T.C.
BURSA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

İKİ KEZ SIKIŞTIRILMIŞ SES SİNYALLERİNİN ANALİZİ VE OTOMATİK OLARAK TESPİT EDİLMESİ

DOKTORA TEZİ

Aykut BÜKER

Elektrik-Elektronik Mühendisliği Anabilim Dalı

Ocak 2023

T.C.
BURSA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

İKİ KEZ SIKIŞTIRILMIŞ SES SİNYALLERİNİN ANALİZİ VE OTOMATİK OLARAK TESPİT EDİLMESİ

DOKTORA TEZİ

Aykut BÜKER
(182331541001)

Elektrik-Elektronik Mühendisliği Anabilim Dalı

Tez Danışmanı: Doç. Dr. Cemal HANİLÇİ

Ocak 2023

BTÜ, Lisansüstü Eğitim Enstitüsü'nün 182331541001 numaralı Doktora Öğrencisi Aykut BÜKER, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "İKİ KEZ SIKIŞTIRILMIŞ SES SİNYALLERİNİN ANALİZİ VE OTOMATİK OLARAK TESPİT EDİLMESİ" başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı :	**Doç. Dr. Cemal HANİLÇİ** Bursa Teknik Üniversitesi	.....
Jüri Üyeleri :	**Prof. Dr. Hakan GÜRKAN** Bursa Teknik Üniversitesi	.....
	Prof. Dr. Vedat TAVŞANOĞLU Işık Üniversitesi	.....
	Doç. Dr. Ersen YILMAZ Bursa Uludağ Üniversitesi	.....
	Dr. Öğr. Üyesi Selma YILMAZYILDIZ KAYAARMA Bursa Teknik Üniversitesi	.....

Teslim Tarihi :
Savunma Tarihi : 30.01.2023

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, Bursa Teknik Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Lisansüstü Eğitim Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

İNTİHAL BEYANI

Bu tezde görsel, işitsel ve yazılı biçimde sunulan tüm bilgi ve sonuçların akademik ve etik kurallara uyularak tarafımdan elde edildiğini, tez içinde yer alan ancak bu çalışmaya özgü olmayan tüm sonuç ve bilgileri tezde kaynak göstererek belgelediğimi, aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim.

Öğrencinin Adı Soyadı: Aykut BÜKER

İmzası :

ÖNSÖZ

Doktora eğitim hayatım boyunca akademisyenliği öğreten, hayatımdaki her sorunda bana yardımcı olan, beraber çalışabildiğim için kendimi şanslı hissettiğim tez danışmanım Doç. Dr. Cemal HANİLÇİ'ye bana kattıklarından dolayı teşekkür ederim.

Lisansüstü eğitimimde bilgisini, desteğini ve zamanını esirgemeyen, tez çalışmamda farklı bakış açıları katan Prof. Dr. Vedat TAVŞANOĞLU'na teşekkürü borç bilirim.

Araştırma görevliliğim boyunca bana tecrübelerini aktaran, bana karşı her zaman anlayışlı ve destekleyici olan Elektrik-Elektronik Mühendisliği bölüm başkanı Prof. Dr. Hakan GÜRKAN ve Elektrik-Elektronik Mühendisliği bölüm başkan yardımcısı Doç. Dr. Mehmet Barış TABAKCIOĞLU'na minnettarım.

Hayatımda her zaman destek olan sevgili ailem ve aileden saydığım dostlarıma teşekkür ederim. Her anlamda bana güvendikleri için tüm zor anlarımda devam edebilme cesaretini bu insanlar sayesinde bulabildim. Giderek büyüyen aileme sevgiler.

Keyifli bir çalışma ortamı sunan, rahat bir şekilde bilgi paylaşımı yapabildiğim iş arkadaşlarıma teşekkürü borç bilirim. Son olarak beni her zaman motive eden, akademisyenliği sevdiren öğrencilerime teşekkür ederim.

Ocak 2023 Aykut BÜKER

İÇİNDEKİLER

Sayfa

KISALTMALAR

AAC : Advanced Audio Coding
AI : Artificial Intelligence
AMR : Adaptive Multi-Rate
ANN : Artificial Neural Network
CELP : Code-Excited Linear Prediction
CNN : Convolutional Neural Network
DCT : Discrete Cosine Transform
DFT : Discrete Fourier Transform
DNN : Deep Neural Network
FLAC : Free Lossless Audio Codec
GMM-UBM : Gaussian Mixture Model Universal Background Model
LP : Linear Prediction
LSTM : Long Short-Term Memory
LTAS : Long-Term Average Spectrum
MDCT : Modified Discrete Cosine Transform
MDSVC : MIT Mobile Device Speaker Verification Corpus
MITD : Multicodec Invdec Tampering Dataset
MLP : Multi Layer Perceptron
MP3 : MPEG-1 Audio Layer III
PCM : Pulse Code Modulation
SNR : Signal-to-Noise Ratio
STFT : Short Time Fourier Transform

SEMBOLLER

b	**:** bias
k	**:** ayrık frekans indisi
n	**:** örnek indisi
N	**:** konuşma çerçevesindeki toplam örnek sayısı
t	**:** çerçeve indisi
T	**:** konuşma sinyalinden elde edilen toplam çerçeve sayısı
x[n,t]	**:** konuşma sinyalinden elde edilen t. çerçeve
X(k,t)	**:** konuşma sinyalinin spektrogramı
w[n]	**:** pencere fonksiyonu
W	**:** ağırlık katsayısı
θ(k,t)	**:** konuşma sinyalinin faz spektrumu

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

İKİ KEZ SIKIŞTIRILMIŞ SES SİNYALLERİNİN ANALİZİ VE OTOMATİK OLARAK TESPİT EDİLMESİ

ÖZET

Bu tez çalışmasında iki kez sıkıştırılmış ses sinyallerinin tespit edilmesi problemi ele alınacaktır. Ses kayıtları mahkemelerde delil olarak yaygın bir şekilde kullanılmaktadır. Bununla birlikte ses kaydının içeriğinin değiştirilip değiştirilmediğini doğrulamak önemli bir problemdir. Ses kaydında manipülasyon yapmak isteyen biri ses sinyalini öncelikle dalga formuna dönüştürmelidir. Dalga formu üzerinde değişiklikler yapıldıktan sonra bu değişiklikleri gizlemek için sıkıştırma işlemi ile aynı formata dönüştürebilir. Bu sebeple iki kez sıkıştırılmış ses sinyallerini tespit etmek oldukça önemlidir. Literatürde yapılan çalışmalarda iki kez sıkıştırılmış ses tespitinde genellikle sıkıştırma için kullanılan ses kodlayıcısı hakkında bilgiler kullanılarak öznitelikler elde edilmiştir. Bu tez çalışmasında kodlayıcı hakkında herhangi bir ön bilgi olmadan genel öznitelikler elde edilmeye çalışılarak daha başarılı iki kez sıkıştırılmış ses sinyalinin tespit edilmesi hedeflenmektedir. Ayrıca literatürde derin sinir ağlarının kullanıldığı çalışma sayısının çok sınırlı olması sebebi ile derin sinir ağlarının hem öznitelik çıkarma hem de sınıflandırma aşamalarındaki güçlerinden faydalanarak başarılı bir sistem geliştirmek hedeflenmektedir. Öznitelikleri derin öğrenme yapılarının eğitim aşamasında otomatik olarak öğrenen ve gürültü, kanal ve kayıt ortamı gibi performansı olumsuz etkileyebilecek değişimlere karşı başarılı, iki kez sıkıştırılmış ses tespit sistemi geliştirilecektir. Farklı derin öğrenme yapıları ve farklı veri kümeleri kullanarak detaylı analiz yapılacaktır. Tek bir veri kümesi ile eğitilen sistemin diğer veri kümelerindeki ses sinyalleri ile test edildiğinde de iki kez sıkıştırılmış sesleri tespit etmede başarılı olması hedeflenmektedir. Kullanılacak veri kümelerinden biri daha önceden farklı kodlayıcılar ile sıkıştırılmış ses sinyalleri içermekte olup bu ses sinyalleri kullanılarak iki kez sıkıştırılmış ses tespitindeki performans incelenecektir. Tez çalışmasının ilk kısmında, dar bant AMR kodlayıcı kullanılarak farklı bit-oranları ile sıkıştırma işlemi yapılarak bir kez ve iki kez sıkıştırılmış ses sinyalleri üretilecektir. Ses sinyallerinin spektrogramları kullanılarak derin öğrenme modelleri ile eğitilecektir. Farklı modeller ve farklı kayıp fonksiyonları kullanılarak sistemin daha yüksek başarı oranı vermesi için çalışmalar yapılmıştır. Literatürde geniş bant AMR kodlayıcı ile iki kez sıkıştırılmış ses tespiti problemi daha önce çalışmadığından, tezin ikinci kısmında geniş bant AMR kodlayıcı ile iki kez sıkıştırılmış ses sinyalleri üretildikten sonra bu veriler ile iki kez sıkıştırılmış ses tespit sistemi incelenecektir.

Anahtar kelimeler: İki kez sıkıştırılmış ses tespiti, Uyarlamalı Çoklu-Oran Kodlayıcı, Konuşma sinyallerinin adli uygulamaları

ANALYSIS AND AUTOMATIC DETECTION OF DOUBLE COMPRESSED AUDIO SIGNALS

SUMMARY

In thesis study, double compressed audio detection problem will be investigated. As a result of widespread usage of audio recording as evidence in courts, this addresses the problem of authenticity of a audio content. In order to tamper an audio file, attacker first needs to decompress it to waveform and then manipulate the waveform which is followed by recompressing the manipulated audio waveform into the original format. For this reason, double compressed audio detection is an important audio forensic problem. Previous studies addressing the double compressed audio detection problem with prior knowledge about encoding and decoding processes of codec. In this study, it is aimed at detecting double compressed audio with extracting general feature when there is no prior knowledge about codec. There are a few studies including deep learning approaches on double compressed audio detection problem. The proposed system is used for both feature extraction and classifier with deep learning methods. The system will be developed to be succesful against noise, channel and recording environment variations. The system will be invegstigated using different deep learning models and datasets detailed. The performance of the system will be developed on cross-database evaluation. One of the dataset including audio with compression history will be investigated on double compressed audio detection problem. In the first part of thesis, single and double compressed audio recordings will be generated with narrow band AMR codec. Audio spectrograms will be used as input of deep learning model. The system will be improved with different deep learning models and different loss function to try on better performance. There is no previous study addressing the double compressed audio detection problem including wide band AMR codec. The system will be developed double compressed audio compressed by wide band AMR codec.

Keywords: Double compressed audio detection, Adaptive Multi-Rate Codec, Audio Forensic

1. GİRİŞ

İletişim teknolojilerindeki ve multimedya işleme sistemlerindeki son gelişmeler, bilgi içeriklerinin (örneğin metin, ses, görüntü ve video) kolayca üretilmesini, düzenlenmesini, paylaşılmasını ve manipüle edilmesini mümkün kılmaktadır. İnternet kullanıcılarının çoğu bu tür içerikleri çevrimiçi olarak paylaşmaktadır. Sonuç olarak, insanlar kendi kişisel medyalarını, çoğunlukla başkalarının kullanımına açık olan çeşitli biçimlerde (ses, görüntü ve video) üretirler. Bu içerikler, ücretsiz olarak kullanılabilen multimedya düzenleme araçlarıyla kolayca değiştirilebilir, düzenlenebilir veya manipüle edilebilir. Dijital formlardaki bilgi içeriklerinin yaygınlaşması ve manipülasyon araçlarına ulaşılabilirliğin kolaylaşmasının bir sonucu olarak, multimedya içerikleri genellikle mahkemelerde veya kolluk kuvvetlerinde delil olarak karşımıza çıkmaktadır. Bu, bir multimedya içeriğinin bütünlüğüne ve orijinalliğine odaklanan multimedya adli tıp (multimedia forensics) problemini işaret etmektedir. [1,2].

Multimedya adli tıp araştırmaları çoğunlukla görüntülerden kamera tanımlama [3], görüntülerdeki gizli mesajların varlığını belirlemeyi amaçlayan görüntü steganalizi [4], görüntü sahteciliği tespiti [5] ve iki kez sıkıştırılmış görüntülerin tespiti [6] gibi görüntü adli tıp uygulamaları üzerine odaklanmaktadır. Benzer sorunlar ses/konuşma sinyallerinin kullanıldığı adli uygulamalar açısından ele alınmış olsa da, ses/konuşma sinyallerinin kullanıldığı çalışmaların sayısı görüntü adli tıp çalışmalarına kıyasla göre çok daha azdır. Kayıt cihazı tanımlama [7], ses steganalizi [8], ses sahteciliği tespiti [9] ve iki kez sıkıştırılmış ses tespiti [10], adli görüntüye benzer temel adli ses problemleridir.

Yukarıda belirtilen ses sinyallerinin kullanıldığı adli uygulamalar arasında, iki kez sıkıştırılmış (double compressed-DC) ses/konuşma sinyali tespiti konusunda yalnızca sınırlı sayıda çalışma mevcuttur. Ancak daha fazla dikkat gerektiren zorlu ve önemli bir problemdir. Çünkü bir ses içeriği, ücretsiz olarak bulunabilen ses düzenleme yazılımları kullanılarak kolayca düzenlenebilir, değiştirilebilir veya manipüle edilebilir ve algısal olarak ayırt edilemeyen değiştirilmiş bir ses üretmek için orijinal

sıkıştırma formatına geri sıkıştırılabilir. Bu nedenle, bu tezde uyarlamalı çok oranlı (Adaptive Multi Rate-AMR) konuşma kodlayıcı ile iki kez sıkıştırılmış ses kayıtlarını tespit etme problemi ele alınmıştır. AMR ses sıkıştırma kodlayıcısı [11], mevcut mobil cihazların (telefon, tablet, kayıt cihazı vb.) çoğunun ses içeriğini AMR formatında depolaması ve AMR sesini açma, düzenleme ve diğer formatlara dönüştürmeye olanak tanıyan ücretsiz olarak kullanılabilen birkaç program olması nedeniyle seçilmiştir. AMR kodlayıcının mobil cihazlarda yaygın olarak kullanılması nedeni ile iki kez sıkıştırılmış AMR ses dosyalarının mahkemelerde delil olarak görünmesi daha olasıdır. Bu sebeple, iki kez sıkıştırılmış AMR sinyallerinin bütünlüğünü ve gerçekliğini doğrulamak için güvenilir ses adli tıp teknikleri gerekmektedir.

İki kez sıkıştırılmış ses tespit problemini ele alan önceki çalışmalar, kullanılan ses sıkıştırma kodlayıcı türüne göre üç gruba ayrılabilir: (i) iki kez sıkıştırılmış Hareketli Görüntü Uzmanlar Birliği (Moving Pictures Expert Group-MPEG)-1 Ses Katmanı III (MP3) kodlayıcı ile sıkıştırılmış ses tespiti, (ii) iki kez sıkıştırılmış MPEG- algılama 2/4 gelişmiş ses kodlama (Advanced Audio Coding-AAC) kodlayıcı ile sıkıştırılmış ses tespiti ve (iii) iki kez sıkıştırılmış AMR ses tespiti. İki kez sıkıştırılmış MP3 ses parçalarını tespit etmek için literatürde yer alan çalışmalarda genellikle değiştirilmiş ayrık kosinüs dönüşümü (Modified Discrete Cosine Transform-MDCT) tabanlı öznitelikler kullanılmıştır. Örneğin, bir kez sıkıştırılmış (single compressed-SC) ve iki kez sıkıştırılmış (double compressed-DC) MP3 sinyallerinden elde edilen nicemlenmiş MDCT katsayılarının küçük değerlerinin sayısı sahte kalitede MP3 dosyaları tespit etmek için karşılaştırılmış ve bu sayıların bir kez sıkıştırılmış ve iki kez sıkıştırılmış MP3 dosyaları arasında önemli ölçüde farklılık gösterdiği tespit edilmiştir [12]. Aynı araştırmacılar, iki kez sıkıştırılmış MP3 tespiti için nicemlenmiş MDCT katsayılarının ilk basamaklarından çıkarılan istatistiksel özellikler ile destek vektör makineleri (support vector machines-DVM) sınıflandırıcısını kullanmayı önermiştir [13]. [14] çalışmasında, MDCT katsayılarından çıkarılan 64 adet istatistiksel özniteliğin, DVM (destek vektör makineleri) sınıflandırıcı kullanılarak iki kez sıkıştırılmış MP3 dosyalarını tespit etmek için kullanılması önerilmiştir. [15, 16]'da araştırmacılar, sorgulanan MP3 dosyasının (muhtemelen iki kez sıkıştırılmış) nicelenmiş MDCT katsayılarının histogramları ile bir kez sıkıştırılmış MP3 dosyası arasında benzerlik ölçüsü kullanmayı önermiştir. Ardından, bunun bir bir kez sıkıştırılmış veya iki kez sıkıştırılmış MP3 ses olup olmadığına karar vermek için benzerlik ölçüsüne bir eşik değer uygulanmıştır. Ma ve ark. aynı bit hızında (bit-rate-

BR) iki kez sıkıştırılmış MP3 dosyalarını algılamak için pencerelenmiş ses çerçevelerinden çıkarılan ölçek faktörlerinin istatistiksel özelliklerini kullanmıştır [17]. MP3 kodlayıcı parametrelerinden çıkarılan istatistiksel özelliklerin [18]'de çoklu MP3 sıkıştırmasını tespit etmek için SVM sınıflandırıcı ile kullanılması önerilmiştir. İki kez sıkıştırılmış MP3 ses algılamaya benzer şekilde, iki kez sıkıştırılmış AAC ses dosyalarını tespit etmek için MDCT katsayılarından, Huffman kod çizelgesi dizinlerinin istatistiklerinden veya ölçek faktörlerinden çıkarılan öznitelikler kullanılmıştır [19, 20, 21]. Genel olarak, ikinci sıkıştırmanın bit hızı birinci sıkıştırmanınkine eşit veya daha büyük olduğunda daha yüksek bir tanıma doğruluğu elde edilmiştir [19, 20, 21].

Neredeyse tüm taşınabilir cihazlar ses kayıtları için AMR kodlayıcıyı kullandığından, iki kez sıkıştırılmış AMR ses tespiti son yıllarda büyük ilgi görmüştür. [10]'da, iki kez sıkıştırılmış AMR tespiti için SVM sınıflandırıcısı ile çeşitli frekans tanım bölgesi istatistiksel özellikleri (alt bant enerji oranı, düşük frekans enerji oranı, iki bantlı özellikler ve uzun vadeli doğrusal öngörülü kodlama) kullanılmıştır. Tanıma doğruluğunun, ikinci sıkıştırma bit hızı birinci sıkıştırma bit hızından yüksek olduğunda daha yüksek olduğu gösterilmiştir. AMR kodlayıcı ile iki kez sıkıştırılmış ses tespiti için derin öğrenme yaklaşımını kullanmaya yönelik ilk girişimde, üç farklı derin sinir ağı (deep neural network-DNN) mimarisi kullanılmıştır [22]. Kullanılan DNN mimarileri iki gizli katmana sahip çok katmanlı algılayıcı (multi-layer perceptron-MLP), yığın otokodlayıcı (stacked auto-encoder-SAE) ve seyreltme katmanlarına sahip MLP yöntemleridir [22]. Söz konusu çalışmada, her 1 saniye uzunluğundaki ses sinyali 400 örnekten oluşan bibiri ile örtüşmeyen kısa çerçevelere bölünmüş ve daha sonra bu ham ses çerçeveleri DNN mimarilerine giriş olarak uygulanmıştır. Test aşamasında karar için oy çokluğu stratejisi kullanılmıştır. Aynı yazarlar iki kez sıkıştırılmış AMR dosyalarının sınıflandırılmasında öznitelik çıkarımı için SAE derin sinir ağı mimarisini ve sınıflandırma yöntemi olarak Gauss karışım modeli-evrensel arka plan modelini (GMM-UBM) sınıflandırıcısı kullanmayı önermiştir [23]. Önceki çalışmalarına benzer şekilde, SAE girişi 400 örnekten oluşan ham ses çerçeveleri olup, SAE gizli katmanlarından çıkarılan öznitelik vektörleri GMM-UBM sınıflandırıcısı ile modellenmiştir. Daha yakın tarihli bir çalışmada, doğrusal öngörü (linear prediction-LP) analizinden çıkarılan istatistiksel öznitelikler, AMR kodlayıcı ile iki kez sıkıştırılmış ses sinyallerinin tespit edilmesi için SVM sınıflandırıcı ile birlikte kullanılmıştır [24]. İki kez sıkıştırılmış AMR ses tespit üzerine

yoğunlaşan bir diğer çalışmada, araştırmacılar önce yedi AMR kodlayıcı parametresini (örn. LP katsayıları, çizgi spektral çiftleri, perde kazancı, perde gecikmesi vb.) çıkarmış ve ardından bu kodlayıcı parametrelerinden 657 adet istatistiksel öznitelik hesaplamıştır [25]. Son olarak öznitelik seçme yöntemi ile öznitelik sayısı azaltılmış ve tespit için SVM sınıflandırıcı kullanılmıştır.

Yukarıda belirtilen literatür özetinden de görüldüğü üzere iki kez sıkıştırılmış ses sinyallerinin otomatik olarak tespit edilmesi problemine odaklanan çalışma sayısı oldukça kısıtlıdır. Üstelik mevcut kısıtlı çalışmaların büyük çoğunluğunda MP3 kodlayıcı ile sıkıştırılmış ses sinyalleri kullanılmıştır. Az sayıda da olsa dar bant AMR kodlayıcı kullanan (örnekleme frekansı 8 kHz) çalışmaların büyük bölümünde AMR kodlayıcı parametreleri (LP katsayıları ve bunlardan türetilen istatistiksel özellikler) öznitelik olarak kullanılmıştır. Bu tezde literatürden farklı olarak yapılan çalışmalar ve tez çalışmasının özgün değerini ortaya koyan başlıca yenilikler şu şekilde özetlenebilir:

1. Dar bant AMR kodlayıcı ile iki kez sıkıştırılmış ses sinyallerinin tespiti için uçtan-uca evrişimsel sinir ağlarının (ESA) kullanılması ve oldukça yüksek başarım değerlerinin elde edilmesi [26].
2. Literatürde yer alan çalışmalardan farklı olarak öznitelik çıkarma aşamasında AMR kodlayıcının kullandığı teknikler hakkında herhangi bir ön bilgiye sahip olmadan elde edilen spektrogram öznitelikleri iki kez sıkıştırılmış ses sinyallerinin tespitinde ilk kez bu tezde kullanılmıştır [26].
3. Derin evrişimsel sinir ağlarının iki kez sıkıştırılmış ses sinyallerinin tespitinde öznitelik çıkarma modülü olarak kullanılması ilk kez bu tez kapsamında yürütülen çalışmalarda önerilmiştir [26].
4. İki kez sıkıştırılmış AMR sinyallerinin tespitine kanal (mikrofon), ortam ve söylenen söz öbeğinin etkisi ilk kez bu tezde analiz edilmiştir [26].
5. İki kez sıkıştırılmış sinyallerin tespit edilmesi çalışmalarında, çapraz veri kümeleri ile değerlendirme (cross-database evaluation) yani sistemin eğitildiği veri kümesinden tamamen farklı bir veri kümesi ile test edilmesi ilk kez bu tez çalışmasında analiz edilmiştir [26,27].
6. Uzun dönem ortalama spektrogram (long-term average spectrum – LTAS) öznitelikleri ilk kez iki kez sıkıştırılmış seslerin tespiti için bu tez kapsamında kullanılmış olup, LTAS özniteliklerinin zamansal olarak bölütlemenin performansı artırdığı gösterilmiştir [27].

7. Klasik DNN çalışmalarında sınfılandırma katmanında kullanılan softmax aktivasyonuna açısal terimlerin eklenmesi ile iki kez sıkıştırılmış seslerin tespit edilebilme başarısının artırıldığı ilk kez bu tez çalışmasında gösterilmiştir [28].
8. Literatürde henüz çalışılmamış olan geniş bant AMR kodlayıcı (örnekleme frekansı 16 kHz) ile sıkıştırılmış seslerin tespit edlmesi problemi, bu problem için paralel DNN yapılarının birlikte kullanılması ve farklı derin özniteliklere bilgi birleştirme (information fusion) yöntemleri uygulanması ilk kez bu tezde çalışılmıştır.

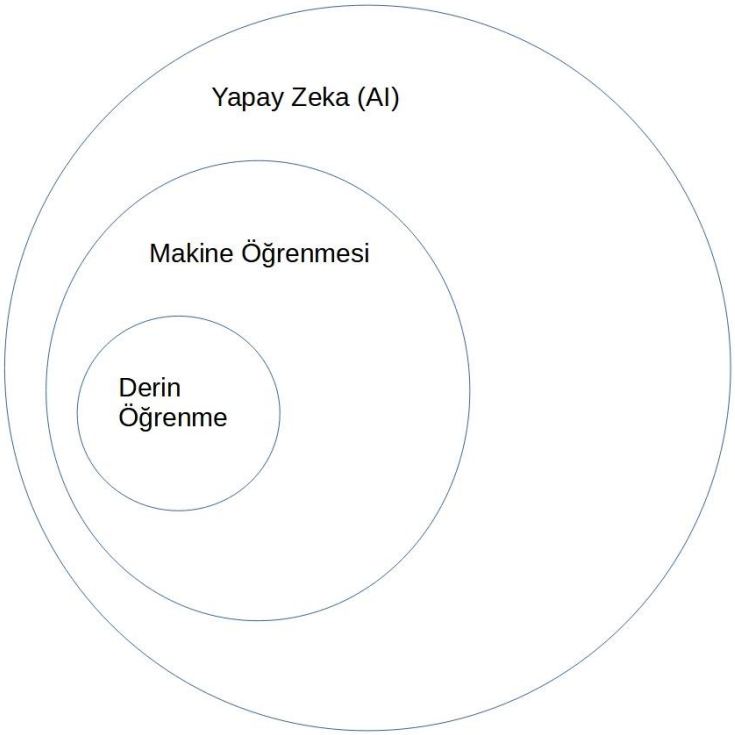
2. MATERYAL ve METOT

2.1 Derin Öğrenme

Derin yapay sinir ağları (Deep Neural Networks-DNN), günümüzde birçok modern yapay zeka uygulamasının temelini oluşturmaktadır [29]. DNN'lerin konuşma tanımaya [30] ve görüntü tanımaya [31] yapılan devrim niteliğinde uygulamasından bu yana, DNN'leri kullanan uygulama sayısı giderek artmaktadır. Derin yapay sinir ağları otonom araçlardan [32], kanser tespitine [33] ve karmaşık oyunlar oynayan [34] birçok alanda kullanılmaktadır. Bu alanlarda birçok DNN, insan doğruluğunu aşabilmektedir. DNN'lerin üstün performansının en önemli sebebi istatistiksel öğrenme yöntemleri kullanarak ham verilerden yüksek seviyeli öznitelikleri öğrenebilmesidir. Bu, uzmanlar tarafından el yordamı ile elde edilen öznitelikler ve kurallar gibi daha önceki yıllarda geleneksel olarak uygulanan yaklaşımlardan farklıdır.

DNN'ler, aynı zamanda derin öğrenme olarak da adlandırılmakta olup, 1950'lerde terimi kullanan bilgisayar bilimcisi John McCarthy tarafından tanımlanan yapay zeka alanının geniş bir parçasıdır. Yapay zeka, insanlar gibi hedeflere ulaşabilen akıllı makinelere sahip olma bilimi ve mühendisliği olarak tanımlanmaktadır. Derin öğrenmenin yapay zeka bütünüyle ilişkisi, Şekil 2.1'de gösterilmektedir. Yapay zekanın içinde, 1959'da Arthur Samuel tarafından tanımlanan olan makine öğrenmesi alt alanı bulunur ve Bu, bilgisayarların programlanmadan kendi kendine öğrenme yeteneğine sahip olma alanı olarak tanımlamaktadır.

Bilim insanları hala beynin nasıl çalıştığı hakkında ayrıntıları keşfetmeye devam ederken, genellikle beyinin ana bilgisayar elemanının nöron olduğu kabul edilir. Ortalama bir insan beyninde yaklaşık 86 milyar nöron vardır. Nöronlar kendilerine giren ve onları terk eden elemanlara sahip dendritlerle birbirlerine bağlıdır. Nöron, dendritler aracılığıyla giriş sinyallerini kabul eder, bu sinyaller üzerinde birtakım hesaplamalar yapar ve axonda bir çıkış sinyali üretir. Bu giriş ve çıkış sinyalleri aktivasyonlar olarak adlandırılır. Bir nöronun axonu ayrılır ve çok sayıda başka nöronların dendritlerine bağlanır. Axonun bir dalı ve dendrit arasındaki bağlantı sinaps olarak adlandırılır. Ortalama bir insan beyninde 10^{14} ile 10^{15} arasında sinaps olduğu tahmin edilmektedir.



Şekil 2.1 : Yapay zeka ve derin öğrenme arasındaki ilişki.

