Derin Sinir Ağları ile Tekrar Saldırılarının Tespiti Replay Spoofing Attack Detection Using Deep Neural Networks

Bekir Bakar ve Cemal Hanilçi
Elektrik-Elektronik Mühendisliği Bölümü
Bursa Teknik Üniversitesi
Bursa, Türkiye
b.bakar@outlook.com, cemal.hanilci@btu.edu.tr

***Özetçe* —Son yıllarda konuşmacı doğrulama (KD) sistemlerine olan ilgi artmış ve KD sistemlerinin kullanımı yaygınlaşmıştır. Bu durum; yanıltma saldırılarının tespitini, gerçek sesin sahte sesten ayırt edilmesi, KD sistemleri için önemli bir araştırma konusuna dönüştürmüştür. Bu çalışmada KD sistemlerine daha önceden kaydedilmiş bir sesin tekrar oynatılması ile gerçekleştirilebilecek tekrar saldırılarının tespiti problemi ele alınmıştır. Mel frekansı kepstrum katsayıları (MFCC) ve uzun dönem ortalama spektrum (LTAS) istatistikleri özniteliklerinin, derin sinir ağları (DNN) sınıflandırıcısı ile tekrar saldırılarının tesit edilmesi gerçekleştirilmiştir. ASVspoof 2017 veritabanı ile yapılan deneylerde: DNN ile LTAS ve MFCC öznitelikleri sınıflandırıldığında, ASVspoof 2017 yarışmasının referans yöntemi olan sabit Q dönüşümü kepstral katsayılarının (CQCC) GMM sınıflandırıcısı ile modellenmesine göre daha iyi performans gösterdiği tespit edilmiştir.**

Anahtar Kelimeler—konuşmacı doğrulama, yanıltma saldırıları, saldırı tespiti, derin sinir ağları

***Abstract*—In recent years, there has been increased interest in speaker verification (SV) systems and their usage has become widespread. This situation made the detecting of spoofing attacks, the discrimination of genuine speech from spoofed speech, an important research area for speaker verification (SV) systems. In this study, detection of replay spoofing attacks where a pre-recorded speech signal is used to gain unauthorized access to ASV systems is studied. Mel frequency cepstral coefficients (MFCC) and long-term average spectrum (LTAS) statistics features are used to detect replay attacks using deep neural network (DNN) classifier. Experimental results using ASVspoof 2017 database show that MFCC and LTAS features with DNN classifier outperforms the Gaussian mixture model (GMM) classifier with constant Q transform cepstral coefficients (CQCC) which is the baseline replay attack detection system of the ASVspoof 2017 challenge.**

Keywords—speaker verification, spoofing attacks, anti-spoofing, deep neural networks

I. GİRİŞ

Konuşmacı doğrulama (KD) [1], verilen bir ses sinyalinin iddia edilen kişiye ait olup olmadığının tespit edilmesini hedefleyen, kimlik kabul veya reddetme işlemidir. Tıpkı parmak izi tanıma, yüz tanıma vb. biyometrik sistemlerde olduğu gibi; KD sistemleri kişisel verilerin güvenliği, çeşitli sistemlere erişim kontrolü, banka işlemleri ve daha bir çok alanda kullanılmaktadır [2]. Teknolojik gelişmeler ve bu alanda yapılan çalışmaların artmasıyla birlikte, sistemlerin kullanım alanı gün geçtikçe artmaktadır.

KD ve diğer biyometrik sistemlerin kullanımının yagınlaşması, sistemlere yapılan yanıltma saldırılarını da beraberinde getirmiştir. Yanıltma saldırısı, bir kişinin (saldırganın) erişim izni olmayan bir sisteme erişme girişimidir [2]. En son geliştirilen KD sistemleri de dahil olmak üzere, sistemlerin yanıltma saldırılarına karşı savunmasız olduğu birçok çalışmada açıkca gösterilmiştir [3].

