*Avrupa Bilim ve Teknoloji Dergisi*
*Özel Sayı, S. 398-406, Ağustos 2020*

***Araştırma Makalesi***

www.ejosat.com ISSN:2148-2683

*European Journal of Science and Technology*
*Special Issue, pp. 398-406, August 2020*

***Research Article***

# Sentetik ve Dönüştürülmüş Konuşmaların Tespitinde Genlik ve Faz Tabanlı Spektral Özniteliklerin Kullanılması

Burak Kasapoğlu[1], Turgay Koç[2]

[1]Süleyman Demirel Üniversitesi, Mühendislik Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, Isparta, Türkiye (ORCID: 0000-0003-3580-0465)
[2]Süleyman Demirel Üniversitesi, Mühendislik Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, Isparta, Türkiye (ORCID: 0000-0002-4846-7772)

*(Bu yayın 26-27 Haziran 2020 tarihinde HORA-2020 kongresinde sözlü olarak sunulmuştur.)*

(**DOI:** 10.31590/ejosat.780650)

**ATIF/REFERENCE:** Kasapoğlu, B. & Koç, T. (2020). Sentetik ve Dönüştürülmüş Konuşmaların Tespitinde Genlik ve Faz Tabanlı Spektral Özniteliklerin Kullanılması. *Avrupa Bilim ve Teknoloji Dergisi*, (Special Issue), 398-406.

**Öz**

Teknolojideki gelişmeyle birlikte güvenlik ihtiyacı bulunan uygulamalarda kişisel erişimi sağlayabilmek amacıyla parmak izi, retina, yüz, ses gibi kişiden kişiye değişiklik gösteren biyometrik sinyallerin kullanımı gün geçtikçe yaygınlaşmaktadır. Bu biyometrik sinyallerden ses yani konuşma sinyalinin hem kişiden kolaylıkla elde edilebilir olması hem de yüksek mobilite sağlaması otomatik konuşmacı doğrulama (Automatic Speaker Verification – ASV) sistemlerini popüler hale getirmektedir. ASV sistemlerinin güvenlik alanlarında yaygınlaşmasıyla birlikte bu sistemleri yanıltmaya yönelik farklı saldırı yöntemleri geliştirilerek bu saldırıların ASV sistemleri için ciddi birer tehdit oluşturduğu gözlenmiştir. Bu çalışmada, ASV sistemlerine en büyük tehdit oluşturan yöntemlerden ikisi olan ses sentezi ve ses dönüştürme yöntemleri kullanılarak ASV sistemlerine yapılan saldırıların tespit edilebilmesi için yeni bir sistem önerilmiştir. Önerilen sistemde, daha önce ses dönüştürme ve ses sentezleme yöntemiyle üretilen sahte seslerin tespit edilebilmesi amacıyla 2015 yılında düzenlenmiş olan ASVSpoof yarışmasında en iyi performansı gösteren genlik spektrumu tabanlı anlık Q kepstral katsayıları (Constant Q Cepstral Coefficients – CQCC) özniteliği ile konuşma sinyalinin ters filtrelenmesiyle elde edilen gırtlak akımına ait faz bilgisi içeren değiştirilmiş grup gecikmesi (Glottal Flow Modified Group Delay – GFMGD) özniteliği birlikte kullanılarak Gauss Karışım Modeli tabanlı sınıflandırma sistemi oluşturulmuştur. Doğrudan gerçek ses parçaları kullanılarak üretilen sahte seslerin sınıflandırılmasında hem CQCC tabanlı temel sistem hem de önerilen sistem için sistem performansları arasında belirgin bir fark görülmeyip her iki sistem de %1'in altında sınıflandırma hatası göstermiştir. Ancak, dalga form filtreleme ile üretilen sahte seslerin sınıflandırılmasında her iki sistem de benzer şekilde diğer saldırı yöntemlerine göre daha zayıf performans göstermiştir. Önerilen sistem, sadece CQCC kullanan temel sistem ile kıyaslandığında özellikle son yıllarda geliştirilmiş olan modern yapay sinir ağları ve ses kodlayıcılar tarafından sentezlenen ya da dönüştürülen konuşma sinyallerine karşı %55'e kadar performans artışı sağlayabilmektedir.

**Anahtar Kelimeler:** Konuşmacı Tanıma Sistemleri, Konuşma İşleme, Sahte Konuşmacı Algılama Sistemleri

# Using Magnitude and Phase Based Spectral Features for Detecting Synthetic and Converted Speech

**Abstract**

With the advancement in technology, the use of biometric signals that differ from person to person such as fingerprint, retina, face, and voice is becoming more popular in order to provide personal access in applications that need security. The fact that among these biometric signals, voice, that is, speech signal can be easily obtained from the person and provides high mobility make automatic speaker verification (ASV) systems popular. Due to the widespread use of ASV systems in security applications, different spoofing attack methods have been developed to mislead these systems and it is observed that these developed spoofing attack methods pose a serious threat to ASV systems. In this study, a new system is proposed to detect the spoofing attacks using speech synthesis and voice conversion methods, which are two of the biggest threats to ASV systems. Proposed system uses Gaussian Mixture Model based classifier using the fusion of magnitude spectrum based constant Q cepstral coefficients (CQCC), that was chosen as best countermeasure feature of ASVSpoof challenge for detection of speech produced with speech synthesis and voice conversion methods, and glottal flow modified group delay (GFMGD) feature, that contains phase spectrum information of glottal flow obtained by applying inverse filtering on speech signal. In the classification of spoof speech produced by using genuine speech signals, due to both systems having classification error below 1%, it is not found any major difference in classification performance between proposed system and CQCC based baseline system. However, in the classification of spoof speech produced by using waveform filtering method both systems similarly performed poorly compared to other attacking methods. On the other hand, the proposed system can provide up to 55% performance increase against speech signals synthesized or converted by modern artificial neural networks and audio vocoders compared to the baseline system using only CQCC.
**Keywords:** Speaker Recognition Systems, Speech Processing, Spoofing Attack Detection Systems.