Sinapsın bir önemli özelliği üstünden geçen sinyali (x_i) ölçekleyebilmesidir. Ölçeklendirme faktörü ağırlık (w_i) olarak adlandırılabilir ve beyinin öğrendiğine inanılan şey sinapslarla ilişkili ağırlıkların değişimidir. Bu nedenle, farklı ağırlıklar farklı bir girişe cevap verir. Öğrenme, bir öğrenme uyarısına cevap olarak ağırlıkların ayarlanmasıdır, ancak beynin organizasyonu (yani programı) değişmez. Bu özellik, beyni makine öğrenimi tarzı bir algoritma için mükemmel bir ilham kaynağı yapar.

Sinir ağları, nöron'un hesaplamasının girdi değerlerinin ağırlıklı toplamını içerdiği fikrinden ilham alır. Bu ağırlıklı toplamlar sinapslar tarafından gerçekleştirilen değer ölçeğiyle ve nöron'daki bu değerlerin birleştirilmesiyle eşleşir. Ayrıca, nöron sadece bu ağırlıklı toplamı çıkartmaz, çünkü bir kaskad nöronlar ile ilişkili hesaplama basit bir lineer cebir işlemi olurdu. Bunun yerine, nöron içinde girdilerin birleştirildiği bir işlevsel işlem gerçekleşir. Bu işlem girdilerin bir eşik değerini aşması sonucu bir nöronun çıktı üretmesine neden olan bir lineer olmayan fonksiyon gibi görünür. Bu nedenle, sinir ağları girdi değerlerinin ağırlıklı toplamına lineer olmayan bir fonksiyon uygular.

Sinir ağının giriş katmanındaki nöronlar bazı değerleri alır ve bunları ağın orta katmanına, ayrıca sıklıkla "gizli katman" olarak adlandırılan katmana iletmektedir. Bir veya daha fazla gizli katmandan gelen ağırlıklı toplamlar nihayetinde çıkış katmanına iletilir ve bu, kullanıcıya ağın sonuçlarını sunmaktadır. Beyin-benzetimi terimlerini sinir ağlarıyla uyumlu hale getirmek için, nöronların çıktıları sıklıkla aktivasyon olarak, sinapslar ise ağırlık olarak adlandırılır.

$$y_j = f\left(\sum_i W_{ij} * x_i + b\right). \qquad (2.1)$$

Denklem 2.1 her bir katmanda yapılan hesaplamayı göstermektedir. W_{ij}, b, x_i ve y_j sırasıyla ağırlıklar, bias değeri, giriş aktivasyonu ve çıkış aktivasyonu olup $f(.)$ lineer olmayan fonksiyonu tanımlamaktadır.

Sinir ağları alanında, birden fazla gizli katmanı olan neural networklerin çalıştığı bir alan olan derin öğrenme bulunur. Bugün, derin öğrenmede kullanılan tipik ağ katman sayıları beşten binlere kadar değişir.

DNN'ler bir makine öğrenimi algoritmasının bir örneği olduğundan, temel program öğrenirken görevlerini yapması için değişmez. DNN'lerin özel durumunda, bu öğrenme ağdaki ağırlıkların (ve bias) değerini belirlemeyi içerir ve bu ağı eğitmek olarak adlandırılır. Bir kez eğitildikten sonra, program, eğitim süreci sırasında belirlenen ağırlıkları kullanarak ağın çıktısını hesaplayarak görevini yapabilir.

Bir ağı eğitirken, ağırlıklar (w_{ij}) genellikle gradian inişi (gradient descent) adı verilen optimizasyon süreci kullanılarak güncellenir. Kaybın her ağırlıkla ilgili gradyanın, kaybın ağırlıkla ilgili parçalı türevine göre çarpımı, ağırlığı güncellemek için kullanılır ($w_{ij}^{t+1} = w_{ij}^{t} - \alpha \frac{\partial L}{\partial w_{ij}}$, α öğrenme oranı olarak adlandırılır.) Bu gradyan kaybı azaltmak için ağırlıkların nasıl değiştirilmesi gerektiğini göstermektedir. Bu süreç iteratif olarak tekrar edilerek genel kaybı azaltır. Gradyanın parçalı türevlerini hesaplamak için geriye yayılım (backpropagation) algoritması kullanılmaktadır. Geriye yayılım algoritması zincir kuralından türetilen bir hesaplamadır ve ağın geriye doğru değerler geçirilerek, kaybın her ağırlık tarafından nasıl etkilendiğini hesaplamasıyla çalışır.

DNN'lerin yaygın bir formu, Evrişimsel Sinir Ağları (CNN) 'dır, bu ağlar birden çok evrişim katmanından oluşur. Bu ağlarda, her katman, girdi verisinin bir sonraki düzeyde yüksek seviyeli bir soyutlamasını üretir, bu soyutlama özelleştirilmiş ancak önemli bilgiyi koruyacak bir öznitelik haritası olarak adlandırılır. Modern ESA'lar, çok derin katman hiyerarşisi kullanarak üstün performans göstermeyi başarabilmektedir. ESA, görüntü işleme [31], konuşma tanıma [35], oyun oynama [34], robotik [36] gibi birçok uygulamada yaygın olarak kullanılmaktadır.

ESA'lardaki her bir evrişim katmanı yüksek boyutlu konvolüsyonlardan oluşmaktadır. Katmanda bulunan giriş aktivasyonları her birine kanal adı verilen iki boyutlu giriş öznitelik haritalarının (input feature maps) bir seti olarak yapılandırılmaktadır. Her kanal, filtrelerin yığınından herhangi bir iki boyutlu filtreyle konvolüsyon işlemi hesaplar, bu iki boyutlu filtrelerin yığını genellikle tek bir üç boyutlu filtre olarak adlandırılır. Her noktadaki konvolüsyon sonuçları, tüm kanallar arasında toplanır. Ek olarak, filtreleme sonuçlarına sabit değer (bias) eklenebilir, ancak bazı yeni ağlar [37] katmanların bazı bölümlerinde kullanımını ortadan kaldırır. Bu hesaplamanın sonucu, bir çıktı öznitelik haritasının (output feature maps) bir kanalını oluşturan çıktı aktivasyonlarıdır. Aynı girdi üzerinde ek üç boyutlu filtreler kullanılarak ek çıktı kanalları oluşturulabilir. Son olarak, birden fazla girdi öznitelik haritaları, filtre ağırlıklarının tekrar kullanımını iyileştirme potansiyeline sahip olarak bir yığın olarak işlenebilir.

ESA'lar, görüntü verilerini işleyerek, verilerdeki belirgin özellikleri bulma amacıyla tasarlanmıştır. Bu nedenle, görüntü analitik, sınıflandırma, tanıma ve benzeri problemlerde oldukça yaygın olarak kullanılmaktadır. Çalışma şekli olarak, ESA'lar öncelikle görüntü verilerini küçük bir parçaya (piksel gruplarına) ayırır ve bu parçalar üzerinde çeşitli filtreler kullanılmaktaadır. Bu filtreler görüntü verilerinin belirgin özelliklerini belirlemeyi amaçlar. Daha sonra, bu filtreler görüntüdeki özellikleri bulmak için kaydırılır ve bu işlem tekrar edilir. Bu aşamada, görüntü verilerindeki belirgin özelliklerin tanımlanması sağlanır. Daha sonra, bu özellikler, fully connected katmanları tarafından sınıflandırılır ve sonuç olarak, görüntü verilerine ait bir etiket (örneğin, "kedi", "köpek", vs.) verilir.

CNN'ler, görüntü verilerinin işlenmesi için oldukça yüksek bir performansa sahiptir ve aynı zamanda eğitimi de oldukça hızlıdır. Bu nedenle, görüntü analitik ve sınıflandırma gibi alanlarda sıklıkla tercih edilir. Örnek olarak, resim sınıflandırma,

nesne tanıma, görüntü segmentasyon, görüntü öğrenme ve benzeri alanlarda kullanılabilir.

LSTM (Long Short-Term Memory), bir tip sinir ağı modelidir ve uzun vadeli bağımlılıkları öğrenmeye ve tahmin etmeye çalışan bir model olarak tanımlanır. Bu model, birbirini izleyen verilerin (örneğin, metin, ses, finansal veriler, vb.) işlenmesinde oldukça yaygın olarak kullanılmaktadır. LSTM'in çalışma prensibi, giriş verilerinin bir sıralama içinde geçmiş verileri hatırlamasına ve bu verileri kullanmasına dayanır. Bu sayede, LSTM, zaman serisi verilerindeki uzun vadeli bağımlılıkları anlamaya ve bu bağımlılıkları tahmin etmeye çalışır.

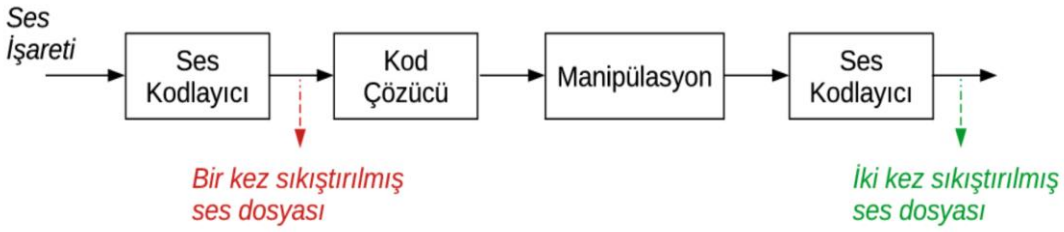
LSTM'in temel bileşeni olan hücre, verilerin sırası boyunca verilerin hatırlanmasına ve verilerin değişmesine dayanır. Bu hücre, giriş, unutma ve çıkış portları gibi bir dizi port ile donatılır ve veriler bu portlar aracılığıyla işlenir. LSTM'in uzun vadeli bağımlılıkları tahmin etme becerisi, metin veya ses verilerinin anlamlı bir şekilde sınıflandırılması, zaman serisi verilerinin tahmin edilmesi gibi uygulamalarda kullanılabilir. Örneğin, metin sınıflandırma, dil modelleme, ses tanıma ve benzeri alanlarda tercih edilmektedir.

2.2 AMR Kodlayıcı ile İki Kez Sıkıştırılmış Ses

AMR ses kodlayıcı, konuşma sinyalleri için optimize edilmiş ve 3. Nesil Ortaklık Projesi (3GPP) [11] tarafından geliştirilmiş bir konuşma sıkıştırma formatıdır. Başlangıçta, 8 kHz örnekleme hızıyla dar bant sinyalleri kodlayan bir dar bant konuşma kodlayıcı olarak geliştirilmiştir. AMR kodlayıcı, konuşma sinyalini sekiz farklı bit hızında (BR) kodlar: $BR \in \{4.75, 5.15, 5.9, 6.7, 7.4, 7.95, 10.2, 12.2\}$ kbps (kbit/s). Konuşma sıkıştırma işlemi her biri 160 örnekten oluşan konuşma çerçevelerine kod uyarımlı doğrusal öngörü (code excited linear prediction-CELP) modeli kullanarak gerçekleştirilir. CELP model parametreleri, 10. dereceden doğrusal öngörü (linear prediction-LP) analizi kullanılarak elde edilir ve LP katsayıları, çizgi spektral çiftlerine (line spectral pairs-LSP) dönüştürülür. Kodlama sırasında CELP model parametreleri kodlanır ve iletilir. Kod çözme aşamasında, önce kodlanan parametrelerin kodu çözülür ve ardından yeniden oluşturulan uyartım (excitation) sinyaline LP sentez filtresi uygulanarak konuşma sinyali sentezlenir [38].

Benzer şekilde geniş bant AMR (AMR-WB) kodlayıcı [39], 3GPP tarafından geliştirilen ve geniş bant konuşma sinyalleri için optimize edilmiş ses sıkıştırma formatıdır. 16 kHz örnekleme frekansını desteklemektedir. Cebirsel kod uyartımlı doğrusal öngörü analizi kullanarak kodlama yapmaktadır. Geniş bant AMR kodlayıcı konuşma sinyallerini dokuz faklı bit hızında ($BR \in$ $6.60, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, 23.85$ kbps) kodlamaktadır.

Şekil 2.2'de iki kez sıkıştırılmış AMR ses dosyası oluşturmanın olası bir yolu gösterilmektedir. Orijinal konuşma dalga biçimi, AMR kodlayıcı kullanılarak kodlanır, böylece bir kez sıkıştırılmış AMR ses kaydı oluşturulur. Bir saldırgan veya dolandırıcı, bir kez sıkıştırılmış AMR dosyası üzerinde oynamayı veya manipüle etmeyi hedefliyorsa, darbe kodu modülasyonu (PCM) ses dalga biçimini elde etmek için önce bir kez sıkıştırılmış AMR sinyali kod çözücü yardımı ile orijinal PCM dalga formuna dönüştürülür. PCM ses dalga formu elde edildikten sonra, herhangi bir ses düzenleme yazılımı yardımı ile konuşma sinyali üzerinde oynama (örneğin, dalga biçimine bir ses segmenti eklemek/silmek) veya manipülasyon (örneğin, konuşmacının cinsiyetini değiştirmek veya konuşma içeriğini bastırmak için dalga biçimine gürültü eklemek) işlemleri gerçekleştirebilir. Son olarak, kurcalamayı/manipülasyonu maskelemek için manipüle edilmiş ses sinyalini AMR kodlayıcı ile yeniden sıkıştırarak iki kez sıkıştırılmış AMR sinyali üretilmiş olur. Bu nedenle, orijinal/gerçek bir AMR sinyali, bir dolandırıcı veya sahtekar gerçekleştirdiği oynama veya manipülasyonun ayak izlerini gizlemek için ikinci kez yeniden sıkıştırmadığı sürece iki kez sıkıştırılmak zorunda değildir. Bu gerçeğe dayanarak, iki kez sıkıştırılmış AMR kaydının tespiti, temel ve önemli bir adli ses uygulamasıdır.



Şekil 2.2 : Ses sinyallerinden iki kez sıkıştırılmış ses sinyallerinin oluşturulma adımları [40].

2.3 Konuşma Öznitelikleri

Spektrogram, bir ses veya konuşma sinyalindeki gizli bilgileri ortaya çıkarmak için kullanılan önemli bir görselleştirme aracı olup; konuşma işleme uygulamalarında yaygın olarak kullanılmaktadır [41]. Konuşma sinyalinin spektral içeriğinin zaman içindeki değişimini gösteren önemli bir analiz yöntemi olan spektrogram gösterimi, sinyalin kısa dönem analizi ile elde edilir. Kısa dönem analizinde, bir konuşma sinyali birbiri ile örtüşen ve her biri toplam N adet örnekten oluşan kısa çerçevelere ($x[n,t]$) bölünür. Burada $n = 0, 1, \ldots, N-1$ örnek indeksidir ve $t = 1, 2, \ldots, T$ çerçeve indeksidir. Her çerçeve daha sonra bir veri inceltme penceresi (genellikle Hamming veya Hanning penceresi) $w[n]$ kullanılarak pencerelenir. Son olarak, her çerçevenin spektrumunu elde etmek için ayrık Fourier dönüşümü (DFT) alınır:

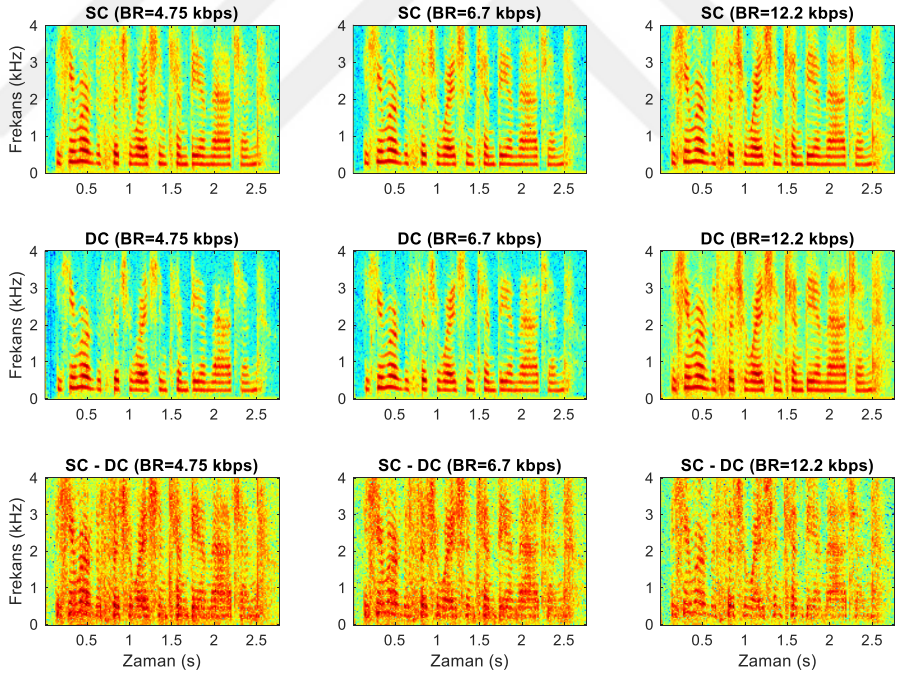
$$X(k,t) = \sum_{n=0}^{N-1} w[n]x[n,t]e^{-j2nk/N} \qquad (2.2)$$

burada k ayrık frekans indeksi olup $x[n,t]$ ses çerçevesinin $[0, N-1]$ aralığının dışında sıfır olduğu varsayılır. Kısa dönem Fourier dönüşümü (STFT) ile elde edilen tüm çerçevelerin logaritmik güç spektrumu ($10log|X(k,t)|^2$), ses sinyalinin spektrogramı olarak bilinir ve bir ısı haritası ile görselleştirilir.

Bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR konuşma sinyalleri arasındaki farkları anlamak için Şekil 2.3'te aynı sinyalin farklı sıkıştırma bit hızları ile sıkıştırılması neticesinde elde edilen sinyallerin spektrogramları analiz edip karşılaştırılmıştır. Şeklin sütunları sırasıyla 4.75, 6.7 ve 12.2 kbps sıkıştırma bit hızı değerlerine karşılık gelmektedir. Şeklin son satırı, her bir sıkıştırma bit hızı için iki kez sıkıştırmadan en çok etkilenen frekans bölgelerini ortaya çıkarmak amacı ile bir kez sıkıştırılmış ve iki kez sıkıştırılmış spektrogramlarının çıkarılmasıyla elde edilen diferansiyel spektrogramları göstermektedir. Şekilden, bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR sinyalleri arasında önemli farklılıkların bit hızından bağımsız olarak yüksek frekanslarda meydana geldiği görülebilir. Bir kez sıkıştırılmış AMR ses dosyaları, yüksek frekanslarda (yaklaşık 2 kHz'in üzerinde) iki kez sıkıştırılmış AMR dosyalarından çok daha yüksek enerji değişimine sahiptir. Bu, diferansiyel spektrogram görüntülerinden (şeklin üçüncü satırı) kolayca doğrulanabilir. Formant

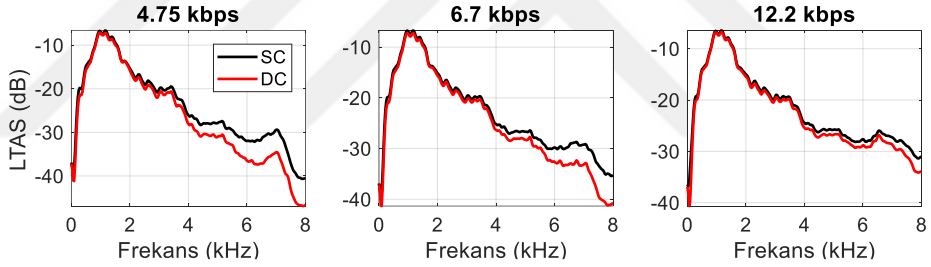
desenleri diferansiyel spektrogramlarda hala görünürken, yaklaşık 2 kHz'in üzerinde gürültü benzeri değişimler gözlemlenebilir. Bir başka ilginç gözlem de, sıkıştırma bit hızı arttıkça, hem bir kez sıkıştırılmış hem de iki kez sıkıştırılmış sinyaller için yüksek frekanslarda daha büyük enerji değişimlerinin meydana gelmesidir. Diğer bit hızı değerleri için de benzer gözlemler geçerlidir. Bununla birlikte, tek bir konuşma sinyalinin spektrogramını görselleştirmenin, iki kez sıkıştırmanın neden olduğu akustik değişiklikler hakkında genel bir açıklama yapmak için yeterli olmayabileceği tartışılabilir. Bu amaçla, uzun dönem ortalama spektrumlar (LTAS) kullanılabilir. LTAS, adli ses biliminde [42] ve bir ses sinyalinin işitilebilirliğini ölçen konuşma anlaşılırlık indeksini hesaplamak için [43] yaygın olarak kullanılmaktadır. LTAS, tüm çerçeveler üzerinden spektrogramın zaman ortalaması alınarak hesaplanır:

$$LTAS = \frac{1}{T}\sum_{t=1}^{T} X(k,t). \qquad (2.3)$$



Şekil 2.3 : 4.75, 6.7 ve 12.2 kbps kullanılarak oluşturulmuş bir kez sıkıştırılmış ve iki kez sıkıştırılmış ses sinyallerinin spektrogramları. Son satır bir kez ve iki kez sıkıştırılmış seslerin spektrogram farklarını belirtmek için verilmiştir [26].

Üç farklı bit hızında sıkıştırılmış 100 adet ses sinyali kullanarak bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR ses dosyalarının ortalama uzun dönem ortalama spektrumları hesaplanmıştır. Her bit hızı için aynı 100 ses sinyali (aynı konuşmacılar ve aynı içerikler) kullanılmıştır. Şekil 2.4, bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR ses dosyaları kullanılarak hesaplanan spektrumları göstermektedir. Spektrogram görüntüleri üzerindeki bulgulara benzer şekilde, hem bir kez sıkıştırılmış hem de iki kez sıkıştırılmış ses gücünün bit hızından bağımsız olarak 2 kHz'in altında benzer eğilimler gösterdiği gözlemlenmektedir. $f > 2$ kHz için, bir kez sıkıştırılmış ve iki kez sıkıştırılmış ses sinyalleri arasındaki farklar artmaktadır. Sıkıştırma bit hızının arttırılması bir kez sıkıştırılmış ve iki kez sıkıştırılmış LTAS grafikleri arasındaki farkları azaltsa da, daha küçük bit hızı değerlerine kıyasla ses sinyallerinin gücü artmaktadır. Bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR ses kayıtlarının spektrogram görüntülerinde ve LTAS grafiklerinde gözlemlenen bu temel farklılıklar, spektrogramı iki kez sıkıştırılmış AMR ses tespiti için ESA ile kullanmak için potansiyel olarak iyi bir temsil haline getirir.



Şekil 2.4 : Bir kez ve iki kez sıkıştırılmış AMR ses kayıtlarında spektrum ortalamasının (LTAS) farklı bit hızlarında karşılaştırılması [26].

Konuşma sinyaline kısa dönem Fourier dönüşümü uygulanarak elde edilen $X(k,t)$ gösterimi polar formda

$$X(k,t) = |X(k,t)|e^{j\theta(k,t)} \qquad (2.4)$$

Şeklinde ifade edilebilir. Daha önce de belirtildiği gibi $|X(k,t)|$ genlik spektrumu spektrogram olarak tanımlanırken, faz bileşeni $\theta(k,t)$ ise sinyalin faz spektrumu veya daha yaygın olarak fazgram olarak tanımlanır ve genlik spektrumunun taşıdığı bilgilerden farklı bilgiler taşıdığı bilinmektedir. Bu nedenle birçok çalışmada genlik spektrumu öznitelikle birlikte kullanılarakperformans artırıcı etkisi gözlenmiştir [44].

Yaygın olarak kullanılan diğer bir spektrum tahmin yöntemi, doğrusal tahmine (LP) dayanmaktadır [xx]. LP analizinde, $s(n)$ ses sinyalinin önceki p örneğinden tahmin edilebileceği varsayılır, $\check{s}(n) = -\sum_{k=1}^{p} a_k s(n-k)$. Burada $s(n)$ orijinal konuşma örneğidir, $\check{s}(n)$ tahmin edilen örnektir ve p tahmin sırasıdır (zaman aralığı). Geleneksel otokorelasyon yöntemi genel olarak enerjiyi en aza indirerek $e(n) = s(n) - \check{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k)$ tahmin katsayılarını $\{\propto_k\}_{k=1}^{p}$ tahmin etmek için kullanılır. Optimum katsayılar

$$a_{opt}^{lp} = -R_{lp}^{-1} r_{lp} \qquad (2.5)$$

denkleminden elde edilir. Burada R_{lp} bir Toeplitz otokorelasyon matrisidir ve r_{lp} bir otokorelasyon vektörüdür. Tahmin katsayıları a_k verildiğinde LP spektrumu şu şekilde elde edilir:

$$S_{LP}(f) = \frac{1}{\left|1 + \sum_{k=1}^{p} a_k e^{-j2\pi f k}\right|^2} \qquad (2.6)$$

2.4 İki Kez Sıkıştırılmış AMR Ses Tespiti

Bir s konuşma sinyali verildiğinde, iki kez sıkıştırılmış AMR ses tespiti, iki hipotez arasında karar vermeyi amaçlayan bir hipotez testi olarak gerçekleştirilebilir:

- H_0 : s sinyali bir kez sıkıştırılmış AMR sinyalidir.
- H_1: s sinyali iki kez sıkıştırılmış AMR sinyalidir.

H_0 ve H_1 hipotezleri arasında karar vermek için minimum hata oranını garanti eden optimum karar kuralı, $p(H_0|s)$ ve $p(H_1|s)$ sonsal olasılıklarına dayanan Bayes karar kuralıdır ve şu şekilde tanımlanır:

$$Karar = \begin{cases} H_0, & p(H_0|s) > p(H_1|s) \\ H_1, & p(H_1|s) > p(H_0|s) \end{cases} \qquad (2.7)$$

Bu nedenle, iki kez sıkıştırılmış konuşma sinyallerini tespit etmek için bir sistem tasarlarken belirlenmesi gereken önemli bileşenlerden biri belirli bir ses sinyali s için $p(H_0|s)$ ve $p(H_1|s)$ sonsal olasılıklarının hesaplanabileceği için güvenilir bir teknik

seçmektir. Denklem 2.7 ile belirtilen karar kuralı, Bayes kuralını takip ederek olabilirlik fonksiyonları ($p(s|H_0)$ ve p$(s|H_1)$) türünden ifade edilebilir ($p\big((H_0|s)\big) = p(s|H_0)p(H_0)$). İstatistiksel örüntü tanıma yöntemleri, her sınıfın (bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR ses sınıfları) eğitim verilerini kullanarak olasılık yoğunluk fonksiyonlarını tahmin etmeyi amaçlar. Bunun için öncelikle bir kez sıkıştırılmış ve iki kez sıkıştırılmış ses sinyallerini ayırt edebilen, uygun bir öznitelik çıkarma yönteminin belirlenmesi gerekmektedir. Daha sonra, olabilirlik fonksiyonlarını tahmin etmek için etkili bir istatistiksel modelleme tekniği gereklidir. Açıkçası, bu iki bileşen (öznitelik çıkarma ve modelleme teknikleri) için kullanılacak yöntemlerin belirlenmesi, iki kez sıkıştırılmış ses tespiti probleminin en önemli aşamaları olup zorlu bir görevdir. Herhangi bir varsayım olmaksızın el yordamı ile iki sınıfı birbirinden ayırt edebilecek öznitelikleri çıkarmak için uygun bir öznitelik çıkarma tekniği aramak ve her sınıf için olabilirlik fonksiyonlarını tahmin etmek için istatistiksel modelleme yöntemi belirlemek yerine, her iki amaç için de (öznitelik çıkarımı ve sonsal olasılıkların belirlenmesi) derin öğrenme yöntemleri kullanılabilir. Makine öğrenimindeki son gelişmelerle birlikte ESA, çeşitli örüntü tanıma görevleri için son teknoloji bir yöntem (state-of-the-art) haline gelmiştir. ESA tabanlı teknikler genellikle öznitelik çıkarma veya sınıflandırma için kullanılmaktadır. Öznitelik çıkarımı için, genellikle ham veriler veya verilerin basit iki boyutlu temsili ESA'ya giriş olarak uygulanır ve ara katmanların çıkışı, giriş verilerinin öznitelik temsili olarak elde edilir. Uçtan uca sınıflandırmaya gelince, ESA giriş verilerini alır ve uygulanan giriş sinyali için tahmin edilen sınıf etiketini döndürür. ESA her iki durum için de (öznitelik çıkarma ve sınıflandırma) güçlü bir araçtır. Çünkü görev hakkında herhangi bir önceden bilgi sahibi olmadan ayırt edici öznitelikleri otomatik olarak öğrenme yeteneğine sahiptir.