Bir KD sistemine yapılması olası saldırı türleri; konuşmacının sesinin sentezlenerek oluşturulması *-ses sentezleme-* [4], çeşitli yazılımlar kulanılarak saldırganın herhangi bir sesi hedef kişinin sesine dönüştürmesi *-ses dönüştürme-* [5], hedef konuşmacının sesinin taklit edilmesi *-taklit-* [6], hedef konuşmacının önceden kaydedilmiş ses kaydının kullanılması *-tekrar oynatma-* [3] şeklinde alt ketegorilere ayrılabilir. Bu saldırı türlerinden taklit saldırısı özel bir yetenek gerektirmesi nedeniyle, karşılaşılması güç bir saldırı türüdür ve önemli bir tehdit olarak algılanmamaktadır. Ses sentezleme ve ses dönüştürme saldırıları kolay erişilebilir birçok yazılım kullanılarak yapılabileceği ve saldırı başarı oranının yüksek olması nedeniyle önemli bir tehdit olarak görülmektedir. Tekrar saldırıları ise; hemen herkesin sahip olduğu basit bir cep telefonu ile yapılabilmesi sebebiyetiyle en önemli yanıltma saldırısı türüdür.

Literatürde bulunan bazı tekrar saldırısı tespit çalışmaları şöyle özetlenebilir: Villalba ve Lleida [7], bilinen başarılı konuşmacı doğrulama tekniklerinden biri olan birleşik etmen analizi (joint factor analysis - JFA) ile konuşmacı doğrulama deneylerinde normal durumda 0.71% eşit hata oranı (EER) değeri elde edilirken, önceden kaydedilmiş (tekrar saldırıları) seslerin yanlış sınama sesleri ile yer değiştirilmesi neticesinde, sistemin tekrar kayıtlarının %68'ini kabul ettiğini ortaya koymuştur [7]. Başarılı yöntemlerin tekrar saldırıları karşısında bu ölçüdeki savunmasızlığı, yeni bir arayışı beraberinde getirmiştir. Bu, tekrar saldırılarının tespit edilebilmesi problemidir. Bu kapsamda, 2017 yılında *otomatik konuşmacı doğrulama yanıltma ve saldırı tespiti (Automatic speaker verification spoofing and countermeasures challenge - ASVspoof 2017)* düzenlenmiştir [8]. Bu yarışma neticesinde yapılan çalışmalardan birinde [9], farklı öznitelikler ile Gauss Karışım Modeli (Gaussian mixture model - GMM), destek vektör makineleri (support vector

978-1-5386-1501-0/18/$31.00 © 2018 IEEE

machines - SVM) ve derin sinir ağları (deep neural networks - DNN) sınıflandırıcıları kullanılarak, tekrar saldırılarının tespit edilmesi problemi ele alınmış olup; öznitelik çıkarmada yüksek frekans bölgesi analizinin ve DNN sınıflandırıcısının daha iyi sonuç verdiği gösterilmiştir. Yüksek frekans bölgesi analizinin, tekrar saldırılarının tespitinde performansı iyileştirdiğini gösteren bir diğer çalışmada [10] ise; öznitelikler 6000-8000 Hz frekans aralığından çıkarıldığında %3.38 EER değeri elde edilmiştir. ASVspoof 2017 yarışmasında verilen referans saldırı tespiti yöntemine (GMM sınıflandırıcısı ve sabit Q dönüşümü kepstral katsayıları) nazaran daha iyi performansın elde eden bir diğer çalışmada [11] ise; derin öğrenme yöntemlerinin karşılaştırılması yapılmıştır. Tüm bu çalışmalar göstermektedir ki; tekrar saldırıları KD sistemleri için önemli bir tehdittir ve DNN sınıflandırıcısı ile geliştirilen sistemlerin performansı bu çalışmada motivasyon kaynağı olmuştur.

