *Geleneksel konuşmadan duygu tanımlama sistemi [21]*

Bu geleneksel yaklaşımlardan sonra KDT sistemlerinde derin öğrenme mimarileri kullanılarak yapılan çalışmalar son yıllarda artış göstermiştir. Derin öğrenme, uzun yıllardır yapay zekâ çalışmalarında ortaya çıkan sorunları çözmede büyük ilerlemeler kaydetmiş ve yüksek boyutlu verilerdeki karmaşık yapıları keşfetmede çok iyi olduğu ortaya çıkmıştır [22]. Derin öğrenme, denetimli ya da denetimsiz olarak nesne tanıma [23], doğal dil işleme [24], özellik çıkartma [25], konuşma tanıma [26] gibi alanlarda kullanılan çok sayıda doğrusal olmayan işlevleri yerine getiren çok katmanlı sinir ağlarından oluşan makine öğrenme çeşididir. Şekil 2' de görüldüğü üzere derin öğrenme klasik makine öğrenim metotlarının aksine resim, video, ses verilerine ait işaretlerden, yerine getirmesi gereken işlevleri otomatik olarak öğrenebilmekte olup, esnek yapısı sayesinde veri büyüklüğüne bağlı olarak tahmin doğruluğu artmaktadır [27].



Şekil 2. *Klasik makine öğrenme yaklaşımı ile derin öğrenme arasındaki fark [27]*

Derin Boltzmann Makinesi-DBM *(Deep Boltzmann Machine-DBM)*, Tekrarlayan Sinir Ağı-TSA *(Recurrent Neural Network- RNN)* çalışması, Özyinelemeli Sinir Ağı-ESA *(Recursive Neural Network-RvNN)*, Derin İnanç Ağı-DİA *(Deep Belief Network-DBN)*, Evrişimsel Sinir Ağları-*ESA (Convolutional Neural Networks-CNN)* ve Otomatik Kodlayıcı-OK *(Auto Encoder-AE)* gibi derin öğrenme teknikleri, konuşmadan duygu tanıma için kullanılan ve tasarlanan sistemin sınıflandırma performansını pozitif yönde etkileyen yöntemlerdendir [28].

2.1. Konuşmadan Duygu Tanımada Özellikler (Features in Speech Emotion Recognition)

Konuşma üzerinden duygu tanıma gerçekleştirecek sistemin tasarımında en önemli konulardan biri farklı duyguları karakterize eden uygun özelliklerin çıkarılmasıdır. Doğru özelliklerin seçimi sınıflandırma performansını etkiler ve duygu tanıma oranını arttıracaktır. Günümüze kadar geliştirilen sistemlerde farklı özellikler kullanılmıştır fakat ayırt edici ve kesin kabul görmüş bir dizi özellik yoktur [29]. Konuşma sinyalleri değişken uzunluklu sürekli sinyallerdir. Konuşma işlemede bu sinyali çerçeve adı verilen küçük parçalara bölmek yaygın bir yaklaşımdır. Her çerçeve içinde sinyalin neredeyse sabit olduğu kabul edilir. Her çerçeveden çıkarılan perde ve enerji gibi prozodik konuşma özellikleri yerel özellikler olarak

adlandırılır. Öte yandan global özellikler bir ifadeden çıkarılan tüm konuşma özelliklerinin ortalama, minimum, maksimum değer ve standart sapma gibi istatistikleri olarak hesaplanır. Bu nedenle gerekli yaklaşıma bağlı olarak global veya yerel özellikleri çıkarılabilir [30]. Araştırmacılar özellik çıkarma işlemi için çalışmalarında openSMILE [31], Praat [32] gibi yazılım araçlarından faydalanmışlardır [33,34].

Konuşmada yer alan global ve yerel özellikler; prozodik özellikler , niteliksel özellikler (qualitative feautres), spektral özellikler ve TEO *(Teager Enerji Operatörü)* olmak üzere dört kategoride gruplandırılabilir [30]. Tablo 2' de bu özelliklere yer verilmiştir.

Tablo 2. *Konuşmadan duygu tanımada özellikler*

Prozodik Özellikler	**Niteliksel Özellikler**	**Spektral Özellikler**	**Teager Enerji Operatörü**
-Temel Frekans [35] -Enerji özellikleri [19] -Perde özellikleri [36] -Süre özellikleri [37]	-Titreme-Jitter [38] -Parıltı-Shimmer [38] -Harmonik Gürültü Oranı (HGO) [39]	*Geleneksel lineer spektral özellikler şunları içerir [42]:* -Doğrusal Tahmin Katsayısı-DTK (Linear Predictor Coefficient -LPC) [40], -Log Frekans Güç Katsayısı-LFGK (Log Frequency Power Coefficient LFPC) [41], *Kepstral (Cepstrum) özellikleri şunları içerir [42] :* - Mel-Frekans Kepstral Katsayısı -MFKK (Mel Frequency Cepstral Coefficient-MFCC) [35] - Doğrusal Öngörü Kepstral Katsayıları-DÖKK (Linear Predictive Cepstral Coefficient- LPCC) [43]	-TEO-FM-Var, TEO-Auto-Env, TEO-CB-Auto-Env -TEO tabanlı özellikler konuşmadaki stresi, tespit etmek için kullanılabilir [44,45] -Konuşma üretiminin doğrusal olmayan hava akımı yapısı hakkında faydalı bilgiler çıkarır [14]

Daha ayırt edici duygu özelliklerini çıkarmak, araştırmacıların konuşma duygusunu tanımadaki ana görevlerinden biri olmuştur. Özellik çıkarma yöntemlerinin farklılığına göre konuşma öznitelikleri manuel (el ile çıkarılmış) özellikler ve öğrenilmiş özellikler olarak sınıflandırılabilir. El ile özelliklerin çıkarılması, tecrübeye dayalı stratejiler kullanılarak dikkatlice tasarlanır ve nasıl çalıştığı ve ne yaptığı daha ayrıntılı olarak açıklanabilir. Derin Sinir Ağları-DSA *(Deep Neural Networks-DNN)* [46,47], ESA [48] gibi farklı derin ağlar tarafından çıkarılan öğrenilmiş özellikler ise tahminde önemli derecede iyi performans gösterir. Bu nedenle, tahminler yapmak için derin özellikleri öğrenmek giderek daha popüler hale gelmiştir [49].

2.2. Veri Setleri (Datasets)

Duygunun otomatik olarak tanımlanması amacıyla önerilen modeller kadar bu modellerin test edildiği veri setleri de çok önemlidir. Çünkü bu veri setlerinin kalitesi duygunun tanınma başarısını etkiler. Veri setlerinde yer alan sinyalin kalitesi kadar çeşitliliği de önemlidir. Örneğin konuşma içeren veri setlerinde

farklı cinsiyetlerde, farklı yaşlarda, farklı dillerde ve farklı duygu gruplarında kayıtların bulunması uygulanacak modelin geçerliliğini güçlendirecektir.

Günümüze kadar görsel, konuşma ve metin içeren veri setleri üzerinde duygu tanımlama amaçlı çalışmalar gerçekleştirilmiştir. Örneğin Berlin Emo-DB *(Berlin Database of Emotional Speech)* [50] gibi bazı veri setleri sadece konuşma kayıtlarından oluşmaktayken, SAVEE *(Surrey Audio-Visual Expressed Emotion)* [51], RAVDESS *(The Ryerson Audio-Visual Database of Emotional Speech and Song)* [52] gibi bazı veri setleri hem görsel hem konuşma kayıtlardan oluşmaktadır. Geliştirilen veri setlerinden bazıları doğrudan erişime açık olarak sunulmuşken, bazı veri setleri araştırmacılar için bir kullanım sözleşmesi maili yoluyla ücretsiz indirilebilecek kısıtlı erişime sahiptir. Bunların dışında ticari amaçlı ücretli kullanıma sunulmuş veri setleri de mevcuttur. Duygusal veri setlerinin diğer bir niteliği de kayıtların oluşturulduğu ortamlardır. Çalışmalarda incelenen veri setlerinin birçoğu [50-52,56] belirli ifadelerin aktörler tarafından uygun ortamlarda kaydedilmesi ile oluşturulurken, AFEW 5.0 *(Acted Facial Expression in the Wild)* [53] ve BAUM-1 s *(BAhçeşehir University Multimodal Database)* [54] gibi veri setleri ise spontane(doğal) kayıtlardan oluşmaktadır.

Bu çalışma kapsamında gerçekleştirilen araştırmalarda ve incelenen yayınlarda derin öğrenme yöntemleri ile üzerinde çalışmalar yapılmış veri setleri Tablo 3'te nitelikleri ile detaylandırılmıştır.

Tablo 3. *Çalışmalarda kullanılan duygusal veri setleri ve barındırdığı duygular, kayıt içerikleri ve kayıt türleri.*