İki kez sıkıştırılmış AMR ses tespiti probleminde ESA kullanmak için, giriş sinyali iki boyutlu bir dizi olarak düzenlenmelidir. Ses sinyalleri tek boyutlu diziler olduğundan, problemin kısıtlamalarına göre sinyalin uygun bir iki boyutlu gösterimi seçilmelidir. Bu nedenle, bir ses sinyalinin spektrogram gösterimi bu amaç için iyi bir adaydır ve konuşmacı tanıma [45] ve konuşma tanıma [46] gibi ESA'ları kullanan çeşitli konuşma işleme uygulamalarında yaygın olarak kullanılmıştır. Bununla birlikte, spektrogram ses sinyalinin pencerelenmiş çerçevelerinden hesaplanır. Bu nedenle, spektrogram bir görüntü olarak ele alınırsa, satırlar ayrık frekans indeksine ve sütunlar çerçeve indeksine karşılık gelmektedir. Farklı sürelerdeki ses sinyallerinden elde edilen toplam

çerçeve sayısı doğal olarak farklı olacaktır. Bu nedenle, ses sinyalinin süresi değiştikçe spektrogram görüntüsünün sütun sayısı da değişecektir. Spektrogramın zaman-frekans şeklini sabitlemek için, genellikle bir ses sinyalinin toplam çerçeve sayısı bir üst sınıra sabitlenir [47, 48]. Bu, spektrogramın zaman ekseni boyunca (eksen çerçevelere karşılık gelir) sabit bir boyutta kesilmesi veya çerçeve sayılarını eşleştirmek için üst sınırdan daha az çerçeve sayısına sahip kısa ses dosyalarının çerçevelerinin birleştirilmesiyle gerçekleştirilir.

2.5 Kayıp Fonksiyonları

Derin öğrenmenin kullanıldığı çoğu sınıflandırma probleminde, genellikle giriş eğitim örneğinin sınıf etiketini tahmin etmek için softmax kaybı kullanılır. İki kez sıkıştırılmış AMR ses tespitinde olduğu gibi iki sınıflı bir sınıflandırma görevi için softmax kaybı şu şekilde tanımlanır:

$$L = -\frac{1}{K}\sum_{i=1}^{K}\log\left(\frac{e^{W_{yi}^{T}x_i+b_{yi}}}{\sum_{j=1}^{2}e^{W_j^{T}x_i+b_j}}\right) \qquad (2.8)$$

burada K, bir eğitim grubundaki (yığın – batch) eğitim örneklerinin sayısı olup, $x_i \in R^d$, d–boyutlu derin öznitelik vektörünü temsil etmektedir. y_i i. sınıf etiketini (iki sınıflı sınıflandırma için $y_i \in \{1,2\}$), b_j j. çıkış biriminin yanlılık terimini (bias) ve W_j sınıflandırma katmanının ağırlık matrisinin ($W \in R^{d\times 2}$) j. sütununu belirtir. Softmax kaybındaki logit terimi ($W^T x$), $W^T x = \| W \| \| x \| cos(\theta)$ şeklinde yazılabilir; burada θ, ağırlık matrisi W ile öznitelik vektörü x arasındaki açıdır. Böylece, (2.8)'de tanımlanan softmax kaybı şu şekilde yeniden yazılabilir:

$$L = -\frac{1}{K}\sum_{i=1}^{K}\log\left(\frac{e^{\|W_{y_i}\|\|x_i\|\cos(\theta_{yi,i})+b_{yi}}}{\sum_{j=1}^{2}e^{\|W_{yj}\|\|x_j\|\cos(\theta_{j,i})+b_j}}\right) \qquad (2.9)$$

burada $\theta_{j,i}$,W_j ve x_i arasındaki açıdır. Bu nedenle, Denklem 2.9'da gösterildiği üzere geleneksel softmax kaybı, ağırlık vektörünün normundan ($\| W_{y_i} \|$) etkilenmektedir. Tüm yanlılık terimlerinin (b_j) 0 olduğunu ve iç çarpım özelliğini kullanarak tüm $j = 1,2$ için $\| Wj \| = 1$ olacak şekilde ağırlıkların normalize edildiği varsayılırsa, değiştirilmiş softmax kayıp işlevi şu şekilde yazılabilir:

$$L = -\frac{1}{K}\sum_{i=1}^{K}\log\left(\frac{e^{\|x_i\|\cos(\theta_{yi,i})}}{\sum_{j=1}^{2} e^{\|x_i\|\cos(\theta_{j,i})}}\right) \quad (2.10)$$

Bu modifikasyonla, ağırlık vektörünün normunun etkisi ortadan kaldırılır ve bir x_i öznitelik vektörünün sonsal olasılığı, derin öznitelik vektörünün normu kullanılarak ve ağırlıklar ile derin öznitelik vektörü arasındaki açı ile belirlenir.

Açısal marjin softmax türevlerinde, kosinüs terimi, açısal sınırı softmax kaybına dahil etmek için daha sıkı bir fonksiyonla değiştirilir. A-Softmax kaybı [49], aşağıdaki gibi çarpımsal bir şekilde modifiye edilmiş softmax kaybına bir açısal marj dahil eder:

$$L = -\frac{1}{K}\sum_{i=1}^{K}\log\left(\frac{e^{\|x_i\|\psi(\theta_{yi,i})}}{e^{\|x_i\|\psi(\theta_{j,i})} + \sum_{j\neq y_i}^{2} e^{\|x_j\|\cos(\theta_{j,i})}}\right) \quad (2.11)$$

burada $\psi(\theta_{yi,i}) < \cos(\theta_{yi,i})$, şu şekilde tanımlanan parçalı monoton azalan açı fonksiyonudur:

$$\psi(\theta_{y_i,i}) = (-1)^k \cos(m\theta_{yi,i}) - 2k \quad (2.12)$$

burada $k \in [0, m-1], m \geq 1$ marjin boyutunu kontrol eden tamsayı parametresidir ve $\theta_{y_i,i} \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}]$ şeklindedir.

AM-Softmax [50], açısal marjı eklemeli (additive) bir şekilde tanıtır ve öznitelik vektörlerini ‖x‖=1 olacak şekilde normalize eder. $\psi(\theta_{y_i,i}) = cos(\theta_{y_i,i} - m)$ burada m yine kosinüs marjının büyüklüğünü kontrol eder. Ardından AM-Softmax kaybı şu şekilde ifade edilir:

$$L = -\frac{1}{K}\sum_{i=1}^{K}\log\left(\frac{e^{s(\cos(\theta_{y_i})-m))}}{e^{s(\cos(\theta_{y_i})-m))} + \sum_{j\neq y_i}^{2} e^{s\cos(\theta_{j,i})}}\right) \quad (2.13)$$

burada s, kosinüs değerini ölçeklendirmek için kullanılan hiperparametredir.

Toplam açısal marj softmax (AAM-Softmax) kaybında [51], $\psi(\theta_{y_i,i}) = cos(\theta_{y_i,i} + m)$ açı fonksiyonu kullanılır. A-Softmax ve AM-Softmax kayıplarının

aksine, AAM-Softmax ağırlık ve özmitelik vektörleri arasına açısal marj cezasını m getirir ve şu şekilde tanımlanır:

$$L = -\frac{1}{K}\sum_{i=1}^{K}\log\left(\frac{e^{s(\cos\left(\theta_{y_i}\right)+m))}}{e^{s(\cos\left(\theta_{y_i}\right)+m))} + \sum_{j\neq y_i}^{2} e^{s\cos\left(\theta_{j,i}\right)}}\right) \qquad (2.14)$$

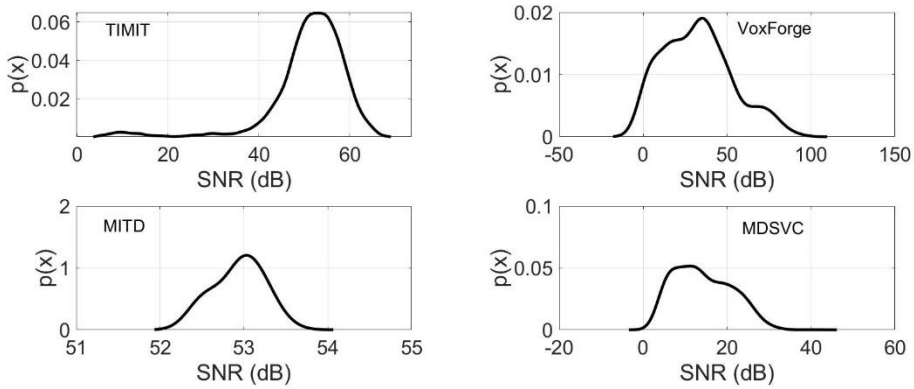
yine s, eğitim sırasında yakınsamayı kolaylaştırmak için kullanılan ölçekleme faktörüdür.

Özetle, geleneksel softmax kaybı yalnızca sınıflandırma hatasını cezalandırmaya odaklandığından, sınıflar arası örneklerin ayrılabilirliğini ve sınıf içi örneklerin kompaktlığını teşvik etmez. Toplamsal veya çarpımsal bir şekilde açısal marj cezasının getirilmesinin, Softmax kaybının [49, 50, 51] eksikliğinin üstesinden geldiği gösterilmiştir. Bu nedenle, açısal marj softmax kayıp türevleri, özellikle veritabanları arası değerlendirme deneylerinde DC AMR ses algılama performansının iyileştirilmesine sezgisel olarak yardımcı olacaktır.

2.6 Veri Kümeleri

İki kez sıkıştırılmış dar bant AMR ses tespiti deneyleri, TIMIT [52], Multicodec Invdec Tampering Dataset (MITD) [53], MIT Mobile Device Speaker Verification Corpus (MDSVC) [54] ve ücretsiz olarak temin edilebilen bir Türkçe ses veri kümesi olan VoxForge olmak üzere dört farklı veri kümesi üzerinde gerçekleştirilmiştir. Bu veri kümeleri, önerilen iki kez sıkıştırılmış AMR tespit sisteminin çeşitli koşullar altındaki performansını araştırmak için seçilmiştir. Örneğin, TIMIT veri kümesi, 630 konuşmacıdan kontrollü koşullar altında herhangi bir kanal veya ortam değişikliği olmaksızın toplanan temiz ses kayıtlarından oluşmaktadır. Bu nedenle, genel olarak ön deneyler ve analizler için TIMIT veri kümesi kullanılmıştır. MITD veri kümesi, Samsung Galaxy S4 cep telefonu kullanılarak kaydedilen iki konuşmacı arasındaki 39 ses kaydı içerir [53]. Ancak, 39 kayıttan 38'i bir orijinal kayıttan oluşturulmuştur. Orijinal kayıt, çeşitli bit hızlarına sahip dört farklı kodlayıcı (MP3, AAC, HE-AAC ve mp3PRO) kullanılarak sıkıştırıldıktan sonra kodlanmış işaretler tekrar çözülerek PCM sinyalleri elde edilmiştir. Bu nedenle, MITD veri kümesi daha önce kodlanmış ve kodu çözülmüş ses kayıtlarını içerir. MDSVC veri kümesi ise üç farklı lokasyonda iki farklı mikrofon kullanılarak toplanan 48 konuşmacının ses kayıtlarını içermektedir [54].

Bununla, farklı mikrofonların ve çeşitli konumların iki kez sıkıştırılmış AMR algılama performansının araştırılması amaçlanmıştır. Son olarak, VoxForge veri kümesi, kontrolsüz koşullar altında toplanan 90 konuşmacının konuşma kayıtlarından oluşmaktadır. Çünkü VoxForge açık bir konuşma veri kümesidir ve katılımcılar kendi kurulumlarını kullanarak (genellikle evlerinde mikrofon donanımlı kişisel bilgisayarlarını kullanarak) seslerini kaydederler. Böylece, VoxForge veri kümesi gürültülü kayıtların yanı sıra mikrofon ve ortam varyasyonlarını da içermektedir. Özetlemek gerekirse, bu dört farklı ses veri kümesini kullanmak, iki kez sıkıştırılmış AMR ses tespit performansının farklı yönlerden incelenmesine yardımcı olacaktır. Örnek olarak, Şekil 2.5 her veri kümesindeki konuşma kayıtlarının tahmini sinyal-gürültü oranlarının (signal-to-noise ratio - SNR) dağılımlarını göstermektedir. Şekilden, TIMIT ve MITD veri kümelerindeki konuşma kayıtlarının büyük çoğunluğunun nispeten yüksek SNR seviyesine sahip olduğu görülmektedir. TIMIT veri kümesinde SNR değeri 20 dB'nin altında olan kayıtlar olmasına rağmen, kayıtların çoğunda SNR değeri 40 dB'yi aşarken, MITD veri setindeki kayıtların SNR değeri 52 dB'nin üzerindedir. TIMIT ve MITD veri kümelerinin aksine, VoxForge veri kümesi çok yüksek (SNR > 70 dB) ve çok düşük (SNR < 0 dB) SNR seviyelerine sahip kayıtlardan oluşur. Ancak kayıtların çoğunda SNR değerleri [0; 50] dB aralığındadır. Son olarak, MDSVC veri kümesindeki kayıtların çoğu, diğer üç veri kümesine kıyasla en düşük SNR değerlerine (0 ile 30 dB arasındaki SNR değerleri) sahiptir. Bu nedenle, her veri kümesi farklı gürültü seviyelerine sahip kayıtlardan oluşur ve bunların iki kez sıkıştırılmış AMR ses algılama üzerindeki davranışlarını araştırmak ilginç olacaktır.



Şekil 2.5 : İki kez sıkıştırılmış AMR tespit deneylerinde kullanılan her bir veri kümesi için tahmini SNR değerlerinin dağılımı [26].

Ön deneyler dışındaki tüm veri kümeleri için deneylerde 1 saniyelik ses kaydı süresi kullanılmıştır. TIMIT veri kümesi üzerinde yapılan ön deney ve analizlerde 5 saniyelik ses kayıtları kullanılmıştır. Bir ses sinyali 5 saniyeden kısaysa, örnekler kopyalanarak içeriği 5 saniyeye uzatılmıştır. TIMIT veri kümesi için toplam 6000 adet 1 saniye uzunluğunda ses sinyali kullanılmıştır. Sinyalin süresi 1 saniyeden uzunsa, 1 s uzunluğunda ses klipleri elde etmek için sinyalin ilk 1 s uzunluğundaki parçası alınmıştır. Benzer şekilde, MDSVC veri kümesinden 1 s süreli toplam 5185 ses sinyali seçilmiştir. Sırasıyla, VoxForge ve MITD veri kümeleri için toplam ses sinyali sayısı, TIMIT ve MDSVC veri kümelerinden önemli ölçüde daha azdır. Ancak sinyallerin süreleri çok daha uzundur. Bu nedenle VoxForge ve MITD veri kümelerindeki her bir ses sinyali 1 saniye uzunluğundaki kayıtlara bölünmüştür. Bu sayede, MITD ve VoxForge veri kümeleri için sırasıyla toplam 21099 ve 4405 ses kaydı elde edilmiştir. Her veri kümesindeki her bir ses sinyali, bir kez sıkıştırılmış ses sinyallerini elde etmek için 4.75 ila 12.2 kbps arasında değişen rastgele seçilmiş sıkıştırma bit hızı (ilk sıkıştırma bit hızı -BR1) ile AMR kodlayıcı [11, 38] kullanılarak sıkıştırılarak her bit hızı için eşit sayıda bir kez sıkıştırılmış ses dosyaları oluşturulmuştur. Daha sonra bir kez sıkıştırılmış AMR ses dosyalarının kodu çözülerek iki kez sıkıştırılmış AMR ses sinyallerini oluşturmak için rastgele seçilmiş bir bit hızı (ikinci sıkıştırma bit hızı - BR2) kullanılarak yeniden sıkıştırılmış ve bu sayede, iki kez sıkıştırılmış ses sinyalleri elde edilmiştir. Her BR1-BR2 kombinasyonu için iki kez sıkıştırılmış AMR ses kayıtlarının sayısı, bit hızı sıkıştırma değerini uniform bir dağılımdan seçtiğimiz için aynıdır. Her bir veri kümesi için, ses kayıtlarının yaklaşık %25'i sistemi eğitmek için geri kalan ses dosyaları ise test için kullanılmıştır. Sınıf dengesizliği sorununu önlemek için eğitim kümesinde aynı sayıda bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR ses kaydı bulunmaktadır. MITD veri kümesi dışındaki tüm veri kümeleri için eğitim ve test kümeleri arasında herhangi bir konuşmacı veya ses kaydı örtüşmesi bulunmamaktadır. Çünkü MITD veri kümesi, iki konuşmacı arasındaki 39 adet tek kanallı konuşma ses kaydından oluştuğundan her kayıtta aynı konuşmacılar yer almaktadır. TIMIT veri kümesinde ise her konuşmacının on kaydı bulunurken, söz öbeklerinden ikisi (SA1 ve SA2 cümleleri) tüm konuşmacılar tarafından seslendirilmiştir. Bu nedenle, tespit sistemimizin bu ifadelere karşı önyargılı olmasını önlemek için SA1 ve SA2 ifadeleri eğitim kümesinden çıkarılmıştır. Bu yüzden, eğitim kümesi konuşmacı başına sekiz kayıt içerirken, TIMIT veri kümesi için test kümesinde her konuşmacı için on ses kaydı olduğu anlamına gelir. Benzer şekilde VoxForge ve

MDSVC veri kümeleri için yapılan deneylerde birbirini dışlayan eğitim ve test kümeleri kullanılmıştır.

Geniş bant AMR kodlayıcı kullanılarak oluşturulan iki kez sıkıştırılmış ses tespiti deneyleri, TIMIT veri kümesi ve sıkıştırma geçmişine sahip TIMIT veri kümesi olmak üzere iki veri kümesinde gerçekleştirilmiştir. Sıkıştırma geçmişine sahip TIMIT veri kümesi, her bir ses sinyali rasgele seçilen sıkıştırma bit hızı kullanılarak farklı kodlayıcılar (MP3, AAC ve FLAC) ile sıkıştırılarak elde edilmiştir. İki veri kümesinin eğitim ve test kümeleri aynı ses dosyalarını içerecek şekilde oluşturulmuştur.

Geniş bant AMR kodlayıcı ile iki kez sıkıştırılmış ses tespiti deneylerinde her iki veri kümesi için 1 saniye uzunluğundaki ses sinyalleri kullanılmıştır. Ses sinyalinin uzunluğu 1 saniyeden uzunsa ses sinyali 1 saniye olacak şekilde kırpılmıştır. Bir kez sıkıştırılmış AMR sinyalleri oluşturulurken veri kümesindeki her bir ses sinyali rasgele seçilen bit hızı ile sıkıştırılmıştır. Bit hızarı düzgün dağılıma sahip olup toplamda 6300 adet bir kez sıkıştırılmış AMR sinyali elde edilmiştir. Sıkıştırılmış AMR sinyallerinin kodu çözülerek PCM dalga formuna dönüştürüldükten sonra rasgele seçilen bit hızı ile sıkıştırılarak iki kez sıkıştırılmış AMR sinyalleri elde edilmiştir. Her bir BR1-BR2 kombinasyonundan eşit sayıda sinyal üretmek düzgün dağılıma sahip ikinci sıkıştırma bit hızları kullanılmıştır. Toplamda 6300 adet iki kez sıkıştırılmış AMR sinyalleri elde edilmiştir. Her bir veri kümesi için, ses kayıtlarının yaklaşık %25'i sistemi eğitmek için geri kalan ses dosyaları ise test için kullanılmıştır.

2.7 Performans Ölçütü

İki kez sıkıştırılmış AMR sesini bir kez sıkıştırılmış AMR kayıtlarından ayırmayı hedeflediğimiz için, iki kez sıkıştırılmış AMR ses sınıfı pozitif sınıf, bir kez sıkıştırılmış AMR ses sınıfı ise negatif sınıf olarak değerlendirilir. Bu nedenle sistem performansı gerçek pozitif (TP), gerçek negatif (TN), yanlış pozitif (FP) ve yanlış negatif (FN) türünden ölçülmektedir. Daha sonra tespit doğruluğu şu şekilde tanımlanır:

$$Tespit\ Oranı\ [\%] = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \qquad (2.15)$$

Deneylerde, sırasıyla bir kez sıkıştırılmış AMR ses tespiti ve iki kez sıkıştırılmış AMR ses tanıma için TNR ve TPR oranlarını rapor edilmiştir. TNR oranı (özgüllük olarak da bilinir), bir kez sıkıştırılmış bir AMR ses denemesinin bir kez sıkıştırılmış AMR ses sınıfına ait olduğunun belirlenme olasılığıdır (a posteriori) ve şu şekilde hesaplanır:

$$TNR[\%] = \frac{TN}{TN + FP} \times 100 \qquad (2.16)$$

Dolayısıyla, TNR oranı değeri ne kadar yüksekse, sistemimizin FP sonuçları üretme olasılığı o kadar düşüktür. Benzer şekilde, TPR oranı (hassasiyet olarak da bilinir), iki kez sıkıştırılmış bir AMR ses denemesinin sistem tarafından bir iki kez sıkıştırılmış AMR ses olduğuna karar verilmesi olasılığıdır ve şu şekilde hesaplanır:

$$TPR[\%] = \frac{TP}{TP + FN} \times 100 \qquad (2.17)$$

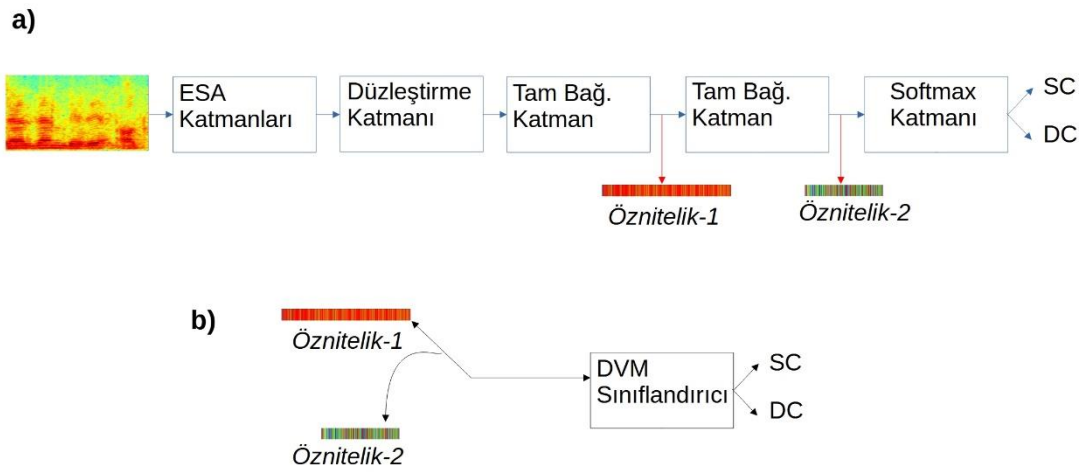
3. DENEYSEL SONUÇLAR

3.1 Dar Bant AMR Kodlayıcı ile İki Kez Sıkıştırılmış Seslerin Tespit Sonuçları

3.1.1 Derin ESA ile iki kez sıkıştırılmış AMR seslerin tespit sonuçları

Daha önce belirtildiği gibi, ön deneyler ve sistem davranışını analiz etmek amacı ile iki kez sıkıştırılmış AMR ses tespiti için TIMIT veri kümesi kullanılmıştır. Bunun için öncelikle TIMIT veri kümesinden 5 saniye uzunluğundaki ses kayıtları kullanılmıştır. Böylece 5 saniyelik her bir ses kaydından elde edilen 257×498 boyutlu spektrogram görüntüsü ESA girişine uygulanmıştır. Bu ön deneylerde her bir sıkıştırma bit hızı değeri için toplam 6000 adet bir kez sıkıştırılmış ve 6000 adet iki kez sıkıştırılmış AMR ses kaydı kullanılmıştır. İki kez sıkıştırılmış AMR ses dosyaları, aynı birinci sıkıştırma bit hızı (BR1) ve ikinci sıkıştırma bit hızı (BR2) kullanılarak oluşturulmuştur.

Şekil 3.1'de uçtan uca bir sınıflandırıcı ve öznitelik çıkarıcı olarak iki kez sıkıştırılmış AMR ses tespiti için ESA mimarisi kullanılmıştır. İki kez sıkıştırılmış AMR ses tespiti için kullanılan ayrıntılı ESA mimarisi Çizelge 3.1'de açıklanmaktadır. Uçtan uca iki kez sıkıştırılmış AMR ses tespit sistemi, giriş ses spektrogramını alır ve giriş ses sinyalinin tahmin edilen sınıf etiketini (bir kez sıkıştırılmış veya iki kez sıkıştırılmış ses) döndürür. Öznitelik çıkarıcıya gelindiğinde (Şekil 3.1(b)), giriş sesinin öznitelik temsilini çıkarmak için ESA kullanılmaktadır. Bir giriş ses spektrogramı verildiğinde, ESA'nin ara katmanlarından iki farklı derin öznitelik çıkarılır. Derin öznitelikler, (i) düzleştirme katmanının çıkışından (Öznitelik-1) ve (ii) son tamamen bağlı katmanın (Öznitelik-2) çıkışından elde edilmiştir. Bu iki derin öznitelik gösterimi daha sonra DVM sınıflandırıcı [55] kullanılarak modellenmiştir. Bu nedenle, ESA'yı iki farklı amaç için kullanmak, ESA'nın her iki görev için (özellik çıkarma veya uçtan uca sınıflandırma) gücünü ortaya çıkarmamıza ve iki kez sıkıştırılmış AMR ses tespiti üzerindeki performanslarını karşılaştırmamıza yardımcı olacaktır.



Şekil 3.1 : Önerilen iki kez sıkıştırılmış AMR ses sinyallerinin tespit sistemi. a) Uçtan uca iki kez sıkıştırılmış AMR tespit sistemi. b) ESA sisteminden elde edilen derin öznitelik vektörleri kullanılarak oluşturulan DVM sistemi. ESA sisteminin girişine ses sinyallerinin spektrogramları uygulanmıştır [26].

Deneylerde kullanılan ESA mimarisi, ağı eğitirken aşırı öğrenmeyi önlemek için her bir evrişim katmanını, bir maksimum havuzlama katmanı ve bir seyreltme katmanının izlediği dört evrişimli katmandan oluşmaktadır. Benzer şekilde, aşırı öğrenmeyi önlemek için birinci evrişim katmanından sonra bir yığın normalleştirme katmanı kullanılmıştır. Düzleştirme katmanı, son evrişimli katman çıkışını bir vektöre dönüştürür, böylece tam bağlantılı katmanlara uygulanabilmektedir. ESA mimarisinin çıkış katmanı, birinci çıkış biriminin bir kez sıkıştırılmış AMR sınıfına ve ikinci çıkış biriminin iki kez sıkıştırılmış AMR sınıfına karşılık geldiği iki birimden oluşan bir softmax katmanıdır. Evrişimli katmanlar için doğrultulmuş doğrusal birim (ReLu) aktivasyon fonksiyonu kullanılırken, tam bağlantılı katmanlarda sigmoid aktivasyon fonksiyonu kullanılmıştır. Ağın optimizasyonu için Adadelta optimizasyon algoritması kullanılmaktadır. Tüm evrişimli katmanlar için 0,25 seyreltme oranı seçilirken, tam bağlanltılı katmanlarda 0,5 seyreltme oranı kullanılmıştır. Tüm bu parametreler, çeşitli sayıda katman, çekirdek boyutu, seyreltme oranları, optimize ediciler, aktivasyon fonksiyonları ve farklı sayıda birime sahip tam bağlantılı katmanların sayısı kullanılarak ilk deneylere göre optimize edilmiştir. Böyle kapsamlı bir ilk analiz ve deneylere dayanarak, Çizelge 3.1'de özetlenen ağın en iyi iki kez sıkıştırılmış AMR ses tespit performansını sağladığı bulunmuştur. Her veri kümesi için, eğitim veya test kümelerinde yer almayan 500 adet bir kez sıkıştırılmış ve 500 adet iki kez sıkıştırılmış ses kaydından oluşan bir geliştirme kümesi, eğitim sırasında

ağ hiperparametrelerini ayarlamak ve erken durdurma eşiğini belirlemek için kullanılmıştır.

Çizelge 3.1 : Deneylerde kullanılan ESA mimarisinin her bir katman parametreleri ve detayları.