# 1. Giriş

Günümüzde insanlar kişisel verilerini saklamak ya da bazı bölgelere belirli kişilerin erişebilmesini sağlamak adına kişiye özel sinyalleri baz alan sistemler kullanmaktadır. Parmak izi, retina, yüz, ses vb., kişiden kişiye değişiklik gösteren bilgilere biyometrik sinyal denir. Konuşma sinyalinin kişiden kolayca elde edilebilmesi kişinin konuşmasındaki biyometrik özellikleri kullanan otomatik konuşmacı doğrulama (Automatic Speaker Verification – ASV) sistemlerinin kullanımını popüler hale getirmektedir. ASV sistemlerinin popülerleşmesiyle birlikte bu sistemlere karşı kötü niyetli kişilerin kullandığı yeni saldırı yöntemleri oluşmakta ve teknolojideki gelişimle birlikte bu saldırı yöntemleri gün geçtikçe daha etkili hale gelmektedir. Son yıllarda ASV sistemlerinin, gelişen saldırı yöntemlerine karşı giderek savunmasız kaldığı görülmüş ve bu alanda çalışan bilim insanları tarafından düzenlenen yarışmalar ve konferanslarla bu konuda farkındalık yaratılarak ilgili çalışmalara hız kazandırılmıştır. Yapılan çalışmalarda genellikle 4 farklı saldırı yöntemi üzerinde durulmaktadır [1]. Bunlar; ses taklidi [2-3], kayıttan çalma [4], ses sentezi [5] ve ses dönüştürme [6] yöntemleridir. Bahsi geçen bu yöntemlerden özellikle ses sentezi, ses dönüştürme ve kayıttan çalma yöntemleri ASV sistemleri için ciddi birer tehdit oluşturmaktadır. Bu yöntemlerden uygulanabilirliği en basit olan kayıttan çalma yöntemidir. Ancak ASV sisteme erişim için kullanılan parolanın değişmesi durumunda bu yöntem efektifliğini kaybeder. Diğer yandan ses sentezi ve ses dönüştürme yöntemleri daha geniş dağarcıklı ses üretilmesini sağlayabildiği için daha büyük tehdit oluşturabilir. Ses sentezi ve ses dönüştürme yöntemleri bazı temel farklar dışında birbirlerine benzerlik göstermektedir. Her iki saldırı tipi için de amaç, kayıttan çalma yönteminden farklı olarak, daha önceden hedef konuşmacıdan elde edilen orijinal ses kayıtlarını işleyerek ASV sistemini yanıltacak sentetik veya dönüştürülmüş konuşma sinyalleri üretmektir.

Sentetik ses tabanlı saldırılar çoğunlukla yazıdan konuşmaya dönüştürme (text-to-speech – TTS) sistemleriyle ilişkili bir tekniktir. Buradaki amaç belirli bir metinden yola çıkarak anlaşılabilir, kulağa doğal gelen bir yapay ses oluşturmaktır. Ses sentezi yöntemi günümüzde araçlardaki navigasyon sistemleri, konuşabilen robotlar gibi uygulamalarda aktif olarak kullanılmaktadır. Bir ses sentezleyici sistem genel olarak iki ana kısımdan oluşur. Bunlardan biri metin analiz kısmı ve diğeri ise dalga formu üretecidir. Sistem girişine gelen metin analiz edilerek ne söylenmek istediği anlaşılır ve konuşulacak dile ait özellikler belirlenir. Daha sonrasında dalga formu üreteci bu özellikleri baz alarak söylenmek istenen metni, konuşmaya dönüştürür. 90'lı yılların ikinci yarısında birim seçme (Unit Selection–US) yöntemiyle doğal ses parçaları kullanılarak sentetik konuşma üreten sistemler popüler olmuştur. Ancak US yönteminin konuşmacıya bağlı olması ve çok büyük miktarda konuşma kaydına ihtiyaç duyulması sebebiyle 2000'li yılların sonuna doğru daha az konuşma kaydı ile daha esnek yapıya sahip istatiksel parametrik konuşma sentezi (statistical parametric speech synthesis – SPSS) [7-10] kullanılmaya başladı. 2010'lu yılların başında derin öğrenme yöntemlerinin de gelişmesi ve popülerleşmesiyle birlikte daha doğal ve yüksek kaliteli ses üretebilen ses sentezleyiciler gerçekleştirilmeye başlanmıştır. Geliştirilen bu modern sistemler konuşma üretmek için Wavenet ve Neural source filter modeli gibi gelişmiş derin yapay sinir ağları ve WORLD gibi ses kodlayıcılar kullanmaktadır [11-12]. ASV sistemlerin ses sentezi yöntemiyle oluşturulan sahte seslere karşı zayıflığı farkedildikten sonra bu konu geniş kapsamlı olarak ilk defa ASVSpoof 2015 yarışmasında ele alınmış ve benzer saldırılara karşı ASV sistemleri koruyabilecek bazı yöntemler geliştirilmeye başlanmıştır. Bugüne kadar ses sentezi yöntemi ile oluşturulmuş seslerin tespit edilebilmesi için genellikle sentezlenen sesler ve gerçek insan sesleri arasındaki farklar incelenmiştir. Örneğin, insan işitme sisteminin faz bilgisine karşı duyarlı olmaması sebebiyle ses sentezleyicilerin genellikle minimum-faz sistem olarak tasarlanması sonucu doğal ses ile sentetik ses arasında faz farklılıklarının meydana geldiği bilinmektedir [13-16]. Bu bilgiden yola çıkılarak gerçek insan sesi ve sentezlenen sesler arasındaki faz spektrumu tabanlı farklar öznitelik olarak kullanılabilir. Faz spektrumu tabanlı özniteliklere ek olarak genlik spektrumu tabanlı öznitelikler de sahte seslerin ayırt edilebilmesi açısından başarılı performans göstermektedir.

ASV sistemlerine yüksek seviyede tehdit oluşturan bir diğer saldırı yöntemi ise ses dönüştürme (voice conversion – VC) saldırılarıdır. Bu tip saldırılarda saldıran kişinin doğal sesi, sistemin geçiş izni vereceği bir hedef konuşmacının sesine dönüştürülmeye çalışılır [17]. VC sistemlerin, sahte konuşma üretmek dışında kullanıldığı popüler alanlar olarak film dublajları, şarkı kayıtlarındaki hataları otomatik olarak gideren "auto tune" sistemleri sayılabilir. VC sistemleri kaynak konuşmacı ile hedef konuşmacının akustik parametreleri arasında oluşturulan bir transfer fonksiyonu ve bu transfer fonksiyonundan elde edilen çıktıları kullanarak konuşma sinyali sentezleyen dalga formu üreteci olmak üzere iki ana kısımdan oluşur. 2000'li yıllarda genellikle Gauss Karışım Modeli (GKM) [18-19] tabanlı VC yöntemleri kullanılıyordu. Fakat bu yöntemler popüler olmasına rağmen konuşmayı aşırı yumuşatması veya aşırı eğitim gibi sebeplerden dolayı seste boğukluğa ve hedef konuşmacı benzerliğinde bozukluklara sebebiyet veren kritik hatalar barındırmaktaydı [20-24]. 2010'lu yılların başlangıcından itibaren akustik parametrelerine ait transfor fonksiyonunun elde edilmesi ve konuşma sentezlenmesi için TTS sistemlerinde olduğu gibi modern derin yapay sinir ağları kullanılmaktadır. Dönüştürülmüş ve sentetik seslerin tespit edilebilmesi amacıyla düzenlenen ASVSpoof 2015 yarışmasında problemin çözümü için mel-frekansı kepstral katsayıları (mel-frequency cepstral coefficients – MFCC), doğrusal frekans kepstral katsayıları da (linear frequency cepstral coefficients – LFCC) ve anlık Q kepstral katsayıları (constant Q cepstral coefficients – CQCC) gibi öznitelikler önerilmiş, bunlar arasında en iyi performans CQCC özniteliği ile elde edilmiştir [25]. CQCC tek başına en iyi performans gösteren öznitelik olmakla birlikte faz tabanlı bir öznitelikle birleştirildiğinde performansının nasıl etkileneceği bir araştırma konusudur. Bu amaçla yapılan çalışmada genlik spektrum tabanlı CQCC ile birlikte konuşmanın kaynağı olan gırtlak akımının (GA) faz bilgisini içeren değiştirilmiş grup gecikmesi (Glottal Flow Modified Group Delay – GFMGD) özniteliği kullanılması durumunda sahte konuşmacı algılama sisteminin performansı incelenmiştir.