Veri seti	**Diller**	**Yıl**	**Erişim Türü**	**Duygular**	**İçerik-Kayıt Bilgileri**	**Tür**
Berlin Database of Emotional Speech (Berlin Emo-DB) [50]	Almanca	2005	Erişime Açık	7 duygu: öfke, nötr, korku, can sıkıntısı, mutluluk, üzüntü, iğrenme	10 aktöre (5 erkek + 5 kadın) ait 10 cümle ile oluşturulmuş 800 kayıt	Ses
Surrey Audio-Visual Expressed Emotion (SAVEE) [51]	İngilizce	2011	Kısıtlı Erişim	7 duygu: öfke, iğrenme, korku, mutluluk, üzüntü, şaşkınlık ve nötr.	4 erkek aktöre ait 120 şer cümle toplam 480 kayıt	Ses/ Görsel
The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [52]	İngilizce	2018	Erişime Açık	7 duygu: sakin, mutlu, üzgün, kızgın, korkmuş, şaşkınlık ve iğrenme	24 profesyonel aktöre (12 erkek, 12 kadın) ait 7356 kayıt (sadece ses + sadece video + ses ve video)	Ses/ Görsel
Acted Facial Expression in the Wild (AFEW5.0) [53]	İngilizce	2015	Kısıtlı Erişim	7 duygu: öfke, sevinç, üzüntü, iğrenme, şaşkınlık, korku ve nötr	3 yorumcu tarafından ifade edilmiş 1645 kayıt	Ses/ Görsel
BAUM-1 s [54]	Türkçe	2013	Kısıtlı Erişim	6 temel duygu: mutluluk, öfke, üzüntü, iğrenme, korku, şaşkınlık. Veri seti bu duygularla birlikte can sıkıntısı ve küçümsemeyi	Veriler 17'si kadın 31 kişiden toplanmıştır. Deneklerin anadili Türkçe ve yaş aralığı 19-65'tir. Ortalama 1,82 saniyelik 1184	Ses/ Görsel

				yansıtan kayıtları içerir. Ayrıca zihinsel birkaç durum da yer almaktadır: Emin olmama (kafa karışıklığı, kararsız), düşünme, konsantre olma ve rahatsız olma.	klipten oluşmaktadır.	
Interactive Emotional Dyadic Motion Capture (IEMOCAP) [55]	İngilizce	2008	Kısıtlı Erişim	5 duygu: mutluluk, öfke, üzüntü, hayal kırıklığı ve nötr.	10 aktöre (5 erkek + 5 kadın) ait 12 saatlik çok modlu veri kaydı	Ses/ Görsel
Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [56]	İngilizce	2014	Erişime Açık	6 duygu: kızgın, iğrenmiş, korkmuş, mutlu, tarafsız ve üzgün	91 aktöre (48 erkek + 43 kadın) ait toplam 7.442 kayıt	Ses/ Görsel
The eNTERFACE'05 Audio-Visual Emotion Database (eNTERFACE) [57]	İngilizce	2005	Erişime Açık	6 duygu: öfke, korku, şaşkınlık, mutluluk, üzüntü ve iğrenme.	42 denekten (%81'i erkek, %19'u kadın) toplam 1166 video kaydı (264 kadın + 902 erkek)	Ses/ Görsel
Chinese Emotional Speech Corpus (CASIA) [58]	Mandarin	2008	Ücretli erişim	5 duygu : öfke, mutluluk, sürpriz, nötr ,üzgün [43]	4 kişiden alınan kayıt ve kişi başına 1500 duygu ifadesi (her duygu için 300 ifade)	Ses
Geneva Multimodal Emotion Portrayal (GEMEP) [59]	Fransızca	2006	Kısıtlı Erişim	12 duygu: pozitif duygular: sevinç, eğlence, gurur, zevk, rahatlama, ilgi. Negatif duygular: öfke, panik, umutsuzluk, gerginlik (irritation), kaygı, üzüntü. 6 ek duygu: Pozitif: Hayranlık, Hassasiyet, Sürpriz Negatif: İğrenme, Aşağılama, Utanç	10 profesyonel aktörden (5 erkek + 5 kadın) alınan 1260 kayıt	Ses/ Görsel

Ryerson Multimedia Lab (RML) Emotion Database [60]	İngilizce, Çince (Mandarin), Urduca, Pencapça, Farsça, İtalyanca.	2005	-	6 duygu: mutluluk, üzüntü, öfke, korku, sürpriz ve iğrenme	6 farklı dil konuşan 8 denekten toplam 500 video kaydı	Ses/ Görsel
Urdu [61]	Urduca	2018	Erişime açık	4 duygu: kızgın, mutlu, üzgün ve nötr.	38 kişiden (27 erkek + 11 kadın) elde edilmiş (TV Şovları) toplam 400 ifade	Ses
EMOVO [62]	İtalyanca	2014	Erişime açık	7 duygu: iğrenme, korku, öfke, sevinç, şaşkınlık, üzüntü, ve nötr	6 aktörden (3 erkek + 3 kadın) elde edilen Her duygu için söylenmiş 14 cümle ve toplam 588 ses kaydına	Ses

3. KONUŞMADAN DUYGU TANIMADA DERİN ÖĞRENME UYGULAMALARI (DEEP LEARNING APPLICATIONS IN SPEECH EMOTION RECOGNITION)

KDT sistemlerinin önemli aşamalarından birisi de sınıflandırma işlemidir. Yapılan çalışmalarda geleneksel sınıflandırıcılar ve derin öğrenme mimarileri olmak üzere gruplandırılabilecek farklı sınıflandırma yöntemleri denenmiş ve uygulanmıştır. Ancak birçok karmaşık problemde olduğu gibi bu sistemlerde de genel kabul görmüş bir yöntem yoktur [29, 63]. Derin öğrenme makine öğreniminde her geçen gün gelişmekte olan bir araştırma alanı olmakla birlikte son yıllarda daha fazla ilgi görmektedir. Geleneksel KDT sistemlerinde Saklı Markov Modeli-SMM *(Hidden Markov Model HMM)*, Gaussian Karışım Modeli-GKM *(Gaussian Mixture Model GMM),* Destek Vektör Makineleri-DVM *(Support Vector Machine-SVM),* Yapay Sinir Ağları-YSA *(Artificial Neural Networks-ANNs)*, K En yakın Komşuluk K-EK *(k-Nearest Neighbor K-NN)* gibi çeşitli sınıflandırıcılar kullanılmıştır [64-66]. Birçok hassas hesaplama ve ön işleme gibi mühendislik özellikleri gerektiren bu yöntemlerde özelliklerdeki herhangi bir değişiklik, tekniğin genel mimarisinin yeniden modellenmesini gerektiriyordu [67].

KDT sistemlerinde derin öğrenme teknikleri, geleneksel yöntemlere göre; özellikleri ve karmaşık yapıları manuel olarak ayarlamaya gerek kalmadan tespit edebilmesi, düşük seviyeli özelliklerin çıkarılması, etiketlenmemiş verilerle başa çıkma yeteneği gibi çeşitli avantajlara sahiptir. Yapılan çalışmalarda DSA ve ESA gibi ileri beslemeli mimariler, görüntü ve video işleme için verimli sonuçlar sağladığı görülmüştür [28]. Öte yandan, ESA, TSA ve Uzun Kısa Süreli Bellek-UKSB *(Long Short-Term Memory-LSTM)* gibi tekrarlayan mimariler, konuşmadan duygu tanıma sistemlerinde yaygın olarak kullanılan derin öğrenme mimarileridir [68].

Bu bölümde KDT uygulamaları için yakın zamanda geliştirilen modelleri öğrenmek adına 2019-2021 yılları arasında gerçekleştirilmiş ve derin öğrenme tekniklerinin kullanıldığı KDT uygulamalarına ait bir literatür araştırması sunulmuştur. Sunulan çalışmalarda derin öğrenme mimarileri konuşma özelliklerinin otomatik olarak öğrenilmesi ve/veya son aşamada duygunun sınıflandırılması amacıyla kullanılmıştır.

Demir ve ark. derin öğrenme yöntemlerini kullanarak sesin duygulara göre ayırt edici özelliklerinin kullanan bir otomatik duygu tanıma sistemi önermişlerdir. Çalışmalarında RAVDESS, SAVEE ve RML

veri setlerini kullanmışlardır. Özellik çıkartma aşamasında çalışma kapsamında, MFKK, DTK, DÖKK, spektral alt-bant ağırlık merkezi, yansıma katsayıları, log-alan oranları katsayıları ve log-enerji özellikleri kullanılmıştır. Sınıflandırma işleminde ise UKSB kullanılmış ve RAVDESS veri setinde %68,8, SAVEE veri setinde %72,13, RML veri setinde %70,35 sınıflandırma doğruluğuna ulaşmışlardır [69].

Meng ve ark. 2019 yılında yayınlanan çalışmalarında konuşmadan duygu tanıma için yeni bir derin öğrenme modeli önermişlerdir. Bu model artık blok ile Genişletilmiş Evrişimsel Sinir Ağı-GESA *(Dilated Convolutional Neural Networks-DCNN with residual block)* ve dikkat mekanizmasına dayalı İki Yönlü Uzun Kısa Süreli Bellekten İY-UKSB *(Bi-directional Long-Short Term Memory-BiLSTM)* oluşan ADRNN çerçevesidir. Modelin avantajları olarak uzun süreli dizileri eğitmek için daha az belleğe ihtiyaç duyması, TSA ile karşılaştırıldığında evrişim işlemini daha kolay hale getirmesi ve hangi içeriğin hatırlanması gerektiği ile ilgili daha faydalı geçmiş bilgilerini kullanma gibi avantajlar belirtmişlerdir. Konuşma sinyalinden elde edilen 3B Log-Mel spektrum özniteliklerini ise sınıflandırmak için geliştirdikleri modelde kullanmışlar ve uygulamalarını Berlin Emo-DB ve IEMOCAP veri setlerinde test etmişlerdir. Sonuç olarak Berlin Emo-DB de konuşmacıya bağlı deneylerde %90,78, konuşmacıdan bağımsız deneylerde %85,39 tanıma doğruluğu, IEMOCAP'de ise konuşmacıya bağlı deneylerde %74,96 konuşmacıdan bağımsız deneylerde ise %69,32 tanıma doğruluğu elde etmişlerdir [70].

Zhao ve ark. tarafından 1 Boyutlu Evrişimsel Sinir Ağı-1B ESA *(1 Dimensional Convolutional Neural Networks-1D CNN)* ve 2 Boyutlu Evrişimsel Sinir Ağı-2B ESA *(2 Dimensional Convolutional Neural Networks-2D CNN)* UKSB ağları üzerine gerçekleştirilen çalışmada Berlin Emo-DB ve IEMOCAP veri setleri kullanılmıştır. Özellikle 2B ESA UKSB ağının, seçilen veri setlerinde DİA ve ESA yöntemlerinden daha iyi performans gösterdiği tespit edilmiştir. Konuşmacıya bağlı ve konuşmacıdan bağımsız deneylerde geleneksel yaklaşımlarla elde edilen %91,6 ve %92,9 doğrulukla karşılaştırıldığında, Berlin Emo-DB'sinde sırasıyla %95,33 ve %95,89 tanıma doğruluğuna ulaşmıştır. Ayrıca IEMOCAP veri setinde konuşmacıya bağlı ve konuşmacıdan bağımsız deneylerde DİA ve ESA tarafından elde edilen %73,78 ve %40,02 doğruluktan çok daha yüksek olan %89,16 ve %52,14 tanıma doğruluğuna ulaşılmıştır [49].