Bu çalışmada iki farklı yöntem kullanılarak, ses sinyalinden öznitelikler çıkarılmıştır. Bu öznitelikler kullanılarak; ses sinyalinin gerçek veya sahte sınıfa ait olup olmadıklarının ayırt edilmesi için derin sinir ağları yöntemiyle bir model oluşturulmuştur. Daha sonra bu modelden elde edilen olasılıklar, sistem performansını tespit etmek için; ses işleme problemlerinde yaygın olarak kullanılan skor hesaplama yöntemine çevrilmiştir. DNN ile tekrar saldırısı tespit sonuçları, ASVspoof 2017 organizasyon komitesi tarafından önerilen referans yöntem olan sabit Q transformu kepstrum katsayılarının GMM sınıflandırıcısı ile modellenmesi (CQCC-GMM) yöntemi sonuçları ile karşılaştırılmıştır. Çalışmanın devamı; tekrar saldırısı tespit yöntemi, tercih edilen öznitelik çıkarma yöntemleri, deneysel çalışmalar, deneysel sonuçlar ve tartışma şeklinde devam etmektedir.

II. Tekrar Saldırılarının Tespiti

Konuşmacı tanıma sistemlerine yapılabilecek tekrar saldırılarının tespiti aslında iki sınıflı bir örüntü tanıma problemidir. Bu sınıflardan biri *gerçek/orijinal* sınıf iken diğeri ise *sahte* sınıftır. Tekrar saldırılarının tespiti ise verilen bir ses sinyalinin orijinal veya daha önceden kaydedilmiş bir ses olup olmadığının tespit edilmesi problemidir.

Saldırı tespiti, aslında iki sınıf arasında karar verme işlemi olup, karar için logaritmik olabilirlik oran (log-likelihood ratio - $\mathcal{LLR}$) skoru kullanılabilir:

$$\mathcal{LLR} = \log p(\mathbf{X}|C_1) - \log p(\mathbf{X}|C_2) \tag{1}$$

Burada, C_1 ve C_2 sırası ile *gerçek/orijinal* ve *tekrar/sahte* sınıflarını temsil eden akustik modellerdir ve her bir sınıfa ait eğitim öznitelikleri kullanılarak bu modeller eğitilebilir. Bu çalışmada kullanılan ASVspoof 2017 veritabanı ile tekrar saldırısı tespiti için organizasyon komitesi tarafından Gauss Karışım Modeli (GMM) sınıflandırıcısı referans modelleme yöntemi olarak önerilmiştir.

GMM yönteminde, her bir sınıf (orijinal ve tekrar), M adet çok boyutlu Gauss yoğunluk fonksiyonunun ağırlıklandırılmış toplamı şeklinde, $p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i p_i(\mathbf{x})$, temsil edilir. Burada $p_i(\mathbf{x})$ ortalaması μ_i, ortak değişinti matrisi $\mathbf{\Sigma}_i$ olan $D-$boyutlu (D:öznitelik vektörlerinin boyutu) Gauss yoğunluk fonksiyonunu belirtmekte olup bütün bir GMM $\lambda = \{w_i,\ \mu_i,\ \mathbf{\Sigma}_i\}_{i=1}^{M}$ şeklinde gösterilir. GMM yönteminin eğitim aşamasında, her bir sınıfın eğitim öznitelik vektörleri $\mathbf{X} = \{\mathbf{x}_1,\ \mathbf{x}_2,\ \ldots,\ \mathbf{x}_N\}$ eğitim öznitelikleri kullanılarak en büyük olabilirlik kriteri kullanılarak beklentinin maksimumlaştırılması algoritması ile GMM parametreleri (ağırlık, ortalama ve ortak değişinti) tahmin edilir. Test aşamasında ise; test ses sinyalinden elde edilen öznitelik vektörleri, $\mathbf{Y} = \{\mathbf{y}_1,\ \ldots,\ \mathbf{y}_T\}$, kullanılarak logaritmik olabilirlik oranı skoru,

$$\mathcal{LLR} = \log p(\mathbf{Y}|\lambda_{\text{gercek}}) - \log p(\mathbf{Y}|\lambda_{\text{tekrar}}) \tag{2}$$