Çalışmanın devamındaki konu akışı; Bölüm 2'de çalışmada kullanılmış olan öznitelikler hakkında genel bilgi ve nasıl elde edileceği hakkında bilgi verilmesi, Bölüm 3'de kullanılan veritabanı hakkında detaylı bilgi ve öznitelikler elde edilirken kullanılan konfigürasyonlar hakkında detaylı bilgi paylaşılması, Bölüm 4'de yapılan çalışmalar sonucunda elde edilen sonuçların paylaşılması, Bölüm 5 'de sonuçların değerlendirilmesi ve yapılacak çıkarımlar ve son olarak Bölüm 6'da çalışmanın geliştirilebilmesi açısından geleceğe yönelik iyileştirme çalışmaları olacak şekilde planlanmıştır.

### 1.1. Sahte Konuşmacı Algılama Sistemi

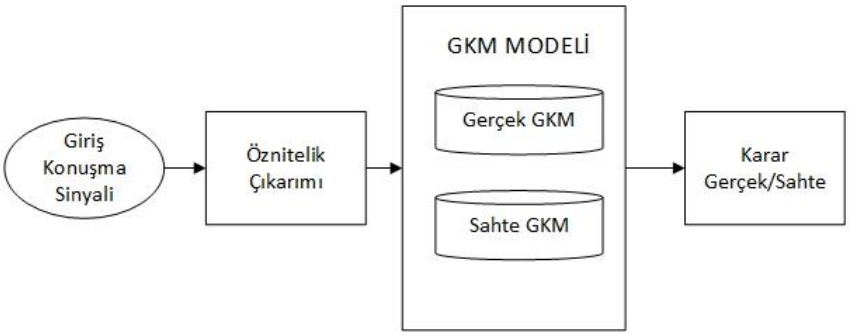
Bu çalışmada geliştirilen sahte konuşmacı algılama sistemi sistem girişi, öznitelik çıkarımı, sınıflandırıcı skor elde edilmesi ve son olarak skorların değerlendirilerek karar verilmesi aşamalarından oluşmaktadır. İlk olarak sistemin eğitim aşamasında mevcut veritabanındaki gerçek ve sahte konuşma kayıtları için ilgili öznitelikler elde edilip bu öznitelikler kullanılarak gerçek ve sahte konuşma kayıtları için ayrı olarak GKM modelleri oluşturulmaktadır. Test aşamasında ise sistem girişine gelen bir konuşma sinyali üzerinden aynı öznitelikler elde edilerek bu bilgiler hem sahte hem de gerçek GKM'lere giriş olarak verilmektedir. Son aşamada ise giriş sinyali için GKM'lerin oluşturdukları olasılık değerleri birbirleriyle kıyaslanmakta ve sinyalin gerçek mi yoksa sahte mi olduğu Bayes Karar Kuralı'na göre belirlenmektedir. Bu amaçla (1) (2) ve (3) numaralı denklemler kullanılarak karar verilir.

$$\theta\,(gerçek) = log(P(\Delta|\rho_{gerçek})) \tag{1}$$

$$\theta\,(sahte) = log(P(\Delta|\rho_{\text{sahte}})) \tag{2}$$

$$\theta\,(final) = \theta\,(gerçek) - \theta\,(sahte) \tag{3}$$

Burada $\Delta$ giriş sinyalinden elde edilen öznitelik vektörünü, $\rho_{\text{gerçek}}$ ve $\rho_{\text{sahte}}$ gerçek ve sahte sesler için oluşturulmuş olan GKM modellerini ve $\theta$ ise elde edilen skoru temsil etmektedir. Geliştirilen sistemle alakalı temel bir blok diyagram Şekil 1.'de gösterilmektedir.




*Şekil 1. Önerilen sahte konuşmacı algılama sistemi*

## 2. Kullanılan Öznitelikler

Bu çalışmada, konuşma sinyalinin anlık Q cepstral katsayıları (CQCC) ve konuşma sinyalinden elde edilen gırtlak akımına (GA) ait değiştirilmiş grup gecikmesi (GFMGD) olmak üzere sırasıyla genlik ve faz spektrumu tabanlı iki farklı öznitelik kullanılmıştır.

### 2.1. Anlık Q Kepstral Katsayıları (Constant Q Cepstral Coefficients – CQCC)

CQCC istenilen herhangi bir frekans bandı için yüksek çözünürlüklü bir Anlık Fourier Dönüşümü (STFT) olarak düşünülebilir. Bu yaklaşım alçak frekanslarda yüksek frekans çözünürlüğü sağlarken yüksek frekanslarda ise yüksek zamansal çözünürlük sağlamaktadır. Ayrık zamanlı bir sinyal olan x[n]'e ait Anlık Q Dönüşümü (CQT), $X^{CQ}$ (k, n) ifadesi (4) numaralı denklem sayesinde elde edilebilir.

$$X^{CQ}(k,n) = \sum_{j=n-\left\lfloor \frac{N_k}{2} \right\rfloor}^{n+\left\lfloor \frac{N_k}{2} \right\rfloor} x(j) a_k^*\left(j - n + \frac{N_k}{2}\right)\ k = 1,2,\ldots.K \tag{4}$$

Burada k frekans indeksi, $a_k^*(n)$, $a_k(n)$'in kompleks konjugesi ve $N_k$ ise değişken pencere uzunluklarını ifade etmektedir. $\lfloor . \rfloor$ notasyonu en yakın küçük tam sayıya yuvarlama anlamına gelmektedir. $a_k(n)$ basit fonksiyonları ise kompleks değerli zaman-frekans atomlarını ifade etmektedir ve (5) numaralı denklem yardımıyla hesaplanabilir.

$$a_k(n) = \frac{1}{C}\left(\frac{n}{N_k}\right) e^{\left[j\left(2\pi n \frac{f_k}{f_s} + \Phi_k\right)\right]} \tag{5}$$

Burada $f_k$, frekans indeksi k'ya karşılık gelen merkez frekansı değerini, $f_s$ örnekleme frekansını, $\Phi_k$ faz offset değerini, $\omega(t)$ pencereleme fonksiyonunu, ve $\beta$ ise her oktavdaki atom sayısını ifade etmektedir. Elde edilen bu bilgiye göre ölçekleme faktörü C, Q faktörü ve değişken pencere uzunlukları $N_k$ aşağıdaki (6) (7) ve (8) numaralı denklemlerle elde edilebilir [28]:

$$C = \sum_{l=-\lfloor N_k/2 \rfloor}^{\lfloor N_k/2 \rfloor} \omega\left(\frac{l + N_k/2}{N_k}\right) \tag{6}$$

$$Q = \frac{f_k}{f_{k+1} - f_k} = (2^{1/B} - 1)^{-1} \quad (7)$$

$$N_k = \frac{f_s}{f_k} Q \quad (8)$$

$X^{CQ}$ (k, n) ifadesini hesapladıktan sonra, [28]'da belirtildiği şekilde x[n]'e ait CQCC (9) numaralı denklemde gösterildiği şekilde elde edilebilir.

$$CQCC(p) = \sum_{l=1}^{L} log|X^{CQ}(l)|^2 \cos\left[\frac{p(l-1/2)\pi}{L}\right] \quad (9)$$

Çalışmalar sırasında kullanılan CQCC özniteliği 0. Kepstral katsayılar dahil olmak üzere 29 statik,29 delta ve 29 delta delta katsayıları kullanılarak toplamda 90 boyutlu olacak şekilde belirlenmiştir.