Xie ve ark. 2019 yılında yaptıkları çalışmada zaman çerçeveleri arasındaki duygusal doygunluk farkından tam olarak yararlanmak için, dikkat temelli UKSB tekrarlayan sinir ağları ile birleştirilmiş çerçeve düzeyinde konuşma özelliklerini kullanan yeni bir yöntem önermişlerdir. Model, çerçeve düzeyinde konuşma özelliğini girdi olarak alır ve 2 katmanlı UKSB aracılığıyla her çerçevenin zamanına karşılık gelen çıkışı elde eder. Model zaman ve özelliklerdeki duygu farkını ayırt etmek için, UKSB çıktısı üzerinde sırasıyla zaman boyutu ve özellik boyutu üzerinde ağırlıklandırma işlemi gerçekleştirir ve iki ağırlıklı sonucu tam bağlantı katmanlarının girdisi olarak oluşturur. Son olarak, softmax katmanının çıktısı ile duygu tanıma işleminin sonucu elde edilir. Geliştirilen model CASIA, eNTERFACE ve GEMEP duygusal veri setlerinde denenmiştir. Hem zaman hem özellik boyutlarında dikkat mekanizmasına dayalı modelin unutma kapısı varyasyonu *(LSTM-(time and feature dimensions)-TF-at)* ile denenen model CASIA veri setinde %92,8, eNTERFACE veri setinde %89,6, GEMEP veri setinde %57,0 Ağırlıksız Ortalama Duyarlılık-AszOD *(Unweighted Avarage Recall-UAR)* sonuçlarına ulaşmıştır [71].

Jalal ve ark. 2020 yılındaki çalışmalarında her ikisi de dikkat mekanizması temelli olan ESA ve İY-UKSB modelleri ile geliştirdikleri KDT uygulamasını sunmuşlardır. 23 boyutlu log-Mel filtre bankası özellikleri ile çalışmışlardır. Önerilen ağ modellerinde dikkat ağırlıklarının büyük ölçüde ünlü seslerine meyilli olduğu ve akustik bağlam ve prozodiden önce/sonradan söze önem verdiği gösterilmiştir. Daha önce varsayıldığı gibi, daha küçük akustik bağlamların duyguları taşımada hayati olduğu da gösterilmiştir. Modellerini IEMOCAP veri setinde test etmişlerdir. Saf akustik verilerde %80,1 ile kendilerinden önceki modellerden daha yüksek ağırlıksız doğruluk oranı elde etmişlerdir [72].

Aouani ve Ayed 2020 yılında yayınladıkları çalışmalarında özellik çıkarma ve sınıflandırma motoru olarak adlandırdıkları iki aşamalı bir konuşmadan duygu tanıma modeli önermişlerdir. Özellik çıkarımı aşamasında 39 katsayılı MFKK, Sıfır Geçiş Oranı-*SGO (Zero Crossing Rate-ZCR)*, HGO ve TEO içeren 42 boyutlu bir öznitelik vektörü çıkarılmıştır. Burada ilgili parametrelerin seçimi için OK yönteminin kullanılmasını önermişler ve sınıflandırma işlemini DVM ile gerçekleştirmişlerdir. Önerilen model RML veri setinde denenmiş ve %74,07 sınıflandırma doğruluğu elde edilmiştir [39].

Issa ve ark. yayınladıkları çalışmada RAVDESS, Berlin Emo-DB, IEMOCAP veri setlerinde uygulamalarını gerçekleştirdikleri yeni bir model sunmuşlardır. Konuşma kayıtlarından MFKK'lar, Mel ölçekli spektrogram, Kromagram, Spektral kontrast özelliği, Tonnetz temsili olmak üzere beş farklı özellik çıkarmışlardır. Bu özellikler kullanılarak 1B ESA ile sınıflandırma işlemi gerçekleştirmişlerdir. Önerdikleri model , 8 sınıflı RAVDESS için %71,61, 7 sınıfta 535 örnekle Berlin Emo-DB için %86,1 , 7 sınıfta 520 örnekle Berlin Emo-DB için %95,71 ve 4 sınıflı IEMOCAP için %64,3 sınıflandırma doğruluğu elde etmişlerdir [73].

Mustaqeem ve Kwon yayınlarında KDT sistemleri için Evrişimsel Uzun Kısa Süreli Bellek-EUKSB *(Convolutional LSTM-ConvLSTM)* modelini önermişlerdir. Önerilen model mimarisi Kapı Yinelemeli Birimler-KYB *(Gated Recurrent Unit-GRU)* leri kullanan Global Özellikleri Öğrenme Bloğu-GÖÖB (*Global Feature Learning Block GFLB),* EUKSB'yi kullanan Yerel Özellikleri Öğrenme Bloğu-YÖÖB *(Local Feature Learning Block-LFLB)* ve bir önişleme adımı ile merkez kayıplarını ve softmax kayıplarını kullanan çok-sınıflı sınıflandırıcı katmanı olmak üzere üç modülden oluşmaktadır. Önerilen sistemi IEMOCAP ve RAVDESS veritabanlarında test etmişler ve sırasıyla %75 ve %80 tanıma oranı elde etmişlerdir [74].

2020 yılında Mustaqeem ve ark. tarafından yapılan çalışmada tanıma doğruluğunu iyileştirmek, model maliyetini ve işlem süresini azaltmak için yeni bir KDT yaklaşımı sunulmuştur. Konuşmadan daha verimli bir dizi seçmek amacıyla Radyal Tabanlı Fonksiyon-RTF *(Radial Based Function-RBF)* tabanlı K-means kümeleme algoritmasını kullanılmış ve bu dizi Kısa Zamanlı Fourier Dönüşümü-KZFD *(Short Time Fourier Transform-STFT)* algoritması ile spektrogramlara dönüştürülmüştür. Daha sonra ResNet fc-1000 katmanlarını kullanarak konuşma sinyali spektrogramlarından özellik çıkarma işlemi gerçekleştirilmiş ve sınıflandırma İY-UKSB ile uygulanmıştır. Önerilen sistem, IEMOCAP, Berlin Emo-DB ve RAVDESS olmak üzere farklı standart veri setleri üzerinden değerlendirilmiştir. Duygu tanıma doğruluğu, IEMOCAP, Berlin Emo-DB ve RAVDESS veri setlerinde sırasıyla %72,25, %85,57 ve %77,02 olarak elde edilmiştir [75].

Anvarjon ve ark. 2020 yılında yayınlanan çalışmalarında düşük hesaplama karmaşıklığına ve yüksek tanıma doğruluğuna sahip yeni bir hafif etkili KDT modeli önermişlerdir. Bu model konuşmadan duygu tanıma için daha fazla ayırt edici güce sahip değiştirilmiş bir örnekleme (pooling) stratejisi ile düz bir dikdörtgen çekirdek (rectangular kernel) kullanarak derin frekans özelliklerini öğrenmek için ESA yaklaşımını kullanır. Önerilen ESA modeli derin frekans özelliklerini analiz ederek gizli duygusal özellikleri tanımak için ağırlıklı olarak konuşma spektrogramlarındaki frekans özelliklerine odaklanmaktadır. Maliyet karmaşıklığını azaltmak için daha az parametre kullanılmıştır. Sistem IEMOCAP ve Berlin Emo-DB veri setlerinde test edilmiştir ve sırasıyla %77,01 ve %92,02 tanıma sonucu edilmiştir [76].

2021 yılında yayınlanan çalışmalarında Li ve ark. KDT için Yönlü Öz Dikkat-YÖD *(Directional Self-Attention-DiSA)* ağı ile İY-UKSB modelini önermişlerdir. Model mimarisinde konuşma sinyali özellikleri sırasıyla ileri ve geri İY-UKSB tarafından çözülür ardından bunlar YÖD mekanizması ile kodlanır. Konuşma sinyaline ait openSmile tarafından çıkarılan düşük seviyeli tanımlayıcı özellikler arasında MFKK, spektral yuvarlanma noktası, spektral akı, spektral merkez, spektral entropi, spektral yayılma, SGO, temel frekans, enerji, enerji entropisi ve bunların birinci dereceden farkı bulunur. Modellerini IEMOCAP ve Berlin Emo-DB veritabanlarında test etmişler ve sırasıyla %61,20 %85,95 ağırlıklı doğruluk elde etmişlerdir [77].

Mustaqeem ve Kwon 2021 yılında yayınlanan çalışmalarında 1B ESA'na dayanan uçta uca gerçek zamanlı bir KDT sistemi önermişlerdir. Önerilen sistem, ham sesli konuşmadan yerel özellikleri öğrenmek için kullanılan GESA'dan, yerel öğrenilmiş özellikleri geliştirme bloğu (Upgrade Features Learning Block-UFLB) için atlama bağlantılı tek boyutlu artık bloklardan (Residual Blocks with a Skip Connection-RBSC) ve sıralama modülünden (Seq_L) oluşur. Sistem RBSC ve Seq_L blokları ile çoklu tip özellikleri öğrenir. Füzyon katmanı ise Tam Bağlantılı Ağ-TBA *(Fully Connected Network – FCN)'*dan geçirilen uzamsal ve zamansal bilgileri ve nihai karar için kullanılan softmax katmanını birleştirmek için kullanılır. Önerdikleri

modeli IEMOCAP ve Berlin Emo-DB veri kümelerini kullanarak değerlendirmişler ve sırasıyla %73 ve %90 olan tanıma doğruluğu elde etmişlerdir [78].

Yusuf ve ark. çalışmalarında KDT sistemlerinde veri azlığı ve çevresel gürültüyü iki önemli sorun olarak belirtmişlerdir. Bu sorunları çözmek üzere Derin Öğrenme Aktarımı-DÖA *(Deep Transfer Learning-DTL)* kullanarak konuşmadan duygu tanıma için yeni bir sağlam çok pencereli spektrogram artırım *(Robust Multi Window Spectrogram ugmentation - RMWSaug)* modelini önermişlerdir. Denemelerinde IEMOCAP veri setini ikiye bölerek kullanmışlardır. AHSN olarak isimlendirilen kısım öfkeli, mutlu, üzgün ve nötr duygusal konuşma örneklerinden oluşmakta, AESN olarak isimlendirilen kısım ise öfke, heyecan, üzgün ve nötr duygusal konuşma örneklerinden oluşmaktadır. Bu veri setlerinde temiz konuşma örnekleri, 15dB,20dB ve 40dB gürültülü konuşma örnekleri üzerinde deneyler gerçekleştirmişlerdir. AHSN veri setinde Ağırlıklı Ortalama Duyarlılık-AOD *(Weighted Avarage Recall-WAvgRc)* değeri temiz konuşma örneklerinde %65,61 , 40db gürültülü örneklerde 65,90 olarak bulunmuş, AESN veri setinde AOD değeri temiz konuşma örneklerinde %62,99, 40db gürültülü örneklerde %62,81 olarak bulunmuştur [79].