	Filtre boyutu	Çıkış boyutu	Parametre sayısı
Giriş		257 x 498	
Evrişim Katmanı	3 x 3	255 x 496 x32	320
Toplu Normalleştirme	-	255 x 496 x 32	128
Maksimum Havuzlama	2 x 2	127 x 248 x 32	-
Seyreltme	0.25	127 x 248 x 32	-
Evrişim Katmanı	3 x 3	125 x 246 x 32	9248
Maksimum Havuzlama	2 x 2	62 x 123 x 32	-
Seyreltme	0.25	62 x 123 x 32	-
Evrişim Katmanı	3 x 3	60 x 121 x 32	9248
Maksimum Havuzlama	2 x 2	30 x 60 x 32	-
Seyreltme	0.25	30 x 60 x 32	-
Evrişim Katmanı	3 x 3	28 x 58 x 32	9248
Maksimum Havuzlama	2 x 2	14 x 29 x 32	-
Seyreltme	0.25	14 x 29 x 32	-
Düzleştirme Katmanı	-	12992	-
Tam Bağlantılı Katman	-	512	6652416
Seyreltme	0.5	512	-
Tam Bağlantılı Katman	-	256	131328
Seyreltme	0.5	256	-
Softmax	-	2	514
Toplam Parametre	-	-	6812450

Ses kayıtlarının %25'i kullanılarak her bir sıkıştırma bit hızı için farklı bir ESA modeli eğitilmiştir (toplamda sekiz farklı ESA modeli eğitilmiştir). Her bir sıkıştırma bit hızı için geri kalan test sinyali için uçtan uca ESA sistemi kullanılarak elde edilen doğruluk değerleri Çizelge 3.2'de özetlenmiştir. Çizelgede, eğitim ve test bit hızları aynı olduğu duruma ait (BR1=BR2) sonuçların raporlanmasının yanı sıra (tablonun köşegen öğeleri), eğitim ve test bit hızı farklı olduğunda elde edilen sonuçlar da verilmiştir (tablonun köşegen dışı öğeleri). Sistemi test etmek için farklı sıkıştırma bit hızı değerlerinin kullanılması, daha önce görülmemiş sıkıştırma bit hızı değeri karşısında önerilen sistemin iki kez sıkıştırılmış AMR ses tespiti performansı hakkında bazı bilgiler verecektir. Tablonun son sütununda her satıra ait ortalama tespit oranları verilmiştir. Tabloda verilen sonuçlardan görüldüğü üzere, uçtan uca ESA sistemi, sıkıştırma bit hızından bağımsız olarak tüm durumlar için iki kez sıkıştırılmış AMR ses tespiti görevinde oldukça yüksek performans göstermektedir. Tespit oranları çoğu

durum için %99'un üzerindedir. Genel olarak, yüksek sıkıştırma bit hızlarında düşük bit hızlarına nazaran biraz daha yüksek tanıma doğruluğu elde edilmiştir. Eğitim ve test kayıtları arasında sıkıştırma bit hızı değerleri açısından bir uyumsuzluk olduğunda (tablonun köşegen dışı öğeleri), özellikle sistem düşük bit hızları kullanılarak sıkıştırılmış ses dosyaları ile eğitildiğinde, eğitim ve test bit hızları arasındaki fark arttıkça tanıma doğruluğu azalmaktadır. Tablonun son sütunundan da görüldüğü gibi uçtan uca sistem, 4.75 kbps durumu dışında ortalama %99'un üzerinde doğruluk değerleri vermektedir. Sistem 4.75 kbps'de sıkıştırılmış ses dosyaları kullanılarak eğitildiğinde, diğer eğitim bit hızı değerlerinden çok daha düşük olan ortalama %97,14 tespit oranı elde edilmiştir.

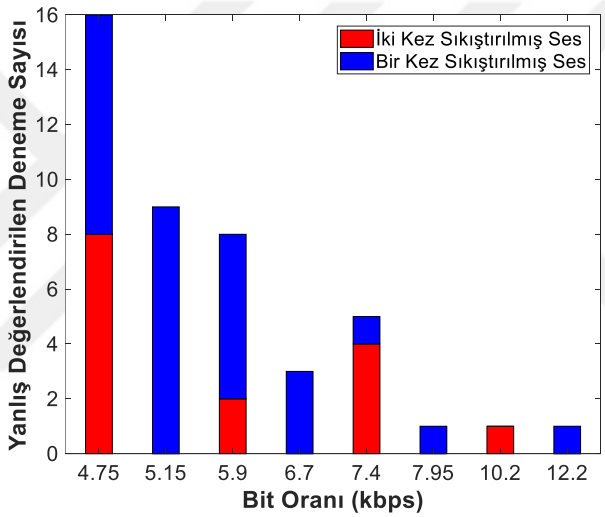
Çizelge 3.2 : Uçtan uca ESA sistemi kullanılarak elde edilen iki kez sıkıştırılmış AMR ses tespit oranları (%). Son satır eğitim kümesinde tüm bit hızlarını içeren sinyaller bulunduğunda elde edilen tespit sonuçlarını göstermektedir. Son sütun ise eğitilen her bir bit hızı için yapılan testlerin ortalamasını göstermektedir.

		Test edilen bit hızı (kbps)								
		4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2	Ort.
Eğitim bit hızı (kbps)	4.75	99.80	99.83	99.80	98.78	97.05	97.23	94.90	89.75	97.14
	5.15	98.93	99.88	99.85	99.86	99.80	99.90	99.71	99.02	99.61
	5.9	98.81	99.71	99.90	99.72	99.42	99.76	99.60	96.91	99.22
	6.7	98.95	99.95	99.98	99.96	99.96	100	99.90	99.62	99.79
	7.4	99.82	99.93	99.95	99.96	99.93	99.91	99.71	98.12	99.66
	7.95	98.63	99.85	99.95	99.88	99.90	99.98	99.85	99.95	99.74
	10.2	99.93	99.97	99.98	99.91	99.92	99.95	99.98	99.97	99.95
	12.2	99.65	99.86	99.91	99.91	99.95	99.98	99.97	99.98	99.90
	Tümü	100	100	100	100	100	100	100	100	100

Test sinyalinin sıkıştırma bit hızı değeri ile ilgili ön bilgi sağlanmadıkça, her bir sıkıştırma bit hızı değeri için farklı bir model eğitmenin makul bir yaklaşım olmadığı iddia edilebilir. Bu nedenle, herhangi bir bit hızında sıkıştırılmış ses sinyallerini tanıyabilen tek bir sistem daha çok tercih edilir. Bu amaçla, sekiz farklı bit hızında bir kez ve iki kez sıkıştırılmış AMR ses sinyallerinden oluşan tüm eğitim verilerini kullanarak tek bir ESA sistemi eğitilmiştir. Elde edilen sonuçlar Çizelge 3.2'nin son satırında özetlenmiştir. Beklendiği gibi, tüm bit hızları için tüm eğitim verilerinin bir havuzda toplanması ve tek bir modelin eğitilmesi, sıkıştırma bit hızı değerinden bağımsız olarak mükemmel tanıma doğruluğu sağlamaktadır. Bu tür bir sistemin tek dezavantajı, eğitim verisi miktarı, her bir bit hızı için ayrı ayrı bir modelin eğitilmesine göre sekiz kat daha fazla olduğu için, eğitim süresinin önemli ölçüde artmasıdır. Ancak

sistemin eğitimi, test aşamasının aksine çevrim dışı gerçekleştirildiğinden bu durum kabul edilebilir olarak değerlendirilmektedir.

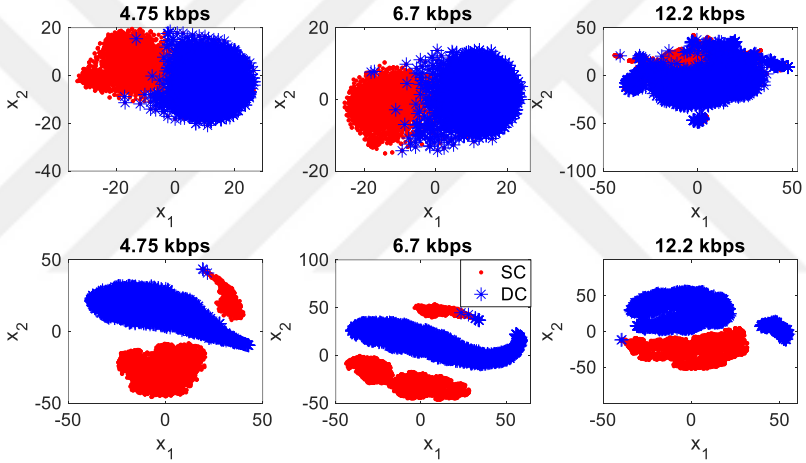
Uçtan uca ESA sistemi için her test bit hızı değeri için sistem tarafından yanlış karar verilen sınamaların sayısı Şekil 3.2'de verilmiştir. Şekil incelendiğinde, sıkıştırma bit hızı arttıkça, yanlış sınıflandırılan deneme sayısının genellikle azaldığı görülmektedir. ESA tespit sistemi, dört bit hızı (5.15, 6.7, 7.95 ve 12.2 kbps) için tüm iki kez sıkıştırılmış AMR ses dosyalarını doğru bir şekilde tespit etmektedir. 5.9 kbps durumunda, sistem yalnızca iki adet iki kez sıkıştırılmış AMR sinyali sınamasını tespit etmede başarısız olurken, 10.2 kbps için sistem yalnızca bir adet iki kez sıkıştırılmış AMR ses denemesini yanlış sınıflandırmaktadır.



Şekil 3.2 : ESA sisteminde her test bit hızı için yanlış sınıflandırılan deneme sayıları [26].

Ön deneylerden ve uçtan uca ESA sisteminin büyük başarısını gözlemledikten sonra, sistemin davranışını daha iyi anlamak için ESA mimarisinin gizli katman çıkışları analiz edilmiştir. Bu amaçla, ESA modelinin ara katmanlarından *Öznitelik-1* ve *Öznitelik-2* olarak adlandırılan iki farklı derin öznitelik vektörü çıkarılmıştır. Öznitelik-1 derin öznitelikleri ESA mimarisinin düzleştirme katmanının çıkışlarından, Öznitelik-2 vektörleri ise son gizli katman çıkışlarından elde edilmiştir. Bu derin öznitelikler daha sonra DVM sınıflandırıcı ile iki kez sıkıştırılmış AMR ses tespitinde kullanılmıştır. Elde edilen derin özniteliklerin ayırt edicilik performansını görselleştirmek için yüksek boyutlu öznitelikleri iki boyutlu uzaya taşımak amacı ile öznitelik vektörlerine t-dağılımlı stokastik komşu yerleştirme (t-SNE) doğrusal olmayan boyut azaltma yöntemi [56] uygulanmıştır. Şekil 3.3, 4.75, 6.7 ve 12.2

kbps'de bir kez ve iki kez sıkıştırılmış AMR ses dosyaları kullanılarak çıkarılan Öznitelik-1 (şeklin üst satırı) ve Öznitelik-2 (şeklin alt satırı) öznitelik vektörlerinin dağılım grafiğini göstermektedir. Dağılım grafikleri oluşturulurken TIMIT veri kümesindeki 6000 adet ses sinyalinin tümü kullanılmıştır. Şekildeki dağılım grafikleri, iki ses sınıfının (bir kez ve iki kez sıkıştırılmış ses sınıfları) her iki öznitelik gösterimi için kolaylıkla ayrıştığını göstermektedir. Ses sinyalleri 12.2 kbps'de sıkıştırılıp düzleştirme katmanından (Öznitelik-1) elde edilen derin öznitelikler kullanıldığında iki sınıf özniteliklerinin büyük oranda örtüştüğü görülmektedir. İki derin öznitelik arasındaki görsel farklılığın muhtemel sebebi, Öznitelik-1 için 12992 boyutlu vektörleri iki boyutlu uzaya indirgenmesi, ikinci durumda ise (Öznitelik-2) 256 boyutlu öznitelik vektörünün iki boyutlu uzaya indirgenmesi gerçeğinden kaynaklanmaktadır.



Şekil 3.3 : 4.75 kbps (ilk sütun), 6.7 kbps (ikinci sütun) ve 12.2 kbps (üçüncü sütun) ile sıkıştırılmış bir kez sıkıştırılmış (SC) ve iki kez sıkıştırılmış (DC) ses dosyalarından elde edilen Öznitelik-1 (üst satır) ve Öznitelik-2 (alt satır) derin öznitelik vektörlerinin dağılımları [26].

Şekil 3.3'te yapılan gözlemler ve çıkarılan derin özniteliklerin iki kez sıkıştırılmış sinyal sınıfını ayırt edici özelliklerinin yüksek olması nedeni ile TIMIT veri kümesi kullanılarak 5 saniye uzunluğundaki ses kayıtları üzerinde yapılan son çalışmada, ESA modelinden elde edilen derin öznitelik vektörleri DVM sınıflandırıcısı ile birlikte kullanılmıştır. ESA mimarisinden çıkarılan iki farklı derin öznitelik gösterimini kullanan DVM deneylerinde, üç farklı çekirdek fonksiyonu (doğrusal, radyal tabanlı işlev (RBF) ve sigmoid çekirdekler) iki kez sıkıştırılmış AMR ses tespit performansını karşılaştırmak için kullanılmıştır. LIBSVM paketi [57], SVM eğitimi ve testi için

kullanılmıştır. Üç farklı çekirdek fonksiyonu kullanan DVM sınıflandırıcı ile elde edilen sonuçlar Çizelge 3.3'te verilmiştir. Çizelgenin son sütunu, her bir çekirdek fonksiyonu için tüm bit hızları üzerinden hesaplanan ortalama tespit oranını göstermektedir.

Çizelge 3.3 : ESA sisteminden elde edilen derin öznitelikler ve DVM sınıflandırıcı kullanılarak elde edilen iki kez sıkıştırılmış AMR ses tespit oranları (%). Öznitelik-1 vektörü düzleştirme katmanından ve Öznitelik-2 son tam bağlntılı katmandan elde edilmiştir.

		Test edilen bit hızı (kbps)								
		4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2	Ort.
Öznitelik-1	Doğrusal	100	99.98	99.92	99.98	99.95	100	100	100	99.98
	RBF	97.25	99.97	99.92	99.98	99.96	100	99.92	100	99.62
	Sigmoid	93.80	99.98	99.92	99.98	99.95	100	100	100	99.20
Öznitelik-2	Doğrusal	99.80	99.95	99.91	99.97	99.95	99.98	99.98	100	99.94
	RBF	99.78	99.91	99.90	100	99.93	99.98	100	100	99.94
	Sigmoid	99.80	99.88	99.90	99.98	99.93	99.98	100	100	99.93

Derin öznitelikler kullanan DVM sınıflandırıcı, test kayıtlarını sıkıştırmak için kullanılan bit hızından bağımsız olarak yüksek algılama performansı sağlamaktadır. Doğrusal çekirdek fonksiyonu çoğu durumda RBF ve Sigmoid çekirdeklerinden üstündür. Bu gözlem beklenen bir durumdur çünkü Şekil 3.3'te çoğu durumda iki sınıfın iki boyutlu uzayda bile neredeyse doğrusal olarak ayrılabilir olduğu gösterilmiştir. İki derin öznitelik temsilini (Öznitelik-1 ve Öznitelik-2) karşılaştırırken, Öznitelik-1, genel olarak 256 boyutlu tam bağlantılı katman çıkışında elde edilen öznitelliklerden (Öznitelik-2) daha iyi tespit oranları vermektedir. Bunun nedeni muhtemelen, DVM sınıflandırıcının genellikle optimum ayırıcı hiperdüzlemi kolayca bulması ve bu nedenle yüksek boyutlu özellik vektörleriyle daha iyi performans göstermesidir [58]. Çizelge 3.3'te verilen DVM sonuçları ile uçtan uca ESA sistemi sonuçları (Çizelge 3.2) karşılaştırıldığında, DVM sınıflandırıcısının ESA sisteminden biraz daha iyi performans gösterdiği görülmektedir. DVM sınıflandırıcısının üstün performansının nedeni, ESA sisteminin aslında son gizli katmandan (Öznitelik-2) çıkarılan 256 boyutlu öznitelikleri kullanarak basit softmax sınıflandırıcısını kullanması olabilir. Bununla birlikte, DVM sınıflandırıcısı, öznitelik vektörlerini bir çekirdek fonksiyonu kullanarak daha yüksek boyutlu uzaya yansıtarak ayırıcı

hiperdüzlemi bulur, böylece iki sınıfın o yüksek boyutlu çekirdek uzayında bir hiperdüzlemle ayrıştırılması daha kolay olur. Öznitelik-1 ile doğrusal çekirdek fonksiyonu (ESA'nın düzleştirme katmanından çıkarılan öznitelikler) diğer çekirdek fonksiyonlarından ve genel olarak Öznitelik-2'den üstün olduğundan, geri kalan DVM deneylerinde doğrusal çekirdek konfigürasyonuna sahip Öznitelik-1 öznitelik vektörleri kullanılacaktır.

Yapılan ön deneyler ve analizlerde, hem uçtan uca ESA sisteminin hem de ESA'nın gizli katmanından çıkarılan derin özniteliklere sahip DVM sınıflandırıcısının iki kez sıkıştırılmış AMR ses algılamada harika performans verdiğini görülmüştür. Ancak ön deneylerde 5 sn uzunluğunda ses kayıtları kullanılmış ve her sıkıştırma bit hızı için farklı bir model eğitilmiştir. İlk olarak, iki kez sıkıştırılmış AMR ses tespitini ele alan önceki çalışmalarda [22, 23], 5 saniye uzunluğundaki ses kayıtları gereksiz yere uzundur ve yalnızca 1 saniye uzunluğundaki ses dosyaları kullanılarak makul tespit oranları elde edilmiştir. İkinci olarak, her bir sıkıştırma bit hızı için ayrı bir model eğitmek alışılmadık bir durumdur ve sıkıştırma bit hızından bağımsız olarak iki kez sıkıştırılmış AMR sesini algılamak için tek bir model kullanmak daha çok tercih edilir. Bu nedenle, sonraki deneylerde 1 saniye uzunluğunda ses kayıtları kullanılmıştır ve iki kez sıkıştırılmış AMR seslerini tespit için tek bir model eğitilmiştir. Tek bir modeli eğitirken, ön deneylerde yapılanın aksine eğitim setindeki iki kez sıkıştırılmış ses dosyalarının sayısı sekiz kat arttırılmamıştır. Eğitim kümesi eşit sayıda bir kez ve iki kez sıkıştırılmış ses sinyallerinden meydana gelecek şekilde oluşturulmuştur. Ayrıca eğitim kümesinde her birinci ve ikinci sıkıştırma bit hızı kombinasyonu için eşit sayıda sinyal kullanılmıştır. Bir saniye uzunluğunda ses kayıtlarını (257×100 boyutunda bir spektrogram görüntüsü) kullanmak, 5 saniye uzunluğundaki ses kayıtlarına (257×498 boyutunda bir spektrogram görüntüsü) kıyasla giriş parametrelerinin sayısını beş kat azalttığından, daha küçük boyutlu giriş verisinde aşırı öğrenmeyi engellemek için üç evrişimli ESA katmanı ardından her biri 512 birimden oluşan iki adet tam bağlantılı katman kullanılmıştır. Geri kalan tüm parametreler (çekirdek boyutları, aktivasyon fonksiyonları vb.) Çizelge 3.1'de gösterilen model ile aynıdır. Bu bölümde TIMIT veri tabanından 1 saniye uzunluğunda ses kayıtları kullanılarak elde edilen tespit sonuçları rapor edilmektedir.

Ses kayıtlarının spektrogramları hesaplanırken, her ses sinyali 10 ms'lik çerçeve kaydırma kullanılarak 25 ms'lik çerçevelere bölünür. Ön-vurgulama işlemi uygulanan

her bir ses çerçevesi daha sonra 200 örneklik bir Hamming penceresi kullanılarak pencerelenir. Daha sonra her pencerelenmiş çerçevenin güç spektrumu daha sonra 512 noktalı DFT kullanılarak hesaplanır. Simetri özelliği nedeniyle, güç spektrumunun yalnızca ilk 257 örneği korunur. 8 kHz'de örneklenen 1 saniye uzunluğundaki bir ses kaydı için bu, 257×100'lük bir spektrogram görüntüsü verir çünkü 10 ms çerçeve kaydırma, saniyede 100 çerçeve kare hızı verir.

TIMIT veri tabanından 1 saniye uzunluğundaki ses kayıtları kullanılarak elde edilen ortalama bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR ses tespit oranları Çizelge 3.4'te verilmiştir. Bir kez sıkıştırılmış ses tespiti için, tabloda sekiz sıkıştırma bit hızı değeri (TNR oranları) için elde edilen tespit oranlarının ortalaması alınmıştır. İki kez sıkıştırılmış AMR ses kayıtlarına gelindiğinde, birinci ve ikinci sıkıştırma bit hızlarının (sırasıyla BR1 ve BR2) olası her bir kombinasyonu dikkate alınmıştır. Bu, 64 olası birinci ve ikinci sıkıştırma bit hızı konfigürasyonuna $(8 \times 8 = 64)$ karşılık gelmektedir. Çizelge 3.4'te bildirilen ortalama iki kez sıkıştırılmış ses tespit oranları, bu 64 tespit oranı değerinin (TP oranları) ortalaması alınarak hesaplanmıştır. Her bir olası BR1 ve BR2 konfigürasyonunun tespit oranını ayrı ayrı vermek yerine ortalama tespit oranlarını bildirmenin arkasındaki mantık, hem ESA hem de DVM sistemlerinin 64 farklı kombinasyonun büyük çoğunluğu için mükemmel performans (%100'lük tespit oranı) sağlamasıdır. Bu nedenle, karışıklığı önlemek için ortalama tespit oranları verilmiştir. Çizelge 3.4'ten görüldüğü gibi, hem ESA hem de DVM sistemleri, %100 ortalama bir kez sıkıştırılmış AMR ses tespit oranı sağlar; bu, her iki sistemin de tüm bit hızları için %100 TNR oranına ulaştığı anlamına gelmektedir. Bu nedenle, derin özniteliklere sahip uçtan uca ESA ve DVM sistemi, bir kez sıkıştırılmış AMR ses kayıtlarını tespit etmede çok güçlüdür. İki kez sıkıştırılmış AMR ses tespiti durumunda, DVM sistemi ortalama tespit oranı bakımından uçtan uca ESA sisteminden biraz daha iyi performans gösterse de, iki sistem arasındaki performans farkı (%99,92'ye karşı %99,94) önemsizdir. Uçtan uca ESA sistemi ile tespit oranı %100'ün altında olan sadece dört durum bulunmuştur. İlginç bir şekilde, bu dört durumun hepsinde BR1 değerinin 4.75 kbps olduğu gözlenmiştir. En düşük iki kez sıkıştırılmış AMR ses tespit oranı (%96,40), hem BR1 hem de BR2 4.75 kbps olduğunda uçtan uca sistem kullanılarak elde edilmiştir. Derin öznitelikleri kullanan DVM sistemi incelendiğinde, 64 olası BR1 ve BR2 kombinasyonu arasından 62 durum için %100 tespit oranı vermektedir. Uçtan uca ESA sistemi ile elde edilen sonuçlara

benzer şekilde BR1 ve BR2 değerleri aynı ve 4.75 kbps olduğunda en düşük iki kez sıkıştırılmış AMR tespit oranı (%97) DVM sistemi ile elde edilmektedir. Özetle hem uçtan uca ESA hem de DVM sistemleri TIMIT veri tabanından alınan 1 saniye uzunluğundaki ses kayıtlarının kullanılması durumunda dahi çok iyi performans göstermektedir.

Çizelge 3.4 : TIMIT veri kümesindeki 1 saniye uzunluğundaki ses kayıtları kullanılarak ESA ve DVM sistemlerinde bir kez sıkıştırılmış AMR ve iki kez sıkıştırılmış AMR tespit oranlarının ortalamaları (%).

	Bir Kez Sıkıştırılmış	İki Kez Sıkıştırılmış
ESA	100	99.92
DVM	100	99.94

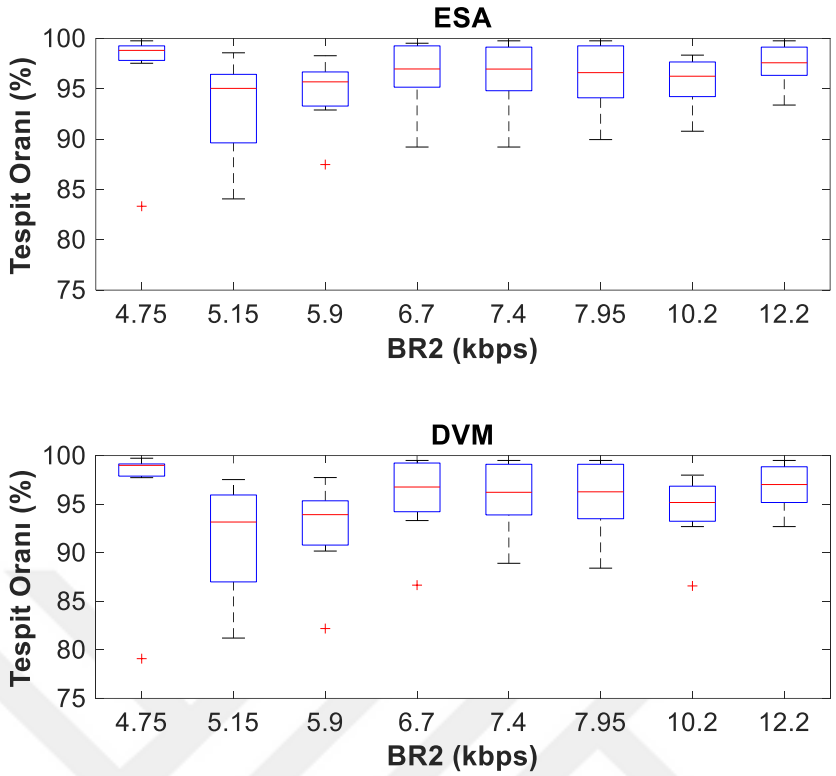
Ardından, iki kez sıkıştırılmış AMR ses tespit deneyleri MITD veri kümesi [51] üzerinde gerçekleştirilmiştir. MITD veri kümesi, çeşitli bit hızları ile farklı ses kodlayıcıları (MP3, AAC, HE-ACC ve mp3PRO) kullanılarak önceden kodlanmış ve kodu çözülmüş ses kayıtlarını içeren bir veri kümesidir. Bu nedenle, MITD veri kümesini kullanan deneyler, önceki ses manipülasyonlarının iki kez sıkıştırılmış AMR ses tespiti üzerindeki etkisini araştırmamıza yardımcı olacaktır. Her kayıt önce 1 saniye uzunluğunda ses kayıtlarına bölünmüştür ve ardından bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR ses dosyaları oluşturulmuştur. Uçtan uca ESA sistemi ve MITD veri kümesindeki derin özniteliklere sahip DVM sınıflandırıcısı kullanılarak elde edilen bir kez sıkıştırılmış AMR ses tespit oranları (TNR oranları) Çizelge 3.5'te gösterilmektedir. Sıkıştırma bit hızı arttıkça, her iki sistem de genel olarak daha iyi tespit oranı sağlamaktadır. Örneğin, uçtan uca sistem kullanılarak 4.75 kbps için %93,39 bir kez sıkıştırılmış AMR ses tespit oranı elde edilirken, sıkıştırma bit hızı 12.2 kbps olduğunda bu oranın %98,67'ye çıktığı gözlenmektedir. ESA sistemi, sıkıştırma bit hızından bağımsız olarak bir kez sıkıştırılmış AMR ses tespit görevinde DVM sınıflandırıcısından daha yüksek performans göstermektedir.

Çizelge 3.5 : MITD veri kümesinde ESA ve DVM sistemleri kullanılarak elde edilen bir kez sıkıştırılmış AMR tespit oranları (%).

	4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2
ESA	93.39	94.71	94.86	96.82	95.79	96.28	98.87	98.67
DVM	85.37	88.60	88.60	91.05	89.58	91.68	96.82	95.94

Şekil 3.4, uçtan uca ESA ve DVM sistemleri için her bir BR2 değeri başına iki kez sıkıştırılmış AMR ses algılama oranlarının (TP oranları) kutu çizimlerini karşılaştırmaktadır. Her bir kutunun yüksekliği, sabit bir BR2 değeri için 4.75 kbps ile 12.2 kbps arasında değişen sekiz farklı BR1 değeri için elde edilen iki kez sıkıştırılmış AMR tespit oranlarının varyasyonları ile belirlenir. Her kutudaki kırmızı çizgiler, sekiz tespit oranının örnek ortancasını temsil ederken, bir kutunun üst ve alt kenarları tespit oranlarının sırasıyla 75. ve 25. yüzdelik dilimlerine karşılık gelir. Şekilden, uçtan uca ESA sisteminin tüm BR2 değerleri için sistematik olarak derin özniteliklerle DVM sınıflandırıcısından daha iyi performans gösterdiğini görebiliriz. İlginç bir şekilde, her iki sistem için de algılama oranları, BR2 4.75'ten 10.2 kbps'ye yükselirken küçük farklılıklarla benzer eğilimler göstermektedir. Ancak, 10.2 ve 12.2 kbps için tespit oranlarında daha büyük farklılıklar gözlenmektedir. BR2'nin 10.2 kbps'ye ulaşmasıyla tespit doğruluğu BR1'in değerinden bağımsız olarak büyük ölçüde düşmektedir. BR1 10.2 veya 12.2 kbps olduğunda, göreceli performans düşüşü diğer BR1 değerlerinden çok daha fazladır. Örneğin BR1 10.2 kbps seçildiğinde uçtan uca ESA sisteminin elde ettiği tespit oranı %94,46'dan %80,19'a düşerken BR2 7.95 kbps'den 10.2 kbps'ye yükselmektedir. Ancak BR1 4.75 kbps olduğunda 7.95 ve 10.2 kbps BR2 değerleri için sırasıyla %99,19 ve %98,16 tespit oranları elde edilmiştir. Uçtan uca ESA sistemine benzer şekilde, BR1 değeri 10.2 kbps'den küçük olduğunda, BR2 değerinden bağımsız olarak DVM sınıflandırıcısı kullanılarak daha iyi tanıma oranları elde edilmektedir. Yine BR2 10.2 kbps'ye ulaştığında, uçtan uca ESA sisteminde olduğu gibi algılama oranları önemli ölçüde düşmektedir.

[23] çalışmasında yazarlar SAE+GMM sisteminin MITD veri setinde ortalama %92,30 tespit oranı verdiğini bildirmişlerdir. Bununla birlikte, bu tez çalışmasında önerilen uçtan uca ESA sistemi ile SAE +GMM sisteminden daha iyi performans gösteren %97,41 ortalama tanıma oranına ulaşılmıştır.



