NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., 2024; 13(1), 350-358

Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences

Araştırma makalesi / Research article

www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

Görsel kelime tabanlı ses sahteciliği tespit yöntemi

Audio forgery detection method based on visual word

Beste Üstübioğlu[1,*], Arda Üstübioğlu[2]

[1] *Karadeniz Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, 61080, Trabzon Türkiye*
[2] *Trabzon Üniversitesi, Yönetim Bilişim Sistemleri Bölümü, 61335, Trabzon, Türkiye*

Öz

Ses kayıtlarının adli olaylarda delil olarak kullanılması halinde, bu kayıtlarının içeriğinin değiştirilmesi suç teşkil etmektedir. Ses kopyala-yapıştır sahteciliği, konuşmanın içeriğini değiştirmek amacıyla yapılan sahteciliklerden en yaygın olanıdır. Bu sahtecilik, sesteki bir kelime ya da kelime grubunun kopyalanıp, aynı sesin içinde herhangi bir konuma yapıştırılmasıyla gerçekleştirilmektedir. Bu çalışmada ses kopyala-yapıştır sahteciliğini tespit etmek için görsel kelimelere dayalı sağlam ve yeni bir yöntem önerilmektedir. Önerilen yöntem, şüpheli ses dosyasındaki sahtecilik ipuçlarını tespit etmek için sesten elde edilen kelimelerin Mel-Spectogram görüntülerini kullanır. Bu amaçla ses dosyası öncelikle perde bazlı ses aktivite algılama (Voice Activity Detection-VAD) yöntemi kullanılarak kelimelere ayrılır. Daha sonra her kelime Mel Spectogram görüntüsüne dönüştürülür. Spectogram görüntüleri arasındaki benzerliği hesaplamak için yapısal farklılık (Structural Difference-DSSIM) kullanılır. Kelime görüntüleri arasındaki DSSIM değerlerine göre sahte kelimeler işaretlenir. Deneysel sonuçlar, önerilen yöntemin diğer çalışmalara kıyasla son işlem operasyonlarına karşı önemli ölçüde yüksek dayanıklılığa sahip olduğunu ve daha yüksek doğruluk değerini verdiğini göstermektedir.

Anahtar kelimeler: Ses sahteciliği, Kopyala-yapıştır, Spectogram, DSSIM.

Abstract

Changing the content of these recordings constitutes a crime if speech recordings are used as evidence in judicial cases. Audio copy-move forgery is the most common forgery made to change the content of the conversation. This forgery is carried out by copying a word or group of words in the speech and pasting it to any position within the same speech. In this study, a robust new method based on visual words is proposed to detect audio copy-move forgery. The proposed method uses Mel-Spectogram images of words extracted from the audio to detect forgery clues in the suspicious audio file. For this purpose, the audio file is first separated into words using the pitch-based VAD method. Each word is then converted into a Mel Spectogram image. DSSIM is used to calculate the similarity between spectrogram images. Forgery segments are marked according to the DSSIM values between word images. Experimental results show that the proposed method has significantly higher robustness to post-processing operations and yields higher accuracy compared to other works.

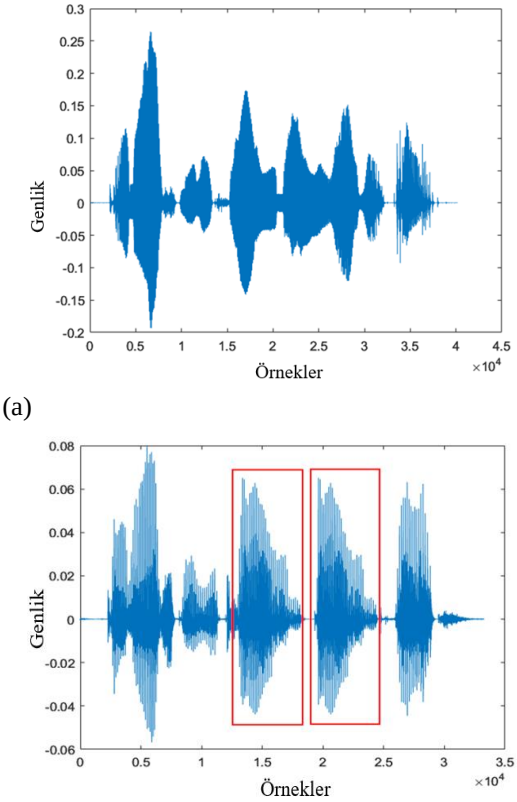
Keywords: Audio forgery, Copy-move, Spectrogram, DSSIM.

1 Giriş

Günümüzde ses teknolojisinde yaşanan hızlı gelişmeler, ses verilerinin çok hızlı bir şekilde üretilmesine, işlenmesine, saklanmasına ve dağıtılmasına olanak sağlamaktadır. Bu durum, çoğu zaman olumlu sonuçlar doğursa da birçok sorunu da beraberinde getirmektedir. Bu sorunlardan en önemlisi ses verisinin güvenilirliğidir. Güvenilirlik, ses verilerinin bütünlüğü ile yani bilgi içeriğinin herhangi bir şekilde yetkisiz değişiklik, silme veya imha tehditlerine karşı korunması ile sağlanmaktadır. Gelişmiş ses düzenleme yazılımlarının kolay kullanılabilirliği sayesinde ses dosyalarının bütünlüğü bozulabilmekte ve saldırganlar tarafından oldukça kolay bir şekilde ses sahteciliği yapılabilmektedir. Ses tanıma alanında [1-3] da sistemi yanıltmak üzere bu tür sahteciliklere oldukça fazla başvurulmaktadır. Bir başka örnek olarak bir dava sırasında mahkemeye delil olarak sunulan bir konuşma kaydının içeriğinde "Olay gecesi Ali oradaydı ama Hasan orada değildi" cümlesi yer almaktadır. Saldırgan tarafından bu cümledeki "orada değildi" cümle parçası kopyalanarak, "oradaydı ama" cümle parçasının üzerine yapıştırılıyor. Böylece ses kaydındaki cümlenin içeriği saldırgan tarafından "Olay gecesi Ali orada değildi, Hasan orada değildi "şeklinde değiştirilmekte ve sahte bir ses kaydı oluşturularak cümlenin anlamı tamamen farklılaştırılmaktadır. Oluşturulan bu sahte kayıt da ilgili davanın gidişatını tamamen tersine çevirebilme durumunu beraberinde getirmektedir. Bu açıdan bakıldığında özellikle adli davalarda ses kayıtlarının doğruluğunun araştırılması büyük önem arz etmektedir. Örnek cümledeki gibi sesteki bir cümle parçasının kopyalanıp aynı sesin başka bir konumuna yapıştırılmasıyla ses kopyala-yapıştır sahteciliği gerçekleştirilmektedir. Şekil 1' de ses kopyala-yapıştır sahteciliği için bir örnek verilmiştir. Şekil 1(a) TIMIT veri tabanından ("sa2.wav") alınan sesi gösterirken, Şekil 1(b)

* Sorumlu yazar / Corresponding author, e-posta / e-mail: bustubioglu@ktu.edu.tr (B. Üstübioğlu)
Geliş / Recieved: 19.09.2023 Kabul / Accepted: 11.12.2023 Yayımlanma / Published: 15.01.2024
doi: 10.28948/ngumuh.1363316

orijinal sesteki dördüncü kelimenin üçüncü kelime üzerine yapıştırılması ile elde edilen sahte sesi göstermektedir. Şekilde kırmızı çerçeveler kopyalanan ve yapıştırılan bölümleri göstermektedir.