Zhang ve ark. 2021 yılında yayınlanan çalışmalarında doğal konuşmadan duygu tanıma için çoklu ESA'lar ile derin çok modlu özellikleri öğrenmek amaçlı bir mimari sunmuşlardır. Önerdikleri yöntem üç adımdan oluşmaktadır. Birinci adımda uygun çok modlu ses temsillerinin oluşturulması, ikinci adımda çoklu ESA'lar ile çok modlu ses özelliklerinin öğrenilmesi (1B, 2B ve 3B ESA'dan öğrenilen derin çok modlu özellikler), üçüncü adımda çok modlu sonuçları puan düzeyinde birleştirme işlemleri gerçekleşir. Geliştirdikleri modellerini AFEW 5.0 ve BAUM-1 s veri setlerinde test etmişlerdir. Sonuç olarak AFEW 5.0 veri seti için %35,77 doğruluk oranı, BAUM-1 s veri seti için ise %44,06 doğruluk oranı elde edilmiştir [80].

İncelenen çalışmaların birçoğunda girdi olarak konuşmaya ait elde edilen spektral özelliklerin kullanılmış ve sınıflandırmada aşamalarında ise ESA, TSA, UKSB gibi mimariler tercih edilmiştir. Yukarıda açıklanan çalışmalara ait kullanılan veri setleri, özellikler, sınıflandırıcı yöntemleri ve elde edilen sonuçların listesi Tablo 4' te özet halinde sunulmuştur.

Tablo 4. *Çalışmaların listesi*

Makale Başlığı	**Kullanılan Veri seti(leri)**	**Özellikler**	**Sınıflandırıcı**	**Sonuçlar**
Deep learning and audio based emotion recognition (2019) [69]	SAVEE, RAVDESS, RML	MFKK, DTK, DÖKK, spektral alt-bant ağırlık merkezi, yansıma katsayıları, log-alan oranları katsayıları ve log-enerji özellikleri	UKSB	RAVDESS veri seti için %68,8 SAVEE veri seti için %72,13 RML veri seti için %70,35 doğruluk oranları
Speech emotion recognition from 3D log-mel spectrograms with deep learning network (2019) [70]	Berlin Emo-DB, IEMOCAP	3-B Log-Mel spektrumları	ADRNN (dikkat mekanizmasına dayalı artık bloklu genişletilmiş ESA ve İY-UKSB)	IEMOCAP veri seti için konuşmacıya bağlı deneylerde %74,96, konuşmacıdan bağımsız deneylerde %69,32. Berlin Emo-DB veri seti için konuşmacıya bağlı deneylerde %90,78, konuşmacıdan bağımsız

				deneylerde %85,39 doğruluk oranları
Speech emotion recognition using deep 1D & 2D CNN LSTM networks (2019) [49]	Berlin Emo-DB, IEMOCAP	ESA ve UKSB ile öğrenilen yerel ve global özellikler	2B ESA-UKSB	Berlin Emo-DB veri seti için konuşmacıya bağlı deneylerde %95,33, konuşmacıdan bağımsız deneylerde %95,89 IEMOCAP veri seti için konuşmacıya bağlı deneylerde %89,16, konuşmacıdan bağımsız deneylerde %52,14 doğruluk oranları
Speech emotion classification using attention-based LSTM (2019) [71]	CASIA, eNTERFACE, GEMEP	openSMILE ComParE özelliklerine dayalı çerçeve düzeyinde özellikler. (bir kısmı: log harmonik gürültü oranı, perde frekansı, F0 zarfı, ses yüksekliği, MFKK)	UKSB	CASIA veri setinde %92,8, eNTERFACE veri setinde %89,6, GEMEP veri setinde %57,0 ağırlıksız ortalama recall (UAR) oranları
Empirical Interpretation of Speech Emotion Perception with Attention Based Model for Speech Emotion Recognition (2020) [72]	IEMOCAP	23 boyutlu log-Mel filtre bankası özellikleri	dikkat mekanizmasına dayalı ESA ve İY-UKSB	%80,1 ağırlıksız doğruluk oranı
Speech emotion recognition with deep learning. (2020) [39]	Ryerson Multimedia Laboratory (RML)	MFKK,SGO, HGO ve TEO	Özellik seçiminde OK, sınıflandırma için DVM	%74,07 doğruluk oranı
Speech emotion recognition with deep convolutional neural networks. (2020) [73]	RAVDESS, Berlin Emo-DB, IEMOCAP	MFKK, Mel ölçekli spektrogram, Kromagram, Spektral kontrast özelliği, Tonnetz sunumu	1B ESA	RAVDESS veri setinde 8 duygu sınıfı için %71,61, Berlin Emo-DB veri setinde 535 örnekli 7 duygu sınıfı için %86,1 520 örnekli uygulamada %95,71

				IEMOCAP veri setinde 4 duygu sınıfı için %64,3 doğruluk oranları
CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network (2020) [74]	IEMOCAP, RAVDESS	EUKSB ile öğrenilen yerel özellikler, KYB ile öğrenilen global özellikler	EUKSB	IEMOCAP veri setinde %75, RAVDESS veri setinde %80 doğruluk oranları
Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM (2020) [75]	IEMOCAP, Berlin Emo-DB, RAVDESS	ESA ile öğrenilen özellikler	İY-UKSB	IEMOCAP veri setinde %72,25, Berlin Emo-DB veri setinde %85,52, RAVDESS veri setinde %77,02 doğruluk oranı
Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features (2020) [76]	IEMOCAP, Berlin Emo-DB,	ESA ile öğrenilen yüksek seviye derin frekans özellikleri	ESA	IEMOCAP veri setinde %77,01, Berlin Emo-DB veri setinde %92,02 ağırlıksız doğruluk oranı
Speech emotion recognition using recurrent neural networks with directional self-attention (2021) [77]	IEMOCAP, Berlin Emo-DB	openSmile tarafından çıkarılan düşük seviyeli tanımlayıcı özellikler (Bir kısmı: MFKK, spektral yuvarlanma noktası, spektral akı, spektral merkez, spektral entropi, spektral yayılma, sıfır geçiş oranı)	YÖD ağı ile İY-UKSB	IEMOCAP veri setinde %61,20, Berlin Emo-DB veri setinde %85,95 ağırlıklı doğruluk oranı, IEMOCAP veri setinde %54,99 , Berlin Emo-DB veri setinde %82,6 ağırlıksız doğruluk oranı
MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach (2021) [78]	IEMOCAP, Berlin Emo-DB	GESA ile öğrenilen yerel özellikler, KYB ile öğrenilen global özellikler	1B GESA	IEMOCAP veri setinde %73, Berlin Emo-DB veri setinde %90 duyarlılık (recall)
RMWSaug: Robust Multi-window Spectrogram Augmentation Approach for Deep Learning	IEMOCAP veri setinin AHSN olarak isimlendirilen kısım öfkeli, mutlu, üzgün ve nötr duygusal	ESA ile öğrenilen özellikler	DÖA	AHSN olarak isimlendirilen kısımda %65,61 , 40db gürültülü örneklerde 65,90 ağırlıklı ortalama recall oranı,

based Speech Emotion Recognition (2021) [79]	konuşma örneklerinden oluşmakta, AESN olarak isimlendirilen kısım ise öfke, heyecan, üzgün ve nötr duygusal konuşma örneklerinden oluşmaktadır.			AESN olarak isimlendirilen kısımda temiz konuşma örneklerinde %62,99, 40db gürültülü örneklerde %62,81 ağırlıklı ortalama recall oranı
Learning deep multimodal affective features for spontaneous speech emotion recognition (2021) [80]	AFEW 5.0, BAUM-1 s	1B, 2B ve 3B ESA ile öğrenilen özelliklerin birleştirilmesi ile oluşturulan özellik vektörü	Duygu sınıflandırması için son aşamada DVM	AFEW 5.0 veri seti için %35,77 doğruluk oranı, BAUM-1 s veri seti için ise %44,06 doğruluk oranı

4. TÜRKÇE DUYGUSAL KONUŞMA VERİ SETLERİ VE DUYGU TANIMA UYGULAMALARI (TURKISH EMOTIONAL SPEECH DATASETS AND EMOTION RECOGNITION APPLICATIONS)

KDT sistemlerinde Türkçe veri setleri üzerinde çalışmalar yapmak da araştırmacılar için dikkat çeken alanlardan biri olmuştur. Bu bölümde Türkçe veri setleri üzerinde gerçekleştirilmiş çalışmalara daha geniş bir tarih aralığında yer verilmiştir. Bununla birlikte derin öğrenme yöntemleri dışında geleneksel yöntemler kullanılarak yapılan çalışmalara da yer verilmiştir.

Türkçe duygusal konuşma veri setleri üzerine yapılan ilk çalışmalardan birisi 2011 yılında Oflazoğlu ve Yıldırım'ın gerçekleştirdiği "Türkçe Duygusal Konuşma Veritabanı" çalışmasıdır. Bu veri setinde 55 adet Türkçe filmden çıkarılmış 5304 tane konuşma sinyali ve bunların metinsel içeriklerinden oluşmaktadır. Konuşma sinyalleri mutlu, şaşkın, üzgün, kızgın, korku, nötr gibi duygulara kategorilendirilmiş hem de değerlik, aktivasyon ve baskınlık olmak üzere 3 boyutlu duygu uzayında etiketlenerek veri seti oluşturulmuştur [81].