şeklinde hesaplanır. Burada λ_{gercek} ve λ_{tekrar} sırası ile gerçek ve tekrar sınıflarına ait GMM'leri belirtmektedir. $p(\mathbf{Y}|\lambda)$ $\mathbf{Y}$ test öznitelik vektörleri için GMM olabilirlik değeri olup şu şekilde hesaplanmaktadır:

$$p(\mathbf{Y}|\lambda) = \prod_{t=1}^{T} \sum_{i=1}^{M} w_i p_i(\mathbf{y}_i|\lambda). \tag{3}$$

Bu çalışmada, referans sınıflandırıcı yöntem olan Gauss karışım modeli tekniğine ek olarak derin sinir ağları (deep neural network - DNN) modelleme yöntemi kullanılmıştır. DNN son yıllarda popüler olarak kullanılan bir yaklaşım olup, farklı çalışmalarda hem sınıflandırma hem de öznitelik çıkarma amacı ile kullanılmıştır.

DNN sınıflandırıcısı ile gerçek ve tekrar seslerini birbirinden ayırt etmek için farklı öznitelikler ile ASVspoof 2017 veri tabanı eğitim kümesini kullanarak bir derin sinir ağı eğitilmiştir. Giriş katmanındaki nöron sayısı, kullanılan özniteliklerin boyutu ile aynı olup, gizli katmanlarda farklı sayılarda nöron kullanılmıştır. Çıkış katmanında ise biri gerçek, diğeri tekrar/sahte sınıfı temsil eden softmax aktivasyon fonksiyonunun kullanıldığı iki nöron kullanılmıştır. Her bir çıkış nöronu, ilgili sınıfın sonsal (posterior) olasılığını temsil ettiğinden, sonsal olasılıklar, logaritmik olabilirlik oranı skoruna şu şekilde dönüştürülmüştür:

$$\mathcal{LLR} = \log p(C_1|\mathbf{X}) - \log p(C_2|\mathbf{X}) \tag{4}$$

III. Öznitelik Çıkarma

DNN yönteminin güçlü yönlerinden biri sınıflandırma için gerekli olan ve ayırt ediciliği yüksek olan öznitelikleri kendisinin çıkarması olsa da; ses sinyallerinin kullanıldığı çalışmaların çoğunda ön-işleme adımında çıkarılan öznitelikler DNN algoritmasına giriş olarak verilmektedir [11]. Bu çalışmada iki tür öznitelik çıkarım yöntemi tercih edilmiştir:

A. Mel-Frekansı Kepstrum Katsayıları

Mel-frekansı kepstrum katsayıları (MFCC) [1], [12], konuşma analizi problemlerinde en yaygın kullanılan öznitelik çıkarma yöntemlerinden biridir. MFCC öznitelikleri, ön vurgulama yapılan ses sinyalinin her biri 25 ms uzunluğunda, 10 ms'lik kısımları birbiri ile örtüşen Hamming pencere ile pencerelenmiş çerçevelerden elde edilir. Pencerelenmiş çerçevelerin ayrık Fourier dönüşümü (DFT) alınarak genlik spektrumları elde edilir ve üçgen süzgeçlerden oluşan süzgeç takımından geçirilir. Logaritmik süzgeç çıkışlarının ayrık kosinüs dönüşümünün alınmasıyla MFCC öznitelikeri elde edilir. Burada öznitelik vektörünün ilk katsayısı logaritmik enerjidir. Bu yöntem için açık kaynak kodlu dil ve konuşmacı tanıma paketi SIDEKIT [13] kullanılmıştır.

B. Uzun Dönem Ortalama Spektrum İstatistikleri

Uzun dönem ortalama spektrum (LTAS) istatistikleri [14] özniteliklerini elde etmek için; s ses sinyali öncelikle ön-vurgulama süzgecinden geçirilerek birbiri ile örtüşen kısa çerçevelere bölünür. Hamming pencere fonksiyonu ile pencerelenen her bir çerçevenin N noktalı ayrık Fourier dönüşümü (AFD) alınarak genlik spektrumu hesaplanır. $S_t(k)$, $t = 1, 2, \ldots, T$, t. çerçevenin Fourier dönüşümünü belirtmek üzere ($k = 1, \ldots, N/2$) uzun dönem ortalama ve varyans istatistikleri elde edilir:

$$\mu(k) = \frac{1}{T}\sum_{t=1}^{T}\log|S_t(k)| \tag{5}$$

$$\sigma^2(k) = \frac{1}{T}\sum_{t=1}^{T}(\log|S_t(k)| - \mu(k))^2 \tag{6}$$

Son olarak, ortalama ve varyans istatistikleri birleştirilerek, uzun dönem spektral istatistikler öznitelik vektörleri elde edilir.