Şekil 1. (a)Orijinal (b) Sahte ses

Kopyala-yapıştır sahteciliğin tespitini zorlaştıran en önemli etkenlerden biri, sahte sesin oluşturulma aşamasında sahtecilik izlerini gizlemek için, saldırganlar tarafından sahte sese uygulanan gürültü ekleme, medyan filtreleme ve sıkıştırma gibi son işlem operasyonlarıdır. Önerilecek sahtecilik tespit yönteminin bu operasyonlara dayanıklı olması gerekmektedir. Literatürde ses kopyala-yapıştır sahteciliğini tespit etmeye yönelik önerilen yöntemler üç sınıfa ayrılabilir: Pencere tabanlı, VAD tabanlı yöntemler ve spectrogram tabanlı.

Pencere tabanlı yöntemlerde, ses dosyası önce örtüşmeyen veya eşit uzunluktaki örtüşen pencerelere bölünmektedir. Daha sonra ses pencerelerinden son işlem operasyonlarına dayanıklı özellikler elde edilir. Pencerelerden elde edilen özellikler arasındaki benzerliğe göre sahte pencereler işaretlenmektedir. Pencere tabanlı ilk çalışma Xiao ve ark. [4] tarafından önerilmiştir. Önerilen çalışmada öncelikle ses dosyası T periyotlu pencerelere bölünmüştür. Pencereler arasındaki benzerliği hesaplamak için hızlı evrişimsel algoritmalardan yararlanılmıştır. Belirli bir eşiğin üzerinde benzerliğe sahip pencereler, sahte pencereler olarak işaretlenmiştir. Bir diğer çalışmada Su ve ark. [5] ses dosyası içerisindeki sesli bölümler içindeki ve arasındaki ses kopyala-yapıştır sahteciliğini tespit etmek için iki ardışık pencere (SW) stratejisi önermiştir. Çalışmada anlık Q kepstral katsayıları (constant Q cepstral coefficients–CQCC) özelliği ile pencerelerden özellik çıkarılmıştır. Farklı kısa pencereler arasındaki benzerlik Pearson korelasyon katsayıları ile hesaplanmıştır. Su ve ark. [6] diğer çalışmasında seste kopyalanan bölümleri tespit etmek için sabit Q spectral çizimlerini ve özelleştirilmiş bir genetik algoritma (GA) ile destek vektör makinesinin (Support Vector Machine-SVM) birleşimini kullanmıştır. Bu amaçla öncelikle CQSS özniteliklerini çıkarmak için kare-büyüklük sabiti Q dönüşümünün logaritmasının ortalamasını aldılar. Daha sonra çıkarılan CQSS özelliklerini optimize etmek için GA' yı kullandılar. Son aşamada ise SVM ile optimize edilen özellikler sınıflandırılıp sahte segmentler işaretlenmiştir.

VAD tabanlı yöntemlerde ses dosyasının sesli bölümleri VAD yöntemleriyle çıkarılmaktadır. Sonrasında sesli bölümlerden elde edilen özellikleri arasında benzerlik hesaplaması yapılmaktadır. Sahte segmentler benzerlik hesabına göre tespit edilmektedir. Wang ve ark. [7], tekil değer ayrıştırma (Singular Value Decomposition-SVD) dönüşümü ve ayrık kosinüs dönüşümü (Discrete Cosinus Transform- DCT) ile ses kopyala-yapıştır sahteciliğini tespit etmek için bir yöntem önermiştir. Önerilen yöntemde ses dosyası VAD yöntemiyle sesli bölümlere ayrıldıktan sonra her ses parçasından DCT katsayıları elde edilmiştir. Daha sonra bu katsayıların kare matrisine SVD uygulanarak özvektörler elde edilmiş ve benzerlik hesaplamasında Öklid uzaklığından (ED) faydalanılmıştır. Yan ve ark. [8] önce ses dosyasından perde dizisini (Pitch) elde etmiştir. Perde dizilerinin benzerliğini hesaplamak için ortalama fark (AD) ve Pearson korelasyon katsayısı (Pearson Correlation Coefficient-PCC) yöntemlerinden faydalanmışlardır. Xie ve ark. [9] sesli bölümlerin her birinden ses perdesi, ayrık Fourier dönüşüm katsayıları (Discrete Fourier Transform-DFT), Mel frekans cepstral katsayıları (Mel-frequency cepstral coefficient-MFCC) ve gamatonlar gibi dört ayrı özellik çıkarmıştır. Bu özelliklerin benzerliklerini PCC ve ADD yöntemlerini kullanarak hesaplamıştır. Bu dört özellikten elde edilen tespit sonuçları, nihai kararı elde etmek için C4 karar ağacıyla birleştirilmiştir. İmran ve ark. [10] yerel ikili örüntü (LBP) yöntemine dayalı bir ses kopyala-yapıştır sahteciliği tespit yöntemi önermiştir. Bu yöntemde de diğer VAD tabanlı yöntemler gibi öncelikle sesli bölümler önerdikleri VAD yöntemiyle elde edilmiştir. Daha sonra, her sesli bölümlerden özellik olarak LBP histogramları üretilmiştir. Histogramların benzerlik hesaplaması MSE ve ER metrikleri ile yapılmıştır. Anh ve ark. [11] fonetik diziye dayalı bir yaklaşım sunmuştur. Yöntemlerinde, çıkarılan sesli bölümlerden fonetik diziler elde edilmiştir. Farklı fonetik diziler arasındaki benzerlik en küçük sapmalarla hesaplanmıştır. Mannepalli ve ark. [12] sesten sesli bölümleri elde ettikten sonra her sesli bölümün MFCC özelliğini çıkarmıştır. MFCC özelliklerinin benzerlik hesaplamasında dinamik zaman bükmesi (Dynamic Time Warping-DTW) yönteminden faydalanılmıştır. Üstübioğlu ve ark. [13] YAAPT yöntemiyle sesten sesli bölümleri çıkarmışlardır. Sonrasında bu sesli bölümlerden MDCT katsayıları elde etmişler ve özellik olarak katsayı matrisinin transpozesinin ortalamasını almışlardır. Sesli bölümlerin

özellikleri arasındaki benzerlikleri ölçmek için Öklid mesafesi (Euclidean Distance-ED) kullanılmıştır. Huang ve diğerleri. [14] sesli bölümlerden DFT katsayılarını çıkarmışlardır. Hesaplama maliyetini azaltmak için, önerdikleri yöntemde bu özellikler sıralanmıştır. Sesli bölümler arasındaki benzerliği hesaplamak için PCC yöntemi kullanılmıştır. Yan ve ark. [15] ses dosyasının sesli bölümlerini normalleştirilmiş düşük frekanslı enerji oranıyla çıkarmıştır. Sesli bölümlerden perde ve formant dizileri çıkarıldıktan sonra DTW yöntemiyle benzerlik hesaplaması yapılmıştır.