Yine Oflazoğlu ve Yıldırım 2013 yılında yayınladıkları çalışmalarında Türkçe konuşmalardan akustik özelliklere göre duygu tespiti gerçekleştirmişlerdir. Bu amaçla, daha önceki çalışmalarında oluşturdukları duygusal konuşma veri tabanını (TurES-Turkish Emotional Speech) üzerinde konuşmaların akustik özelliklerini kullanarak dört temel duygu sınıfının (nötr, üzgün, mutlu ve kızgın) sınıflandırması ve ilk duygu tahminleri gerçekleştirilmiştir. Bu amaçla VAM *(The Vera am Mittag German audio-visual emotional speech database)* [82] veri seti ile çapraz deneyler gerçekleştirmişlerdir. Çalışmada LIBSVM kütüphanesi ile uygulanan radyal temel çekirdek fonksiyon ile DVM (SVM-RBF)'nin ve Weka uygulaması tarafından sağlanan Bayes Ağının (BayesNet) performanslarını değerlendirmişlerdir [83].

Önder ve ark. 2013 yılında yayınladıkları çalışmaları ile BAUM-1 (BAhçeşehir University Multimodal Database-1) olarak isimlendirdikleri görsel işitsel Türkçe duygusal veri setini sunmuşlardır. Veri setinin oyunculuk gösterilerek oluşturulmuş kısmını BAUM-1 a (acted) , doğal (spontane) oluşturulmuş kayıtlardan oluşan kısmını ise BAUM-1 s (spontaneous) olarak etiketlendirmişlerdir. BAUM-1 a mutluluk, üzüntü, kızgınlık, tiksinme, sıkılma, ilgi(merak), korku, kafa karışıklığı, nötr ve şaşırma olmak üzere 10 duygu v içeren 278 video kayıttan oluşmaktadır. BAUM-1 s veri seti ise öfke, korku, iğrenme, mutluluk, üzüntü, sürpriz, can sıkıntısı, aşağılama, kafa karışıklığı, nötr, düşünme hali, konsantre olma durumu ve rahatsız olma olmak üzere 13 duygusal ve zihinsel durum içeren 1222 video kayıttan oluşmaktadır. Kayıtların alındığı denekler 18-66 yaş aralığında 13 kadın, 18 erkek toplam 31 kişiden oluşmaktadır. Aynı zamanda oluşturulan veri seti üzerinde DVM kullanarak duygu tanıma ile ilgili uygulamalar

gerçekleştirmişlerdir. Bu uygulamalar içinde eNTERFACE veri seti de sonuçların karşılaştırılması bakımından kullanılmıştır. BAUM-1 a kayıtlarında yapılan konuşmadan duygu tanıma deneylerinde beş duygu için %71 civarında bir duygu tanıma oranı, eNTERFACE veri setinde %73 bir duygu tanıma oranı elde edilmiş ve bu oranlar uyumlu olarak yorumlanmıştır [54].

Eroğlu Erdem ve ark. 2015 yılında yayınlanan çalışmalarında ilk olarak farklı dillerdeki filmlerden ve TV programlarından görsel-işitsel yüz video klipleri çıkarabilen yarı otomatik bir yöntem sunmuşlardır. Bu yöntemle oluşturdukları görsel-işitsel veri setini BAUM-2 olarak isimlendirmişlerdir. Bu video kayıtlarındaki tepe karelerden elde edilen resimler ile de BAUM-2i olarak isimlendirdikleri bir veri seti oluşturmuşlardır. Kayıtlar 5-73 yaş aralığında, 118'i kadın 168'i erkek olmak üzere 286 aktörden elde edilmiştir. 616'sı Türkçe, 431'i İngilizce olmak üzere toplam 1047 kayıt yer almaktadır. Mutluluk: 248, öfke: 173, üzüntü: 137, iğrenme: 51, sürpriz: 152, korku: 68, aşağılama:49, nötr: 169 olmak üzere 8 farklı duyguyu barındırmaktadır [84].

Kaya ve ark. 2014 yılında yayınladıkları çalışmalarında Meral ve ark. tarafından oluşturulan Boğaziçi Üniversitesi Duygusal Konuşma Veritabanını (BÜ-EE Duygulu Konuşma Veritabanı) [85] kullanmışlardır. Bu veritabanı 11 amatör aktörün 11 duygusal olarak tanımsız cümleyi Stanislavski etkisi kullanarak 4 ayrı duygu durumu (sevinç, olağan, kızgınlık, üzüntü) ile canlandırdığı toplam 484 kayıttan oluşmaktadır. Bu veritabanı üzerinde eğitim, geçerleme ve test kümeleri tanımlanmış, DVM ve Rassal Ormanlar ile referans sonuçlar elde edilmiştir. Test kümesinde en iyi sonuç %64,20 ile doğrusal çekirdekli DVM ile elde edilmiştir [86].

Parlak ve ark. 2014 yılında yayınladıkları çalışmalarında EmoSTAR ismini verdikleri bir veri setini tanıtmışlardır. Televizyon ve internetteki kaynaklardan derlenerek hazırlanmış bir veri setidir. Toplamda 393 olmak üzere kızgın, nötr, mutlu ve üzgün Türkçe ve İngilizce örneklerden oluşmaktadır. Yayınladıkları bu çalışmadan Berlin Emo-DB veri seti ile çapraz testler gerçekleştirmişlerdir. Sınıflandırma işlemlerini ise Naive Bayes ve DVM (Sıralı Minimal Optimizasyon-SMO) algoritmaları kullanarak gerçekleştirmişlerdir. Farklı özellik seçme yöntemlerini de kullanarak yaptıkları denemelerde DVM-SMO sınıflandırıcı ile en yüksek %97,20 başarı oranı ile iyi bir sonuç elde edilmiştir [87].

Oflazoğlu ve Yıldırım 2015 yılında sundukları bir diğer çalışmalarında Türkçe duygusal konuşmaların sınıflandırılmasında ikili sınıflandırma (binary classification) performanslarını hesaplamışlardır. TurES veri setinde yer alan 4 temel duygu sınıfına (mutlu, kızgın, üzgün, nötr) ait konuşma sinyalleri üzerinde başarıları incelenmiştir. Yapılan bu çalışmada konuşma sinyallerine ait akustik özniteliklerden yararlanılmıştır. Öznitelikler OpenSMILE yazılımı ile elde edilmiştir. Duyguların sınıflandırılmasında ise DVM, Bayes Ağı ve WEKA'da yer alan J48 karar ağacı sınıflandırıcıları kullanılmıştır. İkili duygu sınıflarının tahmininde başarım oranı, DVM ve Bayes Ağı ile kızgın-üzgün, kızgın-nötr duyguları için yaklaşık %80 bir ağırlıksız toplam doğruluk olarak hesaplanmıştır. En düşük başarım oranı ise mutlu-nötr duygusu için %64 olarak elde edilmiştir [88].

Korkmaz ve Atasoy 2015 yılında yayınladıkları çalışmalarında DVM ve k-en yakın komşuluk sınıflandırma algoritmaları kullanarak ses sinyali üzerinde mutlu, kızgın, nötr ve üzgün olmak üzere dört farklı duygu için tanıma işlemi gerçekleştirmişlerdir. Çalışmalarında EmoSTAR [50] veri setini kullanmışlardır. Konuşma sinyallerinin duygusal içeriği için mel frekans kepstral katsayıları incelenmiştir. MFKK'nın istatistiksel değerleri ve MFKK'nın enerji, sıfır kepstral katsayısı, birinci türevi, ikinci türevi incelenmiş ve çıkarılan özellikler sırasıyla DVM ve K-EK sınıflandırıcıları kullanılarak analiz edilmiştir. Daha sonra, MFKK üzerinde farklı çerçeve uzunluğu ve kaydırma süresinin sonuçları DVM ve K-EK algoritmaları için incelenmiştir. En iyi sonuçlar MFKK 'nın 12 boyutlu özellik vektörü için elde edilmiştir. Sonuç olarak, EmoSTAR veritabanı için %97,5 bir başarı ortaya çıkmış ve duygu sınıflandırmasında DVM yönteminin K-EK yönteminden daha iyi olduğunu göstermiştir[20].

Zhalehpour ve ark. 2016 yılında yayınlanan çalışmalarında BAUM-1 veri setinde çalışmışlar ve sınıflandırma işlemi için DVM sınıflandırıcı kullanmışlardır. BAUM-1 a veri seti için 8 duygu ve zihinsel durum sınıflandırması sadece ses ile gerçekleştirildiğinde ortalama %63,53 tanıma oranına, sadece video ile gerçekleştirildiğinde ortalama %31,24 tanıma oranına, hem ses hem de video ile gerçekleştirildiğinde ortalama %65,84 tanıma oranına ulaşılmıştır. 5 duygu ile yapılan çalışmalarda sadece ses duygusu sınıflandırması %71,71 oranında olmuştur. Kullanılan modelin doğruluk oranı eNTERFACE veri setinde

ise sadece ses için %72,95 olarak hesaplanmıştır. Bununla birlikte BAUM-1 s veri setinde elde edilen doğruluk oranlarının BAUM-1 a veri setinde elde edilen sonuçlardan çok daha düşük olduğu görülmüştür [89].

Bakır ve Yuzkat 2018 yılında yayınladıkları çalışmada, Türkçe ses formları ve özelliklerin dikkate alındığı hibrit bir model kullanarak otomatik KDT sistemi geliştirmişlerdir. Çalışma için 25 erkek ve 25 kadından yaklaşık 3000 farklı uzunluktaki kelime ve cümle Türkçe ses örneği toplanmış ve özgün bir Türkçe veri seti oluşturulmuştur. Ses örneklerinin özellikleri MFKK ve *Mel Frequency Discrete Wavelet Coefficients (MFDWC)* kullanılarak elde edilmiştir. Ayrıca modelin ilk aşamasında bu ses örneklerinin spektral özellik vektörleri DVM kullanılarak elde edilmiştir. Elde edilen özellik vektörleri, GKM, YSA, Dinamik Zaman Bükme-DZM *(Dynamic Time Warping-DTW),* GKM ve DVM hibrit modelleri ile eğitilmiştir. Hibrit model, DVM ve GMM ile birleştirilerek uygulanmıştır. Sonuçlar dikkate alındığında hibrit model diğer konuşma duygu tanıma yöntemlerine göre daha başarılı sonuçlar vermiştir [90].