IV. Deneysel Çalışmalar

Deneyler sırasında ASVspoof 2017 [15] veritabanı kullanılmıştır. ASVspoof 2017 veritabanı 16 kHz'de örneklenmiş, 16 bit çözünürlüğünde kaydedilmiş seslerden oluşmaktadır. Bu sesler, Tablo I'de gösterildiği gibi eğitim, geliştirme ve değerlendirme olmak üzere birbiri ile örtüşmeyen üç alt kümeye ayrılmıştır. Eğitim verisi; gerçek ve sahte sınıflara ait akustik modellerin eğitilmesi amacı ile, geliştirme kümesi, kullanılan sistemin parametrelerini optimize ederek en iyi saldırı tespiti sistemini oluşturmak, değerlendirme kümesi ise en iyi modelin performansının genelleştirilebilirliği amacıyla kullanılmıştır. Veritabanı ile ilgili daha kapsamlı bilgiye [8] çalışmasından ulaşılabilir.

TABLO I: ASVspoof 2017 veri tabanı

Alt Küme	Konuşmacı Sayısı	Kayıt Sayısı	
		Gerçek	Sentetik
Eğitim	10	1507	1507
Geliştirme	8	760	950
Değerlendirme	24	1294	11987

Sınıfladırıcı olarak Bölüm II'de bahsedilen DNN yöntemi kullanılmıştır. MFCC öznitelikleri için üç ve uzun dönem ortalama spektrum istatistikleri (LTAS) öznitelikleri için ise beş adet gizli katmandan oluşan DNN kullanılmıştır. Hata fonksiyonunu optimize etmek için Stochastic Gradient Descent (SGD) algoritması kullanılmıştır. Her iki öznitelik türü kullanılarak farklı sayılarda nöron ve gizli katman içeren DNN sınınflandırcısı ile yapılan deneylerle ideal parametreler tespit edimiştir. MFCC ve LTAS öznitelik vektörlerinin boyutları ve karakteristikleri benzer olmadığından, öznitelikler için tespit edilen ideal DNN yapısı farklıdır. Burada MFCC için giriş katmanı nöron sayısı; kullanılan öznitelik sayısı (19 adet) ve ilave olarak logaritmik enerji katsayısı olmak üzere 20, her bir gizli katman ise 1024 birimden oluşmaktadır. LTAS öznitelikleri için ise 256 adet ortalama, 256 standart sapma istatistikleri birleştirilerek tek bir öznitelik vektörü oluşturulduğundan giriş katmanında 512, gizli katmanlarda ise 1024 nöron bulunmaktadır. Her iki durum için de, gizli katmanlarda ReLU aktivasyon fonksiyonu kullanılmıştır. KD sistemler için tekrar saldırı tespiti ikili sınıflandırma problemi olduğundan; çıkış katmanlarının nöron sayısı iki ve aktivasyon fonksiyonu softmax olarak belirlenmiştir. İki adet çıkış birimi sırası ile gerçek ve sahte sınıflara karşılık gelmektedir. Bu deneyler; açık kaynak kodlu derin öğrenme kütüphanesi Keras [16] kullanılarak gerçekleştirilmiştir.