Şekil 3.4 : Tüm ikinci sıkıştırma bit hızı değerleri için ESA ve DVM sistemlerinde iki kez sıkıştırılmış AMR ses tespit oranları.

Kayıt cihazının (kanal), kayıt ortamının ve konuşmacıların söylediği cümlelerin iki kez sıkıştırılmış AMR ses tespiti üzerindeki etkisini analiz etmek için bir sonraki deneylerde MDSVC veri kümesi kullanılmıştır. MDSVC veri kümesi, üç konumda (koridor, kavşak ve ofis) iki farklı mikrofon (kulaklık ve dahili) kullanılarak kaydedilen 48 farklı konuşmacıya ait ses kayıtlarını içeren bir veri kümesidir [54]. Kayıtlar iki farklı oturumda gerçekleştirilmiş olup veri tabanında toplam 67 farklı söz öbeği bulunmaktadır. Bu nedenle, MDSVC veri tabanı, kanal, kayıt ortamı ve sözlü ifade gibi farklı parametrelerin iki kez sıkıştırılmış AMR ses tespiti üzerindeki etkilerini araştırmak için iyi bir adaydır. Ancak, veri kümesinin genel performansını araştırmak için öncelikle bu parametreleri dikkate almadan sonuçlar elde edilmiştir. Çizelge 3.6 ve 3.7, sırasıyla uçtan uca ESA sistemi kullanılarak elde edilen bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR ses tespit oranlarını özetlemektedir. Çizelge 3.6'da verilen bir kez sıkıştırılmış AMR tespit sonuçlarından, yedi sıkıştırma bit hızı için tespit oranının %99'un üzerinde olduğu görülmektedir. Bunun tek istisnası, %98,94 algılama oranı sağlayan 7.4 kbps sıkıştırma bit hızı değeridir. İki kez

sıkıştırılmış AMR ses tespiti sonuçları incelendiğinde (Çizelge 3.7), 64 durumdan yalnızca 5'inde %99'un altında bir tespit oranı elde edilmektedir. Uçtan uca sistem çoğu durumda %100 tespit oranı vermektedir. En düşük tespit oranı (%94,76) ses sinyalleri birinci ve ikinci sıkıştırma aşamalarında 4.75 kbps'de sıkıştırıldığında elde edilmiştir. Derin özelliklere sahip DVM sınıflandırıcısı için de benzer gözlemler geçerlidir. Örneğin, uçtan uca ve DVM sistemleri kullanılarak sırasıyla %99,69 ve %99,81'lik ortalama iki kez sıkıştırılmış AMR ses tespit oranları elde edilmiştir. Bu nedenle, her iki sistem de ihmal edilebilir bir farkla ortalama olarak benzer iki kez sıkıştırılmış ses tespit performansı göstermektedir.

Çizelge 3.6 : MDSVC veri kümesinde ESA sistemi kullanılarak elde edilen bir kez sıkıştırılmış AMR tespit oranları (%).

	4.75	5.15	5..9	6.7	7.4	7.95	10.2	12.2
ESA	99.58	99.78	99.77	99.33	98.94	100	99.77	100

Çizelge 3.7 : MDSVC veri kümesinde ESA sistemi kullanılarak elde edilen iki kez sıkıştırılmış AMR tespit oranları (%). İlk sütunda bulunan değerler ilk sıkıştırma bit hızı (BR1) ve ilk satırda bulunan değerler ikinci sıkıştırma bit hızını (BR2) göstermektedir.

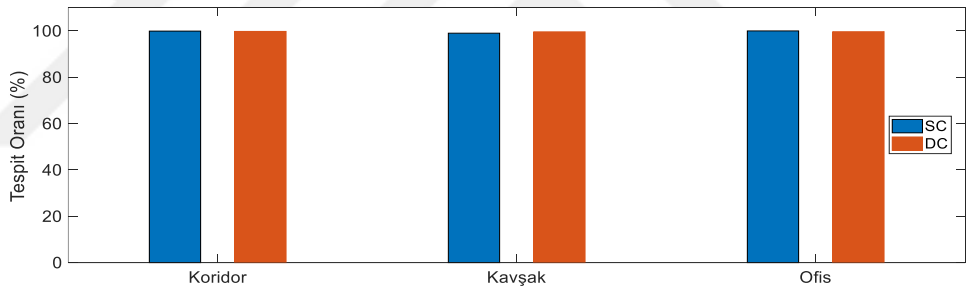
	4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2
4.75	94.76	97.07	99.70	98.53	99.16	99.16	100	99.79
5.15	100	100	100	99.78	99.78	99.78	100	100
5.9	100	100	100	100	100	100	100	99.77
6.7	100	100	100	100	100	100	100	100
7.4	100	100	100	100	100	100	100	100
7.95	100	100	99.78	100	100	100	99.78	100
10.2	100	99.54	99.32	99.09	99.32	99.54	99.77	100
12.2	100	98.96	98.34	100	100	100	99.79	100

Daha sonra, kayıt cihazının (kanal) tespit performansı üzerindeki etkisi analiz edilmiş ve elde edilen sonuçlar Çizelge 3.8'de verilmektedir. Çizelgede özetlenen tespit oranları, tüm sıkıştırma bit hızları için doğrulukların ortalaması alınarak hesaplanmıştır. Çizelgeden, kulaklıklı mikrofonlar ile kaydedilen ses kayıtları kullanılarak elde edilen tespit oranlarının, dahili mikrofon kayıtlarına göre biraz daha iyi olduğunu gözlemlenmektedir. Ancak, kulaklık ve dahili mikrofonlar arasındaki performans farkı göz ardı edilebilecek düzeydedir. Bu, önerilen iki kez sıkıştırılmış AMR ses tespit sisteminin kanal varyasyonlarına karşı dayanıklı olduğunu göstermektedir.

Çizelge 3.8 : MDSVC veri kümesinde yer alan farklı mikrofon tipleri için bir kez ve iki kez sıkıştırılmış AMR tespit oranları (%).

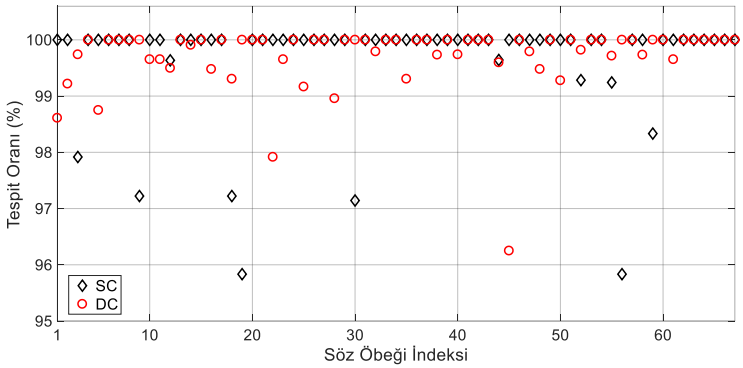
	Kulaklık	Dahili
Bir Kez Sıkıştırılmış AMR	99.89	99.40
İki Kez Sıkıştırılmış AMR	99.71	99.67

Üç farklı konumda kaydedilen ses dosyaları kullanılarak elde edilen ortalama SC ve DC AMR ses tespiti oranları Şekil 3.5'te gösterilmektedir. Algılama oranları, kayıt ortamı değiştikçe önemli ölçüde değişmemektedir. İlginç bir şekilde, koridor ve ofis kayıtları için bir kez sıkıştırılmış AMR ses tespit oranlarının, iki kez sıkıştırılmış AMR ses tespit oranlarından biraz daha yüksek olduğu görülmektedir. Örneğin, koridorda kaydedilen bir kez sıkıştırılmış ses dosyaları ile %99,91 doğruluk oranı elde edilirken, aynı konum için iki kez sıkıştırılmış AMR ses algılama oranı %99,78'dir. Ancak kesişim yeri için bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR dosyaları için sırasıyla %99,02 ve %99,61 ortalama tespit oranları elde edilmiştir.



Şekil 3.5 : Farklı kayıt ortamları için bir kez (SC) ve iki kez sıkıştırılmış (DC) AMR ses tespit oranları (%) [26].

Daha sonra, bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR ses tespit oranlarının kayıt sırasında konuşulan söz öbeğine bağlı olup olmadığını araştırılmıştır. MDSVC veri kümesi, konuşmacılar tarafından 67 farklı cümlenin söylendiği ses dosyalarından oluşmaktadır. Şekil 3.6, her cümle için bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR ses tespit oranlarının ortalamasını göstermektedir. Çoğu cümle için hem bir kez sıkıştırılmış hem de iki kez sıkıştırılmış AMR ses algılama oranlarının %99'un üzerinde olduğu görülmektedir. Yalnızca yedi söz öbeğinde bir kez sıkıştırılmış ses tespit oranı %99'un altındadır. Benzer şekilde iki kez sıkıştırılmış ses tespit oranı 5 söz öbeği için %99'un altındadır. En düşük bir kez sıkıştırılmış AMR tespit oranı %95,83 ile “Eugene Weinstein” söz öbeği için elde edilmiştir. İki kez sıkıştırılmış ses tespitinde ise "Mitchell Peabody" söz öbeği %96,24 ile en düşük tespit oranını verir.



Şekil 3.6 : MDSVC veri kümesindeki söz öbeklerinin bir kez sıkıştırılmış (SC) AMR ve iki kez sıkıştırılmış (DC) AMR ses tespit oranları (%) [26].

İki kez sıkıştırılmış AMR ses tespit deneylerinde kullanılan son veri kümesi VoxForge veri kümesidir. Uçtan uca ESA ve DVM sistemleri kullanılarak VoxForge veri kümesinde elde edilen bir kez sıkıştırılmış AMR tespit oranları Çizelge 3.9'da verilmiştir. Bir kez sıkıştırılmış ses tespit oranları (TNR oranı) 5.9 kbps sıkıştırma bit hızı için en düşük tespit oranı elde edilmiştir. En iyi bir kez sıkıştırılmış ses tespit oranları, ses kayıtları 7.95 kbps ile sıkıştırıldığında elde edilmiştir. Derin özniteliklere sahip DVM sınıflandırıcı, VoxForge veri kümesinde bir kez sıkıştırılmış AMR ses tespiti için uçtan uca sistemden üstündür. Her iki sistem için iki kez sıkıştırılmış AMR ses tespit oranları Çizelge 3.10'da gösterilmektedir. Tablodan uçtan uca ESA sisteminin çoğu durumda DVM sınıflandırıcıdan daha iyi performans gösterdiği gözlemlenmiştir. İlginç bir şekilde, DVM sistemi yalnızca üç durum için biraz daha iyi tespit oranları sağlar ve ikinci sıkıştırma bit hızı BR2'nin tüm bu durumlarda 4.75 kbps olduğu bulunur. En yüksek algılama oranı, BR1 7.4 kbps ve BR2 4.75, 7.4 veya 7.95 kbps olduğunda elde edilmiştir. Hem BR1 hem de BR2 çok düşük olduğunda, minimum tespit oranı elde edilmektedir. Örneğin, ses sinyalleri ilk önce 4.75 kbps'de sıkıştırıldığında, BR2 4.75 kbps ve 5.15 kbps olduğunda sırasıyla %83,33 ve %84,06 algılama oranları elde edilmiştir.

Çizelge 3.9 : VoxForge veri kümesinde ESA ve DVM sistemleri kullanılarak elde edilen bir kez sıkıştırılmış AMR tespit oranları (%).

	Sıkıştırma Bit Hızı							
	4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2
ESA	95.09	94.76	91.13	91.40	95.26	99.50	96.48	97.87
DVM	95.46	96.59	94.47	94.98	97.04	99.74	97.13	98.78

Çizelge 3.10 : VoxForge veri kümesinde ESA ve DVM sistemleri kullanılarak elde edilen iki kez sıkıştırılmış AMR tespit oranları (%).

		BR2							
Sistem	BR1	4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2
ESA	4.75	83.33	84.06	92.89	89.21	89.21	89.95	93.13	93.38
	5.15	98.80	95.00	96.19	95.47	95.47	95.00	96.19	95.71
	5.9	97.53	95.07	95.32	96.30	95.81	95.81	96.30	97.29
	6.7	99.26	97.54	98.28	99.26	99.26	99.26	99.28	99.26
	7.4	99.76	98.57	97.15	99.52	99.76	99.76	98.34	99.76
	7.95	99.26	95.32	96.05	99.26	99.01	99.26	97.04	99.01
	10.2	98.82	93.44	93.67	94.84	94.14	93.20	95.31	96.95
	12.2	98.10	85.81	87.47	97.63	98.10	97.39	90.78	97.87
DVM	4.75	79.09	82.11	90.17	88.66	88.91	88.41	92.69	92.69
	5.15	99.02	94.40	94.64	95.13	94.64	94.40	95.37	94.40
	5.9	97.73	93.21	94.22	95.97	94.47	95.72	94.97	96.73
	6.7	99.24	97.49	97.74	99.24	99.24	99.24	97.99	99.24
	7.4	99.73	97.53	96.05	99.50	99.50	99.50	97.53	99.50
	7.95	98.87	93.11	93.62	99.23	98.97	98.97	96.17	98.46
	10.2	99.04	91.88	91.40	93.31	93.31	92.60	93.79	95.94
	12.2	98.04	81.21	82.19	97.56	97.80	96.82	86.58	97.31

Her veri kümesinin bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR ses tespit performansını karşılaştırmak için Çizelge 3.11'de her veri kümesi ve sistem (derin özelliklere sahip uçtan uca ESA ve DVM sınıflandırıcı) için ortalama tespit oranları Çizelge 3.11'de verilmiştir. Genel olarak, ESA sistemi MDSVC veri kümesindeki iki kez sıkıştırılmış AMR ses dosyaları dışında DVM'den üstündür. Ancak, MDSVC veri kümesinde ESA ve DVM sistemleri arasındaki performans farkı çok düşüktür (%99,69 ve %99,81). En iyi AMR ses tespit performansı, beklendiği gibi her iki sistem için de TIMIT veri kümesinde elde edilmiştir. Bunun nedeni, muhtemelen TIMIT veri kümesinin, daha önce de belirtildiği gibi, herhangi bir kanal veya kayıt cihazı değişkenliği olmaksızın, kontrollü koşullar altında toplanan temiz ses kayıtlarından oluşmasıdır. Uçtan uca ESA sistemi için en düşük tespit oranları VoxForge veri kümesinde elde edilmiştir. Bunun nedeni VoxForge veri kümesindeki ses kayıtlarının TIMIT veri kümesine kıyasla nispeten gürültülü olması ve her kayıt arasında kanal ve ortam varyasyonları getiren standart bir kayıt kurulumu kullanılarak toplanmamış olmalarıdır. Farklı kodlayıcılar (AAC, mp3Pro vb.) kullanılarak önceden kodlanmış ve kodu çözülmüş konuşma dosyalarından oluşan MITD veri kümesi ile DVM sınıflandırıcısı kullanıldığında en düşük tespit performansı elde edimiştir. Kayıtlar daha önce farklı bit hızlarında çeşitli kodlayıcılar kullanılarak kodlama ve kod çözme yoluyla manipüle edildiğinden bu beklenen bir durumdur. Dolayısıyla, MITD veri

kümesinden alınan kayıtlar kullanılarak iki kez sıkıştırılmış AMR ses sinyalinin üretilmesi, toplamda üç kez kodlanan bir ses kaydı üretecektir. Her bir kodlama ve kod çözme işleminin kayıtlar üzerinde kendi izlerini bıraktığı düşünüldüğünde, sistem bu tür ses sinyallerini tanıyamaz. Bununla birlikte, önerilen uçtan uca ESA sistemi, MITD veri kümesi için DVM sınıflandırıcısından önemli ölçüde daha iyi performans göstermektedir. Bu, uçtan uca sistemin iki kez sıkıştırılmış AMR ses tespit görevindeki önceki ses manipülasyonlarına karşı daha gürbüz olduğunu göstermektedir.

Çizelge 3.11 : ESA ve DVM sistemleri kullanılarak her veri kümesi için ortalama bir kez sıkıştırılmış (SC) AMR ve iki kez sıkıştırılmış (DC) AMR ses tespit oranları (%) karşılaştırması.

		TIMIT	MITD	MDSVC	VoxForge
ESA	SC	100	96.17	99.64	95.18
	DC	99.92	97.41	99.69	95.83
DVM	SC	100	90.95	99.72	96.76
	DC	99.84	89.82	99.81	94.87

Bir sonraki aşamada, eğitim ve test veri kümeleri tamamen farklı olduğunda uçtan uca ESA sistemini kullanarak iki kez sıkıştırılmış AMR ses tespiti değerlendirilmiştir. Bu deneylerin gerçekleştirilmesinin arkasındaki motivasyon, istenen bir sistemin, test sinyalinin kayıt koşulları hakkında herhangi bir ön bilgi olmaksızın herhangi bir sıkıştırılmış AMR ses dosyasını tanıyabilmesidir. Bu durumu analiz etmek için, test sinyalleri eğitim veri kümesinden farklı bir veri kümesinden yani sistem tarafından tamamen bilinmeyen başka bir veri kümesinden seçilmiştir. Veri kümeleri arası değerlendirme deneyleri kullanılarak elde edilen ortalama iki kez sıkıştırılmış AMR ses tespit oranları Çizelge 3.12'de özetlenmiştir. Çizelgeden, veri kümelerinin çoğu için (MITD hariç) uçtan uca ESA sisteminin, sistem tamamen farklı bir veri kümesi kullanarak eğitildiğinde makul algılama oranları verdiği görülmektedir. Eğitim ve test veri kümeleri arasında uyumsuzluk olması durumunda, MITD veri kümesideki ses kayıtları AMR dosyaları oluşturulmadan önce sıkıştırma geçmişine sahip olduğu için beklenildiği gibi çok düşük performans göstermiştir. Bu sonuçlar, uçtan uca sistemin potansiyel olarak AMR kodlayıcı ile sıkıştırılmış ses dosyalarını algılamak için iyi bir aday olduğunu göstermektedir. Tüm veri kümelerinin (Çizelge 3.12'nin son satırı) tüm eğitim verileri kullanılarak tek bir model eğitildiğinde makul tespit oranları elde edilse de, tespit oranları, eğitim ve test kayıtları aynı veri kümesinden geldiğinde elde edilen doğruluklara kıyasla düşüktür.

Çizelge 3.12 : Veri kümeleri arası iki kez sıkıştırılmış ses sinyallerinin doğruluk oranları.

	Test Veri Kümesi			
Eğitim	MDSVC	MITD	VoxForge	TIMIT
MDSVC	99.68	11.41	96.43	96.65
MITD	44.89	97.35	42.57	17.27
VoxForge	99.22	10.63	95.69	94.04
TIMIT	99.29	10.44	94.04	99.93
Tüm	99.78	90.49	96.42	99.80

Son olarak, bu tez çalışmasında elde edilen sonuçlar daha önce çeşitli çalışmalarda bildirilen sonuçlarla karşılaştırılmıştır. Tez çalışmalarında iki kez sıkıştrılmış AMR ses tespiti amacı ile kullanılan dört farklı veri kümesinden TIMIT veri kümesi önceki çalışmalarda yaygın olarak kullanılmıştır. Adil bir karşılaştırma yapabilmek için sadece iki kez sıkıştırılmış AMR ses tespiti için TIMIT veri kümesinin kullanıldığı çalışmalar seçilmiştir. Eğitim ve test kayıt sayısı, kayıt süresi, kullanılan öznitelikler ve sınıflandırıcılar ile literatürde ve çalışmalarda bildirilen ortalama tanıma doğrulukları Çizelge 3.13'te özetlenmiştir. Çizelgedeki eğitim ve test dosyalarının sayısı sistemi eğitmek ve test etmek için kullanılan toplam SC ve DC AMR ses dosyalarının sayısına karşılık gelmektedir. Çizelge 3.13'te yer alan ortalama doğruluk sütunu, tüm BR'ler üzerinden ortalaması alınan doğruluk değerlerine karşılık gelmektedir. Daha önce bildirilen sonuçlarla karşılaştırıldığında, bu tez kapsamında önerilen sistemlerin (uçtan uca ESA ve DVM), TIMIT veritabanını kullanarak DC AMR ses tespitini ele alan mevcut çalışmalardan daha iyi performans gösterdiği görülmektedir. Tablodan, önceki çalışmaların çoğunda sınıflandırıcıyı eğitmek için daha fazla sayıda sinyal ve test için daha az sayıda ses sinyali kullanmasına rağmen, bu tezde elde edilen sonuçların literatürde elde edilen sonuçlardan daha yüksek olduğu görülmektedir. Tam uzunluktaki TIMIT kayıtlarını [24, 25, 40] kullanan önceki çalışmalarda bildirilen sonuçlar ile yalnızca 1 saniye uzunluğundaki ses kayıtları kullanılarak elde edilen sonuçlarımız karşılaştırıldığında, hem uçtan uca ESA sistemi hem de derin özelliklere sahip DVM sınıflandırıcı, önceki çalışmalarda önerilen sistemlerden önemli ölçüde daha iyi performans göstermektedir. Bu çalışmada kullanılan ESA ve DVM sistemi, spektrogram görüntüsünün ses sinyali hakkında hem spektral hem de zamansal bilgiyi birlikte içermesi nedeniyle önceki çalışmalardan daha iyi performans göstermektedir. Böylece, ESA, DC AMR ses algılama görevi için en ayırt edici bilgileri öğrenirken, önceki çalışmaların çoğunda yalnızca klasik yöntemlerle elde edilen zamansal veya spektral öznitelikler kullanılmıştır.

Çizelge 3.13 : İki kez sıkıştırılmış AMR ses tespit performansının geçmiş çalışmalarla karşılaştırılması.

Çalışma	Eğitim dosya Sayısı	Test dosya Sayısı	Süre	Sınıflandırıcı	Öznitelikler	Tespit oranı(%)
[10]	8820	3780	5-10 s	DVM	Frekans domeni istatistiksel öznitelikleri	84.86
[22]	-	6000	1 s	Oy çokluğu	Yığın otokodlayıcı	91.10
[23]	3000	9000	1 s	UBM-GMM	Yığın otokodlayıcı	98.28
[24]	8820	3780	1-8 s	DVM	Doğrusal öngörü analizi ile elde edilen istatistiksel öznitelikler	93.47
[40]	6300	6300	1-8 s	DNN	LTAS	89.12
[25]	8820	3780	1-8 s	DVM	Amr kodlayıcı parametreleri	99.18
Bu tez	3000	3000	5 s	ESA	STFT	99.92
Bu tez	3000	3000	5 s	DVM	Derin öznitelikler	99.97
Bu tez	3000	3000	1 s	ESA	STFT	99.92
Bu tez	3000	3000	1 s	DVM	Derin öznitelikler	99.84

3.1.2 Spektral tabanlı öznitelikler ile iki kez sıkıştırılmış AMR tespit sonuçları

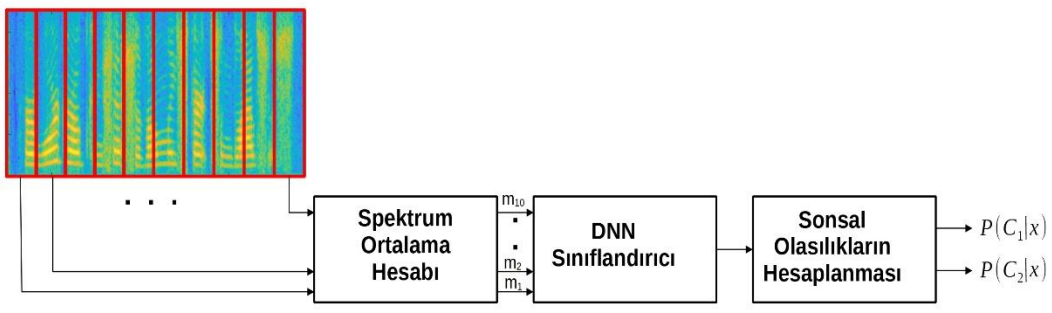
Spektrogram ve zamansal bölütleme öznitelikleri kullanılarak iki kez sıkıştırılmış AMR ses tespit sonuçları TIMIT ve MIT Mobile Devices Speaker Verification Corpus (MDSVC) olmak üzere iki farklı veri kümesinde incelenmiştir. TIMIT veri kümesi 630 konuşmacıdan 6300 söz öbeği içermektedir. TIMIT veri kümesindeki ses kayıtlarının süresi 1-7 saniye arasında değişmektedir. MDSVC veri kümesi ise 48 konuşmacıdan alınan 5184 adet bir saniye uzunluğundaki ses kaydından oluşur (bir konuşma bir saniyeden uzunsa, bir saniye uzunluğunda bir ses kaydı elde etmek için kırpılır). İki veri kümesindeki tüm ses dosyaları, bir kez sıkıştırılmış AMR sesi elde etmek için 4.75 kbps ila 12.2 kbps arasında değişen rastgele seçilen birinci sıkıştırma bit hızı (BR1) ile AMR kodlayıcı kullanılarak sıkıştırılmıştır. Ardından, bir kez sıkıştırılmış AMR ses sinyallerinin kodu çözülerek ve iki kez sıkıştırılmış AMR ses dosyalarını oluşturmak için rastgele seçilen bir ikinci sıkıştırma bit hızı (BR2) kullanarak yeniden sıkıştırılmıştır. Bununla TIMIT veri kümesi için toplam 12600 bir kez sıkıştırılmış (SC) ve iki kez sıkıştırılmış (DC) AMR ses sinyali ($6300\,SC + 6300\,DC = 12600$) ve MDSVC veri tabanı için ise 10368 adet SC ve DC AMR sinyali elde edilmiştir. Bir kez sıkıştırılmış ve iki kez sıkıştırılmış ses sinyalleri oluşturulurken, her bir bit hızı için aynı sayıda ses dosyası elde etmek üzere her bir bit hızı kombinasyonu (BR1 ve BR2), her iki veri kümesi için düzgün dağılımlı olarak seçilmiştir. Her veri kümesi için, ses sinyallerinin %25'i modeli eğitmek için ve sinyallerin kalan %75'i ise sistem performansını test etmek için kullanılmıştır. Eğitim kümesi, hem bir kez sıkıştırılmış hem de iki kez sıkıştırılmış AMR ses sınıfları için olası tüm bit hızı kombinasyonlarını içerecek şekilde oluşturulmuştur. Sınıf dengesizliği sorununu önlemek için eğitim kümesi aynı sayıda bir kez sıkıştırılmış ve iki kez sıkıştırılmış ses dosyasından oluşmaktadır.

Öznitelikleri çıkarırken, önce her bir bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR ses sinyaline ait spektrogramlar oluşturulmuştur. Her ses sinyali, 10 ms çerçeve kaydırma ile 25 ms uzunluğunda çerçevelere bölünerek 512 örnekten oluşan bir Hamming pencere ile pencerelenmiştir. Daha sonra pencerelenmiş çerçevelerin 512 noktalı DFT'si alınarak spektrogram elde edilmiştir. DFT'nin simetri özelliği nedeniyle ilk 256 DFT örneği korunur. Bu, bir saniye uzunluğundaki bir ses sinyali için 256×100 boyutunda bir spektrogramla sonuçlanır. Spektrogram ortalaması öznitelikleri (m), tüm çerçeveler üzerinden spektrogramın ortalaması alınarak elde

edilir. Zamansal bölütleme öznitelikleri için, her bir ses sinyalinin spektrogramını her biri 10 çerçeveden (m) oluşan 10 eşit bölüme ayrılmış ve ardından her bölüm (m_i) için ortalama spektrum hesaplanmıştır. Özetle, bir ses sinyali verildiğinde, 256 katsayıdan oluşan tek bir spektrogram öznitelik vektörü çıkarılır, oysa zamansal bölütleme her biri 256 katsayıya sahip on öznitelik vektörü $([m_1\ m_2 \ldots m_{10}]^T)$ verir.

İki kez sıkıştırılmış AMR ses tespit deneylerinde, sırasıyla 512 ve 128 birimli tam bağlantılı iki gizli katmana sahip basit bir DNN kullanılmıştır. Hem spektrogram ortalaması hem de zamansal bölütleme yöntemleri için 256 boyutlu öznitelik vektörü kullandığımız için giriş katmanı 256 birimden oluşmaktadır. Birinci gizli katmandan sonra yığın (batch) normalizasyon uygulanmış ve aşırı öğrenmeyi önlemek için her gizli katmanda 0,25 seyreltme oranı ile seyreltme katmanları kullanılmıştır. Çıkış katmanı, her biri sırasıyla bir kez sıkıştırılmış (sınıf etiketi 0) ve iki kez sıkıştırılmış AMR ses (sınıf etiketi +1) sınıflarını temsil eden 2 birime sahiptir. Gizli katmanlarda sigmoid aktivasyonları kullanılırken, sınıflandırmayı gerçekleştirmek için çıkış katmanında softmax aktivasyonu kullanılmıştır. Ağ, 0,01 başlangıç öğrenme oranıyla stokastik gradyan iniş (SGD) iyileştirici kullanılarak eğitilmiştir. Öğrenme oranı her 30 çevrimden (epoch) sonra 0,7 oranında azaltılmıştır.