Canpolat ve ark. 2020 yılında yayınlanan çalışmalarında Türkçe dilinde "Türkçe Ses-Duygu Veri Seti (TurEV-DB)" olarak isimlendirdikleri veri setinin geliştirilmesini açıklamışlardır. Veri setinde KZFD spektrogram resimleri, seçilmiş Türkçe kelimelerden oluşan ses dosyaları ve openSMILE ile elde edilen özellikler de yer almaktadır. TurEV-DB veri seti 408 adeti sakin (nötr), 487 adeti sinirli,483 adeti üzgün, 357 adeti mutlu olmak üzere dört farklı duygu durumu barındırmaktadır. Bu dört duygu tipinin (kızgın, sakin, mutlu ve üzgün) üç farklı frekans bandında bir kelimelik seslendirmeleri yer almaktadır. Çalışmalarında bu üç farklı frekans bandında uyguladıkları ESA ve DVM yöntemlerinin kullanıldığı KDT modelini de sunmuşlardır. Model, TurEV kullanılarak eğitilmiş ve doğrulama çalışmaları yapılmıştır. Sonuçlar, modelin duygu tanımada kullanılabilir bir yapıya sahip olduğuna işaret etmektedir [91].

Bir diğer çalışmada Özsönmez ve ark. Türkçe TurES ve TurEV-DB veri setlerini kullanarak duygu sınıflandırma işlemi gerçekleştirmişlerdir. Veri setleri üzerinde özellik çıkarımı için MFKK, Göreceli Spektra-(PLP-RASTA), DTK, Mel Spektrogram yöntemlerini kullanmışlardır. Temel bileşenler analizi yöntemi ile özellik seçimi yapılmış, sınıflandırma ise derin öğrenme modelleri ile gerçekleştirmişlerdir. En iyi sonuç TurEV-DB veri seti üzerinde bütün özellik tiplerinin kullanıldığı; 100 nöron, 1 ara katman, Uyarlanabilir Moment *(Adaptive Moment)*-ADAM optimize edici ve 80 döngü parametresi ile oluşturulan %92,2 doğruluk oranına sahip modele aittir. TurES veri seti üzerinde de Sentetik Azınlık Aşırı-Örnekleme (SMOTE) ve Temel Bileşenler Analizi işlemlerinin uygulandığı ve 250 nöronlu, 2 ara katmanlı Stokastik Gradyan İniş *(Stochastic Gradient Descent)*-SGD optimize edici ve 80 döngülü model oluşturulduğunda önceki çalışmalardan daha yüksek olarak %85,5 ortalama doğruluk oranı elde edilmiştir [92].

Yukarıda açıklanan çalışmalara ait kullanılan veri setleri, özellikler, sınıflandırıcı yöntemleri ve elde edilen sonuçların listesi Tablo 5' te özet halinde sunulmuştur

Tablo 5. *Türkçe veri setleri ile yapılan çalışmaların listesi*

Makale Başlığı	**Kullanılan Veri seti(leri)**	**Özellikler**	**Sınıflandırıcı**	**Sonuçlar**
Recognizing emotion from Turkish speech using acoustic features (2013) [83]	TurES, VAM	openSmile tarafından çıkarılan düşük seviyeli tanımlayıcı özellikler: temel frekans (F0), ses yüksekliği, seslendirme olasılığı, 0-14 MFKK, Mel frekans bantlarının 0 ila 7 logaritmik gücü (logMelFreqBand), 8 DTK katsayısından	DVM, Bayes Ağı	DVM sınıflandırıcı ile ağırlıklı ortalama duyarlılık: %57,5 ağırlıksız ortalama duyarlılık: %43,3 Bayes Ağı sınıflandırıcı ile ağırlıklı ortalama duyarlılık: %51,8 ağırlıksız ortalama duyarlılık: %45,5

		(lspFreq) hesaplanan 0 ila 7 satır spektral çift frekansları ve ses kalitesi (parlaklık-shimmer ve jitter-titreme)		
A Turkish audio-visual emotional database (2013) [54]	BAUM-1, eNTERFACE	MFKK ve algısal doğrusal kestirime dayalı göreli spektral katsayıları (RASTA-PLP). MFKK için 12, RASTA-PLP için 13 katsayı	DVM	BAUM-1 a veri seti için %71 doğruluk, eNTERFACE veri seti için %73
Protocol And Baseline For Experiments On Bogazici University Turkish Emotional Speech Corpus (2014) [86]	BÜ/EE (Boğaziçi Üniversitesi, Elektrik Elektronik Mühendisliği) Duygulu Konuşma Veritabanı	openSmile tarafından çıkarılan düşük seviyeli tanımlayıcı özellikler: (bazıları) temel frekans (F0), MFKK, enerji	DVM, Rassal Ormanlar (RO)	DVM %64,20, RO %52,27 doğruluk
A cross-corpus experiment in speech emotion recognition (2014) [87]	EmoSTAR, Berlin Emo-DB	openSmile ile elde edilen konfigürasyonlardaki özellik sayıları *-Emobase.conf dosyası* : 988 (26 LLD + 26 delta)*19 fonksiyon *-Emo_large.conf dosyası*: 6669 (57 LLD + 57 delta+ 57 delta-delta)*39 fonksiyon	(Weka ile uygulanan) Naive Bayes (NB), SMO ve Bagging (Bag)	4 duygu için (kızgın, mutlu, nötr, üzgün) en yüksek başarım: *Berlin Emo-DB veriseti* için emobase.conf özellikleri ile SMO sınıflandırıcıda %*92,33*, *EmoSTAR veriseti* için emo_large.conf özellikler ile SMO sınıflandırıcıda *%96,94*
Binary classification performances of emotion classes for Turkish Emotional Speech (2015) [88]	TurES	openSmile ile elde edilen *Emo_large.conf dosyası*: 6669 öznitelik. Bazı akustik özellikler: [0-12] mel frekansı kepstrum katsayısı, temel frekans (F0) ve zarfı, seslilik olasılığı, logaritmik	DVM, Bayes Ağı, Karar Ağaçları	İkili duygu çiftlerinin sınıflandırmasında kızgın-üzgün için ağırlıksız doğruluk DVM ile %76, Bayes Ağı ile %79, Karar Ağacı (J48) ile %70

		enerji, sıfır geçiş oranı		
Emotion recognition from speech signal using mel-frequency cepstral coefficients (2015) [20]	EmoSTAR	MFKK	DVM, K-EK	12 boyutlu MFFK vektörü için DVM ile %97,5, K-EK için %93,3 doğruluk
BAUM-1: A spontaneous audio-visual face database of affective and mental states (2016) [89]	BAUM-1, eNTERFACE	MFKK ve algısal doğrusal kestirime dayalı göreli spektral katsayıları (RASTA-PLP). MFKK için 12, RASTA-PLP için 13 katsayı	DVM	Sadece ses için: eNTERFACE veri setinde 6 duygu ile %72,95 doğruluk, BAUM-1a veri setinde 5 duygu ile %71,71 doğruluk
Speech emotion classification and recognition with different methods for Turkish language (2018) [90]	5 duygu (kızgın, korku, üzgün, mutlu, nötr) için yaklaşık 3000 kayıt barındıran özgün bir veri seti.	MFDWC, MFKK, DÖKK	DVM, GKM	Doğrusal çekirdek fonksiyonu ile erkek kayıtlar için %76,78 kadın kayıtlar için %79,85. Polinom çekirdek fonksiyonu ile erkek kayıtlar için %80,67, kadın kayıtlar için %81,37 doğruluk
Turkish Emotion Voice Database (TurEV-DB) (2020) [91]	TurEV-DB	openSmile ile elde edilen özellikler: (bazıları) F0, seslendirme olasılığı ve ses yüksekliği	ESA, DVM	0-8000 HZ frekans bandında Örn: Kızgın duygusu için, 0,84 kesinlik, 0,76 duyarlılık, 0,80 F1 skor
Optimal Classifier Selection in Turkish Speech Emotion Detection (2021) [92]	TurES, TurEV-DB	MFKK, PLP-RASTA, DTK, Mel-Spektrogram,	YSA	TurEV-DB veri setinde %92,2 doğruluk, TurES veri setinde %92,2 doğruluk

5. SONUÇLAR VE TARTIŞMA (DISCUSSION AND CONCLUSIONS)

Bu çalışma derin öğrenme mimarileri kullanılarak gerçekleştirilmiş konuşmadan duygu tanıma uygulamalarını kapsayacak nitelikte hazırlanmıştır. Derin öğrenme uygulamalarının pek çok araştırma alanına dahil olmasıyla birlikte elde ettiği popülerlik KDT sistemlerine de yansımış ve bu literatür araştırması çalışmasının motivasyon kaynağı olmuştur. Bununla birlikte diğer çalışmalardan farklı olarak

Türkçe veri setleri ile yapılmış çalışmalar ayrı bir bölüm olarak incelenmiştir. Farklı veri setleri üzerinden denenmiş yine farklı modellerden elde edilen sonuçlar sunulmuş ve bu yönüyle KDT alanında araştırmalar ve uygulamalar gerçekleştirecek araştırmacılara katkı sunması hedeflenmiştir.

KDT ses işleme ile ilgili bir çalışma alanı olmakla birlikte insan konuşmasının işlenmesinin yanı sıra duygu ve duygunun ifadesinin ses üzerinden elde edilmesi amacı ile ortaya çıkan ek zorluklara sahiptir. Çünkü konuşmada olduğu gibi duyguların ifadesinde de kültüre ve dile dayalı etkiler vardır. KDT uygulamalarının zorlu yönlerinden bir diğeri ise gerçek hayat uygulamalarına uyarlanmasıdır. Bunun sebeplerinden birisi ise veri setlerinde yer alan kayıtların genellikle profesyonel veya amatör aktörlerin gerçekleştirdiği performanslardan elde edilmesi ve dil çeşitliliğin az olmasıdır. KDT sistemlerinin gerçek hayat uygulamalarında ve gerçek zamanlı olarak yer alması sürecine aktörlük yapılmadan, doğal ve gürültülü ortamlardan elde edilen ve dillere ait farklı şivelerin kayıtlarından oluşan veri seti sayısının artırılması katkı sunacaktır.