## 2.2. Gırtlak Akımının Değiştirilmiş Grup Gecikmesi (Glottal Flow Modified Group Delay - GFMGD)

Konuşma sinyali, havanın gırtlaktan geçerken ses tellerinin açılıp kapanması sonucu meydana gelen gırtlak akımı veya ağız boşluğu yani vokal boşluk içerisindeki sıkıştırmalarda meydana gelen türbülanslar sonucu oluşup dil, çene ve diş gibi artikülatörlerin oluşturduğu akustik rezonatörün spektral olarak şekillendirmesi ve dudakların etkisiyle son halini alır. Vokal boşluğun bu etkisi dinamik konuşma için doğrusal ve zamanla değişen bir sistem olarak düşünülür. Ancak konuşma sırasında artikülatörlerin yavaş hareketinden dolayı yaklaşık 20ms-40ms aralığında doğrusal ve zamanla değişmeyen bir sistem olarak kabul edilir. Bu nedenle konuşma sinyalinin küçük parçalarının doğrusal ve zamanla değişmeyen bir sistem ile üretildiği varsayılarak konuşma sinyali üzerindeki vokal boşluğun akustik etkisini ortadan kaldıracak bir ters filtre elde edilebilirse ses tellerinin arasındaki gırtlak akımı dalga şekli elde edilebilir. Konuşmanın doğrusal kaynak süzgeç modeli (10) numaralı denklemdeki gibi ifade edilir.

$$s[n] + \sum_{k=1}^{N} a_k s[n-k] = u_g[n] \quad (10)$$

Burada, $s[n]$ konuşma sinyali, $a_k$'lar vocal boşluk filtresinin parametreleri ve $u_g[n]$ ise gırtlak akımını göstermektedir. Bu modelde dudaklar ayrıca 1. dereceden bir yüksek geçiren vurgu filtresi ile modellenir. Eğer dudakların etkisi (10) numaralı denklem içine alınırsa $u_g[n]$ GA'nın yaklaşık türevini gösterir [29]. Vokal boşluk, her ses için farklı şekilde olduğundan dolayı rezonanslar değişim göstermektedir. Bu nedenle $a_k$'lar her ses için farklı değer almaktadır. Ayrıca vokal boşluk kişiden kişiye değiştiği için aynı ses türlerinde bile farklılık gösterir. Benzer biçimde vokal boşluğun etkisi çıkarıldığında elde edilen gırtlak akımı sesten sese ve kişiden kişiye değişmektedir. Bu nedenle gırtlak akımı konuşmacı tanıma amacıyla kullanılabilir. Konuşma üzerinden GA sinyalinin elde edilebilmesi için $a_k$'ların doğru olarak belirlenmesi gerekir. $a_k$'lar belirlendikten sonra $U_g(Z) = S(Z)A(Z)$ biçiminde ters filtreleme yapılarak $u_g[n]$ elde edilir. Burada ters sistem transfer fonksiyonu tamamen sıfırlardan oluşup $A(Z) = \sum_{k=0}^{N} a_k z^{-k}$ biçiminde ifade edilir ($a_0 = 1$). Filtre katsayılarının belirlenmesi için literatürde bir çok yöntem önerilmiştir. Bu yöntemler ses tellerinin kapanma anları bilgisinin var olması durumunda en iyi performansı göstermektedirler. Eğer ses tellerinin kapanma anlarının bilgisi var ise ses tellerinin kapalı olduğu durumda GA sıfır olduğundan $a_k$'lar kovaryans analizi ile tespit edilebilir. Eğer bu bilgi mevcut değilse GA ve vokal boşluk modelleri kullanılarak kestirim gerçekleştirilebilir. Fakat bu durumda filtre katsayılarının tespitinde hatalar olabilmektedir ve bunun sonucunda beklenilenden farklı sonuçlar gözlenebilir. Filtre katsayılarının en doğru biçimde elde edilebilmesi için ses tellerinin kapanma ve açılma anlarının işaretlenmesi daha uygundur. Fakat bu çalışmada kullanılan ASVSpoof 2019 veritabanındaki veri sayısının çok fazla olması sebebiyle ses telleri kapanma bilgisini kullanmayan ve literatürde yaygın olarak tercih edilen Yinelemeli Adaptif Ters Filtreleme (IAIF) algoritması GA sinyalinin kestirilmesinde kullanılmıştır [26]. IAIF algoritması iki aşamadan oluşur. İlk aşamada birinci dereceden IIR bir filtre kullanılarak glottal akımın katkısı ve vokal boşluğun transfer fonksiyonu spectral olarak elde edilir ve konuşma sinyali bu bilgiler kullanılarak ters filtrelenir. İkinci aşamada ise yine aynı adımlar uygulanır fakat bu sefer ilk adımda kullanılan doğrusal tahmin katsayısı p yerine farklı bir doğrusal tahmin katsayısı q kullanılır. IAIF için 8 kHz örnekleme frekansında kullanılan doğrusal tahmin katsayıları p=10, q=6 ve 16kHz örnekleme frekansında kullanılan doğrusal tahmin katsayıları ise p=18, q=10'dur [26-27].

Konuşma sinyali üzerinden GA sinyali elde edildikten sonra faz tabanlı bir bilgi elde edebilmek için değiştirilmiş grup gecikmesi (MGD) kullanılmıştır. GFMGD sinyali GA'nın MGD'si hesaplanarak elde edilir. MGD'nin hesaplaması için ilk olarak grup gecikmesi (GD) hesaplanmalıdır. GD fonksiyonu faz spektrumunun lineerliğini temsil etmekle birlikte, spektrumun $\omega$ açısal frekansına göre negatif türevi cinsinden ifade edilebilir [30-31]. MGD fonksiyonu (11) numaralı denklemde gösterildiği gibi hesaplanabilir.

$$\tau(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + X_I(e^{j\omega})Y_I(e^{j\omega})}{|X(e^{j\omega})|^2} \quad (11)$$

Burada $X(e^{j\omega})$ve $Y(e^{j\omega})$sırasıyla $u_g[n]$ ve $n.u_g[n]$ sinyallerine ait Short-Time-Fourier-Transform (STFT) spektrumunu, $X_R(e^{j\omega}), Y_R(e^{j\omega}), X_I(e^{j\omega})$ ve $Y_I(e^{j\omega})$ ise spektrumlara ait reel ve imajiner kısımları ifade etmektedir.