Spectrogram tabanlı yöntemlerde ses dosyaları spectrogram görüntüsüne dönüştürülür. Özellik çıkarma aşamasında, VAD ve pencere tabanlı yöntemlerden farklı olarak, giriş verisi ses yerine görüntü olduğundan bu yöntemler, ses özellik çıkarım yöntemleri yerine görüntü özellik çıkarım yöntemlerini kullanır. Özellik çıkarımından sonra, spectrogram görüntüsü üzerinde elde edilen işaretlemeler sese izdüşürülür ve kopyalanan bölümler ses üzerinde işaretlenir. Üstübioğlu ve ark. sahte segmentleri tespit etmek için literatürde ilk kez Mel spectrogramı ile derin öğrenme yöntemini kullanmıştır [16]. Çalışmada önerilen CNN mimarisi, şüpheli Mel spectrogram görüntülerini iki sınıfa ayırmktadır: orijinal ve sahte. Üstubioğlu ve ark. [17] diğer çalışmasında Mel-spectrogram görüntüsündeki anahtar noktaları çıkarmak için SIFT yöntemini kullanmıştır. Görüntü kanallarından elde edilen anahtar noktalar özellik vektörleri aracılığıyla eşleştirilmiş ve anahtar noktaları merkez olarak belirlenen görüntü alt blokları sahte bloklar olarak işaretlenmiştir. Üstubioğlu ve ark. [18] bir başka çalışmalarında şüpheli sesi görselleştirmek için süper çözünürlüklü spectrogram görüntülerini kullanmıştır. Ardından, BRIEF yöntemiyle spectrogram görüntüsünden anahtar noktalar ve tanımlayıcıları çıkarılmıştır. İlgili tanımlayıcıları kümeleme yaklaşımıyla eşleştirmek için OPTICS yöntemi kullanılmıştır. Daha sonra yöntemde, spectrogram görüntüsündeki kümelerde yer alan anahtar noktaların konumuna bağlı olarak kopyalanmış bölümler işaretlenmiştir.

Literatürde ses kopyala-taşı sahteciliğinin tespiti için yapılan bu çalışmalar irdelendiğinde, tüm çalışmalarda genel bir algoritmik yaklaşım uygulandığı görülmüştür. Bu yaklaşımda ses dosyası öncelikle pencerelere veya sesli kısımlara bölünür. Daha sonra bu pencerelerden veya sesli bölümlerden, son işlem operasyonlarına dayanıklı bir ses özellik çıkarım yöntemiyle özellikler elde edilir. Elde edilen özelliklerin benzerliğine göre kopyalanıp yapıştırılan bölümler işaretlenir. Daha önceki çalışmalarımızda Üstübioğlu ve ark. [16-18], literatürdeki bu çalışmalardan farklı olarak ses verileri görüntüye dönüştürülmüş (Mel-spectrogram, Yüksek çözünürlüklü spectrogram) ve giriş verisi olarak ses yerine görüntü alınmıştır. Ses bir görüntü ile temsil edildiğinden diğer çalışmalardan farklı olarak ses özelliği çıkarma yöntemleri yerine görüntüden özellik çıkarma yöntemleri kullanılmıştır. Önerilen bu çalışmada, önceki çalışmalarımızdan farklı olarak spectrogram görüntüsü sesin tamamından oluşturulmamış, sesten elde edilen sesli kısımlar, spectrogram görüntülerine dönüştürülmüştür. Böylece önerilen yöntemde sesin tamamına karşılık gelen spectrogram görüntüsünü çıkarmak yerine, seslendirilen her bir kısma karşılık gelen spectrogram görüntüsü çıkarılarak daha detaylı bilgi içeren bir görüntü elde edilmiştir.

Bu çalışmada son işlem operasyonlarından bağımsız, görsel kelime tabanlı yeni bir ses sahteciliği tespit yöntemi önermektedir. Önerilen yöntem, şüpheli ses dosyasındaki sahtecilik ipuçlarını tespit etmek için sesin sesli bölümlerininin (kelime) Mel spectogram görüntülerini kullanır. Bu amaçla ses dosyası öncelikle ses perdesine dayalı bir VAD yöntemi kullanılarak sesli parçalara bölünmektedir. Daha sonra her sesli bölüm için bir Mel spectogram görüntüsü oluşturulur. Mel spectogram görüntüleri arasındaki benzerliğin hesaplanmasında DSSIM metriği kullanılmaktadır. Mel spectogram görüntüleri arasındaki DSSIM değerlerine göre sahte kelimeler işaretlenmektedir. Önerilen yöntemlerin katkıları şu şekilde özetlenebilir:

- Literatürde ilk kez, ses dosyasının sesli bölümlerinden elde edilen Mel spectogram görüntüleri kullanılarak ses kopyala yapıştır sahteciliğinin olası ipuçları araştırılmıştır.
- Sesli bölümlerden çıkarılan Mel spectogram görüntülerinin saldırılara karşı dayanıklılığı sayesinde önerilen ses kopyala-yapıştır sahteciliği tespit yöntemi, saldırılardan bağımsız bir yöntemdir.
- Deneysel sonuçlar, önerilen yöntemin, ses kopyala-yapıştır sahteciliğinin tespitinde literatürdeki diğer çalışmalara göre üstün performans gösterdiğini kanıtlamaktadır.

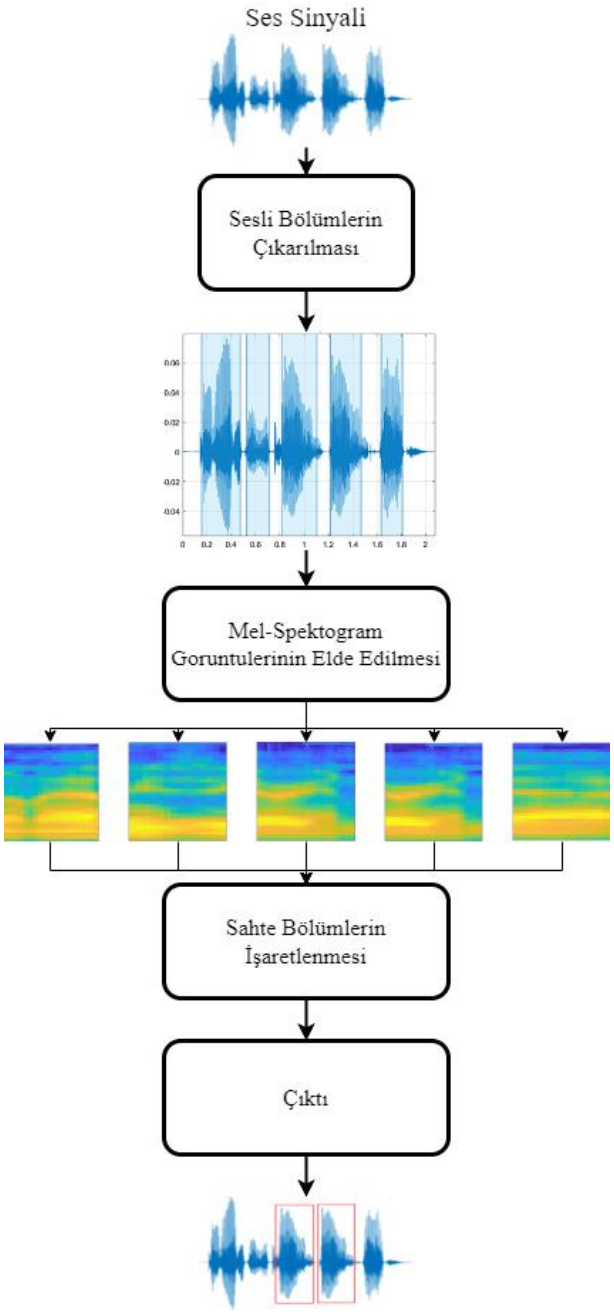
Makalenin geri kalanı şu şekilde düzenlenmiştir: Önerilen ses kopyala-yapıştır sahteciliği yönteminin detayları Materyal ve Metot bölümünde verilmiştir. Bulgular ve Tartışma bölümünde ise deneysel metodoloji sonuçlarının sunumunu ve analizi gösterilmektedir. Sonuç bölümünde ise çalışma sonlandırılmıştır.