KDT uygulamalarında özellik çıkarma ve duygunun sınıflandırılması aşamalarında kabul görmüş bir yöntem olmamakla birlikte son yıllarda derin öğrenme modelleri ile gerçekleştirilen uygulamalarda bir artış olduğu görülmüştür. Derin öğrenme modellerinin uygulandığı çalışmalarda ham konuşma dalga formunu [49,74, 78], spektrogram [75,76,79,91] veya log-mel spektrogramları [49,70] gibi farklı girdi türleri kullanılmıştır. Tablo 6'da derin öğrenme modellerinin kullanıldığı KDT uygulamalarında elde edilen sonuçların test edildikleri veri seti bazında gruplandırılması yer almaktadır. Burada özellikle TSA mimarilerinin ve ESA modellerinin sık uygulanan yöntemler olduğu bununla birlikte dikkat mekanizmasına dayalı olarak geliştirilen modellerde de başarı oranlarının %80'lerin üzerine çıktığı görülmüştür [49,50,72].

Ayrıca geliştirilen modeller farklı veri setlerinde test edildiğinde duygunun sınıflandırmasında farklı doğruluklar ortaya çıkmış ve bir veri setinde elde edilen yüksek doğruluk oranına farklı bir veri setinde ulaşılamadığı görülmüştür. Örneğin [49]'da yapılan çalışmada kullanılan model Berlin Emo-DB veri setinde IEMOCAP veri setinden daha yüksek doğruluk oranı elde etmiştir. Yine tespiti yapılabilecek bir diğer durum ise konuşmacıya bağlı ve konuşmacıdan bağımsız yapılan deneylerde konuşmacıya bağlı deneylerin performansının daha iyi olduğudur. Bu durumda ESA, UKSB modellerinin bir arada kullanılması daha iyi sonuçlar elde edilmesini sağlamıştır.

Tablo 6. *Çalışmaların veri seti bazında gruplandırılması*

Veri seti	**Çalışma**	**Kullanılan Model**	**Sonuçlar**
IEMOCAP	[49]	2B ESA-UKSB	Konuşmacıya bağlı deneylerde %89,16, konuşmacıdan bağımsız deneylerde %52,14 doğruluk
	[70]	ADRNN	Konuşmacıya bağlı deneylerde %74,96, konuşmacıdan bağımsız deneylerde %69,32 doğruluk
	[72]	ESA, İY-UKSB	%80,1 ağırlıksız doğruluk oranı
	[73]	1B ESA	%64,3 doğruluk
	[74]	EUKSB	%75 doğruluk
	[75]	İY-UKSB	%72,25 doğruluk
	[76]	ESA	%77,01 doğruluk
	[77]	UKSB - YÖD	%61,20 ağırlıklı doğruluk, %54,99 ağırlıksız doğruluk
	[78]	1B GESA	%73 duyarlılık
	[79]	DÖA	AHSN olarak isimlendirilen kısımda %65,61 , 40db gürültülü örneklerde 65,90 ağırlıklı ortalama recall oranı, AESN olarak isimlendirilen kısımda temiz konuşma örneklerinde %62,99, 40db gürültülü örneklerde %62,81 ağırlıklı ortalama recall oranı
Berlin Emo-DB	[49]	2B ESA-UKSB	Konuşmacıya bağlı deneylerde %95,33, konuşmacıdan bağımsız deneylerde %95,89

	[70]	ADRNN	Konuşmacıya bağlı deneylerde %90,78, konuşmacıdan bağımsız deneylerde %85,39 doğruluk
	[73]	1B ESA	535 örnekli 7 duygu sınıfı için %86,1 520 örnekli uygulamada %95,71 doğruluk
	[75]	İY-UKSB	%85,52
	[76]	ESA	%92,02 ağırlıksız doğruluk oranı
	[77]	UKSB - YÖD	%85,95 ağırlıklı doğruluk oranı, %82,6 ağırlıksız doğruluk oranı
	[78]	1B GESA	%90 duyarlılık
RAVDESS	[69]	UKSB	%68,8 doğruluk
	[73]	1B ESA	%71,61 doğruluk
	[74]	EUKSB	%80 doğruluk
	[75]	İY-UKSB	%77,02 doğruluk
RML	[69]	UKSB	%70,35 doğruluk
	[39]	Özellik seçiminde OK, sınıflandırma için DVM	%74,07 doğruluk
SAVEE	[69]	UKSB	%72,13 doğruluk
CASIA	[71]	UKSB	%92,8 ağırlıksız ortalama duyarlılık
eNTERFACE			%89,6 UAR
GEMEP			%57,0 UAR
AFEW 5.0	[80]	DVM	%35,77 doğruluk
BAUM-1 s			%44,06 doğruluk

Bir diğer yandan Türkçe veri setlerinin kullanıldığı çalışmalar son zamanlarda artış gösterse de yaygın kullanılan veri setleri ile kıyaslandığında hala yeterli düzeyde olmadığı söylenebilir. Bu alanda yapılacak çalışmaların sayısının artmasına Türkçe veri setlerine erişim imkanının Berlin Emo-DB, RAVDESS gibi veri setleri kadar kolay olması katkı sunabilir. İleriye dönük olarak yapılacak çalışmalarda ise konuşmaya ait özellikleri ve duygu sınıflandırma için kullanılan modelleri çeşitlendirmek ve hibrit modeller üretmek, modelin başarımı için Türkçe veri setleri arasında çapraz uygulamalar gerçekleştirmek Türkçe dili için KDT uygulamalarını güçlendirecektir.

Sonuç olarak, KDT uzun yıllardır araştırmalar gerçekleştirilen bir alan olmuştur. Uygulamaları eğitim, güvenlik, sağlık, pazarlama, IoT, sanal gerçeklik gibi birçok alanda hayatımızda yer alabilecek ve potansiyeli hala keşfetmeye açık güncel bir araştırma alanı olmaya da devam etmektedir.

KAYNAKLAR (REFERENCES)

[1] Duygu kelimesinin tanımı. Türk Dil Kurumu TDK, https://sozluk.gov.tr/ Erişim tarihi: 20/03/2022

[2] Sibel, S. Ü. (2013). Örgütlerde duygusal zeka. Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 16(29), 213-242.

[3] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

[4] Li, X., & Lin, R. (2021, December). Speech Emotion Recognition for Power Customer Service. In 2021 7th International Conference on Computer and Communications (ICCC) (pp. 514-518). IEEE.

[5] Simcock, G., McLoughlin, L. T., De Regt, T., Broadhouse, K. M., Beaudequin, D., Lagopoulos, J., & Hermens, D. F. (2020). Associations between facial emotion recognition and mental health in early adolescence. International journal of environmental research and public health, 17(1), 330.

[6] Saste, S. T., & Jagdale, S. M. (2017, April). Emotion recognition from speech using MFCC and DWT for security system. In 2017 international conference of electronics, communication and aerospace technology (ICECA) (Vol. 1, pp. 701-704). IEEE

[7] Yang, D., Alsadoon, A., Prasad, P. C., Singh, A. K., & Elchouemi, A. (2018). An emotion recognition model based on facial recognition in virtual learning environment. Procedia Computer Science, 125, 2-10.

[8] Er, M. B., & Harun, Ç. İ. Ğ. (2020). Türk Müziği Uyaranları Kullanılarak İnsan Duygularının Makine Öğrenmesi Yöntemi İle Tanınması. Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji, 8(2), 458-474

[9] Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. Image and Vision Computing, 65, 3-14

[10] Nasukawa, T., & Yi, J. (2003, October). Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd international conference on Knowledge capture (pp. 70-77).

[11] Rhanoui, M., Mikram, M., Yousfi, S., & Barzali, S. (2019). A CNN-BiLSTM Model for Document-Level Sentiment Analysis. Machine Learning and Knowledge Extraction, 1(3), 832-847.

[12] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, *18*(1), 32-80.

[13] Busso, C., Lee, S., & Narayanan, S. (2009). Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE transactions on audio, speech, and language processing, 17(4), 582-596.]

[14] Wu, S., Falk, T. H., & Chan, W. Y. (2011). Automatic speech emotion recognition using modulation spectral features. Speech communication, 53(5), 768-785.

[15] Jayalekshmi, J., & Mathew, T. (2017, July). Facial expression recognition and emotion classification system for sentiment analysis. In 2017 International Conference on Networks & Advances in Computational Technologies (NetACT) (pp. 1-8). IEEE.

[16] Wu, T., Peng, J., Zhang, W., Zhang, H., Tan, S., Yi, F., ... & Huang, Y. (2022). Video sentiment analysis with bimodal information-augmented multi-head attention. Knowledge-Based Systems, 235, 107676.

[17] Zadeh, A. (2015). Micro-opinion Sentiment Intensity Analysis and Summarization in Online Videos. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI ’15).

[18] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250.

[19] Koolagudi, S. G., Kumar, N., & Rao, K. S. (2011, February). Speech emotion recognition using segmental level prosodic analysis. In 2011 international conference on devices and communications (ICDeCom) (pp. 1-5). IEEE

[20] Korkmaz, O. E., & Atasoy, A. (2015, November). Emotion recognition from speech signal using mel-frequency cepstral coefficients. In 2015 9th International Conference on Electrical and Electronics Engineering (ELECO) (pp. 1254-1257). IEEE.

[21] Ingale, A. B., & Chaudhari, D. S. (2012). Speech emotion recognition. International Journal of Soft Computing and Engineering (IJSCE), 2(1), 235-238.

[22] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[23] Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems, 30(11), 3212-3232.

[24] Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 13(3), 55-75.

[25] Liang, H., Sun, X., Sun, Y., & Gao, Y. (2017). Text feature extraction based on deep learning: a review. EURASIP journal on wireless communications and networking, 2017(1), 1-12.

[26] Yao, K., Yu, D., Seide, F., Su, H., Deng, L., & Gong, Y. (2012, December). Adaptation of context-dependent deep neural networks for automatic speech recognition. In 2012 IEEE Spoken Language Technology Workshop (SLT) (pp. 366-369). IEEE.

[27] Aravindpai Pai, "CNN vs. RNN vs. ANN – Analyzing 3 Types of Neural Networks in Deep Learning" https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/ Erişim Tarihi: 21/02/2022

[28] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. IEEE Access, 7, 117327-117345.

[29] Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116, 56-76.

[30] El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern recognition, 44(3), 572-587.