Daha düzgün bir faz spektrumu yapısı elde edilebilmesi adına ilgili denklemde $|X(e^{j\omega})|^2$ yerine medyan filtre ile yumuşatılmış $|S(e^{j\omega})|^2$ile birlikte iki adet optimizasyon parametresi rho(ρ) ve gamma (γ) kullanılmıştır. (12) ve (13) numaralı denklemlerde rho ve gamma optimizasyon parametreleri kullanılarak yapılan hesaplamalar ifade edilmektedir.

$$\tau_{\rho}(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + X_I(e^{j\omega})Y_I(e^{j\omega})}{|S(e^{j\omega})|^{2\rho}} \tag{12}$$

$$\tau_{\gamma,\rho}(e^{j\omega}) = \frac{\tau_{\rho}(e^{j\omega})}{|\tau_{\rho}(e^{j\omega})|}|\tau_{\rho}(e^{j\omega})|^{\gamma} \tag{13}$$

$\tau_{\gamma,\rho}(e^{j\omega})$ değiştirilmiş grup gecikmesi spektrumu olmak üzere, veri setine göre optimize edilmiş rho ve gamma katsayıları sırasıyla 0.4 ve 0.6 olarak seçilmiştir. MGD spektrumu elde edildikten sonra spektrum üzerinde ayrık kosinüs dönüşümü uygulanarak 0. Katsayı hariç olmak üzere ilk 12 statik, delta ve delta-delta cepstral katsayıları kullanılarak 36 boyutlu GFMGD özniteliği kullanılmıştır.

# 3. Yapılan Çalışmalar

## 3.1. Veritabanı

Bu çalışmada sentetik ve dönüştürülmüş konuşma veri kaynağı olarak ASVSpoof 2019 lojik erişim senaryosuna ait veritabanı kullanılmıştır. Lojik erişim senaryosu için oluşturulan veriseti, 3 ayrı alt setten oluşmaktadır. Bu alt setler sırasıyla eğitim, test ve doğrulama setleri olarak sayılabilir. Eğitim, geliştirme ve doğrulama setleri içerisinde bulunan gerçek ve sahte ses kayıtlarının sayısı Tablo 1.'de paylaşılmıştır.

*Tablo 1. ASVSpoof 2019 lojik erişim senaryosu alt verisetlerindeki gerçek ve sahte ses kaydı sayıları*

| **Veri Seti** | **#Toplam Ses Kaydı** | **#Gerçek Ses Kaydı** | **#Sahte Ses Kaydı** |
|---|---|---|---|
| Eğitim | 25380 | 2580 | 22800 |
| Geliştirme | 24844 | 2548 | 22296 |
| Doğrulama | 71183 | 7355 | 63828 |

Tablo-1 incelendiğinde veritabanının yaklaşık 120 bin farklı ses kaydından oluştuğu görülmektedir. Bu veritabanı farklı yazıdan-konuşmaya ve konuşma sentezi yöntemlerine ait konfigürasyonları barındıran A01 'den A19 'a kadar numaralandırılmış algoritmalar kullanılmıştır. Bu algoritmalardan ilk 6 tanesi eğitim ve geliştirme setleri için, kalan 13 tanesi ise doğrulama seti için kullanılmış olup detaylar Tablo – 2'de paylaşılmıştır. Doğrulama veritabanı oluşturulurken daha önce eğitim ve geliştirme veritabanında bulunan A04 ve A06 algoritmalarına büyük ölçüde benzeyen A16 ve A19 algoritmalarına ek olarak daha önce sistemin karşılaşmamış olacağı 11 farklı algoritma kullanılmıştır. Eğitim, geliştirme ve doğrulama veritabanlarının oluşturulması için kullanılan algoritmalar hakkında detaylı bilgiler [32-33]'de bulunmaktadır.

*Tablo 2. Lojik erişim senaryosu oluşturulurken kullanılan veritabanı konfigürasyonları*

| **Lojik Erişim Senaryosu Sahte Ses Oluşturma Algoritmaları** | | | |
|---|---|---|---|
| **Veritabanı** | **Algoritma Numarası** | **Saldırı Tipi** | **Dalga Formu Üreteci** |
| EĞİTİM + GELİŞTİRME | A01 | TTS | Yapay sinir ağı dalga formu modeli |
| | A02 | TTS | Ses kodlayıcı |
| | A03 | TTS | Ses kodlayıcı |
| | A04 | TTS | Dalga formu uc uca ekleme |
| | A05 | VC | Ses kodlayıcı |
| | A06 | VC | Spektral filtreleme |
| DOĞRULAMA | A07 | TTS | Ses kodlayıcı +GAN |
| | A08 | TTS | Yapay sinir ağı dalga formu modeli |
| | A09 | TTS | Ses kodlayıcı |
| | A10 | TTS | Yapay sinir ağı dalga formu modeli |
| | A11 | TTS | Griffin-Lim Algoritması |
| | A12 | TTS | Yapay sinir ağı dalga formu modeli |
| | A13 | TTS_VC | Dalga formu uc uca ekleme + dalga formu filtreleme |
| | A14 | TTS_VC | Ses kodlayıcı |
| | A15 | TTS_VC | Yapay sinir ağı dalga formu modeli |
| | A16 | TTS | Dalga formu uc uca ekleme |
| | A17 | VC | Dalga formu filtreleme |
| | A18 | VC | Ses kodlayıcı |
| | A19 | VC | Spektral Filtreleme |

## 3.2. Öznitelikleri Elde Edilişi

GFMGD öznitelikleri elde edilirken, 25ms uzunluğunda ve 8.5ms ötelemeli hamming pencereleme fonksiyonu kullanılmıştır. Pencerelenen sinyal üzerinden elde edilen cepstral katsayılardan 0. katsayı dahil olmak üzere ilk 12 statik, delta ve delta-delta cepstral katsayıları kullanılarak 36 boyutlu GFMGD öznitelik vektörü elde edilmiştir. Ek olarak Bölüm 2.2'de bahsedilen rho ve gamma parametreleri veritabanına göre değişiklik gösterebilen optimizasyon parametreleri olup, ilgili veritabanı için her iki parametre de 0'dan 1'e kadar taranarak geliştirme seti için optimum koşulda sırasıyla 0.4 ve 0.6 olarak seçilmiştir. Bu çalışmada kullanılan diğer bir öznitelik olan CQCC için ise 0. katsayılar dahil olmak üzere ilk 29 statik, delta ve delta-delta cepstral katsayıları kullanılarak toplamda 90 boyutlu CQCC öznitelik vektörü temel sistemde olduğu gibi elde edilmiştir. CQCC hesaplamasında CQT giriş parametreleri olarak, Fmax=8kHz maksimum frekans, Fmin=16Hz minimum frekans, B=96 oktav başına atom sayısı, d=16 örnekleme periyodu olarak belirlenmiştir. Yapılan bütün çalışmalarda hem gerçek hem de sahte ses kayıtları için sınıflandırma amacıyla 1024 noktalı GKM modeli kullanılmıştır. ASVSpoof 2019 temel sistemine ek olarak Covarep Toolbox [34] ve VLFeat Toolbox [35] kütüphanelerinden yararlanılmıştır.