Aşırı eğitimi (over training) engellemek amacıyla, öğrenme oranı (learning rate) düşük seçilmiş ve her bir katmandan sonra 0.75 değerinde dropout [17] eklenmiştir. Öğrenme oranının düşük olması nedeniyle en iyi modele ulaşmak için epoch (devir) sayısı yüksek (10000) seçilmiştir. Diğer parametrelerin sisteme etkisini azaltmak [18] ve hafıza problemlerini aşmak için; eğitim verisi gruplara ayrıştırılarak eğitilmiştir (batch training). Eğitilen modelin, geliştirme verisi kullanılarak belirli bir frekansta validasyonu yapılmış ve validasyon skoru bir önceki sonuçtan iyi olduğu tespit edilen model en iyi model olarak kaydedilmiştir. Eğitim zamanından tasarruf amacı ile, validasyon skoru azalmadığında veya artmaya başladığında 10000 epoch sayısı beklenmeden eğitim işlemi sonlandırılmıştır (early stopping).

ASVspoof 2017 yarışmasında tekrar saldırılarının tespiti problemi için performans metriği olarak eşit hata oranı (equal error rate - EER) kullanılmıştır [8], [19]. EER değeri, yanlış kabul ve yanlış ret oranlarının birbirine eşit olduğu eşik değerdeki hata oranına karşılık gelmektedir. Ayrıca EER metriğine ilaveten, önerilen sistemin bütün performansını görüntülemek amacıyla sezim hata ödünleşimi (detection error trade-off - DET) [20] eğrileri de verilmiştir.

V. Deneysel Sonuçlar

Deneyler sonucunda MFCC ve LTAS öznitelikleri ile geliştirme ve değerlendirme kümesi için elde edilen sonuçlar Tablo II'de verilmiştir. LTAS öznitelikleri ile geliştirme kümesinde $\%4.55$, değerlendirme kümesinde ise $\%18.10$ EER değerleri elde edilmiştir. MFCC öznitelikleri için ise sırasıyla $\%18.78$ ve $\%24.81$ EER değerleri elde edilmiştir. Sonuçlara göre LTAS özniteliklerinin bu sınıflandırıcı yapısı için, MFCC özniteliklerine göre daha iyi bir seçim olduğu açıkça görülmektedir. Sistem performansının daha iyi gözlemlenebilmesi amacıyla; ASVspoof 2017 yarışmasının temel sistemi olan CQCC-GMM [8], [19] sonuçları da tabloda verilmektedir. Görüldüğü gibi, geliştirme kümesinde LTAS öznitelikleri ve DNN sınıflandırıcısı referans sistemden yaklaşık 2.5 kat daha iyi başarım göstermektedir. Fakat MFCC öznitelikleri DNN sınıflandırıcısı ile birlikte kullanıldığında geliştirme kümesinde referans sisteme nazaran daha yüksek hata oranı elde edilmiştir.

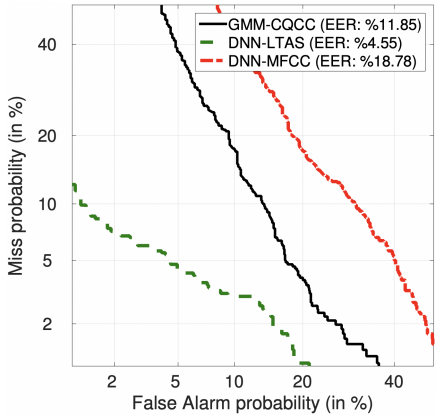
TABLO II: ASVspoof 2017 veri tabanı ile elde edilen EER değerleri

Öznitelik	Sistem	Alt Küme	
		Geliştirme	Değerlendirme
MFCC	DNN	18.78	24.81
LTAS	DNN	4.55	18.10
CQCC	GMM	11.85	30.00

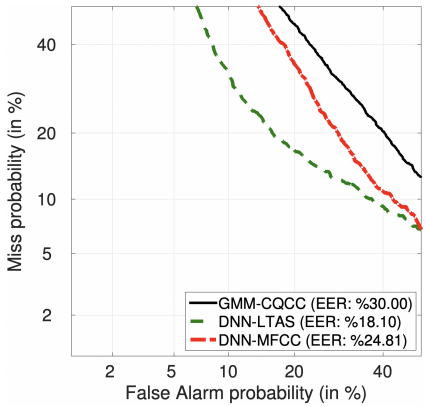
Değerlendirme kümesi sonuçları göz önüne alındığında, önerilen sistemin (LTAS öznitelikleri ve DNN sınıflandırıcısı)

referans sisteme (GMM sınıflandırıcısı ve CQCC öznitelikleri) göre daha yüksek performans gösterdiği görülmektedir. Benzer şekilde MFCC öznitelikleri ile de referans sistemden daha iyi performans elde edilmiştir.