Daha önce bahsedildiği gibi, öznitelik çıkarım yöntemi olarak zamansal bölütleme kullanıldığında toplam on adet 256 boyutlu öznitelik vektörü çıkarılmıştır. Bu nedenle, bir ses sinyalinin verilen test özellikleri, Şekil 3.7'de gösterildiği gibi ses sinyalinin nihai olasılığını elde etmek için tüm zamansal bölümlerin sonsal olasılıkları tahmin edilir ve ardından tüm bölümler üzerinden ortalaması alınmıştır.



Şekil 3.7 : Zamansal bölütleme problem için oluşturulan iki kez sıkıştırılmış AMR tespit sistemi [27].

Önceki bölümde iki kez sıkıştırılmış ses sinyallerinin tespiti için sinyallerin zaman-frekans özelliklerini birlikte barındıran spektrogram gösterimleri derin ESA mimarisine giriş olarak uygulanmıştı. Bu bölümde ise sinyallerin spektrogram gösterimlerinin zaman değişkeni üzerinden ortalamaları alınarak elde edilen LTAS vektörleri ile iki kez sıkıştırılmış ses sinyallerinin tespiti için kullanılacaktır. Deneylerde, öncelikle TIMIT veri kümesinde LTAS ve zamansal bölütleme özniteliklerini kullanarak iki kez sıkıştırılmış AMR ses tespit performansını araştırılmıştır. LTAS ve zamansal bölütleme öznitelikleri kullanılarak elde edilen her birinci ve ikinci sıkıştırma bit hızı (BR1 ve BR2) kombinasyonları için tespit oranları Çizelge 3.14 ve Çizelge 3.15'te özetlenmiştir. Çizelgelerden, sıkıştırma bit hızı (BR1 veya BR2) arttıkça tespit oranının azaldığı gözlemlenmektedir. Sıkıştırma bit hızının arttırılması ses kalitesini iyileştirdiğinden ve bu nedenle iki kez sıkıştırılmış (DC) sesi bir kez sıkıştırılmış (SC) sinyallerden ayırt etmek daha zor hale geldiğinden bu performans düşüşü beklenen bir durumdur. Bir diğer ilginç gözlem ise BR1 ve BR2 arasındaki fark arttıkça daha düşük tespit doğruluğu elde edilmesidir. En yüksek doğruluk, hem BR1 hem de BR2 nispeten düşük olduğunda (4.75-6.7 kbps) elde edilmiştir. Öznitelik çıkarma yönteminden bağımsız olarak, BR2 değeri 7.95 kbps'den 10.2 kbps'ye yükseldiğinde algılama performansı önemli ölçüde düşmektedir. Örneğin, BR1 değeri 12.2 kbps ve BR2 değeri 7.95 kbps olduğunda %62,80 algılama doğruluk değeri elde edilirken, LTAS öznitelikleri ile BR2 10.2 kbps'ye çıkarıldığında %51,59'a düşmektedir. LTAS ve zamansal bölütleme öznitelikleri karşılaştırıldığında, zamansal bölütleme özniteliklerinin, BR1-BR2 kombinasyonlarının büyük çoğunluğu için LTAS özniteliklerinden daha yüksek tespit oranları sağladığı görülmektedir. Örneğin, birinci ve ikinci sıkıştırma için 12.2 ve 4.75 kbps bit hızı değerleri kullanıldığında, LTAS öznitelikleri %77,79 tespit oranı verirken, %86,40 tespit oranı sağlayan zamansal bölütleme ile %11 performans artışı elde edilmektedir. İlginç bir şekilde, hem BR1 hem de BR2 değerleri oldukça yüksek olduğunda (10.2 ve 12.2 kbps) zamansal bölütleme öznitelikleri ile LTAS özniteliklerinden daha düşük tespit oranları elde edilmektedir. Bunun muhtemel nedeni, ses dosyaları çok yüksek bit hızlarında sıkıştırıldığında iki kez sıkıştırılmış ses dosyalarının tespitinin daha zor bir problem haline gelmesidir.

Çizelge 3.14 : TIMIT veri kümesinde spektrum ortalaması yöntemi ile elde edilen öznitelikler kullanılarak oluşturulan iki kez sıkıştırılmış ses tespit sisteminin tespit oranları (%).

				BR2				
BR1	4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2
4.75	94.80	95.80	93.00	91.39	90.20	88.20	80.19	78.39
5.15	92.79	93.19	90.79	87.40	86.19	85.39	76.39	76.80
5.9	95.80	96.60	93.00	91.79	89.60	88.40	78.60	76.59
6.7	93.80	92.59	89.20	84.79	85.00	82.80	70.59	68.80
7.4	90.60	90.60	87.59	85.79	83.20	81.80	71.39	70.99
7.95	89.60	89.20	82.59	80.00	78.39	74.40	62.99	61.79
10.2	86.19	86.79	79.60	74.40	72.79	70.20	59.39	56.99
12.2	77.79	81.59	72.79	69.80	67.40	62.80	51.59	50.99

Çizelge 3.15 : TIMIT veri kümesinde zamansal bölütleme yöntemi ile elde edilen öznitelikler kullanılarak oluşturulan iki kez sıkıştırılmış ses tespit sisteminin tespit oranları (%).

				BR2				
BR1	4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2
4.75	96.00	97.40	95.60	93.40	91.40	88.80	84.00	82.60
5.15	96.60	97.40	94.80	91.00	89.80	89.40	78.40	79.40
5.9	99.00	99.60	98.20	97.00	93.20	92.80	77.80	78.00
6.7	98.40	98.60	94.60	89.00	88.80	85.40	74.20	71.20
7.4	96.00	96.20	91.60	89.40	88.60	86.40	74.40	71.20
7.95	93.00	95.60	85.40	81.00	77.20	73.60	57.80	56.40
10.2	91.80	92.20	83.40	80.00	76.20	75.20	59.80	56.20
12.2	86.40	85.60	74.00	70.40	69.00	65.20	51.00	45.20

TIMIT veri kümesi ile yapılan deneylerin ardından iki kez sıkıştırılmış AMR ses tespit deneyleri, MDSVC veri kümesi üzerinde gerçekleştirilmiştir. Çizelge 3.16 ve Çizelge 3.17, sırası ile LTAS ve zamansal bölütleme öznitelikleri kullanılarak elde edilen tespit oranlarını özetlemektedir. Çizelgelerden ilk olarak MDSVC veri kümesi (Çizelge 3.16 ve Çizelge 3.17) kullanılarak elde edilen tespit doğruluk değerlerinin TIMIT veri kümesi (Çizelge 3.14 ve Çizelge 3.15) kullanılarak elde edilenlerden daha yüksek olduğu görülmektedir. Örneğin ses sinyallerini sıkıştırmada kullanılan BR1 ve BR2 değerlerinin her ikisi de 12.2 kbps olarak seçildiğinde TIMIT veri kümesinde %50,99 tespit oranı elde edilirken, MDSVC veri kümesinde bu oran %74,48 olmaktadır. Bunun nedeni muhtemelen TIMIT veri tabanının 630 farklı konuşmacıya ait ses kayıtları içermesi ve eğitim ve test kümelerinin konuşmacı açısından ayrık olması dolayısı ile konuşmacı değişkenliğinin problemi daha zor hale getirmesidir. Bununla birlikte, MDSVC veri kümesinde, TIMIT veri kümesine kıyasla oldukça az olan

toplam 48 farklı konuşmacı bulunmaktadır. Bu nedenle, konuşmacı değişkenliğinin daha kısıtlı olması nedeniyle tespit oranları daha yüksektir. TIMIT veri kümesi kullanılarak elde edilen sonuçlara benzer şekilde sıkıştırma bit hızları (BR1 ve BR2) arttıkça tespit oranlarında düşüş görülmektedir. Bununla birlikte, tespit oranındaki azalma, TIMIT veri kümesinde gözlemlenenlerden daha düşüktür. TIMIT veri kümesi kullanılarak elde edilen sonuçların aksine, zamansal bölütleme öznitelikleri, tüm BR1-BR2 kombinasyonları için sistematik olarak LTAS özniteliklerinden daha iyi performans göstermektedir. Her iki veri kümesi ile elde edilen sonuçlar, basit bir DSA sınıflandırıcılı spektral özniteliklerin iki kez sıkıştırılmış AMR ses tespitinde ümit verici bir performans verdiğini ve bu nedenle daha gelişmiş öznitelik çıkarma tekniklerinin veya AMR kodlayıcının kodlama sürecine dayanan özniteliklerin makul sonuçlar elde etmek için gerekli olmadığını göstermektedir.

Çizelge 3.16 : MDSVC veri kümesinde spektrum ortalaması yöntemi ile elde edilen öznitelikler kullanılarak oluşturulan iki kez sıkıştırılmış ses tespit sisteminin tespit oranları (%).

				BR2				
BR1	4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2
4.75	97.28	96.86	96.65	96.86	96.86	95.81	95.60	94.56
5.15	98.46	98.02	97.80	97.36	97.80	96.92	96.05	96.05
5.9	97.32	97.09	96.87	97.76	96.87	94.19	96.42	93.30
6.7	96.23	95.57	96.68	96.23	94.46	91.59	92.25	91.81
7.4	96.63	96.42	96.00	90.74	94.74	92.01	92.85	90.12
7.95	91.36	92.65	93.30	90.92	89.63	83.80	86.82	84.66
10.2	93.01	93.01	91.89	90.99	90.76	84.45	88.06	85.36
12.2	88.79	87.34	85.89	84.64	82.36	76.14	78.83	74.48

Çizelge 3.17 : MDSVC veri kümesinde zamansal bölütleme yöntemi ile elde edilen öznitelikler kullanılarak oluşturulan iki kez sıkıştırılmış ses tespit sisteminin doğruluk sonuçları (%).

				BR2				
BR1	4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2
4.75	99.58	99.58	99.58	99.58	99.16	98.95	98.95	98.74
5.15	99.78	99.56	99.78	99.78	99.56	98.68	97.80	98.68
5.9	99.55	99.55	99.10	99.33	98.88	97.99	97.09	96.87
6.7	99.33	99.33	99.33	98.67	98.00	96.90	96.68	95.79
7.4	98.52	98.52	98.52	97.68	97.68	94.95	95.58	94.74
7.95	97.40	97.84	96.97	95.03	93.52	90.92	91.79	90.06
10.2	98.64	98.64	98.42	97.74	95.94	92.56	93.01	90.99
12.2	95.43	95.64	94.19	92.94	90.04	82.98	85.89	82.15

3.1.3 İki kez sıkıştırılmış AMR ses tespitinde açısal softmax ve çeşitleri ile elde edilen sonuçlar

İki kez sıkıştırılmış AMR ses tespit deneylerinde ücretsiz olarak temin edilebilen VoxForge açık konuşma veri setinden elde edilen bir Türkçe ses veri kümesi kullanılmıştır. VoxForge veri kümesi 90 farklı konuşmacıdan alınan ses kayıtlarını içermektedir ve kayıtlar kontrolsüz koşullar altında toplanmıştır. Bunun nedeni, VoxForge'un açık bir konuşma veri tabanı olması nedeniyle, her katılımcının kendi ses örneklerini kendi kayıt kurulumunu kullanarak kaydetmesidir. Bu, kayıtlarda mikrofon ve ortam değişikliklerini ortaya çıkarır ve kayıtların oldukça gürültülü olmasına neden olur. Böylece, VoxForge veri seti, çeşitli kayıt koşulları altında iki kez sıkıştırılmış AMR ses tespit performansının araştırılmasına yardımcı olur.

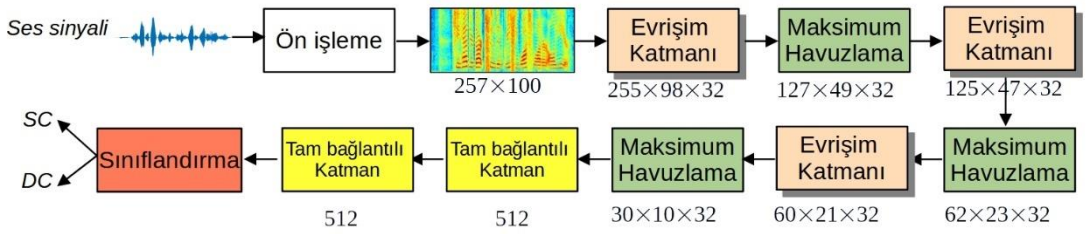
Deneylerde bir saniye uzunluğundaki ses sinyalleri kullanılmıştır. Bu amaçla, veri kümesindeki her bir ses sinyalini bir saniye uzunluğundaki ses kayıtlarına bölerek toplam 4405 ses kaydı elde edilmiştir. AMR kodlayıcı, konuşma sinyalini sekiz farklı bit hızında kodladığından, burada $BR \in \{4.75, 5.15, 5.9, 6.7, 7.4, 7.95, 10.2, 12.2\}$ kbps, her ses sinyali önce bir kez sıkıştırılmış ses sinyalini elde etmek için 4.75 kbps ile 12.2 kbps arasında değişen rastgele seçilmiş bir sıkıştırma bit hızında (ilk sıkıştırma bit hızı -BR1) sıkıştırılmıştır. Birinci sıkıştırma bit hızı değerleri (BR1), her bit hızı için aynı sayıda bir kez sıkıştırılmış ses dosyası elde etmek üzere düzgün dağılımdan seçilerek belirlenmiştir. Daha sonra bir kez sıkıştırılmış ses sinyallerinin kodu çözülerek iki kez sıkıştırılmış AMR ses dosyalarını oluşturmak için rastgele seçilen bir bit hızı (ikinci sıkıştırma bit hızı - BR2) kullanılarak yeniden sıkıştırılmıştır. Bu, toplam 4405 adet bir kez sıkıştırılmış ve 4405 adet iki kez sıkıştırılmış ses kaydıyla sonuçlanır. İki kez sıkıştırılmış ses kayıtları tüm olası BR1-BR2 kombinasyonlarından oluşur (toplam 8 × 8 = 64 BR1-BR2 kombinasyonu) ve her bir BR1-BR2 çifti için kayıt sayısı aynıdır. Eğitim seti, ses kayıtlarının %25'ini içermektedir ve ses dosyalarının geri kalan %75'i test için kullanılmıştır. Tespit sisteminin konuşmacıya doğru önyargılı olmasını önlemek için eğitim ve test setleri konuşmacı açısından ayrıktır. Sınıf dengesizliği sorununu önlemek için eğitim seti aynı sayıda bir kez sıkıştırılmış ve iki kez sıkıştırılmış sinyali içermektedir.

Veri kümeleri arası değerlendirme deneylerinde, sistemin eğitimi ve test edilmesi için farklı ses veri kümeleri kullanılmıştır. Bu amaçla, yukarıda belirtilen VoxForge veri kümesi kullanılarak eğitilen ağ, TIMIT veri kümesinden gelen ses sinyalleri

kullanılarak test edilmiştir. Veri kümeleri arası deneylerde 630 farklı konuşmacıdan toplam 6000 bir saniye uzunluğunda ses sinyali kullanılmıştır. Önceki paragrafta açıklandığı gibi, TIMIT veri setinden 6000 SC ve 6000 DC ses sinyalleri elde edilmiş ve bu sinyaller sadece test için kullanılmıştır. Bu nedenle, veri kümeleri arası değerlendirme deneyleri, sistem tarafından tamamen bilinmeyen ses kayıtlarına karşı sistem test edildiğinde, önerilen sınıflandırıcının ve kayıp fonksiyonlarının performansını araştırmamıza yardımcı olacaktır.

Ses spektrogramları, her 10 ms'de bir 25 ms'lik çerçeveler (her çerçeve 200 örnekten oluşur) kullanılarak hesaplanmıştır. Her ses çerçevesine 0.95 parametreli ön vurgu filtresi uygulandı ve ardından 200 örneklik bir Hamming penceresi kullanılarak pencerelenmiştir. Ardından, güç spektrumunu elde etmek için her pencerelenmiş çerçevenin 512 noktalı DFT'si hesaplanmıştır. Güç spektrumlarının ilk 257 örneği, DFT'nin simetri özelliğinden dolayı korunmuştur. Bu nedenle, 8 kHz'de örneklenen bir saniye uzunluğundaki ses kaydı için 257×100 boyutunda bir spektrogram görüntüsü elde edilmiştir.

İki kez sıkıştırılmış AMR ses tespit deneyleri, Şekil 3.8'de gösterildiği gibi ESA kullanılarak gerçekleştirilmiştir. Her evrişimli ve tam bağlı katmandan sonra 0,25 oranında seyreltme katmanları kullanılmıştır. Son evrişimli katmanı, evrişimli katman çıkışını tam bağlantılı katmana beslemek için bir düzleştirme katmanı takip eder. Performanslarını karşılaştırmak için çıkış katmanında geleneksel softmax ve açısal marjin softmax varyantları (A-Softmax, AM-Softmax ve AAM-Softmax) kullanılmıştır. Açısal marjin Softmax kayıplarının (s ve m) hiperparametreleri, bir dizi ön deney kullanılarak optimize edilerek belirlenmiş ve s = 10 ve m = 0,4 değerleri kullanılmıştır. Ağ, 0,001'lik bir öğrenme oranı kullanılarak Adam iyileştirici ile eğitilmiştir. Sırasıyla SC ve DC AMR ses tespiti için performans kriteri olarak gerçek negatif (TNR) ve gerçek pozitif (TPR) oranları kullanılmıştır.



Şekil 3.8 : Softmax ve açısal çeşitleri ile iki kez sıkıştırılmış AMR tespitinde kullanılan ESA sistemi [28].

Önceki iki bölümde gerçekleştirilen deneyler neticesinde iki kez sıkıştırılmış ses sinyallerinin tespiti için sinyallerin spektrogram özniteliklerinin derin ESA modelinin kullanılmasının, LTAS öznitelikleri ile basit bir tam bağlantılı katmanlardan oluşan DSA yönteminden daha yüksek performans gösterdiği görülmüştür. Klasik DSA yöntemleri sınıflandırıcı katmanında Softmax aktivasyonu kullanmaktadır. Tez çalışmasının bu bölümünde spektrogram özniteliklerini giriş olarak kullanan derin ESA yönteminin son gizli katmanında kullanılan Softmax kayıp fonksiyonu yerine açısal Softmax fonksiyonlarının iki kez sıkıştırılmış seslerin tespitine etkisi incelenecektir. İlk deneyler dört farklı kayıp fonksiyonu ve VoxForge veri kümesi kullanılarak çeşitli bit hızlarında bir kez sıkıştırılmış AMR ses sinyallerinin tespiti üzerinde gerçekleştirilmiştir. Çizelge 3.18, bir kez sıkıştırılmış AMR tespit deneylerinde elde edilen TNR(%) oranlarını göstermektedir. Softmax fonksiyonu kullanılarak yüksek bit hızı değerleri için makul performans elde edilmesine rağmen, sıkıştırma bit hızı azaldıkça tespit oranlarının düştüğünü gözlemlenmektedir.

Çizelge 3.18 : Farklı bit hızları ve kayıp fonksiyonları için bir kez sıkıştırılmış ses sinyallerinin tespit oranları (%).

Bit Hızı	Softmax	A-Softmax	AM-Softmax	AAM -Softmax
4.75	95.09	100	91.17	100
5.15	94.76	100	94.28	100
5.9	91.13	100	90.64	100
6.7	91.40	100	93.85	100
7.4	95.26	100	95.40	100
7.95	99.50	100	99.01	100
10.2	96.48	100	92.50	100
12.2	97.87	100	97.39	100

Çizelgede gösterilen sonuçlar, geleneksel softmax kayıp fonksiyonunun bir kez sıkıştırılmış ses sinyallerinin saptanmasında ümit verici sonuçlar vermesine rağmen, kosinüs fonksiyonuna (A-Softmax ve AAM-Softmax) açısal marjini dahil etmenin

mükemmel performans (%100 tespit oranı) sağladığını göstermektedir. İlginç bir şekilde, AM-Softmax fonksiyonu, yalnızca 6.7 kbps için klasik softmax fonksiyonundan daha iyi tespit oranı sağlamaktadır. Bu sonuçlar, açısal mesafenin ceza terimlerine çarpımsal (A-Softmax) veya toplamsal (AAM-Softmax) şekilde eklenmesinin, sıkıştırma bit hızı değerinden bağımsız olarak bir kez sıkıştırılmış AMR ses tespit performansını artırdığını göstermektedir.

İki kez sıkıştırılmış AMR ses tespit deneylerinde, ilk önce Çizelge 3.19'daki standart softmax fonksiyonu kullanılarak elde edilen sonuçlar verilmiştir. Tablonun satırları birinci sıkıştırma bit hızına (BR1) karşılık gelirken, sütunlar ikinci sıkıştırma bit hızını (BR2) temsil etmektedir. Tablonun son satırı ve sütunu, sırasıyla tüm BR1 ve BR2 değerleri üzerinden ortalaması alınan tanıma oranlarını göstermektedir. Çizelgeden, sıkıştırma bit hızı (BR1 veya BR2) arttıkça, genel olarak daha yüksek tespit oranı elde edildiği gözlemlenmiştir. Örneğin hem BR1 hem de BR2 değeri 4.75 kbps olduğunda sistem %83,33 tespit oranı verirken, BR2 4.75 kbps'den 12.2 kbps'ye çıktığında yaklaşık %12'lik bir performans artışı (tespit oranı %83,33'ten %93,38'e yükselir) gözlenmektedir. Benzer şekilde, sinyaller ilk sıkıştırma aşamasında 4.75 kbps yerine 12.2 kbps kullanılarak sıkıştırıldığında tespit oranı %83,33'ten %98,10'a yükselmektedir. BR1-BR2 kombinasyonlarının çoğu için tespit oranları %95'in üzerinde olmasına rağmen, beklendiği gibi bazı durumlarda özellikle çok düşük bit hızlarında oldukça düşük tespit oranları gözlenmektedir. Bunun muhtemel nedeni, bir ses sinyalinin ilk sıkıştırmada çok düşük bir bit hızında sıkıştırılmasının, ikinci sıkıştırmada daha yüksek bir bit hızında yeniden sıkıştırılsa bile algısal olarak tanınabilen izler ortaya çıkarmasıdır. Bu durum, Çizelge 3.19'da verilen ortalama tespit oranları karşılaştırılarak doğrulanabilir. Ortalama oranlardan, aşırı durumlar (4.75 kbps ve 12.2 kbps) dışında, BR1 belirli bir sıkıştırmada sabitlendiğinde daha yüksek ortalama tespit oranlarının elde edildiği görülmektedir. Bit hızı BR1 veya BR2; 4.75'ten 12.2 kbps'ye çıktığı için performansta sistematik bir iyileşme gözlemlenmemektedir. Bunun nedeni muhtemelen deneylerde kullanılan ses sinyallerinin tek bir kayıt kurulumu kullanılarak toplanmamış olmasıdır. Bu nedenle, tespit oranlarındaki dalgalanmaları gözlemlemek mantıklıdır.

Çizelge 3.19 : VoxForge veri kümesi için Softmax kayıp fonksiyonu kullanılan sistemde iki kez sıkıştırılmış ses sinyallerinin tespit oranları (%).

	4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2	Ort
4.75	83.33	84.06	92.89	89.21	89.21	89.95	93.13	93.38	89.39
5.15	98.80	95.00	96.19	95.47	95.47	95.00	96.19	95.71	95.97
5.9	97.53	95.07	95.32	96.30	95.81	95.81	96.30	97.29	96.17
6.7	99.26	97.54	98.28	99.26	99.26	99.26	99.28	99.26	98.80
7.4	99.76	98.57	97.15	99.52	99.76	99.76	98.34	99.76	99.07
7.95	99.26	95.32	96.05	99.26	99.01	99.26	97.04	99.01	98.02
10.2	98.82	93.44	93.67	94.84	94.14	93.20	95.31	96.95	95.04
12.2	98.10	85.81	87.47	97.63	98.10	97.39	90.78	97.87	94.14
Ort	98.85	93.10	94.62	96.43	96.34	96.20	95.67	97.40	95.83

Bütün parametreler (veri kümesi, giriş öznitelikleri, eğitim tur sayısı, yığın boyutu vb.) sabit kalmak üzere sadece softmax fonksiyonu yerine AM-Softmax kayıp fonksiyonu kullanılması durumunda elde edilen sonuçlar Çizelge 3.20'de verilmektedir. Çizelgede gösterilen sonuçlardan, softmax kayıp fonksiyonu kullanıldığında BR1 ve BR2 değerlerinin her ikisi de 4.75 kbps olduğunda %83,33 tespit oranı elde edilirken, AM-Softmax kayıp fonksiyonu kullanılması yaklaşık %8 performans artışı sağladığı ve %89,70 tespit oranı verdiği görülmektedir. Benzer şekilde, sinyaller 4.75 kbps (BR1) ve ardından 5.15 kbps (BR2) ile sıkıştırıldığında algılama oranı %84,06'dan %89,95'e yükselmektedir. Bir kez sıkıştırılmış AMR ses tespitinin (Çizelge 3.18) aksine, Çizelge 3.20'de verilen sonuçlar, AM-Softmax'ın altmış dört durumun kırk iki BR1-BR2 kombinasyonunda standart softmax kayıp fonksiyonundan daha iyi performans sergilediğini göstermektedir. Bu, bir kosinüs mesafe cezasının getirilmesinin, bir kez sıkıştırılmış ses sinyali tespitinde geleneksel softmax kaybına göre herhangi bir performans iyileştirmesi getirmemesine rağmen, iki kez sıkıştırılmış AMR ses tespit performansını iyileştirdiğini gösterir.

Çizelge 3.20 : AM-Softmax kayıp fonksiyonu kullanılan sistemde iki kez sıkıştırılmış ses sinyallerinin tespit oranları (%).

	4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2
4.75	89.70	89.95	95.58	93.87	93.62	94.36	96.07	95.83
5.15	98.57	97.14	96.90	96.66	95.95	95.71	97.14	96.19
5.9	97.53	96.79	96.05	96.79	96.05	96.79	96.55	97.04
6.7	99.26	99.01	98.28	99.26	99.26	99.26	98.77	99.26
7.4	99.52	98.34	98.34	99.76	99.76	99.76	98.57	99.76
7.95	99.01	96.30	96.30	99.26	99.01	99.26	97.53	99.01
10.2	98.82	96.48	94.37	95.08	95.08	94.61	96.25	97.42
12.2	98.34	89.83	87.70	97.87	98.10	97.87	90.30	97.63

A-Softmax ve AAM-Softmax kayıp fonksiyonları kullanılarak elde edilen iki kez sıkıştırılmış AMR ses tespit oranları sırasıyla Çizelge 3.21 ve 3.22'de özetlenmiştir. Çizelgelerden görüldüğü gibi, sıkıştırma bit hızlarından (BR1 veya BR2) bağımsız olarak, her iki kayıp işlevi de her durumda mükemmel iki kez sıkıştırılmış ses tespit performansı sağlamıştır. Bu, derin öznitelik vektörü ile son tam bağlantılı katmanın ağırlıkları arasına açısal bir mesafe (toplamsal veya çarpımsal) eklemenin, sınıflandırıcının ayırt etme gücünü artırdığını ve geleneksel softmax kaybından daha iyi performans sergilediğini göstermektedir.

Çizelge 3.21 : A-Softmax kayıp fonksiyonu kullanılan sistemde iki kez sıkıştırılmış ses sinyallerinin doğruluk oranları.

				BR2				
BR1	4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2
4.75	100	100	100	100	100	100	100	100
5.15	100	100	100	100	100	100	100	100
5.9	100	100	100	100	100	100	100	100
6.7	100	100	100	100	100	100	100	100
7.4	100	100	100	100	100	100	100	100
7.95	100	100	100	100	100	100	100	100
10.2	100	100	100	100	100	100	100	100
12.2	100	100	100	100	100	100	100	100

Çizelge 3.22 : AAM-Softmax kayıp fonksiyonu kullanılan sistemde iki kez sıkıştırılmış ses sinyallerinin doğruluk oranları.

	4.75	5.15	5.9	6.7	7.4	7.95	10.2	12.2
4.75	100	100	100	100	100	100	100	100
5.15	100	100	100	100	100	100	100	100
5.9	100	100	100	100	100	100	100	100
6.7	100	100	100	100	100	100	100	100
7.4	100	100	100	100	100	100	100	100
7.95	100	100	100	100	100	100	100	100
10.2	100	100	100	100	100	100	100	100
12.2	100	100	100	100	100	100	100	100

Son olarak, veri kümeleri arası iki kez sıkıştırılmış AMR ses tespit deneyleri gerçekleştirilmiştir. Veri kümeleri arası değerlendirme deneylerinde, ESA sistemi VoxForge veritabanı kullanılarak eğittilmiş olup TIMIT veri kümesi yalnızca test için kullanılmıştır. Bu nedenle sistem, eğitim sırasında görünmeyen, tamamen bilinmeyen bir dizi ses sinyali kullanılarak test edilmiştir. Bununla, iki kez sıkıştırılmış AMR ses tespitinde farklı kayıp fonksiyonlarının genelleme kabiliyetini araştırmak amaçlanmıştır. Genel tespit oranları (tüm BR1-BR2 çiftleri için TPR ve TNR

oranlarının ortalaması alınarak hesaplanmıştır) Çizelge 3.23'te özetlenmiştir. İlk olarak, TIMIT veri kümesi hem eğitim hem de test için kullanıldığında ortalama %99,93 tespit oranı elde edilmiştir. Ancak, softmax kayıp fonksiyonuyla VoxForge veri kümesinde eğitilip TIMIT veri kümesinde test edildiğinde tespit oranı %94,04'e düşmektedir. Açısal mesafe kayıp fonksiyonları, test için farklı bir veri kümesi kullanıldığında tespit performansını önemli ölçüde artırmaktadır. Örneğin, AM-Softmax, veri kümesi uyumsuzluğu durumunda (VoxForge eğitim-TIMIT test durumu), eşleşen veri kümesi durumundan (VoxForge-VoxForge durumu) daha iyi bir tespit oranı sağlar. Bunun nedeni, softmax kaybında sınıflar arasındaki açısal mesafenin dahil edilmesinin, daha ayırt edici özniteliklerin öğrenilmesini sağlaması ve dolayısıyla uyumsuzluk durumunda sistem performansını artırmasıdır. Sırasıyla A-Softmax ve AAM-Softmax kayıp fonksiyonları, hem eşleşen hem de uyumsuz veri kümesi koşulları için mükemmel algılama oranları (%100) sağlamaktadır. Bu sonuçlar, veri kümesi uyuşmazlığı koşulları altında bir kez sıkıştırılmış AMR ses dosyalarını iki kez sıkıştırılmış AMR ses sinyallerinden ayırmada açısal mesafe kayıp fonksiyonlarının gücünü ortaya koymaktadır.