[31] Eyben, F., Wöllmer, M., & Schuller, B. (2010, October). Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia (pp. 1459-1462

[32] Boersma, P. & Weenink, D. (1992–2022):Praat: doing phonetics by computer [Computer program]. https://www.fon.hum.uva.nl/paul/praat.html Erişim tarihi: 20/05/2022

[33] Chen, S., Jin, Q., Li, X., Yang, G., & Xu, J. (2014, September). Speech emotion classification using acoustic features. In The 9th International Symposium on Chinese Spoken Language Processing (pp. 579-583). IEEE.

[34] Jacob, A. (2016, April). Speech emotion recognition based on minimal voice quality features. In 2016 International conference on communication and signal processing (ICCSP) (pp. 0886-0890). IEEE.

[35] Zhou, Y., Sun, Y., Zhang, J., & Yan, Y. (2009, December). Speech emotion recognition using both spectral and prosodic features. In 2009 international conference on information engineering and computer science (pp. 1-4). IEEE.

[36] Wang, Y., Du, S., & Zhan, Y. (2008, October). Adaptive and optimal classification of speech emotion recognition. In 2008 fourth international conference on natural computation (Vol. 5, pp. 407-411). IEEE.

[37] Rao, K. S., Koolagudi, S. G., & Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. International journal of speech technology, 16(2), 143-160.

[38] Li, X., Tao, J., Johnson, M. T., Soltis, J., Savage, A., Leong, K. M., & Newman, J. D. (2007, April). Stress and emotion classification using jitter and shimmer features. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07 (Vol. 4, pp. IV-1081). IEEE.

[39] Aouani, H., & Ayed, Y. B. (2020). Speech emotion recognition with deep learning. Procedia Computer Science, 176, 251-260.

[40] Pathak, S., & Kulkarni, A. (2011, April). Recognizing emotions from speech. In 2011 3rd International Conference on Electronics Computer Technology (Vol. 4, pp. 107-109). IEEE.

[41] Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. Speech communication, 41(4), 603-623.

[42] Jiang, P., Fu, H., Tao, H., Lei, P., & Zhao, L. (2019). Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. IEEE Access, 7, 90368-90377.

[43] Jain, M., Narayan, S., Balaji, P., Bhowmick, A., & Muthu, R. K. (2020). Speech emotion recognition using support vector machine. arXiv preprint arXiv:2002.07590.

[44] Zhou, G., Hansen, J. H., & Kaiser, J. F. (2001). Nonlinear feature based classification of speech under stress. *IEEE Transactions on speech and audio processing*, *9*(3), 201-216.

[45] Bandela, S. R., & Kumar, T. K. (2017, July). Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[46] Mairesse, F., Polifroni, J., & Di Fabbrizio, G. (2012, March). Can prosody inform sentiment analysis? experiments on short spoken reviews. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5093-5096). IEEE.

[47] Shen, Q., Wang, Z., & Sun, Y. (2017, October). Sentiment analysis of movie reviews based on cnn-blstm. In *International Conference on Intelligence Science* (pp. 164-171). Springer, Cham.

[48] Rosas, V. P., Mihalcea, R., & Morency, L. P. (2013). Multimodal sentiment analysis of spanish online videos. IEEE Intelligent Systems, 28(3), 38-45.

[49] Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical signal processing and control, 47, 312-323.

[50] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005, September). A database of German emotional speech. In Interspeech (Vol. 5, pp. 1517-1520).

[51] Haq, S. U. (2011). Audio visual expressed emotion classification. University of Surrey (United Kingdom).

[52] 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.

[53] Dhall, A., Ramana Murthy, O. V., Goecke, R., Joshi, J., & Gedeon, T. (2015, November). Video and image based emotion recognition challenges in the wild: Emotiw 2015. In Proceedings of the 2015 ACM on international conference on multimodal interaction (pp. 423-426).

[54] Önder, O., Zhalehpour, S., & Erdem, Ç. E. (2013, April). A Turkish audio-visual emotional database. In 2013 21st Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[55] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., N., Lee, S. & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335-359.

[56] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing, 5(4), 377-390.

[57] Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006, April). The eNTERFACE'05 audio-visual emotion database. In 22nd International Conference on Data Engineering Workshops (ICDEW'06) (pp. 8-8). IEEE.

[58] China Linguistic Data Consortium http://www.chineseldc.org Erişim Tarihi: 25/03/2022

[59] Bänziger, T., Pirker, H., & Scherer, K. (2006, May). GEMEP-GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions. In Proceedings of LREC (Vol. 6, pp. 15-019).

[60] Wang, Y., & Guan, L. (2005, March). Recognizing human emotion from audiovisual information. In Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. (Vol. 2, pp. ii-1125). IEEE.

[61] Latif, S., Qayyum, A., Usman, M., & Qadir, J. (2018, December). Cross lingual speech emotion recognition: Urdu vs. western languages. In 2018 International Conference on Frontiers of Information Technology (FIT) (pp. 88-93). IEEE.

[62] Costantini, G., Iaderola, I., Paoloni, A., & Todisco, M. (2014). EMOVO corpus: an Italian emotional speech database. In International Conference on Language Resources and Evaluation (LREC 2014) (pp. 3501-3504). European Language Resources Association (ELRA).

[63] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. IEEE Access, 9, 47795-47814.

[64] Wang, X., Chen, X., & Cao, C. (2020). Human emotion recognition by optimally fusing facial expression and speech feature. Signal Processing: Image Communication, 84, 115831.

[65] Zehra, W., Javed, A. R., Jalil, Z., Khan, H. U., & Gadekallu, T. R. (2021). Cross corpus multi-lingual speech emotion recognition using ensemble learning. Complex & Intelligent Systems, 7(4), 1845-1854

[66] Demircan, S., & Kahramanli, H. (2018). Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech. Neural Computing and Applications, 29(8), 59-66.

[67] Ganapathy, A. (2016). Speech Emotion Recognition Using Deep Learning Techniques. ABC Journal of Advanced Research, 5(2), 113-122.

[68] Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. Sensors, 21(4), 1249.

[69] Demir, A., Atila, O., & Şengür, A. (2019, September). Deep learning and audio based emotion recognition. In *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)* (pp. 1-6). IEEE.

[70] Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech emotion recognition from 3D log-mel spectrograms with deep learning network. IEEE access, 7, 125868-125881.

[71] Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., & Schuller, B. (2019). Speech emotion classification using attention-based LSTM. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(11), 1675-1685.

[72] Jalal, M. A., Milner, R., & Hain, T. (2020, October). Empirical Interpretation of Speech Emotion Perception with Attention Based Model for Speech Emotion Recognition. In INTERSPEECH (pp. 4113-4117).

[73] Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, 101894.

[74] Mustaqeem, Kwon, S. (2020). CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network. Mathematics, 8(12), 2133.

[75] Mustaqeem, Sajjad, M., & Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access, 8, 79861-79875.

[76] Anvarjon, T., Mustaqeem, & Kwon, S. (2020). Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. Sensors, 20(18), 5212.

[77] Li, D., Liu, J., Yang, Z., Sun, L., & Wang, Z. (2021). Speech emotion recognition using recurrent neural networks with directional self-attention. Expert Systems with Applications, 173, 114683.

[78] Mustaqeem, & Kwon, S. (2021). MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. Expert Systems with Applications, 167, 114177.

[79] Yusuf, S. M., Adedokun, E. A., Muazu, M. B., Umoh, I. J., & Ibrahim, A. A. (2021, October). RMWSaug: Robust Multi-window Spectrogram Augmentation Approach for Deep Learning based Speech Emotion Recognition. In 2021 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-6). IEEE.

[80] Zhang, S., Tao, X., Chuang, Y., & Zhao, X. (2021). Learning deep multimodal affective features for spontaneous speech emotion recognition. Speech Communication, 127, 73-81.

[81] Oflazoglu, Ç., & Yildirim, S. (2011, April). Turkish emotional speech database. In 2011 IEEE 19th Signal Processing and Communications Applications Conference (SIU) (pp. 1153-1156). IEEE.

[82] Grimm, M., Kroschel, K., & Narayanan, S. (2008, June). The Vera am Mittag German audio-visual emotional speech database. In 2008 IEEE international conference on multimedia and expo (pp. 865-868). IEEE.

[83] Oflazoglu, C., & Yildirim, S. (2013). Recognizing emotion from Turkish speech using acoustic features. EURASIP Journal on Audio, Speech, and Music Processing, 2013(1), 1-11.

[84] Eroglu Erdem, C., Turan, C., & Aydin, Z. (2015). BAUM-2: a multilingual audio-visual affective face database. Multimedia tools and applications, 74(18), 7429-7459.

[85] Meral, H. M., Ekenel, H. K., & Ozsoy, A. (2003). Analysis of emotion in Turkish. In XVII National Conference on Turkish Linguistics.

[86] Kaya, H., Salah, A. A., Gürgen, S. F., & Ekenel, H. (2014, April). Protocol and baseline for experiments on Bogazici University Turkish emotional speech corpus. In 2014 22nd Signal Processing and Communications Applications Conference (SIU) (pp. 1698-1701). IEEE.

[87] Parlak, C., Diri, B., & Gürgen, F. (2014, September). A cross-corpus experiment in speech emotion recognition. In SLAM@ INTERSPEECH (pp. 58-61).

[88] Oflazoglu, Ç., & Yıldırım, S. (2015, May). Binary classification performances of emotion classes for Turkish Emotional Speech. In 2015 23nd Signal Processing and Communications Applications Conference (SIU) (pp. 2353-2356). IEEE.

[89] Zhalehpour, S., Onder, O., Akhtar, Z., & Erdem, C. E. (2016). BAUM-1: A spontaneous audio-visual face database of affective and mental states. IEEE Transactions on Affective Computing, 8(3), 300-313.

[90] Bakır, C., & Yuzkat, M. (2018). Speech emotion classification and recognition with different methods for Turkish language. Balkan Journal of Electrical and Computer Engineering, 6(2), 122-128.

[91] Canpolat, S. F., Ormanoğlu, Z., & Zeyrek, D. (2020, May). Turkish Emotion Voice Database (TurEV-DB). In Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL) (pp. 368-375).

[92] Özsönmez, D. B., Acarman, T., & Parlak, İ. B. (2021, June). Optimal Classifier Selection in Turkish Speech Emotion Detection. In 2021 29th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.