2 Materyal ve metot

Çalışmada görsel kelime tabanlı yeni bir ses kopyala-yapıştır sahteciliği tespit yöntemi önerilmektedir. Bu bölümde önerilen yöntemin tüm detayları sunulacaktır. Şekil 2' de görüldüğü gibi, önerilen yöntem üç aşamadan oluşmaktadır: Sesli bölümlerin çıkarılması, sesli bölümlerden Mel spectogram görüntülerinin elde edilmesi, Mel spectogram görüntüleri arasındaki benzerliğe göre sahte segmentlerin işaretlenmesi. İlk aşamada, ses dosyasının sesli bölümlerini çıkarmak için perdeye dayalı VAD tekniğini [13] kullanılmaktadır. Daha sonra elde edilen her bir sesli bölüm ikinci aşamada Mel spectogram görüntüsüne dönüştürülmektedir. Son aşamada ise Mel spectogram görüntüleri arasındaki benzerlik hesabında korelasyon kullanılmaktadır. Tüm görüntüler arasında hesaplanan DSSIM değerleri kaydedilir. Görüntüler arasında elde edilen DSSIM değerlerine göre de sahte kelimeler işaretlenmektedir. Aşağıdaki alt bölümlerde, önerilen yöntemin her aşaması detaylandırılacaktır.

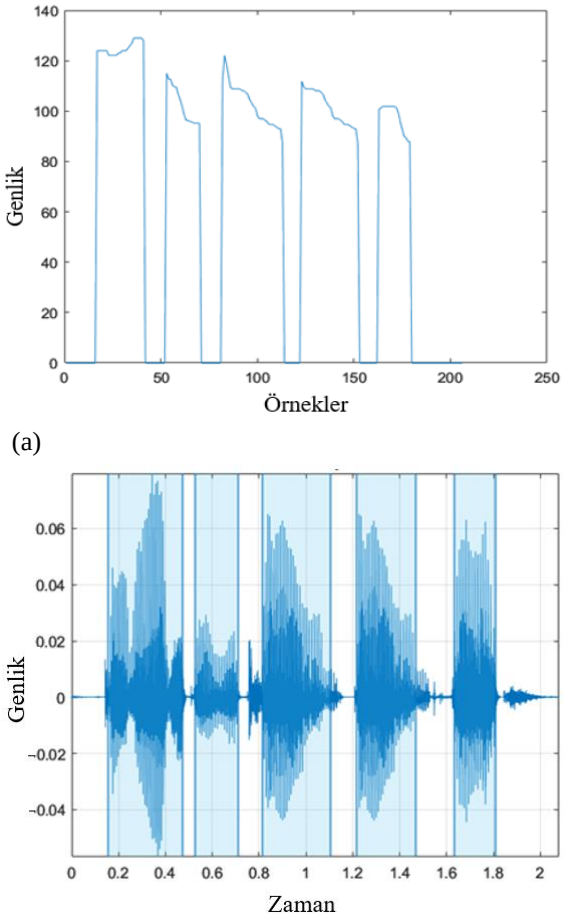
2.1 *Sesli bölümlerin çıkarılması*

Önerilen yöntemde, ses dosyasından sesli bölümlerin çıkarılmasında [13]'da önerilen perde dizisi (Pitch) tabanlı VAD algoritması kullanılmıştır.



Şekil 2. Önerilen yöntemin akış diyagramı

Perde dizisi, temel frekansı ifade eden ve vokal ifadenin titreşim frekansını veren bir kavramdır. Bir kişi aynı sesli bölümleri iki kez söylese bile bu sesli bölümlerin perde dizileri birbirinden farklı olacaktır [19]. YAAPT popüler bir perde dizisi çıkarma yöntemidir. YAAPT yöntemi Ön İşleme, Spectral bilgiyi kullanarak perde dizisi tahmini, Perde dizisi aday tahmini ve Dinamik Programlama ile son perde dizisi belirleme gibi dört aşamadan oluşmaktadır. YAAPT yöntemi ile perde dizisi elde edildikten sonra dizideki sıfırdan büyük frekans değerleri konuşmada sesli bölümler olarak işaretlenir. Şekil 3(a)' da Şekil 1(b)' deki sahte sesten çıkarılan perde dizisi verilmiştir. Şekil 3(b)'deki mavi çizgiler, sesteki perde dizilerine göre elde edilen sesli bölümlerin sınırlarını temsil etmektedir. Şekilden de görülebileceği gibi sahte konuşma, elde edilen perde dizisine göre beş sesli bölüme ayrılmıştır. Üçüncü ve dördüncü bölümler kopyalanıp yapıştırılmış sahte bölümlerdir.



(a)

(b)

Şekil 3. (a)Perde dizisi (b)Ayrılmış sesli bölümler

2.2 *Sesli bölümlerden mel spectogram görüntülerinin elde edilmesi*

Önerilen yöntemde bir önceki aşamada elde edilen sesli bölümler Mel spectogram görüntüsüne dönüştürülmektedir. Sesli bölümlerinin spectrogramını oluşturmak için Kısa Zamanlı Hızlı Fourier dönüşümünü (STFT) kullanılmıştır. Bunun için ses sinyali öncelikle 30 ms' lik karelere bölünmüş ve her kare Hamming penceresiyle çarpılmıştır. Daha sonra Denklem (1)' e göre tüm alt çerçevelere FFT uygulanmıştır.

$$S(f,t) = \sum_{n=0}^{N-1} w_n x_t(n) \exp(-j2\pi(f/f_s)n) \quad f = k f_s/N \qquad (1)$$

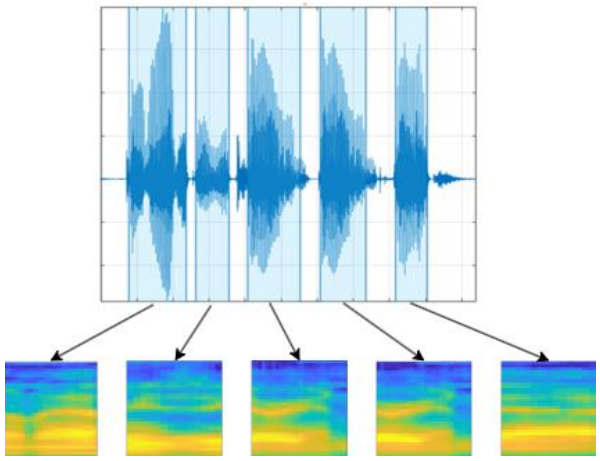
Burada w_n hamming window pencere, *x(n)* orijinal ses sinyali, *f*, *k*=1, 2,…*N*/2+1 için frekans aralığı, *N* her bir çerçevedeki örnek sayısıdır.