## 3.3. Performans Değerlendirmesi

### 3.3.1. Eşit Hata Oranı (Equal Error Rate – EER)

Sahte konuşmacı algılama sistemlerinin performansını değerlendirmek amacıyla ASVSpoof tarafından bilinen hata hesaplama yöntemlerinin dışında farklı bir performans kriteri olan eşit hata oranı (Equal-Error Rate-EER) kullanılmaktadır. EER hesaplanırken, gerçek ve sahte ses gauss karışımlarına gelen örnek bir noktanın gerçek ya da sahte olduğuna karar vermek için bir θ eşik değeri belirlenmelidir. Bu eşik değerinin özelliği Pmiss (Test sinyalinin sahte sese ait olduğu durumda gerçek ses olarak algılanma olasılığı) ve Pfa (Test sinyalinin gerçek sese ait olduğu durumda sahte ses olarak algılanma olasılığı) olasılıklarının birbirine eşit olduğu eşiği ifade etmesidir. EER hesaplanması için Pmiss ve Pfa olasılıkları (14) ve (15) numaralı denklemlerdeki gibi hesaplanabilir [36].

$$P_{miss}(\theta) = \frac{(skor \leq \theta \; olan \; gerçek \; sese \; ait \; skor \; sayısı)}{Toplam \; gerçek \; sese \; ait \; skor \; sayısı} \tag{14}$$

$$P_{fa}(\theta) = \frac{(skor > \theta \; olan \; sahte \; sese \; ait \; skor \; sayısı)}{Toplam \; sahte \; sese \; ait \; skor \; sayısı} \tag{15}$$

Burada $P_{miss}$ ve $P_{fa}$ değerleri θ eşik değerine göre artan ya da azalan olasılık fonksiyonlarıdır. Eğer bu iki olasılığın birbirine eşit olduğu $\theta_{EER}$ değeri eşik değer olarak kabul edilirse bu durumda EER değeri (16) numaralı denklemde belirtildiği gibi bulunabilir.

$$EER = P_{fa}(\theta_{EER}) = P_{miss}(\theta_{EER}) \tag{16}$$

### 3.3.2. Minimum Tandem Tespit Maliyet Fonksiyonu (minimum t-DCF)

ASVSpoof2019'un bir diğer performans değerlendirme kriteri minimum t-DCF olarak tanımlanmıştır. Minimum t-DCF sahte konuşmacı algılama sistemi ile bir ASV sisteminin ardışık olarak kullanılması durumundaki kaskat sistemin performansını ölçmek amacıyla kullanılmaktadır. t-DCF fonksiyonunun hesaplanışı (17) numaralı denklemde gösterilmiştir.

$$tDCF(\theta) = C_1 P_{miss}(\theta) + C_2 P_{fa}(\theta) \tag{17}$$

Burada $C_1$ ve $C_2$ sistemin hatalı karar vermesinin maliyeti, önsel olasılıklar, hatalı tespit olasılıklarına göre belirlenen katsayılardır. Katsayıların hesaplanışı (18) ve (19) numaralı denklemlerde gösterilmektedir.

$$C_1 = \pi_{tar}(C_{miss}^{cm} - C_{miss}^{asv} P_{miss}^{asv}) - \pi_{non} C_{fa}^{asv} P_{fa}^{asv} \tag{18}$$

$$C_2 = C_{fa}^{cm} \pi_{sahte}\left(1 - P_{miss,sahte}^{asv}\right) \tag{19}$$

Burada $C_{miss}^{cm}$ ,$C_{miss}^{asv}$ sırasıyla sahte konuşmacı algılama ve ASV sistemlerinin sahte sesi kaçırma maliyetlerini, $C_{fa}^{cm}$, $C_{fa}^{asv}$ sırasıyla gerçek sesi sahte ses olarak algılama maliyetlerini, $P_{miss}^{asv}$,$P_{fa}^{asv}$ ise sırasıyla ASV sistemin kaçırma ve hatalı alarm verme olasılıklarını, $\pi_{tar}$, $\pi_{non}$ ve $\pi_{sahte}$ ise hedef konuşmacıya ait olan, hedef konuşmacıya olmayan ve sahte sesler için ön olasılık katsayılarını ifade etmektedir. ASVSpoof 2019 için ön olasılık katsayıları sırasıyla $\pi_{tar} = 0{,}9405$, $\pi_{non} = 0{,}0095$ ve $\pi_{sahte} = 0{,}05$ olarak maliyetler ise $C_{miss}^{cm} = 1$ $C_{miss}^{asv} = 1$ $C_{fa}^{cm} = 10$ $C_{fa}^{asv} = 10$ olarak belirlenmiştir[37].

# 4. Çıktılar

Önerilen sahte konuşmacı algılama sistemi ASVSpoof 2019 veritabanı üzerinde test edilerek geliştirme ve doğrulama veritabanları üzerinden elde edilen performanslar algoritma bazlı olarak Tablo – 3'te gösterilmiştir. A01'den A19'a kadar olan algoritmalardan A01-A06 aralığındakiler geliştirme veritabanı, A07-A19 aralığı ise doğrulama veritabanı için kullanılmıştır. Geliştirme veritabanından elde edilen sonuçlara göre A01 ve A02 yöntemleriyle oluşturulan sahte sesler hatasız olarak tespit edilmiş, A03 için önerilen sistem hatasız çalışırken temel sistemde bir miktar hata görülmektedir. A04 için ise bu durumun tersi gözlenmiştir. A05'de ise önerilen sistem, temel sisteme göre yaklaşık %30'luk bir performans artışıyla EER değerini %0.94'den %0.67'ye düşürmektedir. Minimum t-DCF değeri de benzer şekilde etkilenmiştir. A06'da ise EER'de bir değişiklik gözlenmezken t-DCF ise temel sistemden küçük bir farkla 0.0011'den 0.0020'ye artmıştır. Geliştirme veritabanı üzerinde EER ortalamaları alındığında önerilen sistem %0.27, temel sistem ise %0.43 EER performansı göstermektedir. Benzer biçimde önerilen sistemin t-DCF değeri 0.0092 iken temel sisteminki ise 0.0123 olarak hesaplanmıştır. Doğrulama veriseti üzerinde ise A07, A08, A09, A11 ve A16 algoritmaları için her iki sistem de benzer şekilde iyi performans göstermektedir. Temel sistem belirtilen algoritmalar için ortalamada %0.052 EER, önerilen sistem ise %0.044 EER değerine sahiptir. Bunun tersi olarak A10, A13 ve A17 algoritmaları için her iki sistem de %10'un üzerinde EER performansı göstermiştir. Bu yöntemler için temel sistem ortalamada %20.36 EER değerine sahipken önerilen sistem %18.44 EER ile yaklaşık %10 performans artışı sağlamıştır. Bu durumlardan farklı olarak A14 algoritması için temel sistem %10.85 EER değerine sahipken önerilen sistem %10'un altına düşerek %4.72 EER performansı göstermektedir. Diğer taraftan A18 algoritması için temel sistem %3.81 EER değerine sahipken önerilen sistem %10'un üzerine çıkarak %12.48 EER performansı göstermektedir. Önerilen sistemin sağladığı en iyi performans artışı ise A14 ve A15 algoritmaları için görülmektedir. Temel sistem A14 için EER değeri %10'un üstünde kalarak %10.85 EER ve A15 için %1.26 EER performansları göstermekteyken önerilen sistem aynı algoritma için sırasıyla %4.72 ve %0.59 EER performansları göstererek ortalama %55'lik bir performans artışı sağlamıştır. Diğer bir taraftan A17 algoritması için her iki sistem de %10'un üzerinde olmak üzere sırasıyla temel sistem %19.62 ve önerilen sistem %24.04 EER göstermiştir. A18 algoritması incelendiğinde ise temel sistem %3.81'lik bir EER performansı gösterirken önerilen sistem %12.48'lik bir EER ile temel sistemle kıyaslandığında en zayıf performansını göstermiştir.