Çizelge 3.23 : VoxForge veri kümesi ile eğitilen sistemin TIMIT veri kümesi ile test edilmesi.

	Test Veri Kümesi	
Kayıp Fonksiyonu	VoxForge	TIMIT
Softmax	95.69	94.04
A-Softmax	100	100
AM Softmax	96.44	98.02
AAM Softmax	100	100

3.2 Geniş Bant AMR Kodlayıcı ile İki Kez Sıkıştırılmış Seslerin Tespit Sonuçları

Önceki bölümde tez çalışmaları kapsamında dar bant konuşma sinyallerinin (örnekleme frekansı 8 kHz) AMR kodlayıcı ile iki kez sıkıştırılması durumunda farklı derin öğrenme yöntemleri ile tespit edilmesi problemi ele alınmıştı. Bu bölümde geniş bant AMR kodlayıcı (örnekleme frekansı 16 kHz) ile iki kez sıkıştırılmış seslerin tespiti çalışmalarına ilişkin sonuçlar verilecektir.

Bir kez sıkıştırılmış ses sinyalleri oluşturulurken veri kümelerindeki her ses sinyali 6.60 ile 23.85 kbps arasında değişen rasgele seçilmiş sıkıştırma bit hızı (birinci

sıkıştırma bit hızı - BR1) ile geniş bant AMR kodlayıcı kullanılarak sıkıştırılmıştır. Daha sonra bir kez sıkıştırılmış AMR ses sinyallerinin kodu çözülerek iki kez sıkıştırılmış AMR ses sinyallerini oluşturmak için rasgele seçilen sıkıştırma bit hızı (ikinci sıkıştırma bit hızı – BR2) kullanılarak yeniden sıkıştırılmıştır. Bir kez ve iki kez sıkıştırılmış AMR ses sinyalleri oluşturulurken kullanılan birinci ve ikinci sıkıştırma bit hızları eşit sayıda kullanıldığı için her bir BR1-BR2 kombinasyonuna sahip iki kez sıkıştırılmış AMR ses sinyallerinin sayısı eşittir. Deneylerde veri kümelerinin %25'i eğitim kümesi için kullanılmıştır ve geri kalan ses sinyalleri ile sistem test edilmiştir. Eğitim kümesinde sınıf dengesizliğini engellemek için bir kez ve iki kez sıkıştırılmış AMR ses sinyalleri eşit sayıda dosya içermektedir. Ayrıca sistemin konuşmacıya bağımlı olmaması adına konuşmacıya ait tüm ses kayıtları sadece eğitim kümesinde ya da sadece test kümesinde olacak şekilde oluşturulmuştur. TIMIT veri kümesinde her konuşmacıya ait ortak iki söz öbeği (SA1 ve SA2) içeren kayıtlar bulunmaktadır. Oluşturulan sistemin konuşmaya bağlı tespit yapmasını engellemek için bu kayıtlar eğitim kümesinden çıkarılmıştır.

Geniş bant AMR kodlayıcı ile oluşturulan iki kez sıkıştırılmış AMR ses tespit sistemleri farklı öznitelikler kullanılarak performansları incelenmiştir. Ses sinyallerinin hem spektral hem de faz tabanlı öznitelikleri hesaplanmıştır. Ses sinyallerinin spektrogramlarını hesaplamak için her bir ses sinyali 10 ms'lik kaydırmalar kullanılarak 25 ms'lik çerçevelere bölünmektedir. Ardından 400 örnekli bir Hamming penceresi kullanılarak pencerelenir. Her pencereli çerçevenin güç spektrumu 512 noktalı AFD kullanılarak hesaplanır. Her çerçevenin güç spektrumu simetri özelliği sebebiyle ilk 257 örneği ile elde edilir. Logaritmik güç spektrumu hesaplanarak spektrogram elde edilmektedir. LP spektrogram hesaplanırken birbiri ile örtüşmeyen her pencerelenmiş çerçeve için doğrusal öngörü analizi ile çerçevelerin ilk 20 doğrusal öngörü katsayısı hesaplanıp 512 noktalı AFD uygulanmaktadır. Fourier dönüşümünün simerti özelliği dolayısıyla ilk 257 örnek korunur.

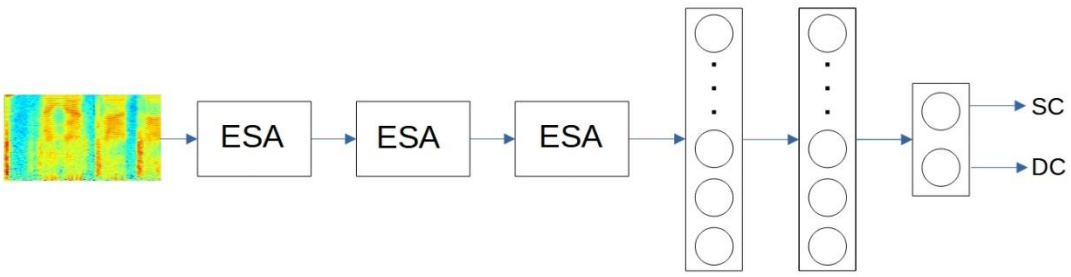
Faz tabanlı öznitelikler hesaplanırken kısa dönem Fourier dönüşümü kullanılmıştır. Fourier dönüşümü ile elde edilen $X(k,t)$ sinyali kutupsal formda yazılarak $\theta(k,t)$ faz spektrumu elde edilmektedir. Faz spektrumundaki süreksizlik noktalarını gidermek amacıyla faz açma (phase unwrapping) uygulanmıştır. Bir diğer faz tabanlı öznitelik oluşturulurken faz açma işlemi uygulanan faz spektrumuna ayrık kosinüs dönüşümü uygulanmıştır.

3.2.1 ESA ile iki kez sıkıştırılmış ses tespit sonuçları

Geniş bant AMR kodlayıcı kullanılarak iki kez sıkıştırılmış AMR seslerin tespiti çalışmalarında öncelikle ESA sistemine giriş olarak spektrogram ve faz tabanlı öznitelikler uygulanarak sistem performansı incelenmiştir. Şekil 3.9'da iki kez sıkıştırılmış AMR seslerin tespiti için kullanılan ESA modeli verilmektedir. Önerilen modelde 3 adet evrişim katmanı bulunmakta olup her evrişim katmanından sonra 2×2 boyutunda maksimum havuzlama katmanı ve 0,5 seyreltme oranına sahip seyreltme katmanı bulunmaktadır. Evrişim katmanlarında 64 adet filtre kullanılmaktadır. Son evrişim katman çıkışına düzleştirme katmanı kullanılarak tek boyuta indirgenerek tam bağlantılı katmanlara uygulanmıştır. Tam bağlantılı katmanların ardından 0,5 seyreltme oranında seyreltme katmanı kullanılmıştır. Çıkış katmanı bir kez sıkıştırılmış AMR sinyali (SC) ve iki kez sıkıştırılmış AMR sinyali (DC) şeklinde iki sınıftan oluşmaktadır. Çizelge 3.24'te TIMIT veri kümesi kullanılarak gerçekleştirilen deneysel sonuçlar gösterilmektedir. Tablodaki sonuçlardan AFD spektrogram öznitelikleri kullanıldığında hem bir kez hem de iki kez sıkıştırılmış ses sinyali tespitinde diğer özniteliklerden daha başarılı sonuçlar elde edilmiştir. Faz tabanlı öznitelikler incelendiğinde faz özniteliklerine DCT uygulandığı durumda elde edilen öznitelik diğer faz tabanlı öznitelikten daha başarılı olup, iki kez sıkıştırılmış sesleri tespit etmede daha yüksek performans göstermektedir.

Çizelge 3.24 : TIMIT veri kümesinde farklı öznitelikler için bir kez (TNR[%]) ve iki kez (TPR[%]) sıkıştırılmış ses sinyallerinin doğruluk oranları.

Öznitelik	TNR[%]	TPR[%]
Spektrogram	70.16	81.38
LP Spektrogram	65.06	74.39
Faz	63.9	60.57
Faz (DCT)	66.08	78.15



Şekil 3.9 : Geniş bant AMR kodlayıcı ile iki kez sıkıştırılmış ses tespitinde kullanılan ESA modeli.

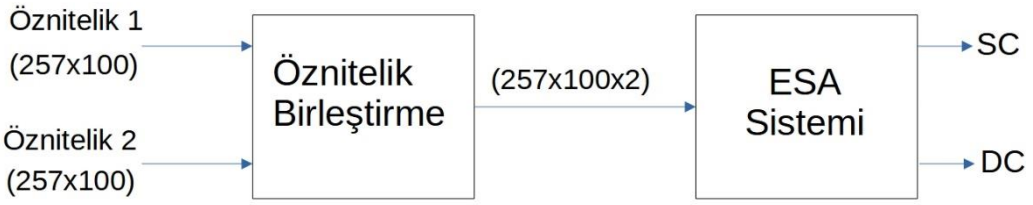
Ardından, iki kez sıkıştırılmış AMR ses tespiti için geliştirilen ESA sisteminin performansını incelemek için sıkıştırma geçmişine sahip TIMIT veri kümesi ile eğitim ve test gerçekleştirilmiştir. Bu amaçla TIMIT veri kümesindeki konuşma sinyalleri önce farklı kodlayıcılar ile (AAC, MP3, FLAC vb.) kodlanıp, kod çözüldükten sonra bu kez AMR kodlayıcı ile sıkıştırılmıştır. Çizelge 3.25, sıkıştırma geçmişine sahip TIMIT veri kümesi ile farklı öznitelikler kullanılarak elde edilen iki kez sıkıştırılmış AMR ses tespit sonuçlarını vermektedir. Sonuçlar incelendiğinde bir kez sıkıştırılmış ses tespitinde DCT ile elde edilen faz tabanlı öznitelik, iki kez sıkıştırılmış ses tespitinde ise AFD spektrogram daha başarılı performans sergilemiştir. Genel olarak bakıldığında orijinal TIMIT veri kümesi ile (sıkıştırma geçmişi olmayan) elde edilen sonuçlar sıkıştırma geçmişine sahip TIMIT veri kümesine göre spektrogram tabanlı özniteliklerde daha başarılıdır. Ancak faz öznitelikleri sıkıştırma geçmişine sahip verilerde daha yüksek performans göstermektedir.

Çizelge 3.25 : Sıkıştırma geçmişine sahip TIMIT veri kümesinde farklı öznitelikler için (TNR[%]) ve iki kez (TPR[%]) sıkıştırılmış ses sinyallerinin doğruluk oranları.

Öznitelik	TNR[%]	TPR[%]
Spektrogram	65.35	81.09
LP Spektrogram	65.03	72.17
Faz	58.84	64.81
Faz (DCT)	76.98	66.45

Spektrogram ve faz tabanlı özniteliklerin iki kez sıkıştırılmış seslerin tespitinde gösterdikleri performans farklılıkları dikkate alındığında, bu özniteliklerin birlikte kullanılmasının (öznitelik birleştirme) yani bu özniteliklerin birbirlerini tamamlayıcı bilgiler içerip içermediğinin analiz edilmesi gerektiğini ortaya çıkarmaktadır. Bu motivasyonla deneysel çalışmaların ikinci aşamasında faz tabanlı ve spektrogram tabanlı öznitelikleri birlikte kullanarak iki kez sıkıştırılmış ses tespitinde iki özniteliğin birlikte etkisi gözlemlenmiştir. İki özniteliğin birlikte kullanıldığı model Şekil 3.10'da verilmektedir. Hem spektrogram tabanlı öznitelikler, hem de faz tabanlı öznitelikler aynı boyuta (257×100) sahip oldukları için bu öznitelikler birleştirilerek ESA sisteminin girişine iki kanal olacak şekilde $(257 \times 100 \times 2)$ uygulanmıştır. Kullanılan ESA sistemi geniş bant AMR kodlayıcı için kullanılan ESA sistemi ile aynı parametreler ile oluşturulmuş olup yalnızca giriş boyutunda değişiklik yapılmıştır. Çizelge 3.26 ve Çizelge 3.27 sırasıyla orijinal TIMIT ve sıkıştırma geçmişine sahip TIMIT veri kümelerinde spektrogram ve faz tabanlı özniteliklerin sisteme birlikte

uygulandığında elde edilen bir kez (TNR[%]) ve iki kez sıkıştırılmış (TPR[%]) ses tespit oranlarını göstermektedir.



Şekil 3.10 : Geniş bant AMR kodlayıcı ile iki kez sıkıştırılmış ses tespitinde kullanılan iki kanallı ESA modeli.

Çizelge 3.26 : TIMIT veri kümesinde spektrogram ve faz tabanlı öznitelikler iki kanallı giriş olarak uygulandığında bir kez (TNR[%]) ve iki kez (TPR[%]) sıkıştırılmış ses sinyallerinin doğruluk oranları.

Kanal 1	Kanal 2	TNR(%)	TPR(%)
LP Spektrogram	Faz	71.95	73.12
LP Spektrogram	Faz (DCT)	68.14	93.06
Spektrogram	Faz	76.04	76.01
Spektrogram	Faz (DCT)	83.38	88.74

Çizelgeler incelendiğinde spektrogram tabanlı öznitelikler karşılaştırıldığında spektrogram genel olarak daha başarılı performans sergilemektedir. Benzer şekilde faz tabanlı öznitelikler incelendiğinde DCT ile elde edilen faz tabanlı öznitelik daha başarılı performans sergilemektedir. Spekrogram ve DCT ile elde edilen faz tabanlı öznitelik sisteme birlikte uygulandığında en iyi tespit oranı elde edilmiştir.

Çizelge 3.24'te en iyi tespit oranları hem bir kez sıkıştırılmış ses tespitinde (%70,16), hem de iki kez sıkıştırılmış ses tespitinde (%81,38) spektrogram kullanılarak elde edilmiştir. Diğer yandan Çizelge 3.26 incelendiğinde girişe spektrogram ve DCT uygulanmış faz öznitelikleri uygulandığında hem bir kez sıkıştırılmış ses tespitinde (%83,38), hem de iki kez sıkıştırılmış ses tespitinde (%88,74) daha üstün performans elde edilmiştir. Geniş bant AMR ile iki kez sıkıştırılmış ses tespit deneylerinde spektrogram tabanlı özniteliklerin yanında faz tabanlı öznitelikler kullanmak problem çözümüne ve daha yüksek performans elde etmeye yardımcı olacaktır.

Çizelge 3.27 : Sıkıştırma geçmişine sahip TIMIT veri kümesinde spektrogram ve faz tabanlı öznitelikler iki kanallı giriş olarak uygulandığında bir kez (TNR[%]) ve iki kez (TPR[%]) sıkıştırılmış ses sinyallerinin doğruluk oranları.

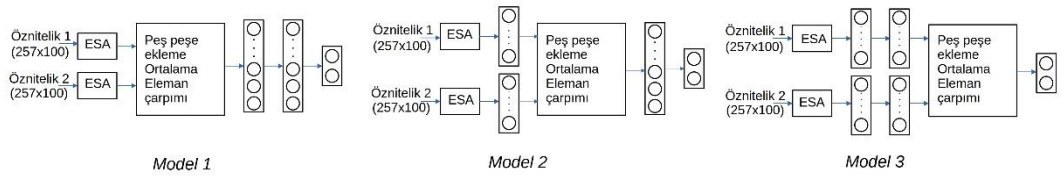
Kanal 1	Kanal 2	TNR(%)	TPR(%)
LP Spektrogram	Faz	55.08	82.20
LP Spektrogram	Faz (DCT)	87.12	77.89
Spektrogram	Faz	77.04	73.94
Spektrogram	Faz (DCT)	87.42	79.05

Benzer şekilde Çizelge 3.25 ve 3.27 incelendiğinde, tek girişli sistemde bir kez sıkıştırılmış ses tespitinde en yüksek doğruluk değeri %76,98 ile DCT uygulanmış faz özntielikleri ile elde edilirken bu öznitelikler spektrogram öznitelikleri ile birlikte kullanıldığında tespit oranı %87,42'ye ulaşmaktadır. Benzer şekilde iki kez sıkıştırılmış ses tespitinde en yüksek performans (%81,09) spektrogram öznitelikleri ile elde edilirken iki girişli sistemde doğruluk oranı %82,20'ye ulaşmaktadır. Bu sonuçlardan geniş bant AMR ile oluşturulan iki kez sıkıştırılmış seslerin tespit edilmesinde iki öznitelik türünü (spektrogram ve faz) birlikte kullanmak daha üstün performans elde etmeyi sağlamaktadır.

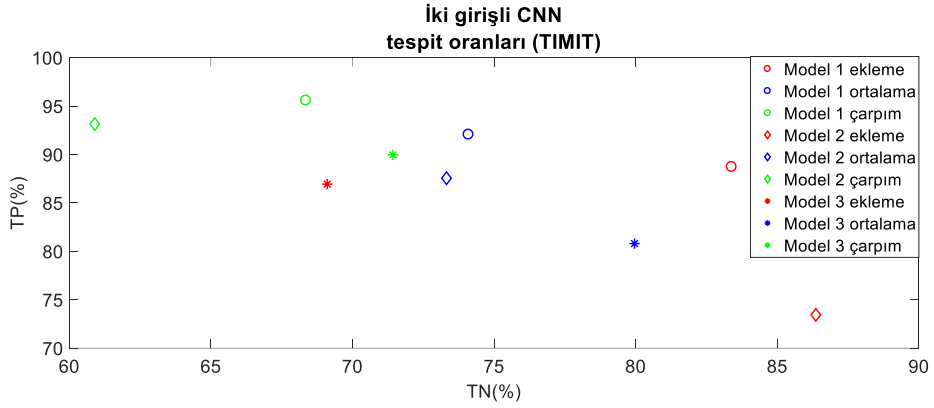
Ardından, iki özniteliği tek bir giriş olarak kullanmak yerine paralel yapıda özdeş ESA katmanları kullanarak iki girişli bir sistemde 3 adet model oluşturulmuştur. Bu modeller paralel olarak özdeş katmanlar kullanıp belirli bir operasyonla (peş peşe ekleme, ortalama alma ve eleman çarpımı) birleştirilir ve sonrasında tam bağlantılı katmanlarla ardışık yapıya dönüşmektedir. Şekil 3.11'de paralel yapıda oluşturulan modellerin detayları verilmektedir. *Model-1*, paralel özdeş üç adet ESA katmanı içermektedir ve birleştirme işlemi düzleştirme katman çıkışları kullanılarak gerçekleştirilir. Birleştirme işlemi sonrasında iki adet tam bağlantılı katman ve çıkış katmanı bulunmaktadır. *Model-2* için paralel yapıda özdeş üç adet ESA katmanı ve bir adet tam bağlantılı katman bulunmaktadır ve tam bağlantılı katman çıkışları birleştirilmiştir. Birleştirme işleminden sonra bir adet tam bağlantılı katman ve çıkış katmanı bulunmaktadır. Paralel yapıda özdeş üç adet ESA katmanı ve iki adet tam bağlantılı katman içeren *Model-3* için birleştirme işlemi ikinci tam bağlantılı katman çıkışları ile gerçekleştirilir.

Gerçekleştirilen ön deneylerde muhtemel bütün öznitelik ikilileri paralel modellere giriş olarak uygulanmış ve spektrogram ile DCT uygulanmış faz öznitelik ikilisi en yüksek performansı veren ikili olarak belirlenmiştir. Dolayısı ile bu öznitelik ikilisi ile

elde edilen sonuçlar raporlanmıştır. TIMIT veri kümesi kullanılarak geliştirilen iki girişli paralel ESA sistemlerinin iki kez sıkıştırılmış ses tespit oranları Şekil 3.12'de verilmiştir. Şekil incelendiğinde, bir kez sıkıştırılmış ses tespitinde en yüksek tespit oranı olan %86,36'a ulaşılırken (Model 2 – peş peşe ekleme işlemi), iki kez sıkıştırılmış ses tespitinde %73,44 tespit oranı elde edilmiştir. Benzer şekilde iki kez sıkıştırılmış ses tespitinde en yüksek tespit oranı %95,60 elde edilirken (Model 1 – eleman çarpımı), bir kez sıkıştırılmış tespit oranı %68,35 olarak bulunmuştur. Modeller birbirleri ile karşılaştırıldığında, iki kez sıkıştırılmış ses tespitinde (TP) Model 1'in diğer modellerden daha başarılı olduğu gözlemlenmiştir. Ayrıca iki kez sıkıştırılmış ses tespitinde derin öznitelikleri birleştirirken eleman çarpımı kullanmak performansı arttırmıştır. Bir kez sıkıştırılmış ses tespitinde peş peşe ekleme işlemi kullanıldığında diğer birleştirme işlemlerinden daha yüksek tespit oranları elde edilmiştir. Dolayısı ile iki kez sıkıştırılmış ses tespitinde öznitelikleri girişte birleştirmek yerine, ESA modeli tarafından öğrenilen derin özniteliklerin birleştirilmesinin daha etkili olduğu ortaya çıkmıştır.

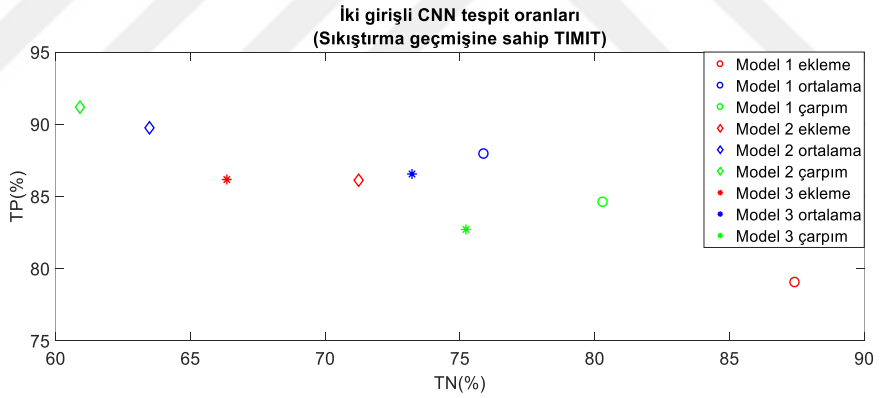


Şekil 3.11 : TIMIT veri kümesinde farklı model ve farklı birleştirme işlemleri için iki kez sıkıştırılmış ses tespit oranları (%).



Şekil 3.12 : TIMIT veri kümesinde farklı model ve farklı birleştirme işlemleri için iki kez sıkıştırılmış ses tespit oranları (%).

Paralel ESA sisteminde farklı modeller ve farklı birleştirme işlemlerinin bir kez sıkıştırılmış ve iki kez sıkıştırılmış ses tespitinde etkisi incelenmiştir. Sıkıştırma geçmişine sahip TIMIT veri kümesi kullanılarak geliştirilen iki girişli paralel ESA sistemlerinin iki kez sıkıştırılmış ses tespit oranları Şekil 3.13'te verilmiştir. Şekil incelendiğinde, bir kez sıkıştırılmış ses tespitinde Model 1 ve peş peşe ekleme işlemi kullanıldığında %87,42 tespit oranı elde edilmiştir. Aynı sistem ile iki kez sıkıştırılmış ses tespitinde %79,05 tespit oranı elde edilmiştir. Fakat Model 2 ve eleman çarpımı kullanıldığında iki kez sıkıştırılmış ses tespitinde %91,18 oranına ulaşılırken bir kez sıkıştırılmış tespit oranı %60,90 olarak bulunmuştur. Modeller karşılaştırıldığında, Model 2 iki kez sıkıştırılmış ses tespitinde, Model 1 ise bir kez sıkıştırılmış ses tespitinde en başarılı modeller olarak gözlemlenmiştir. Ayrıca peş peşe ekleme işlemi bir kez sıkıştırılmış ses tespitinde daha başarılı olurken, ortalama işlemi ile iki kez sıkıştırılmış ses tespitinde daha başarılı sonuçlar elde edilmiştir. Orijinal TIMIT (sıkıştırma geçmişi olmayan) veri kümesi ile paralel ESA sistemi kullanılarak elde edilen sonuçlarda olduğu gibi, paralel sistem sıkıştırma geçmişi olan veriler ile de yüksek doğruluk vermektedir.



Şekil 3.13 : Sıkıştırma geçmişine sahip TIMIT veri kümesinde farklı model ve farklı birleştirme işlemleri için iki kez sıkıştırılmış ses tespit oranları (%).

3.2.2 Uzun kısa süreli bellek ile iki kez sıkıştırılmış ses tespit sonuçları

Her ne kadar önceki bölümlerde kullanılan ESA sistemleri geniş bant AMR kodlayıcı ile iki kez sıkıştırılmış seslerin tespitinde yüksek performans sağlasa da zaman-serisi (time series) verilerinde oldukça büyük öneme sahip uzun-kısa süreli bellek sistemi kullanılarak sonuçların elde edilmesi ve karşılaştırılması önem arz etmektedir. Çünkü konuşma sinyalleri zaman-serisi verilerine gösterilebilecek en önemli örnektir. Bu nedenle bu bölümde LSTM ile elde edilen sonuçlar raporlanacaktır.

İki kez sıkıştırılmış ses tespit sisteminde kullanılan LSTM modeli 4 adet LSTM katmanından oluşmaktadır. LSTM katmanlarında ezberlemeyi önlemek için 0,2 oranında seyreltme kullanılmıştır. LSTM katman çıkışları 32 birimden oluşmaktadır ve son LSTM katmanının çıkışı iki adet tam bağlantılı katmana aktarılmaktadır. Tam bağlantılı katmanların ardından 0,5 oranında seyreltme katmanı uygulanmaktadır. Çıkış katmanı 2 birimden oluşup sinyalin bir kez veya iki kez sıkıştırıldığında karar vermektedir. LSTM katmanlarında aktivasyon fonksiyonu olarak hiperbolik tanjant, tam bağlantılı katmanlarda sigmoid ve çıkış katmanında softmax fonksiyonu kullanılmıştır.

Uzun kısa süreli bellek ile iki kez sıkıştırılmış ses tespit sistem performansı orijinal TIMIT veri kümesi (sıkıştırma geçmişi olmayan) ve sıkıştırma geçmişine sahip TIMIT veri kümesi ile test edilmiştir. Çizelge 3.28 TIMIT veri kümesi kullanılarak elde edilen iki kez sıkıştırılmış ses tespit sisteminin sonuçlarını vermektedir. Çizelge incelendiğinde spektrogram tabanlı özniteliklerin üstün performans gösterdiği gözlemlenmiştir.

Çizelge 3.28 : TIMIT veri kümesinde farklı öznitelikler için bir kez (TNR[%]) ve iki kez (TP[%]) sıkıştırılmış ses sinyallerinin tespit oranları.

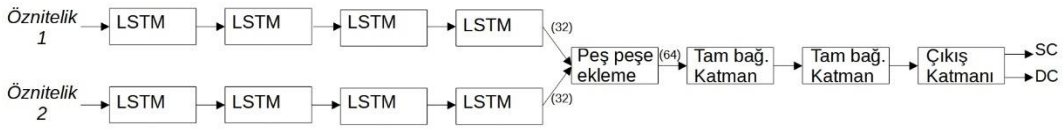
Öznitelik	TNR[%]	TPR[%]
Spektrogram	85.25	77.33
LP Spektrogram	79.04	83.95
Faz	70.61	70.82
Faz (DCT)	73.27	67.72

Benzer şekilde sıkıştırma geçmişine sahip TIMIT veri kümesi kullanılarak elde edilen iki kez sıkıştırılmış ses tespit sisteminin farklı öznitelikler için tespit oranları Çizelge 3.29'de verilmektedir. TIMIT veri kümesindeki sonuçlara benzer şekilde spektrogram tabanlı öznitelikler daha başarılı tespit oranları vermektedir. İnginç bir şekilde faz özniteliği bütün öznitelikler karşılaştırıldığında iki kez sıkıştırılmış ses tespitinde spektrogramdan sonra en başarılı tespit oranına (%82,44) ulaşırken bir kez sıkıştırılmış ses tespitinde en düşük tespit oranına (%58,18) ulaşmaktadır. Çizelge 3.28 ve Çizelge 3.29 karşılaştırıldığında orijinal TIMIT veri kümesi ortalama tespit oranı (TPR+TNR/2) bakımından daha başarılı performans sergilemektedir. Bu beklenen bir durumdur çünkü sıkıştırma geçmişine sahip sinyallerin tespit edilmesi daha zor bir problemdir.