Spectrogram gösteriminde genlik spectrogramı, güç spectrogramı, faz spectrogramı ve log genlik spectrogramı gibi birçok farklı yöntem kullanılmaktadır. Bu yöntemlerde frekanslar eşit aralıklarla temsil edilmektedir. Ancak bazı yöntemler aynı zamanda insan duyu sisteminin özelliklerini de kullanır ve düşük frekanslı bileşenlere daha duyarlı, yüksek frekanslı bileşenlere ise daha az duyarlıdır. Mel spectrogramı, insan duyu sistemini temel alan ve giriş sesinin zaman-frekans gösteriminde kullanılan yöntemlerden biridir. Mel spectogram, ses sinyalinin her çerçevesindeki frekans spectrumuna Mel ölçeğinin uygulanması ile elde edilmektedir. Mel spectrogram katsayıları Denklem (2)'ye göre hesaplanır.

$$S_{mel}(f,t) = \sum_{l=0}^{L-1} m_k(l)|S(l,t)|^2 \qquad (2)$$

Burada L frekans bileşeni sayısı, $m_k(l)$ mel filtre bankasının k.filtresidir.

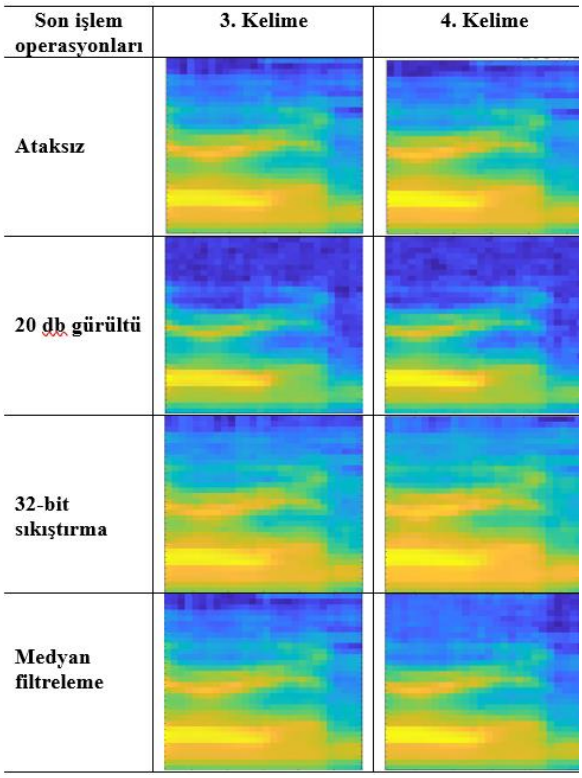
Şekil 4, Şekil 3(b)'de verilen mavi ile ayrılmış beş segmentten elde edilen Mel spectogram görüntülerini sunmaktadır. Şekilde görüldüğü gibi 3. ve 4. segmentlere karşılık gelen Mel spectogram görüntüleri aynıdır.



Şekil 4. Sesli bölümlerden oluşturulan mel spectogram görüntüleri

Önerilen yöntemde Mel spectogram görüntüleri seçilmesinin bir nedeni bu görüntülerin son işlem operasyonlarına olan dayanıklılığıdır. Bu dayanıklılığı göstermek için Şekil 5' te 3. ve 4. sahte segment görüntülerine atak uygulama sonucunda elde edilen görüntüler verilmiştir.

Şekil 5'te görüldüğü gibi segmentlerin ataksız Mel spectogram görüntüleri ile sıkıştırma, gürültü ekleme ve medyan filtreleme uygulanan görüntüleri oldukça benzerdir. Bu da mel spectogram özelliğinin son işlem operasyonlarına oldukça dayanıklı olduğunu göstermektedir.



Şekil 5. Sahte sese atak uygulama sonucunda oluşan mel spectogram görüntüleri

2.3 Sahte bölümlerin işaretlenmesi

Önerilen yöntem, Mel spectogram görüntüleri arasındaki benzerliği hesaplamak için Yapısal Farklılık metriğini (DSSIM) kullanmaktadır. Yapısal Benzerlikten (SSIM) çıkarılan Yapısal Farklılık (DSSIM) metriği bir uzaklık metriği olup, x ve y görüntüleri arasında Denklem (3)' teki gibi hesaplanmaktadır.

$$DSSIM(x,y) = (1 - SSIM(x,y))/2 \qquad (3)$$

Buradaki SSIM metriği, iki görüntünün benzerliğini hesaplamak için kullanılan algısal bir ölçümdür [20]. SSIM bir görüntüden üç özelliği çıkarır: parlaklık, kontrast ve yapı. Bu özelliklerin birleşimini görsel bilgi olarak kullanır. İki görüntü arasında hesaplanan SSIM değeri -1 ile 1 arasında değişmektedir. 1 değeri, benzerliği hesaplanan iki görüntünün aynı olduğunu, -1 değeri ise bu görüntülerin çok farklı olduğunu göstermektedir. Temel olarak, iki x ve y görüntüsü arasındaki yapısal benzerlik indeksi, bir mesafe ölçümünü tanımlar.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_x\sigma_y + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \qquad (4)$$

Burada μ_x ve μ_y sırasıyla görüntü ortalamalarını gösterirken $\sigma_x$ve σ_y standard sapmayı göstermektedir. c_1 ve c_2 sabitlerinin değerleri ise 0,01 ve 0,03 olarak ayarlanmıştır. Bu şekilde ses dosyasındaki sesli bölümlerden çıkarılan Mel

spectogram görüntüleri arasındaki benzerlik hesaplanmaktadır. Hesaplanan benzerlik değerlerinden, belirlenen benzerlik eşiği *t*' den büyük olan değeri veren kelime görüntüleri sahte segmentler olarak işaretlenmektedir. Önerilen çalışmada SSIM metriği yerine DSSIM metriği kullanılmıştır. Çünkü DSSIM değerleri 0 ile 1 arasında değiştiği için eşik değeri daha kolay seçilmiştir. Tablo 1, Şekil 4'te verilen Mel spectogram görüntülerinden elde edilen DSSIM değerlerini göstermektedir. Kalın ile gösterilen değerler eşik değerinden küçük olduğu için bu değerleri veren x ve y kelimeleri sahte kelimeler olarak tespit edilmiştir.