*Tablo 3. Temel Sistem ve geliştirilen sisteme ait genel ve algoritma bazlı performanslar*

| CQCC Tabanlı Temel Sistem ve Geliştirilen Sisteme Ait Sınıflandırma Performansları | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Kullanılan Öznitelik | CQCC-GKM | | CQCC+GFMGD-GKM | | CQCC-GKM | | CQCC+GFMGD-GKM | |
| | Geliştirme Veritabanı | | | | Doğrulama Veritabanı | | | |
| Saldırı Algoritması | min-tDCF | EER | min-tDCF | EER | min-tDCF | EER | min-tDCF | EER |
| Genel Performans | **0.0123** | **0.43** | **0.0092** | **0.27** | 0.2366 | 9.57 | 0.2366 | 9.25 |
| A01 | 0.0000 | 0.00 | 0.0000 | 0.00 | X | X | X | X |
| A02 | 0.0000 | 0.00 | 0.0000 | 0.00 | X | X | X | X |
| A03 | **0.0020** | **0.08** | **0.0000** | **0.00** | X | X | X | X |
| A04 | 0.0000 | 0.00 | 0.0025 | 0.11 | X | X | X | X |
| A05 | **0.0261** | **0.94** | **0.0184** | **0.67** | X | X | X | X |
| A06 | 0.0011 | 0.03 | 0.0020 | 0.03 | X | X | X | X |
| A07 | X | X | X | X | 0.0000 | 0.00 | 0.0005 | 0.02 |
| A08 | X | X | X | X | **0.0007** | **0.04** | **0.0005** | **0.02** |
| A09 | X | X | X | X | **0.0060** | **0.14** | **0.0037** | **0.06** |
| A10 | X | X | X | X | **0.4149** | **15.16** | **0.2883** | **10.32** |
| A11 | X | X | X | X | **0.0020** | **0.08** | **0.0015** | **0.06** |
| A12 | X | X | X | X | **0.1160** | **4.74** | **0.0997** | **4.07** |
| A13 | X | X | X | X | **0.6729** | **26.15** | **0.5445** | **20.98** |
| A14 | X | X | X | X | **0.2629** | **10.85** | **0.1211** | **4.72** |
| A15 | X | X | X | X | **0.0344** | **1.26** | **0.0159** | **0.59** |
| A16 | X | X | X | X | 0.0000 | 0.00 | 0.0017 | 0.06 |
| A17 | X | X | X | X | 0.9820 | 19.62 | 0.9882 | 24.04 |
| A18 | X | X | X | X | 0.2818 | 3.81 | 0.7092 | 12.48 |
| A19 | X | X | X | X | 0.0014 | 0.04 | 0.0084 | 0.26 |

EER'ye ait yukarıdaki gözlemler genellikle t-DCF değerlerine de yansımaktadır. Fakat t-DCF'deki değişim her zaman EER'deki değişim miktarı kadar olmamaktadır. Örneğin doğrulama veritabanı üzerindeki genel EER ortalaması incelendiğinde önerilen sistem %9.25'lik bir EER ve temel sistem %9.57'lik bir EER'ye sahip olmasına rağmen minimum t-DCF değerlerinde bir değişiklik olmayıp her iki sistem için de 0.2366 olarak bulunmuştur.

Yukarıda açıklanan sonuçlar yakından incelenirse GFMGD özniteliği sisteme dahil edildiğinde özellikle A17 ve A18 gibi sahte ses üretmek için direkt olarak gırtlak akımının kendisinden yararlanan algoritmalar dışında faz tabanlı özniteliklerin yapay sinir ağı tabanlı sistemler kullanılarak üretilmiş olan sahte seslerin tespiti konusunda sistem performansını olumlu yönde etkilediği gözlenmiştir. Bunun sebebi olarak bu tip algoritmalarla üretilen sahte seslerde, konuşmanın kaynağı olan glottal akıma ait faz

bilgisinin korunamaması ya da yeteri kadar iyi taklit edilememesi düşünülebilir. Diğer taraftan sahte ses üretme algoritmalarından bazıları dalga form uc uca ekleme olarak bilinen ve hedef konuşmacıya ait gerçek konuşma kesitlerinin uygun biçimde bir araya getirilmesi yöntemini yani doğrudan hedef konuşmacının gerçek sesini kullanmaktadır. Bu tip durumlarda önerilen sistem, CQCC'nin tek başına kullanıldığı durumlarda elde edilen performansa göre daha zayıf bir performans gösterse de %1'in altında bir EER değerine sahip olduğundan bu zaafiyetin sınırlı olduğu gözlenmektedir. Önerilen sistem, A17 ve A18'de olduğu gibi konuşma sinyalinin ters filtrelenmesi sonucu elde edilen gırtlak akımı üzerinde yapılan spektral değişiklikler sonucunda saldırıyı yapan kişiye ait GA'nın hedef konuşmacıya ait GA'ya benzetilmesi durumunda diğer yöntemlere göre daha zayıf performans göstermektedir. Aynı durumlar için temel sistem de benzer bir davranış göstermektedir. Bu noktada, hedef konuşmacıların gerçek sesi ile gerçek GA kullanılarak üretilmiş olan sahte sesler arasındaki spektral farkların, gerek genlik spektrumu kullanan temel sistem gerekse hem genlik hem de faz spektrumu kullanan önerilen sistem tarafından kolaylıkla ayırt edilemediği anlaşılmaktadır. Genel olarak önerilen sistem, modern yapay sinir ağları tabanlı sistemler kullanılarak sentezlenen ya da dönüştürülen sahte seslere karşı temel sisteme göre daha yüksek sınıflandırma performansı göstermektedir.

# 5. Sonuç

Bu çalışmada ASV sistemlerine tehdit oluşturan sentetik ve dönüştürülmüş seslerin tespit edilmesi amaçlanmıştır. Günümüzde konuşmacının gerçek sesini kullanarak sahte ses üreten dalga form uc uca ekleme ve dalga form filtreleme gibi klasik yöntemlere ek olarak son dönemde, ses dönüştürmede daha yüksek konuşma kalitesine sahip, WaveNet gibi yapay sinir ağları ve WORLD gibi ses kodlayıcıları kullanan yöntemler ASV sistemlerini aldatmak amacıyla kullanılmaktadır. Bu amaçla, ASVSpoof 2019 için anlık Q kepstral katsayıları (Constant Q Cepstral Coefficients – CQCC) ile geliştirilmiş temel sisteme gırtlak akımının faz bilgisini içeren değiştirilmiş grup gecikmesi (Glottal Flow Modified Group Delay -GFMGD) özniteliği dahil edilerek yeni bir sistem önerilmiştir. Yapay sinir ağı tabanlı sistemler ve ses kodlayıcılar tarafından sentezlenen veya dönüştürülen sahte seslere karşı faz bilgisi de kullanan önerilen sistem, temel sisteme göre %55'e kadar daha iyi sınıflandırma performansı sağlayabilmektedir. Doğrudan konuşma sinyalinin kendisi kullanılarak üretilen sahte seslerde ise her iki sistemin de %1'in altında EER performansı göstermesi sonucunda sistemler arasında belirgin bir fark gözlenmemiştir. Hem temel sistem hem de önerilen sistem gerçek konuşmanın ters filtrelenmesiyle elde edilen gırtlak akımı bilgisinin kullanılmasıyla üretilen sahte seslerde diğer yöntemlere göre daha zayıf performans göstermektedir.