Son olarak Şekil 1 ve Şekil 2'de her iki öznitelik türü (MFCC ve LTAS) ve referans sistem (GMM-CQCC) ile sırası ile geliştirme ve değerlendirme alt kümeleri için elde edilen DET eğrileri verilmiştir.



Şekil 1: Geliştirme kümesi için elde edilen DET grafiği.



Şekil 2: Değerlendirme kümesi için elde edilen DET grafiği.

VI. Tartışma

Bu çalışmada; daha önceden kaydedilen bir sesin tekrar oynatılması yöntemiyle, KD sistemlerine yapılması olası saldırıların tespit edilmesi problemi ele alınmıştır. MFCC ve uzun dönem ortalama spektrum (LTAS) istatistikleri yöntemleriyle elde edilen öznitelikler, DNN kullanılarak sınıflandırılmıştır. MFCC özniteliklerinin vektör boyutunun çeşitli yöntemlerle büyütülerek kullanılması DNN sistemler için daha iyi sonuç verebileceği düşünülmektedir. Ayrıca DNN yapısının ASVspoof 2017 yarışmasında belirlenen referans sisteme (CQCC - GMM) nazaran daha iyi olduğu gözlemlenmiştir. Farklı öznitelik yöntemleri de kullanılarak daha iyi bir DNN yapısı ile iyi sonuçlar alınabileceği gelecek çalışmalar için motivasyon kaynağıdır.

Teşekkür

Bu çalışma TÜBİTAK (proje numarası 115E916) tarafından desteklenmiştir.

Kaynaklar

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, pp. 12–40, Jan. 2010.

[2] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics systems under spoofing attack: An evaluation methodology and lessons learned," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 20–30, Sept. 2015.

[3] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, Feb. 2015.

[4] P. L. D. Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of hmm-based synthetic speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2280–2290, Oct. 2012.

[5] Y. Stylianou, "Voice transformation: A survey," in *Proc. ICASSP*, Apr. 2009, pp. 3585–3588.

[6] R. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication*, vol. 72, pp. 13–31, Sept. 2015.

[7] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *Proc. FALA*, 2010, pp. 131–134.

[8] T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Sahidullah, M. Todisco, and H. Delgado, "Asvspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training*, vol. 10, no. 1508, p. 1508, 2017.

[9] L. Li, Y. Chen, D. Wang, and F. Zheng, "A study on replay attack and anti-spoofing for automatic speaker verification," in *Proc. INTERSPEECH*, 2017, pp. 92–96.

[10] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Galka, "Audio replay attack detection using high-frequency features." in *Proc. INTERSPEECH*, 2017, pp. 27–31.

[11] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using dnn for channel discrimination," in *Proc. INTERSPEECH*, 2017, pp. 97–101.

[12] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[13] "An extensible speaker identification SIDEKIT in python," http://www-lium.univ-lemans.fr/sidekit/, accessed:2016.

[14] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "Presentation attack detection using long-term spectral statistics for trustworthy speaker verification," in *Proc. BIOSIG*, Sept. 2016, pp. 1–6.

[15] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamäki, D. Thomsen, A. Sarkar, Z. H. Tan, H. Delgado, M. Todisco, N. Evans, V. Hautamäki, and K. A. Lee, "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proc. ICASSP*, March 2017, pp. 5395–5399.

[16] F. Chollet *et al.*, "Keras," https://github.com/keras-team/keras, 2015.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, Jan. 2014.

[18] M. Li, T. Zhang, Y. Chen, and A. J. Smola, "Efficient mini-batch training for stochastic optimization," in *Proc. KDD*, 2014, pp. 661–670.

[19] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. INTERSPEECH*, 2017, pp. 2–6.

[20] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. EUROSPEECH*, 1997.