Tablo 1. Segmentler arasındaki DSSIM değerleri

DSSIM	Segment çiftleri	
0.2257	1	2
0.2043	1	3
0.2078	1	4
0.1617	1	5
0.1994	2	3
0.19741	2	4
0.1948	2	5
0.06553	**3**	**4**
0.1828	3	5
0.1842	4	5

3 Bulgular ve tartışma

Çalışmanın bu bölümünde önerilen yöntemin ve literatürdeki benzer çalışmaların performans değerlendirme sonuçlarına yer verilecektir. Sonuçlar ses kopyala-yapıştır sahteciliği ile oluşturulmuş sahte ses veri tabanı [13] üzerinde elde edilmiştir. Bu veri tabanı iki saniyeden altı saniyeye kadar İngilizce konuşmalardan oluşan TIMIT konuşma veri tabanındaki [21] sesler kullanılarak oluşturulmuştur. Sahte sesler, ses dosyalarındaki rastgele bir sesli bölüm kopyalanıp aynı ses dosyasındaki herhangi bir konuma yapıştırılarak oluşturulmuştur. Her sahte segmentin süresi yaklaşık olarak 0,2 saniye ile 0,6 saniye arasındadır. Toplamda 368 adet sahte ses bulunmaktadır. Önerilen yöntemin son işlem operasyonlarına dayanıklılığını göstermek için bu seslere 30 dB ve 20 dB'lik gürültü ekleme, medyan filtreleme ve 32 kbps ve 64 kbps atakları uygulanmıştır. Böylece veri setinde ataklı seslerde dahil toplam 2208 adet sahte ses bulunmaktadır.

Önerilen yöntemi diğer yöntemlerle kapsamlı bir şekilde karşılaştırmak için bu çalışmada Doğruluk, Kesinlik (Precision), Duyarlılık (TPR, Recall) ve F skor metrikleri kullanılmıştır. Doğruluk, doğru tespit edilen sahte seslerin ile doğru tespit edilen orijinal seslerinin toplamının toplam ses sayısına oranıdır. Doğruluk, Denklem (5)' e göre hesaplanmaktadır.

$$Doğruluk = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

Burada TP, sahte ses olarak tespit edilen sahte ses sayısıdır; TN orijinal ses olarak tespit edilen orijinal seslerin sayısıdır; FP sahte ses olarak tespit edilen orijinal seslerin sayısıdır; FN, orijinal sesler olarak tespit edilen sahte seslerin sayısıdır. Hassasiyet, doğru şekilde tespit edilen sahte seslerin toplam tespit edilen sahte seslere oranını gösterirken; Kesinlik, doğru şekilde tespit edilen sahte seslerin tüm sahte seslere oranını gösterir. F-ölçüsü kesinlik ve duyarlılığın ağırlıklı ortalamasıdır. Hassasiyet, Duyarlılık ve F-skor ölçümleri sırasıyla Denklem (6), (7) ve (8)' de verilmiştir.

$$Hassasiyet = \frac{TP}{TP + FP} \quad (6)$$

$$Duyarlılık = \frac{TP}{TP + FN} \quad (7)$$

$$F\_skor = \frac{2xPrecisionxRecall}{Precision + Recall} \quad (8)$$

Yüksek bir F-skor değeri, hassasiyet ve duyarlılık açısından daha iyi bir genel performansı göstermektedir.

Sahte ses veri tabanı kullanılarak ve bu metriklerle gerçekleştirilecek ilk deneyde önerilen yöntemin ataklara karşı dayanıklılığı test edilmiştir. Bu amaçla her bir atak için önerilen yöntemin performansı değerlendirilmiştir. Tablo 2, ataklı ve ataksız sesler ile önerilen yöntemden elde edilen Doğruluk, Duyarlılık, Hassasiyet ve F-skor değerlerini göstermektedir.

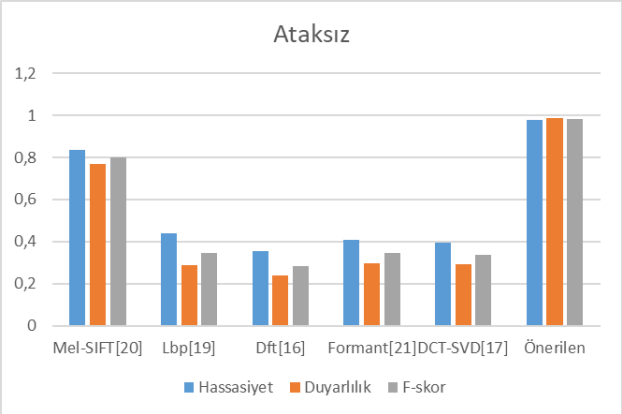
Tablo 2. Önerilen yöntemin tespit sonuçları

	Doğruluk	Hassasiyet	Duyarlılık	F1 skor
Ataksız	0.9758	0.9785	0.9891	0.9838
30db	0.9758	0.9785	0.9891	0.9838
20db	0.9677	0.9785	0.9891	0.9783
Medyan	0.9194	0.9767	0.913	0.9438
32 bit	0.9597	0.978	0.9674	0.9727
64 bit	0.9617	0.9781	0.9701	0.9741

Tablo 1' den görülebileceği gibi, önerilen yöntem sıkıştırma, medyan filtreleme ve gürültü ekleme sonrası işleme işlemlerine karşı oldukça dayanıklıdır. Önerilen yöntemin F1-puan değerleri incelendiğinde 0,94 ve üzerinde olduğu görülmektedir. Bununla birlikte beş atak sonucunda elde edilen F1 skoru değerleri birbirine oldukça yakındır. Atağa rağmen elde edilen skorların yakın olması da önerilen yöntemin atak bağımsız bir yöntem olduğunu göstermektedir.

Bir başka deney olarak önerilen ses kopyala-yapıştır sahtecilik tespit yöntemi literatürdeki diğer çalışmalarla karşılaştırılmıştır. Bu çalışmalar Lbp [10], Dft [14], Formant [15], DCT-SVD [7] ve Mel-SIFT [17] olarak adlandırılmıştır. Bu çalışmalardan LBP, Formant ve DCT-

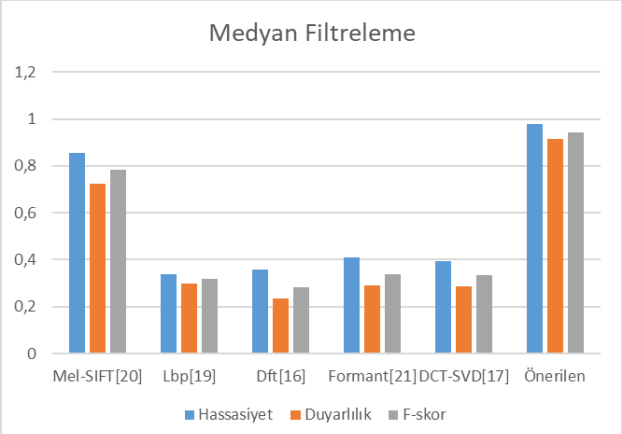
SVD yöntemleri VAD tabanlı yöntemler oldukları için çalışmalarda belirtilen ilgili VAD yöntemleri uygulanmıştır. Şekil 6 ataksız seslerle Mel-SIFT, Lbp, Dft, Formant, DCT-SVD ve önerilen yöntemden elde edilen ortalama Hassasiyet, Duyarlılık ve F-skor sonuçlarını göstermektedir.