# 6. Teşekkür

Bu araştırmada yer alan tüm/kısmi nümerik hesaplamalar TÜBİTAK ULAKBİM, Yüksek Başarım ve Grid Hesaplama Merkezi'nde (TRUBA kaynaklarında) gerçekleştirilmiştir. TÜBİTAK ULAKBİM'e çalışmalarımız sırasında TRUBA kaynaklarını paylaştığı için teşekkür ederiz.

# Kaynakça

[1]. Z. Wu, P. L. D. Leon, C. Demiroglu, vd., "Antispoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 4, pp. 768–783, Nisan 2016.

[2]. R. G. Hautamäki, T. Kinnunen, vd., "Automatic versus human speaker verification: The case of voice mimicry, Speech Commun.," vol. 72, pp. 13–31, 2015.

[3]. Y. W. Lau, M. Wagner, vd., "Vulnerability of speaker verification to voice mimicking," in Proc. Int. Symp. Intell. Multimedia, Video Speech Process., pp. 145-148, Ekim 2004.

[4]. J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in Proc. IEEE Int. Carnahan Conf. Secur. Technol. (ICCST), pp. 1-8, Ekim 2011.

[5]. P. L. De Leon, M. Pucher, vd., "Evaluation of speaker verification security and detection of HMM-based synthetic speech," IEEE Trans. Audio Speech Lang. Process., vol. 20, no. 8, pp. 2280–2290, Ekim 2012.

[6]. Z. Wu and H. Li, "Voice conversion versus speaker verification: An overview," APSIPA Trans. Signal Inf. Process., vol. 3, p. e17, 2014.

[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis." In Proc. Eurospeech, pp. 2347–2350, 1999.

[8] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang. "USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method." In Proc. the Blizzard Challenge Workshop, 2006.

[9] A.W. Black. "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling." In Proc. Interspeech, pages 1762–1765, 2006.

[10] H. Zen, T. Toda, M. Nakamura, and K. Tokuda. "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005." IEICE Trans. Inf. Syst., E90-D(1): 325–333, 2007.

[11] H. Ze, A. Senior, and M. Schuster. "Statistical parametric speech synthesis using deep neural networks." In Proc. ICASSP, pages 7962–7966, Mayıs 2013.

[12] Z. H. Ling, L. Deng, and D. Yu. "Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis." IEEE Transactions on Audio, Speech, and Language Processing, 21(10):2129–2139, Ekim 2013.

[13] T. F. Quatieri. "Discrete-Time Speech Signal Processing: Principles and Practice." Prentice- Hall, Inc., 2002.

[14] L. R. Rabiner, R. W. Schafer, "Theory and Applications of Digital Speech Processing (1st edition).", Prentice- Hall, Inc., 1975
[15] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga. "Evaluation of speaker verification security and detection of HMM-based synthetic speech. Audio, Speech, and Language Processing," IEEE Transactions on, 20(8):2280–2290, Ekim 2012.
[16] Z.Wu, E.S. Chng, and H. Li. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In Proc. Interspeech, 2012.
[17] Y. Stylianou. "Voice transformation: a survey." In Proc. ICASSP, pp. 3585–3588, 2009.
[18] A. Kain and M.W. Macon. "Spectral voice conversion for text-to-speech synthesis." In Proc. ICASSP, volume 1, pp. 285–288, 1998.
[19] Y. Stylianou, O. Capp´e, and E. Moulines. "Continuous probabilistic transform for voice conversion." Speech and Audio Processing, IEEE Transactions on, 6(2):131–142, 1998.
[20] V. Popa, H. Silen, J. Nurminen, and M. Gabbouj. "Local linear transformation for voice conversion." In Proc. ICASSP, pp. 4517–4520. IEEE, 2012.
[21] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu. "Voice conversion with smoothed GMM and MAP adaptation." In Proc. EUROSPEECH, pp. 2413–2416, 2003.
[22] H.-T. Hwang, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen. "A study of mutual information for GMM-based spectral conversion." In Proc. Interspeech, 2012.
[23] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj. "Voice conversion using partial least squares regression." Audio, Speech, and Language Processing, IEEE Transactions on, 18(5):912–921, 2010.
[24] N. Pilkington, H. Zen, and M. Gales. "Gaussian process experts for voice conversion." In Proc. Interspeech, 2011.
[25] Kamble, M. R., Sailor, H. B., Patil, H. A., & Li, H. "Advances in anti-spoofing: from the perspective of ASVspoof challenges." APSIPA Transactions on Signal and Information Processing, 9., 2020.
[26]. P. Alku, E. Vilkman, U. K. Laine, "Analysis of glottal waveform in different phonation types using the new IAIF-method." In Proc. 12th Int. Congress Phonetic Sciences, Vol. 4, pp. 362-365, Ağustos 1991.
[27]. N.P. Narendra, M. Airaksinen, B. Story, P. Alku, "Estimation of the glottal source from coded telephone speech using deep neural networks." Speech Communication, vol. 106, pp. 95-104., 2019
[28]. M. Todisco, H. Delgado, N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification." Computer Speech & Language, vol. 45, pp. 516-535, 2017
[29] Quatieri, T., "Discrete-Time Speech Signal Processing: Principles and Practice." Prentice Hall PTR, pp. 111–174., 2001
[30]. Z. Wu, E.S. Chng, vd., "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in Proc. of Interspeech, 2012.
[31]. L.D. Alsteris & K.K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," Digital Signal Processing, vol. 17, no. 3, pp. 578–616, 2007.
[32] Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., ... & Lee, K. A., "Asvspoof 2019: Future horizons in spoofed and fake audio detection." arXiv:1904.05441, 2019.
[33]. Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., ... & Juvela, L., "ASVspoof 2019: a large-scale public database of synthetic, converted and replayed speech.", arXiv, arXiv-1911, 2019
[34] Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S., "COVAREP—A collaborative voice analysis repository for speech technologies." In 2014 IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP), pp. 960-964. IEEE, Mayıs 2014
[35] Vedaldi, A., & Fulkerson, B., "VLFeat: An open and portable library of computer vision algorithms." In Proceedings of the 18th ACM international conference on Multimedia, pp. 1469-1472, Ekim 2010.
[36]. T. Kinnunen, M. Sahidullah, vd., "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc., pp. 2–6., 2017.
[37]. T. Kinnunen, K. Lee, vd., "t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in Proc. Odyssey, Les Sables d'Olonne, Fransa, Haziran 2018.