Çizelge 3.29 : Sıkıştırma geçmişine sahip TIMIT veri kümesinde farklı öznitelikler için (TNR[%]) ve iki kez (TNR[%]) sıkıştırılmış ses sinyallerinin tespit oranları.

Öznitelik	TNR[%]	TPR[%]
Spektrogram	70.80	87.90
LP Spektrogram	74.88	81.96
Faz	58.18	82.44
Faz (DCT)	64.27	70.42

Ardından, uzun kısa süreli bellek ile iki kez sıkıştırılmış ses tespit sistemi spektrogram tabanlı öznitelikler ile birlikte faz tabanlı öznitelikler kullanılarak tasarlanmıştır. İki öznitelik aynı parametrelere sahip LSTM katmanlarına giriş olarak uygulanmıştır. Şekil 3.14'te kullanılan model özetlenmekte olup paralel olarak birbirini takip eden dört adet LSTM katmanı ile birlikte iki adet tam bağlantılı katman bulunmaktadır. Son LSTM katmanlarından elde edilen tek boyutlu vektörler peş peşe eklenerek birleştirilip tam bağlantılı katmanlara uygulanmıştır. Çıkış katmanı bir kez ve iki kez sıkıştırılmış olmak üzere iki birim içermektedir. Oluşturulan LSTM modelinde model parametreleri (LSTM çıkış boyutu, seyreltme oranı, tam bağlantılı katman çıkış boyutu ve aktivasyon fonksiyonları) tek girişli LSTM modelinde kullanılan parametreler seçilerek oluşturulmuştur.



Şekil 3.14 : TIMIT veri kümesinde farklı model ve farklı birleştirme işlemleri için iki kez sıkıştırılmış ses tespit oranları (%).

Çizelge 3.30 ve Çizelge 3.31 sırasıyla TIMIT veri kümesi ve sıkıştırıma geçmişine sahip TIMIT veri kümesinde iki girişe sahip iki kez sıkıştırılmış AMR ses tespit sisteminin sonuçlarını göstermektedir. Çizelgeler incelendiğinde, iki özniteliğin birlikte kullanılması iki kez sıkıştırılmış AMR ses tespitinde fayda sağlamaktadır. Paralel LSTM modeli kullanıldığında orijinal TIMIT veri kümesinde en başarılı tespit oranı spektrogram tabanlı iki öznitelik birlikte kullanıldığında elde edilmiştir. Benzer şekilde Çizelge 3.31 incelendiğinde de bu iki spektrogram tabanlı öznitelik kullanıldığında tespit oranında başarı elde edilmiştir.

Çizelge 3.30 : TIMIT veri kümesinde spektrogram ve faz tabanlı öznitelikler iki giriş olarak uygulandığında bir kez (TNR[%]) ve iki kez (TPR[%]) sıkıştırılmış ses sinyallerinin tespit oranları.

LSTM Giriş 1	LSTM Giriş 2	TNR(%)	TPR(%)
Spektrogram	Spektrogram	74.88	80.69
LP Spektrogram	LP Spektrogram	82.70	76.09
Faz	Faz	63.05	62.96
Faz (DCT)	Faz (DCT)	73.98	72.78
Spektrogram	Faz	76.27	85.91
Spektrogram	Faz (DCT)	72.82	87.37
Spektrogram	LP Spektrogram	84.14	84.14
LP Spektrogram	Faz	69.51	87.79
LP Spektrogram	Faz (DCT)	69.29	80.01
Faz	Faz (DCT)	65.58	80.46

Çizelge 3.31 : Sıkıştırma geçmişine sahip TIMIT veri kümesinde spektrogram ve faz tabanlı öznitelikler iki kanallı giriş olarak uygulandığında bir kez (TNR[%]) ve iki kez (TPR[%]) sıkıştırılmış ses sinyallerinin doğruluk oranları.

LSTM Giriş 1	LSTM Giriş 2	TNR(%)	TPR(%)
Spektrogram	Spektrogram	69.06	88.61
LP Spektrogram	LP Spektrogram	70.53	83.50
Faz	Faz	63.11	68.28
Faz (DCT)	Faz (DCT)	71.82	72.72
Spektrogram	Faz	69.61	84.43
Spektrogram	Faz (DCT)	68.40	78.63
Spektrogram	LP Spektrogram	72.11	86.78
LP Spektrogram	Faz	71.69	83.21
LP Spektrogram	Faz (DCT)	63.95	75.16
Faz	Faz (DCT)	72.14	71.08

4. TARTIŞMA ve BULGULAR

Bu tez çalışmasında AMR konuşma sıkıştırma kodlayıcısı ile iki kez sıkıştırılmış dar bant ve geniş bant konuşma sinyallerinin derin öğrenme teknikleri ile otomatik olarak tespit edilmesi problemi ele alınmıştır. Tez çalışmasında öncelikle iki kez sıkıştırılmış ses tespiti için ESA kullanılması önerilmiştir. ESA iki farklı amaç için kullanılmıştır: (i) ses dosyasını bir giriş olarak alan ve giriş sinyalinin SC veya DC olup olmadığını çıkış olarak veren uçtan uca bir iki kez sıkıştırılmış AMR sinyali tespit sistemi olarak kullanılması ve (ii) ESA modelinin farklı gizli katman çıkışlarından elde edilen derin özniteliklerin elde edildiği ve daha sonra bu özniteliklerin DVM tabanlı iki kez sıkıştırılmış AMR ses tespiti için kullanıldığı bir öznitelik çıkarıcı olarak kullanılmasıdır. İlk olarak, bir ses sinyalinin iki kez sıkıştırılmasının ses spektrumunun yüksek frekans bölgesini oldukça etkilediği gösterilmiştir ve bu nedenle konuşma sinyalinin spektrogramının ESA için bir giriş görüntüsü olarak kullanılması önerilmiştir. Her bir sıkıştırma BR'si için toplam 8000 adet denemenin kullanıldığı TIMIT veri tabanından alınan 5 saniyelik ses kayıtları üzerinde yapılan ön deneyler, uçtan uca ESA sisteminin DC konuşma sinyallerinin tespit edilmesinde umut verici sonuçlar verdiğini göstermiştir. Genel olarak sistemin eğitimi ve test edilmesi için aynı bit hızı kullanıldığında daha yüksek doğruluk oranına ulaşıldığı gösterilmiştir. Eğitim ve test bit hızı arasında uyumsuzluk olması durumunda, hafif bir performans düşüşü gözlenmiştir. Özellikle sistem, test bit hızından çok daha düşük bit hızı kullanılarak eğitildiğinde, performans azalmasının daha büyük olduğu gözlenmiştir. Öznitelik çıkarıcı olarak ESA ve sınıflandırma için DVM kullanıldığında, düzleştirme katmanının çıkışından elde edilen özniteliklerin, son tam bağlantılı gizli katmandan çıkarılan özniteliklerden biraz daha iyi performans gösterdiği bulunmuştur. Her iki durumda da, tespit oranlarının uçtan uca ESA sisteminden biraz daha yüksek olduğu görülmüştür. Genel olarak, doğrusal çekirdek fonksiyonu kullanmak, DVM sınıflandırıcılı RBF ve sigmoid çekirdek fonksiyonlarından daha iyi performans göstermiştir.

5 saniye uzunluğundaki ses kayıtları üzerinde yapılan ön deneylerden sonra önerilen ESA sistemi, dört farklı veri seti kullanılarak 1 saniye uzunluğundaki ses kayıtları üzerinde test edilmiştir. TIMIT veri kümesi kullanılarak yapılan deneylerde önerilen uçtan uca ESA sisteminin mükemmel bir kez sıkıştırılmış AMR ses tespit performansı (%100 doğruluk) verdiği ve %99,92 iki kez sıkıştırılmış AMR tespit doğruluğu verdiği görülmüştür. Önerilen sistem, önceden manipüle edilmiş ses kayıtlarından oluşan bir veri seti (MITD veri kümesi) kullanılarak test edildiğinde, bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR sinyalleri için sırasıyla %96,17 ve %97,41 ortalama tanıma oranları elde edilmiştir. İki farklı mikrofon kullanılarak üç farklı konumda kaydedilen konuşma sinyallerinden oluşan bir veri seti (MDSVC veri kümesi) üzerinde yapılan deneyler, önerilen ESA sisteminin bir kez sıkıştırılmış ve iki kez sıkıştırılmış AMR kayıtları için sırasıyla %99,64 ve %99,69 ortalama algılama oranları sağladığını göstermiştir. ESA sisteminin mikrofon türünden, konumundan, ve söylenen cümleden bağımsız olarak mükemmel performans sağlayabildiği görülmüştür. En düşük algılama oranları, kanal değişkenliği ve arka plan gürültüsü ile kontrolsüz koşullar altında kaydedilen ses kayıtlarından oluşan VoxForge veri seti kullanıldığında elde edilmiştir. Genel olarak, uçtan uca ESA sisteminin hem bir kez sıkıştırılmış hem de iki kez sıkıştırılmış AMR ses sinyalini tespit etmede derin özelliklerle DVM sınıflandırıcısından daha iyi performans gösterdiği gösterilmiştir.

İki kez sıkıştırılmış AMR ses tespiti için veri tabanları arası değerlendirme deneyleri gerçekleştirilmiş ve eğitim ve test veri kümeleri arasında bir uyumsuzluk olması durumunda MITD veri kümesi dışında, önerilen sistemin %95'in üzerinde ortalama doğruluk oranıyla makul derecede iyi tespit performansı verdiği ortaya çıkarılmıştır. Ancak, önceden değiştirilmiş/manipüle edilmiş bir ses, AMR kodlayıcıyı kullanılarak sıkıştırıldığında ve bu sinyal farklı bir veri kümesi kullanılarak eğitilen sistemi test etmek için kullanıldığında, algılama oranlarının çok düşük olduğu bulunmuştur. Bu nedenle, iki kez sıkıştırılmış AMR ses algılamasını ele alan gelecekteki çalışmalar bu soruna odaklanmalıdır.

TIMIT veritabanını kullanan önceki çalışmalarla karşılaştırıldığında, önerilen çalışmada elde edilen tespit oranlarının mevcut çalışmalardan daha iyi olduğu gösterilmiştir.

Ardından, ESA tabanlı iki kez sıkıştırılmış AMR ses tespit sisteminde açısal softmax kayıp fonksiyonlarının (A-Softmax, AM-Softmax ve AAM-Softmax) kullanılması önerilmiştir. Açısal marj softmax kaybı başlangıçta yüz doğrulama problemi için önerilmiş olsa da, bildiğimiz kadarıyla bunlar daha önce iki kez sıkıştırılmış AMR ses tespiti problemine uyarlanmamıştır. Deneysel sonuçlar, açısal marjin kayıp fonksiyonlarının hem bir kez sıkıştırılmış hem de iki kez sıkıştırılmış AMR ses sinyallerinin tespitinde normal softmax'tan önemli ölçüde daha iyi performans gösterdiğini gösterdi. Sıkıştırma bit hızı değerinden bağımsız olarak, önerilen kayıp fonksiyonlarının hem bir kez sıkıştırılmış hem de iki kez sıkıştırılmış ses kayıtlarını mükemmel şekilde algıladığı (%100 algılama oranı) gösterilmiştir. A-Softmax ve AAM-Softmax kayıp fonksiyonlarının veritabanları arası değerlendirmede önemli bir etkiye sahip olduğu ve her ikisinin de tüm bit hızı değerleri için %100 algılama oranlarına ulaştığı görülmüştür.

Son olarak, iki kez sıkıştırılmış AMR ses tespiti için bir spektral özellik çıkarma yöntemi önerilmiştir. Önerilen zamansal bölütleme öznitelik vektörleri, sinyalin STFT'sinin tek biçimli parçalara bölünmesi ve daha sonra zaman boyutu boyunca her bir zaman frekans parçasının ortalamasının alınmasıyla çıkarılmıştır. Önerilen öznitelik vektörleri, tamamen bağlantılı iki gizli katmandan oluşan basit bir DNN kullanılarak sınıflandırılmış ve zamansal bölümlendirme özniteliklerinin, hem TIMIT hem de MDSVC veri kümelerinde daha önce önerilen spektrum ortalaması (LTAS) öznitelik vektöründen daha iyi performans gösterdiği gösterilmiştir. Sıkıştırma bit hızı arttıkça iki kez sıkıştırılmış AMR ses kayıtlarını tespit etmenin daha zor hale geldiği gösterilmiştir.

Dar bant AMR ile elde edilen sonuçların ardından geniş bant AMR ile iki kez sıkıştırılmış ses tespiti deneyleri gerçekleştirilmiştir. Geniş bant AMR deneylerinde kullanılmak üzere dört farklı öznitelik kullanılmıştır. Geniş bant AMR ile iki kez sıkıştırılmış tespitinde TIMIT ve sıkıştırma geçmişine sahip TIMIT veri kümeleri kullanılmıştır. Sıkıştırma geçmişine sahip TIMIT veri kümesi oluşturulurken faklı kodlayıcılar (FLAC, AAC, MP3) kullanılarak rasgele seçilen sıkıştırma bit hızları ile sıkıştırılmıştır. Veri kümelerindeki her ses sinyali geniş bant AMR kodlayıcı ile rasgele seçilmiş bit hızı ile sıkıştırılarak bir kez sıkıştırılmış AMR sinyalleri elde edilmiştir. Sıkıştırılmış sinyallerin kodu çözülerek PCM dalga formuna dönüştürülmüştür. İki kez sıkıştırılmış AMR sinyallerini elde etmek için dalga formuna dönüştürülen sinyal rasgele seçilen bit hızı ile sıkıştırılmıştır. Seçilen bit hızları düzgün dağılımlı olup her iki veri kümesi 6300 adet bir kez sıkıştırılmış ve 6300 adet iki kez sıkıştırılmış AMR sinyali içermektedir.

Öncelikle iki kez sıkıştırılmış ses tespitinde ESA sistemi kullanılmış olup orijinal TIMIT veri kümesinde (sıkıştırma geçmişine sahip olmayan) AFD spektrogram özniteliği hem bir kez sıkıştırılmış ses tespitinde (TNR), hem de iki kez sıkıştırılmış ses tespitinde (TPR) en iyi tespit sonuçları elde edilmiştir. Sıkıştırma geçmişine sahip TIMIT veri kümesinde ise AFD spektrogram özniteliği iki kez sıkıştırılmış ses tespitinde en iyi tespit oranına ulaşırken (%81,09) DCT uygulanmış faz tabanlı öznitelik bir kez sıkıştırılmış geniş bant ses tespitinde en iyi tespit oranına ulaşmıştır.

Ardından spektrogram tabanlı öznitelikler (257×100) ve faz tabanlı öznitelikler (257×100) birleştirilerek ESA sisteminin girişine tek bir giriş olarak $(257 \times 100 \times 2)$ uygulanmıştır. Kullanılan ESA sistemi için her iki veri kümesinde de AFD spektrogram ve DCT uygulanmış faz öznitelikleri beraber kullanıldığında diğer öznitelik çiftlerinden daha başarılı olmuştur. Ayrıca tek öznitelik kullanulan ESA sisteminin sonuçları ile karşılaştırıldığında hem bir kez sıkıştırılmış hem de iki kez sıkıştırılmış geniş bant ses tespitinde performansın arttığı gözlemlenmiştir.

Spektrogram tabanlı ve faz tabanlı özniteliklerin birlikte kullanıldığı durumlarda tek bir öznitelik kullanılan durumlara göre daha iyi performans elde edildiğinden bir sonraki çalışmada paralel yapıda ESA sistemi kullanılmıştır. Paralel yapılar birbiri ile özdeş olup derin öznitelik vektörlerinin birleştirildiği katman çıkışlarına göre 3 farklı model oluşturulmuştur. Düzleştirme katman çıkışında birleştirme işlemi Model 1 olarak adlandırılmış olup Model 2 ve Model 3 sırasıyla ilk gizli katman çıkışı ve ikinci gizli katman çıkışlarının birleştirilmesi ile oluşturulmuştur. Birleştirme işlemi olarak peş peşe ekleme, ortalama ve eleman çarpımı işlemleri uygulanmıştır. AFD spektrogram ve DCT uygulanmış faz öznitelikleri diğer öznitelik çiftlerinden daha başarılı performans sergilemiştir ve yalnızca bu öznitelik çiftine ait sonuçlar verilmiştir.

Geniş bant AMR kodlayıcı ile iki kez sıkıştırılmış ses tespitinde bir sonraki çalışmada LSTM modeli kullanılmıştır. Bir kez ve iki kez sıkıştırılmış ses tespitinde her iki veri kümesinde de spektrogram tabanlı öznitelikler daha iyi performans sergilerken sıkıştırma geçmişine sahip TIMIT veri kümesinde tüm özniteliklerde iki kez sıkıştırılmış ses tespit oranları bir kez sıkıştırılmış ses tespit oranlarına göre daha başarılıdır. LSTM model sonuçları tek girişli ESA sistemi ile karşılaştırıldığında LSTM modelinin daha üstün performans sergilediği gözlemlenmiştir.

Ardından spektrogram tabanlı ve faz tabanlı özniteliklerini birlikte kullanmanın iki kez sıkıştırılmış ses tespitindeki etkisini incelemek için paralel yapıda LSTM modeli uygulanmıştır. Özdeş paralel LSTM katmanlarından oluşan sistem, son LSTM katman çıkışları peş peşe eklenerek tam bağlantılı katmanlara uygulanarak geliştirilmiştir. Sonuçlar incelendiğinde spektrogram tabanlı öznitelikler birlikte kullanıldığında her iki veri kümesinde de tek girişli LSTM sistem sonuçlarından daha iyi tespit oranlarına ulaşılmıştır. İki öznitelik kullanılarak geliştirilen sistemlerin iki kez sıkıştırılmış ses tespitinden başarıya ulaştığı gözlemlenmiştir.

KAYNAKLAR

[1] Stamm, M. C., Wu, M., & Liu, K. R. (2013). Information forensics: An overview of the first decade. IEEE access, 1, 167-200.

[2] Maher, R. C. (2009). Audio forensic examination. IEEE Signal Processing Magazine, 26(2), 84-94.

[3] Lukas, J., Fridrich, J., & Goljan, M. (2006). Digital camera identification from sensor pattern noise. IEEE Transactions on Information Forensics and Security, 1(2), 205-214.

[4] Boroumand, M., Chen, M., & Fridrich, J. (2018). Deep residual network for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, 14(5), 1181-1193.

[5] De Carvalho, T. J., Riess, C., Angelopoulou, E., Pedrini, H., & de Rezende Rocha, A. (2013). Exposing digital image forgeries by illumination color classification. IEEE Transactions on Information Forensics and Security, 8(7), 1182-1194.

[6] Huang, F., Huang, J., & Shi, Y. Q. (2010). Detecting double JPEG compression with the same quantization matrix. IEEE Transactions on Information Forensics and Security, 5(4), 848-856.

[7] Hanilci, C., Ertas, F., Ertas, T., & Eskidere, Ö. (2011). Recognition of brand and models of cell-phones from recorded speech signals. IEEE Transactions on Information Forensics and Security, 7(2), 625-634.

[8] Ghasemzadeh, H., & Kayvanrad, M. H. (2018). Comprehensive review of audio steganalysis methods. IET Signal Processing, 12(6), 673-687.

[9] Imran, M., Ali, Z., Bakhsh, S. T., & Akram, S. (2017). Blind detection of copy-move forgery in digital audio forensics. IEEE Access, 5, 12843-12855.

[10] Shen, Y., Jia, J., & Cai, L. (2012). Detecting double compressed AMR-format audio recordings. In Proc. 10th Phonetics Conf. China (PCC), China (pp. 1-5).

[11] Adaptive Multi-Rate (AMR) Speech Codec. Erişim 15 Ekim 2022, https://www.3gpp.org/ftp/Specs/archive/26_series/26.073/26073-f00.zip

[12] Yang, R., Shi, Y. Q., & Huang, J. (2009, September). Defeating fake-quality MP3. In Proceedings of the 11th ACM Workshop on Multimedia and Security (pp. 117-124).

[13] Yang, R., Shi, Y. Q., & Huang, J. (2010, January). Detecting double compression of audio signal. In Media Forensics and Security II (Vol. 7541, pp. 200-209). SPIE.

[14] Liu, Q., Sung, A. H., & Qiao, M. (2010). Detection of double MP3 compression. Cognitive Computation, 2(4), 291-296.

[15] Bianchi, T., De Rosa, A., Fontani, M., Rocciolo, G., & Piva, A. (2013, June). Detection and classification of double compressed MP3 audio tracks. In Proceedings of the first ACM workshop on Information hiding and multimedia security (pp. 159-164).

[16] Bianchi, T., Rosa, A. D., Fontani, M., Rocciolo, G., & Piva, A. (2014). Detection and localization of double compression in MP3 audio tracks. EURASIP Journal on information Security, 2014(1), 1-14.

[17] Ma, P., Wang, R., Yan, D., & Jin, C. (2014). Detecting double-compressed MP3 with the Same Bit-rate. J. Softw., 9(10), 2522-2527.

[18] Yan, D., Wang, R., Zhou, J., Jin, C., & Wang, Z. (2018). Compression history detection for MP3 audio. KSII Transactions on Internet and Information Systems (TIIS), 12(2), 662-675.

[19] Jin, C., Wang, R., Yan, D., Ma, P., & Zhou, J. (2016). An efficient algorithm for double compressed AAC audio detection. Multimedia Tools and Applications, 75(8), 4815-4832.

[20] Huang, Q., Wang, R., Yan, D., & Zhang, J. (2018, June). AAC audio compression detection based on QMDCT coefficient. In International Conference on Cloud Computing and Security (pp. 347-359). Springer, Cham.

[21] Huang, Q., Wang, R., Yan, D., & Zhang, J. (2018). AAC double compression audio detection algorithm based on the difference of scale factor. Information, 9(7), 161.

[22] Luo, D., Yang, R., & Huang, J. (2014, May). Detecting double compressed AMR audio using deep learning. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2669-2673). IEEE.

[23] Luo, D., Yang, R., Li, B., & Huang, J. (2016). Detection of double compressed AMR audio using stacked autoencoder. IEEE Transactions on Information Forensics and Security, 12(2), 432-444.

[24] Sampaio, J. F. P., & NASCIMENTO, F. A. D. O. (2019). Double compressed AMR audio detection using linear prediction coefficients and support vector machine. In Congresso Brasileiro de Automática-CBA (Vol. 1, No. 1).

[25] Sampaio, J. F., & Francisco, A. D. O. (2020). Detection of AMR double compression using compressed-domain speech features. Forensic Science International: Digital Investigation, 33, 200907.

[26] Büker, A., & Hanilçi, C. (2021). Deep convolutional neural networks for double compressed AMR audio detection. IET Signal Processing, 15(4), 265-280.

[27] Büker, A., & Hanilci, C. (2021, November). Double Compressed AMR Audio Detection Using Spectral Features With Temporal Segmentation. In 2021 13th International Conference on Electrical and Electronics Engineering (ELECO) (pp. 284-288). IEEE.

[28] Büker, A., & Hanilçi, C. (2021, June). Angular Margin Softmax Loss and Its Variants for Double Compressed AMR Audio Detection. In Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security (pp. 45-50).

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

[30] Deng, L., Li, J., Huang, J. T., Yao, K., Yu, D., Seide, F., ... & Acero, A. (2013, May). Recent advances in deep learning for speech research at Microsoft. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 8604-8608). IEEE.

[31] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

[32] Chen, C., Seff, A., Kornhauser, A., & Xiao, J. (2015). Deepdriving: Learning affordance for direct perception in autonomous driving. In Proceedings of the IEEE international conference on computer vision (pp. 2722-2730).

[33] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. nature, 542(7639), 115-118.

[34] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. nature, 529(7587), 484-489.

[35] Sainath, T. N., Kingsbury, B., Mohamed, A. R., Dahl, G. E., Saon, G., Soltau, H., ... & Ramabhadran, B. (2013, December). Improvements to deep convolutional neural networks for LVCSR. In 2013 IEEE workshop on automatic speech recognition and understanding (pp. 315-320). IEEE.

[36] Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. The Journal of Machine Learning Research, 17(1), 1334-1373.

[37] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[38] AMR speech Codec: ETSI TS 126 090 V5.0.0 (2002-06). Erişim 15 Ekim 2022 https://www.etsi.org/deliver/etsi_ts/126000_126099/126090/05.00.00_60/ts_126090v050000p.pdf.

[39] Adaptive Multi-Rate Wideband (AMR-WB) Speech Codec. Erişim 15 Ekim 2022,https://www.3gpp.org/ftp/Specs/archive/26_series/26.204/26204-h00.zip

[40] Büker, A., & Hanilçi, C. (2019, November). Double compressed AMR audio detection using long-term features and deep neural networks. In 2019 11th International Conference on Electrical and Electronics Engineering (ELECO) (pp. 590-594). Ieee.

[41] Rabiner, L., & Schafer, R. (2010). Theory and applications of digital speech processing. Prentice Hall Press.

[42] Grigoras, C. (2010, June). Statistical tools for multimedia forensics. In Audio engineering society conference: 39th international conference: audio forensics: practices and challenges. Audio Engineering Society.

[43] Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., ... & Ludvigsen, C. (1994). An international comparison of long-term average speech spectra. The journal of the acoustical society of America, 96(4), 2108-2120.

[44] Saratxaga, I., Sanchez, J., Wu, Z., Hernaez, I., & Navas, E. (2016). Synthetic speech detection using phase information. Speech Communication, 81, 30-41.

[45] An, N. N., Thanh, N. Q., & Liu, Y. (2019). Deep CNNs with self-attention for speaker identification. IEEE access, 7, 85327-85337.

[46] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing, 22(10), 1533-1545.

[47] Himawan, I., Madikeri, S., Motlicek, P., Cernak, M., Sridharan, S., & Fookes, C. (2019). Voice presentation attack detection using convolutional neural networks. In Handbook of Biometric Anti-Spoofing (pp. 391-415). Springer, Cham.

[48] Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., & Shchemelinin, V. (2017, August). Audio replay attack detection with deep learning frameworks. In Interspeech (pp. 82-86).

[49] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 212-220).

[50] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., ... & Liu, W. (2018). Cosface: Large margin cosine loss for deep face recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5265-5274).

[51] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4690-4699).

[52] TIMIT Acoustic-Phonetic Continuous Speech Corpus. Erişim 25 Ekim 2022, https://catalog.ldc.upenn.edu/LDC93S1

[53] Gärtner, D., Cuccovillo, L., Mann, S., & Aichroth, P. (2014, June). A multi-codec audio dataset for codec analysis and tampering detection. In Audio engineering society conference: 54th international conference: audio forensics. Audio Engineering Society.

[54] Woo, R. H., Park, A., & Hazen, T. J. (2006, June). The MIT mobile device speaker verification corpus: data collection and preliminary experiments. In 2006 IEEE Odyssey-The Speaker and Language Recognition Workshop (pp. 1-6). IEEE.

[55] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121-167.

[56] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).

[57] Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27.

[58] Müller, K. R., Mika, S., Tsuda, K., & Schölkopf, K. (2018). An introduction to kernel-based learning algorithms. In Handbook of Neural Network Signal Processing (pp. 4-1). CRC Press.

ÖZGEÇMİŞ

Ad-Soyad **:** Aykut BÜKER

Doğum Tarihi ve Yeri:

E-posta **:**

ÖĞRENİM DURUMU:

- **Lisans** **:** 2015, Yıldız Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, Elektronik ve Haberleşme Bölümü

MESLEKİ DENEYİM VE ÖDÜLLER:

- Araştırma Görevlisi – Bursa Teknik Üniversitesi (08.2018 - …)

TEZDEN TÜRETİLEN ESERLER, SUNUMLAR VE PATENTLER:

- Büker, A., & Hanilçi, C. (2021). Deep convolutional neural networks for double compressed AMR audio detection. IET Signal Processing, 15(4), 265-280.
- Büker, A., & Hanilçi, C. (2021, June). Angular Margin Softmax Loss and Its Variants for Double Compressed AMR Audio Detection. In Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security (pp. 45-50).
- Büker, A., & Hanilci, C. (2021, November). Double Compressed AMR Audio Detection Using Spectral Features With Temporal Segmentation. In 2021 13th International Conference on Electrical and Electronics Engineering (ELECO) (pp. 284-288). IEEE.

DİĞER ESERLER, SUNUMLAR VE PATENTLER:

- Büker, A., & Hanilçi, C. (2019, November). Double compressed AMR audio detection using long-term features and deep neural networks. In *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)* (pp. 590-594). Ieee.
- Bursiyer : 1003 Program kodlu, 118R071 nolu Tübitak Projesi, “Yanıltma Saldırılarını Tespit Edebilen Türkçe Konuşmacı Doğrulama Sistemi”, 15/11/2019-15/11/2021.