Şekil 6. Ataksız seslerde yöntemlerin tespit sonuçları

Şekil 6'da görülebileceği gibi, önerilen yöntem Mel-SIFT, Lbp, Dft, Formant, DCT-SVD yöntemlerine göre oldukça yüksek performans göstermiştir. Önerilen yöntemden 0.98 F-skor değeri elde edilirken, önerilen yönteme en yakın F-skor değeri Mel-SIFT yöntemiyle elde edilmiştir. Mel-SIFT yöntemi dışındaki yöntemlerin performansı oldukça düşüktür.

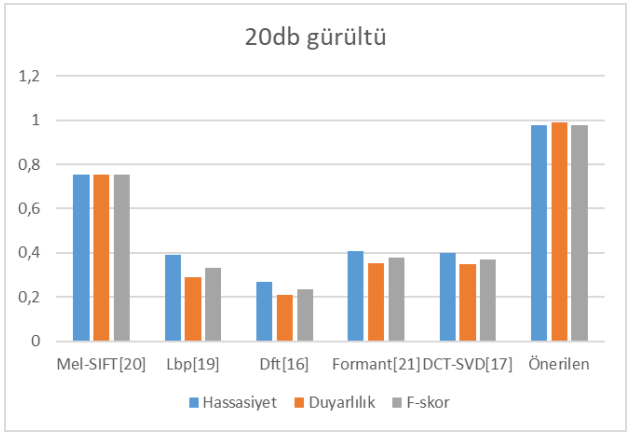
Yöntemlerin medyan filtreleme atağına karşı dayanıklılığı da analiz edilmiştir. Şekil 7, medyan filtreleme sonucunda MelSIFT, Lbp, Dft, Formant, DCT-SVD ve önerilen yönteme ait ortalama Hassasiyet, Duyarlılık ve F-skor sonuçlarını sunmaktadır.



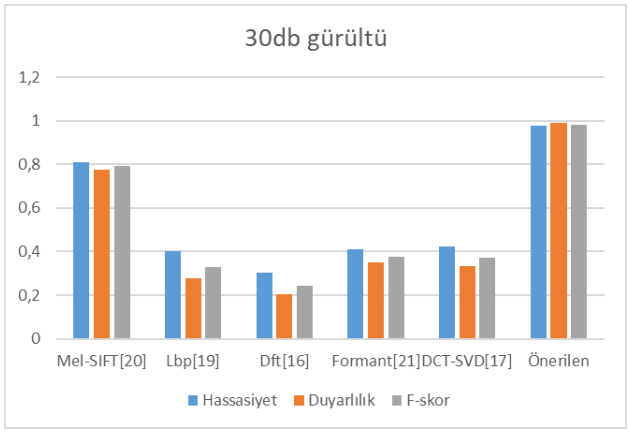
Şekil 7. Medyan filtreleme uygulanmış seslerde yöntemlerin tespit sonuçları

Şekil 7' de verilen değerler incelendiğinde en yüksek F-skor değeri 0.94' ü önerilen yöntem verirken, en düşük değer 0.28'i DFT yöntemi vermektedir. Arada oldukça büyük bir fark olduğu görülmektedir. Bu sonuçta önerilen yöntemin medyan filtreleme saldırısına karşı diğer yöntemlere göre çok daha sağlam olduğunu göstermektedir.

Gürültü atağı değerlendirmek için de önerilen yöntem ve diğer yöntemler hem 20 db hem de 30 db'lik gürültülü seslerle test edilmiştir. Elde edilen performans ölçüm sonuçları, 20db gürültü için Şekil 8(a)'da ve 30db gürültü için Şekil 8(b)'de verilmiştir.



(a)



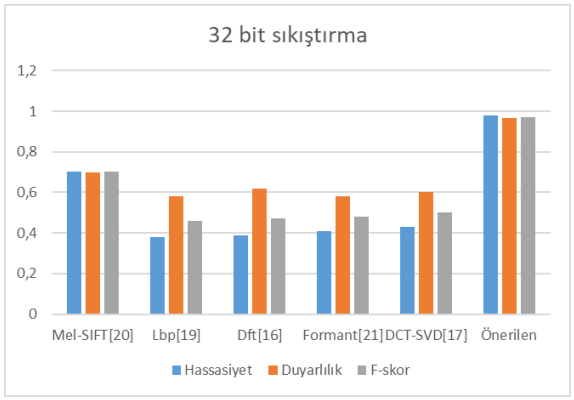
(b)

Şekil 8. Tespit sonuçları (a) 20db (b) 30db gürültü eklenmiş seslerde

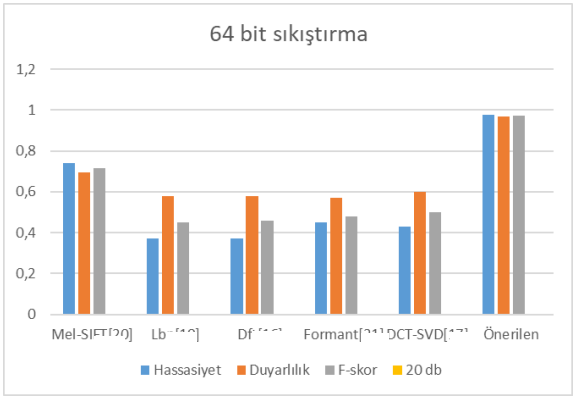
Şekil 8'den de görülebileceği gibi önerilen yöntem için 30db gürültü eklenmesi sonucunda elde edilen Hassasiyet, Duyarlılık ve F-skor değerleri sırasıyla 0.97, 0.98 ve 0.98'dir. Bu değerler oldukça yüksek değerlerdir. 20 db'lik gürültü ekleme için elde edilen Hassasiyet, Duyarlılık ve F-skor değerleri sırasıyla 0.97, 0.98 ve 0.97' dir. Gürültü miktarı artmasına rağmen değerlerin neredeyse değişmediği ve oldukça yüksek olduğu görülmektedir. Ayrıca önerilen yöntem dışında 20db ve 30db gürültü için elde edilen en yüksek F1 skoru değerleri sırasıyla 0.75 ve 0.79 olup Mel-SIFT yöntemiyle elde edilmiştir. Bu deneyden elde edilen sonuçlara göre önerilen ses kopyala-taşı sahtecilik tespit yöntemi gürültüye karşı oldukça dayanıklı ve yüksek doğruluk vermektedir.

Son olarak ise sıkıştırma atağı analiz edilmiştir. Yöntemlerden elde edilen performans metrikleri 32 kbps sıkıştırma için Şekil 9(a)'da ve 64 kbps sıkıştırma için Şekil 9(b)'de verilmektedir. Şekil 9' dan görüldüğü gibi Lbp yöntemi en düşük ortalama değerleri verirken, önerilen

yöntem hem 32 kbps sıkıştırmada hem de 64 kbps sıkıştırmada en yüksek ortalama değerleri vermektedir. Önerilen yöntemle F-skor değerleri sırasıyla 32 ve 64 bit sıkıştırma için 0.97 olarak elde edilmiştir. Farklı sıkıştırma değerlerine rağmen F-skor değeri değişmemiştir. Mel-SIFT yöntemi önerilen yönteme en yakın F-skor değerini vermiştir. Tüm analiz sonuçları değerlendirildiğinde, ses dosyalarına çeşitli ataklar uygulansa dahi önerilen yöntem, literatürdeki diğer yöntemlerle karşılaştırıldığında oldukça yüksek bir performans göstermektedir.



(a)



(b)

Şekil 9. Tespit sonuçları (a)32 bit sıkıştırma (b) 64 bit sıkıştırma uygulanmış seslerde

4 Sonuçlar

Bu çalışmada ses kopyala-yapıştır sahteciliğini tespit etmek için yeni bir yaklaşım önerilmektedir. Giriş verilerinin ses dosyası olması nedeniyle bu alanda önerilen yöntemlerde 1D özellik çıkarımı ve benzerlik hesaplama yöntemlerini kullanmaktadır. 2D özellik çıkarma ve benzerlik hesaplama yöntemlerinin çeşitliliği göz önüne alındığında, bu durum ses sahteciliği tespit alanını sınırlamaktadır. Bu amaçla ses dosyasındaki her sesli bölümden Mel spectogram görüntüleri çıkarılmıştır. Mel spectogram görüntülerinin benzerlik hesaplaması için DSSIM metriğinden faydalanılmıştır. Önerilen yöntem, TIMIT veri setinden oluşturulan sahte ses veri seti ile alandaki yöntemlerle karşılaştırılmıştır. Deneysel sonuçlar, önerilen ses kopyala-yapıştır sahteciliği tespit yönteminin ilgili alanda üstün performans sağladığını, etkili ve atak bağımsız olduğunu göstermektedir.

Çıkar çatışması

Yazarlar çıkar çatışması olmadığını beyan etmektedir.

Benzerlik oranı (iThenticate): %9

Kaynaklar

[1] S. Keser, Ö. N. Gerek, E. Seke, M. B. Gülmezoğlu, A subspace based progressive coding method for speech compression. Speech Communication, 94, 50-61, 2017. https://doi.org/10.1016/j.specom.2017.09.002

[2] S. Keser, R. Edizkan, Phonem-based isolated Turkish word recognition with subspace classifier. In 2009 IEEE 17th Signal Processing and Communications Applications Conference, pp. 93-96, Antalya, Turkey 2009.

[3] O. F. Çıplak, S. Kevser, Gerçek zamanlı ses tanıma ile robot kolu kontrolü. Avrupa Bilim ve Teknoloji Dergisi, 31, 34-39, 2021. https://doi.org/10.31590/ejosat.969608

[4] J. N. Xiao, Y. Z Jia, E. D Fu., Z. Huang, Y. Li, & S. P. Shi, Audio authenticity: Duplicated audio segment detection in waveform audio file. Journal of Shanghai Jiaotong University (Science), 19, 392-397, 2014. https://doi.org/10.1007/s12204-014-1515-5

[5] Z. Su, M. Li, G. Zhang, Q. Wu, & Y. Wang, Robust audio copy-move forgery detection on short forged slices using sliding window. Journal of Information Security and Applications, 75, 103507, 2023. https://doi.org/10.1016/j.jisa.2023.103507

[6] Z. Su, M. Li, G. Zhang, Q. Wu, M. Li, W. Zhang & X. Yao, Robust Audio Copy-Move Forgery Detection Using Constant Q Spectral Sketches and GA-SVM. IEEE Transactions on Dependable and Secure Computing, 2022. https://doi.org/10.1109/TDSC.2022.3215280

[7] F. Wang, C. Li, L. Tian, An algorithm of detecting audio copy-move forgery based on DCT and SVD. In: 2017 IEEE 17th International Conference on Communication Technology (ICCT). IEEE, pp. 1652–1657, Chengdu, China, 2017.

[8] Q. Yan, R. Yang, J.Huang, Copy-move detection of audio recording with pitch similarity. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 1782–1786, Brisbane, Australia, 2015.

[9] Z. Xie, W. Lu, X. Liu, Y. Xue, Y. Yeung, Copy-move detection of digital audio based on multifeature decision. Journal of Information Security and Applications, 43, 37-46, 2018. https://doi.org/10.1016/j.jisa.2018.10.003

[10] M. Imran, Z. Ali, S.T. Bakhsh, S. Akram, Blind detection of copy-move forgery in digital audio forensics. IEEE Access 5, 12843–12855, 2017. https://doi.org/10.1109/ACCESS.2017.2717842

[11] N.T. Anh, H.T.T. Hang, G. Chen, One approach in the time domain in detecting copy-move of speech recordings with the similar magnitude. International Journal of Engineering and Applied Sciences (IJEAS), 6(4), 9–11, 2019. https://dx.doi.org/10.31873/IJEAS/6.4.2019.05

[12] K. Mannepalli, P. Krishna, K. Krishna, Copy and move detection in audio recordings using dynamic time warping algorithm. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9(2), 2244–2249, 2019. http://doi.org/10.35940/ijitee.B6678.129219

[13] B. Ustubioglu, B. Kucukugurlu, G. Ulutas, Robust copy-move detection in digital audio forensics based on pitch and modifed discrete cosine transform. Multimedia Tools and Applications, 81, 27149–27185, 2022. https://doi.org/10.1007/s11042-022-13035-3

[14] X. Huang, Zi Liu, W. Lu, H. Liu, S. Xiang, Fast and effective copy-move detection of digital audio based on auto segment. In: Digital forensics and forensic investigations: breakthroughs in research and practice. IGI Global, pp. 127–142, 2020.

[15] Q. Yan, R. Yang, J. Huang, Robust copy–move detection of speech recording using similarities of pitch and formant. IEEE Transactions on Information Forensics and Security, 14(9), 2331–2341, 2019. https://doi.org/10.1109/TIFS.2019.2895965

[16] A. Ustubioglu, B. Ustubioglu, G. Ulutas, Mel spectrogram-based audio forgery detection using CNN. Signal, Image and Video Processing, 17, 2211 - 2219, 2022. https://doi.org/10.1007/s11760-022-02436-4

[17] B. Ustubioglu, G. Tahaoglu, G. Ulutas, Detection of audio copy-move-forgery with novel feature matching on Mel spectrogram. Expert Systems with Applications 213, 118963, 2023. https://doi.org/10.1016/j.eswa.2022.118963

[18] B. Ustubioglu, G. Tahaoglu, G. Ulutas, A. Ustubioglu & M. Kilic, Audio forgery detection and localization with super-resolution spectrogram and keypoint-based clustering approach. The Journal of Supercomputing, 80, 486-518, 2023. https://doi.org/10.1007/s11227-023-05504-9

[19] A. Stephen, H. Hu, A spectral/temporal method for robust fundamental frequency tracking, The Journal of the Acoustical Society of America, 123:6, pp:4559-4571, 2008. https://doi.org/10.1121/1.2916590

[20] Z. Wang, A.C, Bovik Sheikh, H.R., Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing. 13:4, 600–612, 2004. https://doi.org/10.1109/TIP.2003.819861

[21] TIMIT Acoustic-Phonetic Continuous Speech Corpus, https://catalog.ldc.upenn.edu/LDC93s1